การใช้เทคโนโลยีการหาลำดับเบสรุ่นใหม่ทั่วเอ็กโซมเพื่อหาการกลายพันธุ์ที่เกี่ยวข้องกับโรคพันธุกรรม ๔ โรค

นางสาววิภา พันธ์มณฑา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาศาสตร์การแพทย์
คณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

USING WHOLE EXOME SEQUENCING TO IDENTIFY MUTATIONS OF FOUR DIFFERENT HUMAN DISEASES

Miss Wipa Panmontha

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Medical Science
Faculty of Medicine
Chulalongkorn University
Academic Year 2015

Thesis Title	USING WHOLE EXOME SEQUENCING TO IDENTIFY MUTATIONS OF FOUR DIFFERENT HUMAN DISEASES
By	Miss Wipa Panmontha
Field of Study	Medical Science
Thesis Advisor	Professor Kanya Suphapeetiporn
Thesis Co-Advisor	Professor Vorasuk Shotelersuk

Accepted by the Faculty of Medicine, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

....Dean of the Faculty of Medicine

(ProfessorSuttipong Wacharasindhu)

THESIS COMMITTEE

....Chairman

(Professor Apiwat Mutirangura)

....Thesis Advisor

(Professor Kanya Suphapeetiporn)

....Thesis Co-Advisor

(Professor Vorasuk Shotelersuk)

....Examiner

(Assistant Professor Sunchai Payungporn)

....Examiner

(Assistant Professor Pawinee Rerknimitr)

....External Examiner

(Dr.Surasawadee Ausavarat)

วิภา พันธ์มณฑา : การใช้เทคโนโลยีการหาลำดับเบสรุ่นใหม่ทั่วเอ็กโซมเพื่อหาการกลายพันธุ์ที่เกี่ยวข้องกับโรคพันธุกรรม ๔ โรค (USING WHOLE EXOME SEQUENCING TO IDENTIFY MUTATIONS OF FOUR DIFFERENT HUMAN DISEASES) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ศ. ดร. พญ. กัญญา ศุภปีติพร, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ศ. นพ. วรศักดิ์ โชติเลอศักดิ์, 114 หน้า.

Whole exome sequencing (WES) เป็นการประยุกต์ใช้เทคโนโลยีการหาลำดับเบสรุ่นใหม่ (Next generation sequencing, NGS) เป็นการหาลำดับเบสเฉพาะส่วนที่ต้องการ ได้แก่ coding sequence, splice site, non-coding RNA และ highly conserved regions ซึ่งเป็นประมาณร้อยละ 1 ของจีโนม และประมาณร้อยละ 85 ของการกลายพันธุ์ซึ่งส่งผลกระทบให้เกิดโรคอยู่ในบริเวณดังกล่าวนี้ โดยการศึกษาวิจัยนี้ มีวัตถุประสงค์เพื่อหาการกลายพันธุ์ในยีนที่เป็นสาเหตุของโรคจำนวนสองโรคที่มีการถ่ายทอดทางพันธุกรรมที่เกิดจากยีนเดียว โรคแรก คือ โรคแฟมิเลียลโคมิโดน (familial comedones) เป็นโรคทางผิวหนังที่มีการถ่ายทอดแบบยีนเด่นบนออโตโซม มีครอบครัวชาวไทยสองครอบครัวที่พบเป็นโรคนี้และเข้าร่วมการศึกษา ผู้วิจัยศึกษาครอบครัวแรกโดยใช้ whole exome sequencing ร่วมกับการศึกษาลิงค์เกจทั่วจีโนมโดยการใช้ SNPs array พบการกลายพันธุ์ที่สำคัญหนึ่งตำแหน่ง คือ c.84_85insT ในยีน *PSENEN* สำหรับครอบครัวที่สอง พบการกลายพันธุ์ชนิดและตำแหน่งเดียวกันกับที่พบในครอบครัวแรก ผลการศึกษาการแสดงออกระดับอาร์เอ็นเอของยีน *PSENEN* พบว่า มีระดับของอาร์เอ็นเอมากกว่าในผู้ป่วยเมื่อเทียบกับกลุ่มควบคุมที่ไม่เป็นโรค อีกโรคหนึ่งที่ผู้วิจัยทำการศึกษา คือ กลุ่มอาการพัฒนาการช้าที่ไม่ทราบสาเหตุโดยพบในสมาชิกสองคนในครอบครัวที่ไม่มีการแต่งงานในเครือญาติ ซึ่งคาดว่ามีการถ่ายทอดพันธุกรรมแบบยีนด้อยบนออโตโซม ผู้วิจัยได้ทำ whole exome sequencing จากดีเอ็นเอของผู้ป่วยสองรายที่เป็นพี่น้องกันและพ่อและแม่ของผู้ป่วย อย่างไรก็ตามไม่พบยีนที่เกี่ยวข้องกับการเกิดโรคในครอบครัวนี้ ไม่ว่าจะเป็น copy number variations, single nucleotide variants หรือ indels โดยสรุปการใช้ WES นำไปสู่การค้นพบยีนใหม่ที่เป็นสาเหตุของโรคแฟมิเลียลโคมิโดน แต่อย่างไรก็ตามการหายีนที่เป็นสาเหตุของโรคที่มีสาเหตุทางพันธุรรมที่หลากหลายอย่างเช่นกลุ่มอาการพัฒนาการช้าที่ไม่ทราบสาเหตุยังคงเป็นความท้าทาย

สาขาวิชา วิทยาศาสตร์การแพทย์ ลายมือชื่อนิสิต ....

ปีการศึกษา 2558 ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

ลายมือชื่อ อ.ที่ปรึกษาร่วม ....

5474160730 : MAJOR MEDICAL SCIENCE

KEYWORDS: NEXT GENERATION SEQUENCING / WHOLE EXOME SEQUENCING / FAMILIAL COMEDONES / INTELLECTUAL DISABILITY / MUTATIONS

WIPA PANMONTHA: USING WHOLE EXOME SEQUENCING TO IDENTIFY MUTATIONS OF FOUR DIFFERENT HUMAN DISEASES. ADVISOR: PROF. KANYA SUPHAPEETIPORN, CO-ADVISOR: PROF. VORASUK SHOTELERSUK, 114 pp.

Whole exome sequencing (WES) is an application of the next generation sequencing (NGS). With this technique, the target regions such as coding sequences, splice site, non-coding RNA and highly conserved regions which are about 1 percent of the genome harboring about 85 percent of mutations with large effects on disease-related traits are sequenced. Here, two different Mendelian disorders were studied. The first is familial comedones, a rare autosomal dominant skin disorder. Two unrelated families affected with familial comedones were included. WES combined with whole genome linkage analysis using a single nucleotide polymorphism (SNP) array was conducted in the first family which identified a heterozygous mutation, c.84_85insT in the *PSENEN* gene. This mutation was also identified in the second family. Quantitative real-time PCR indicated increased expression of *PSENEN* mRNA in the patients. Another disease included in this study is an undiagnosed syndrome with intellectual disability in a non-consanguineous family. Two siblings were affected. Exome sequencing was performed in both patients and their parents. Neither pathogenic copy number variations nor SNVs/indels were identified. In summary, conducting WES led us to identify a novel gene underlying familial comedones. However, using WES to find a gene underlying a disease with genetic heterogeneity as intellectual disability remains a challenge.

Field of Study: Medical Science Student's Signature ....

Academic Year: 2015 Advisor's Signature ....

Co-Advisor's Signature ....

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Kanya Suphapeetiporn and Prof. Vorasuk Shotelersuk for the continuous support of my Ph.D study, for their guidance, patience, motivation, and immense knowledge. They advised and helped me in conducting research and writing this thesis. They are the best advisors and mentors for my Ph.D study.

Besides my advisors, I would like to thank all of my thesis committee: Prof. Apiwat Mutirangura, Asst. Prof. Sunchai Payungporn, Asst. Prof. Pawinee Rerknimitr and Dr. Surasawadee Ausavarat, for their comments and also for the challenging questions which widened my research from various perspectives.

My sincere thanks also go to Prof. Matthew State, Asst. Prof. Jeremy Willsey and members of their group at the State's lab, University of California, San Francisco, who provided me an opportunity to join their fantastic team as an intern, and who gave access to the laboratory and research facilities. I have learned many things from them. Without their superior knowledge, experience and continued support, it would be very hard for me to conduct this research.

I would like to thank all the patients and their families for participation in the study. In particular, I am grateful to Dr. Pawinee Rerknimitr and the Dermatology Unit, Department of Medicine for kind supports and helps. Nevertheless, I express my gratitude toward my family and friends who encourage me to pursue all of my goals. I also thank Dr. Patra Yeetong and everyone in the Center of Excellence for Medical Genetics for discussions, supports and advices.

This study was supported by the Royal Golden Jubilee Ph.D. Program to W.P. (grant No. PHD/0011/2554), the Center of excellence for Medical Genetics, and Chulalongkorn University, the Ratchadapiseksomphot Endowment Fund of Chulalongkorn University (RES560530177-HR) and the Thailand Research Fund (RTA5680003, RSA5480022).

CONTENTS

Page

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Autosomal Dominant
ADID	Autosomal Dominant Intellectual Disability
AR	Autosomal Recessive
ARID	Autosomal Recessive Intellectual Disability
AS	*De novo* assembly method
ASD	Autism Spectrum Disorder
ASR	Allele Size Range
bp	Base pair
cDNA	Complementary DNA
CNP	Copy Number Polymorphic
CNV	Copy-number variation
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
ExAC	The Exome Aggregation Consortium
FECD	Fuchs' Endothelial Corneal Dystrophy
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
HGMD	The Human Gene Mutation Database
HS	Hydradenitis Suppurativa
IBD	Identity-by-descent
ID	Intellectual disability
IQ	Intelligence quotient
Kb	Kilobase
LoF	Loss of function
MAF.	Minor allele frequency
Mb	Million base

mRNA	Messenger RNA
NGS	Next Generation Sequencing
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
qRT-PCR	Quantitative real-time PCR
RC	Read count method
RNA	Ribonucleic acid
RP	Read-pair method
SGS	Shprintzen-Goldberg Syndrome
SNV	Single Nucleotide Variant
SNP	Single Nucleotide Polymorphism
SR	Split-read method
SV	Structural Variant
SVD	Singular Value Decomposition method
TD-PCR	Touchdown PCR
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

CHAPTER I
INTRODUCTION

Background and Rationale

Next generation sequencing or NGS is the new technology of sequencing. Since 2005, it was developed from the principle of shotgun sequencing that was used in the human genome project. The principle of NGS is “read the DNA templates randomly along the entire genome” (1). It is composed of several steps including template preparation, reading and imaging, and data analysis (2). The NGS overcomes the conventional method by its costs and volumes of sequence data. It can sequence 3 million base pairs of the whole human genome faster and cheaper than the conventional method up to 200 times (3, 4).

In addition, NGS can select the region of interest to sequence such as exons. Sequencing of all exons or Whole Exome Sequencing (WES) focuses on only protein coding sequences. Although WES captures only 1% of human genome to sequence, it is the most obvious useful application. WES can be used to identify genetic defects underlying single gene disorders. About 85% of the mutations with large effects have been detected in 1% of the sequence (5). Moreover, WES has more advantages than whole genome sequencing in manner of saving costs, increasing sequencing accuracy because of its deeper coverage, decreasing turnaround time and the outcome data adequately handled for functionally interpretable with bioinformatic analysis (5, 6).

Combination of WES with traditional methods such as linkage analysis is the approach that has the greatest impact for autosomal dominant inheritance disorder (4). However, the NGS has something to be considered. Firstly, NGS produces massive data, so it challenges for data storage, analyses, management and interpretation. Another consideration is about the base calling errors because its base calling error rates are greater than Sanger sequencing (1, 3, 7).

In this study, two different Mendelian disorders were included. The first disorder is a skin disease named “familial comedones syndrome” (8). We identified two families with this syndrome. The first family is a large one with 12 affected individuals with variable severity. The pedigree shows an autosomal dominant inheritance. All affected individuals have numerous comedones on face, neck, anterior thorax, posterior thorax, elbows and knees. Some pits are filled with black materials. Some affected individuals have inflammatory nodules and abscesses. The second family with familial comedones has two affected individuals. Only the proband and his mother were examined.

The second disease is an undiagnosed syndrome with intellectual disability (ID). The two affected individuals were siblings. Their parents were non-consanguineous. Besides intellectual disability, they had many other features such as skeletal dysplasia, arachnodactyly, and hyperlaxity. No other family members were affected. A possible mode of inheritance is autosomal recessive.

Until now, the causative gene and molecular mechanism for familial comedones and the undiagnosed syndrome with intellectual disability are unknown. Identification of the genetic defects underlying genetic diseases is the way to obtain new knowledge, give accurate diagnosis and lead to appropriate treatment and counseling. In addition, it could lead to a better understanding of the molecular pathology or the biology of disease resulting in an effective treatment and novel therapy.

Here, I proposed to use NGS technology as a tool to identify human disease genes in these two diseases. Two different strategies were used to analyze an autosomal dominant disorder and an autosomal recessive disorder.

Research questions

1. Which chromosomal region is linked to familial comedones in family I?
2. Which gene in the linked region is the causative gene for familial comedones?
3. Which gene is responsible for the undiagnosed syndrome with intellectual disability?

Objectives

1. To identify the genes responsible for the two different diseases using whole exome sequencing with different strategies
2. To identify the gene responsible for familial comedones using whole exome sequencing together with linkage analysis
3. To identify the gene responsible for the undiagnosed syndrome with intellectual disability using whole exome sequencing.

Hypotheses

1. There is/are a chromosomal region(s) linked to familial comedones in family I.
2. There is a gene responsible for familial comedones in the chromosomal region linked to familial comedones in family I. This gene might also underlie familial comedones in family II.
3. The undiagnosed syndrome with ID has a genetic cause which could be either CNVs or SNVs/indels.
4. The causative mutation(s) for the undiagnosed syndrome with ID will likely be compound heterozygous, homozygous, or *de novo (*gonadal mosaicism*)*.

Research design

Descriptive and *in vitro* studies

Key words

Next generation sequencing, Whole exome sequencing, Familial comedones, Intellectual disability, Mutations

Ethical consideration

The local Ethics Committee has approved this study. Written informed consent was obtained from all patients or their parents who participated in this study.

Expected benefit

1. To provide further understanding into the molecular basis of the two diseases and better insight into the diseases mechanism leading to more effective treatment of the disorders.
2. To provide accurate information and appropriate counseling for families with the diseases.

Research methodology

Familial comedones syndromes

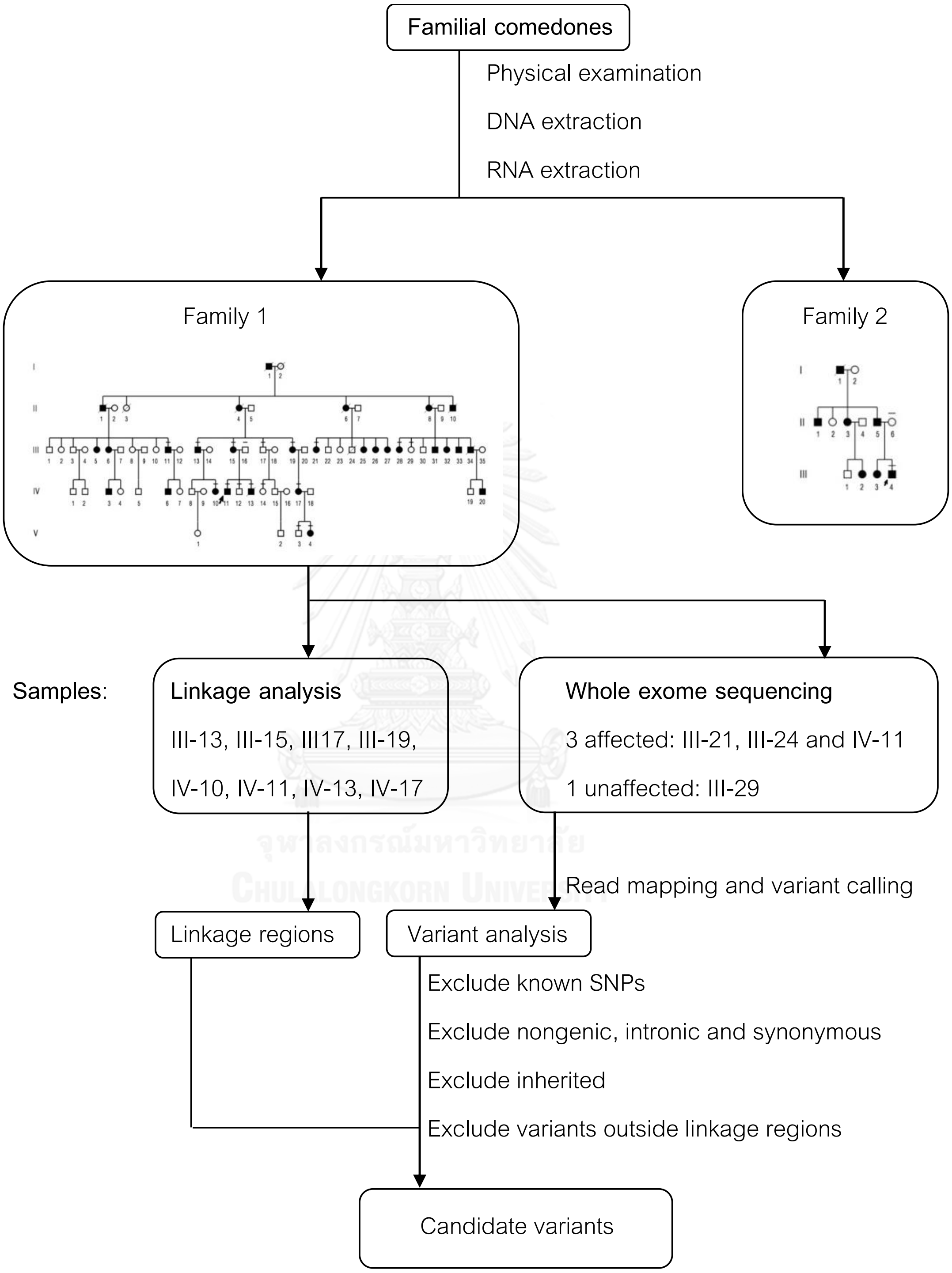
After physical examination and histological examination were obtained, probands of family I and II were diagnosed with familial comedones syndromes. DNA and RNA were extracted from leukocytes of family members from both families.

To identify a causative gene for familial comedones, combination of whole genome linkage analysis using SNPs array and whole exome sequencing were carried out using DNA samples from family members of family I. The candidate variants located in the identified linkage regions were selected for PCR and Sanger sequencing to confirm the existence of the mutation. To support pathogenicity of the mutation, cosegregation analysis was performed by sequencing all family members and also 100 unrelated Thai controls. Mutation analysis of the identified causative gene was performed in family II by sequencing of the entire coding sequences. cDNA of the proband was sequenced for the mutation. Quantitative RT-PCR using TaqMan expression assay was performed with RNA extracted from proband's leukocytes to measure mRNA expression levels of the causative gene.

Intellectual disability syndrome

We have excluded chromosomal abnormalities in the affected siblings by karyotyping. To identify a causative gene, whole exome sequencing of all four members in the family were carried out. CNVs and SNVs/indels can be detected by whole exome sequencing using different strategies. We first searched for SNVs/indels in the *FUCA1* and *SKI* genes which are known to cause syndromes that shared some phenotypes with the siblings. We then filtered for genes containing compound heterozygous, homozygous and also *de novo* variants shared in the siblings.

Conceptual framework



Familial comedones

Physical examination

DNA extraction

RNA extraction

Family 1

Family 2

Samples:

Linkage analysis

III-13, III-15, III17, III-19, IV-10, IV-11, IV-13, IV-17

Whole exome sequencing

3 affected: III-21, III-24 and IV-11

1 unaffected: III-29

Read mapping and variant calling

Linkage regions

Variant analysis

Exclude known SNPs

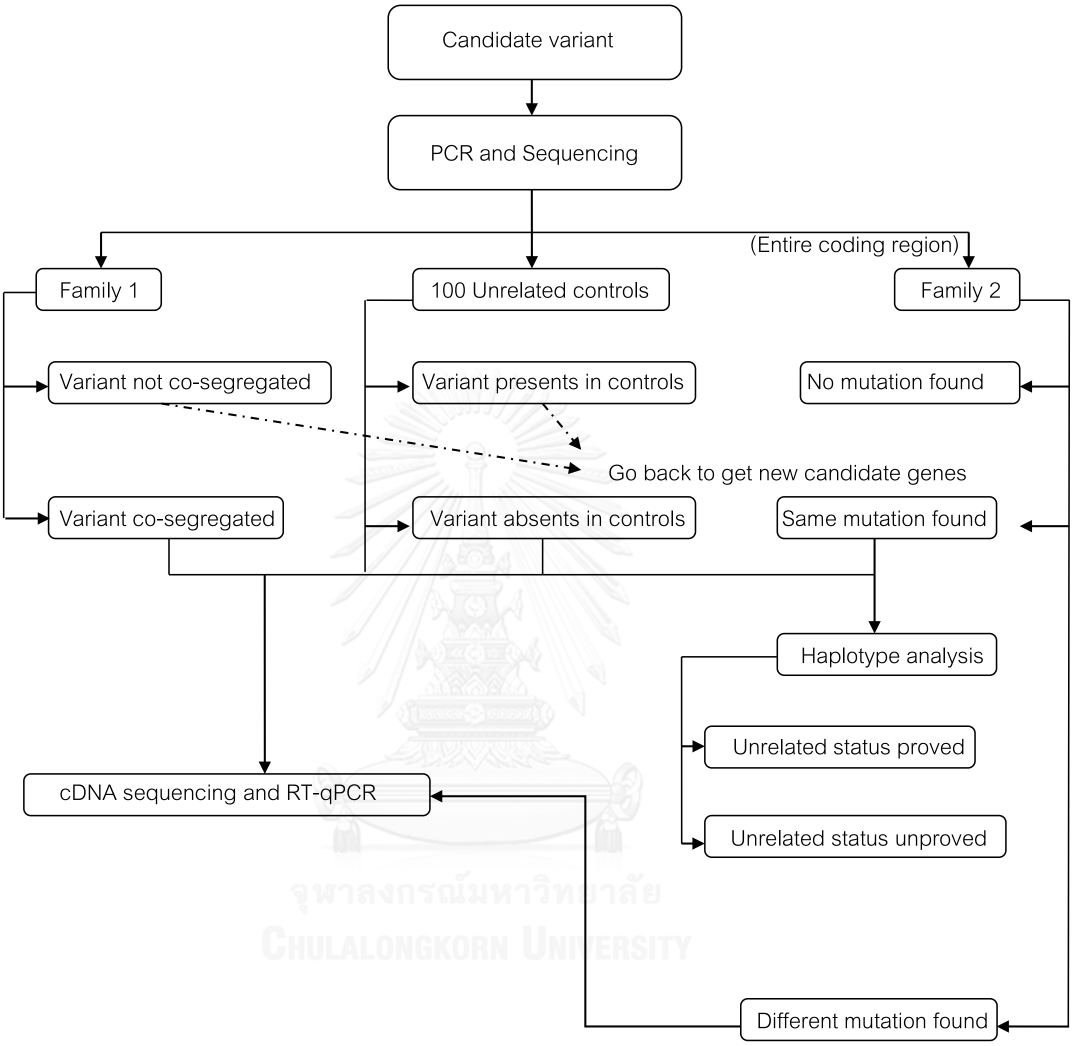
Exclude nongenic, intronic and synonymous

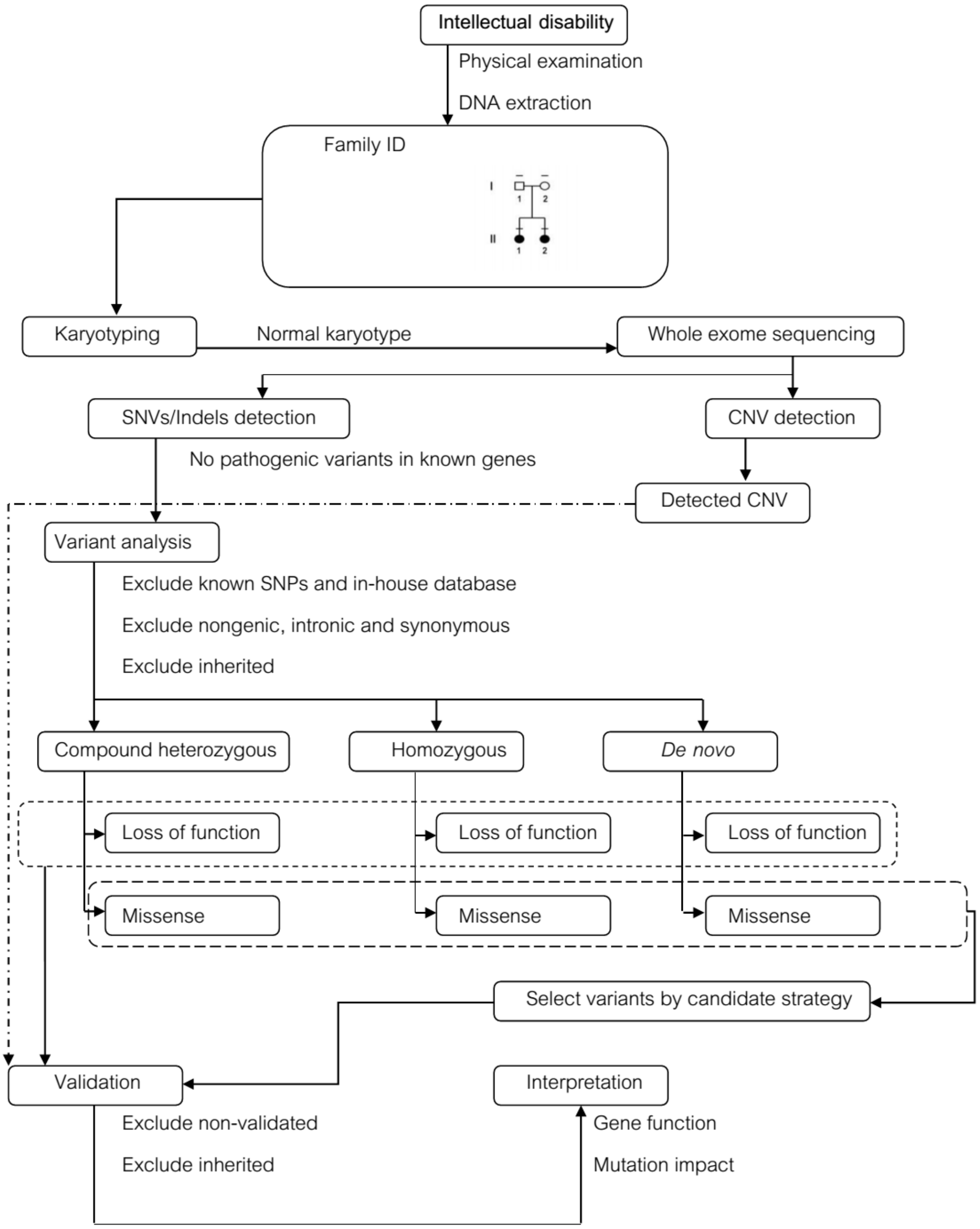
Exclude inherited

Exclude variants outside linkage regions

Candidate variants

-Continued-


Candidate variant
PCR and Sequencing
(Entire coding region)
Family 1
100 Unrelated controls
Family 2
Variant not co-segregated
Variant presents in controls
No mutation found
Go back to get new candidate genes
Variant co-segregated
Variant absents in controls
Same mutation found
Haplotype analysis
Unrelated status proved
Unrelated status unproved
cDNA sequencing and RT-qPCR
Different mutation found


Intellectual disability
Physical examination
DNA extraction
Family ID
I
1
2
II
1
2
Karyotyping
Normal karyotype
Whole exome sequencing
SNVs/Indels detection
CNV detection
No pathogenic variants in known genes
Detected CNV
Variant analysis
Exclude known SNPs and in-house database
Exclude nongenic, intronic and synonymous
Exclude inherited
Compound heterozygous
Homozygous
De novo
Loss of function
Loss of function
Loss of function
Missense
Missense
Missense
Select variants by candidate strategy
Validation
Interpretation
Exclude non-validated
Exclude inherited
Gene function
Mutation impact

CHAPTER II
REVIEW OF RELATED LITERATURE

Next generation sequencing (NGS)

Initial DNA sequencing methods were introduced by Sanger, Maxam and Gilbert in 1977 (9, 10). This sequencing technique has been developed and reached the capacity of 2 kilo base pairs (Kbps). Shotgun sequencing was developed to sequence longer DNA fragments during human genome project (1).

Massively parallel sequencing platforms called the ‘next-generation sequencing’ were introduced in 2004 and the first NGS instrument was on market in a year later (4). DNA templates were read randomly along the genome. There are many platforms for NGS such as 454 pyrosequence, Illumina sequence, and Solid sequence. All of them are based on the same steps including template preparation, sequencing and imaging, and data analysis. In the template preparation step, DNA templates are randomly broken into small fragmented DNA. Then the fragments are conjugated with adaptors that are specific with primers used for amplification. After clonal amplification, each with identical template will be sequenced. Every cycle of adding probes or nucleotides, it has the observed signal to image (2). The number of continuous sequence base or read length provides 50-500 (1). Then the NGS reads will be aligned to the reference sequence (2). Sequencing reads that can be assembled and matched with a reference sequence are called “mappable reads”. Since the read length from NGS is much shorter than that from Sanger sequencing, sufficient coverage is required for accurate assembly of the genomic sequence (1).

To date, these next generation sequencing technologies have been applied in a variety of contexts, such as whole genome sequencing (WGS), whole exome sequencing (WES), target sequencing, transcriptome sequencing, methylome, chromatin immunoprecipitation sequencing, and small RNA sequencing (4).

Whole-exome sequencing (WES)

Whole-exome sequencing is the targeted sequencing of the protein coding. Subset of the genome including exons, non-coding RNAs, highly conserved regions of the genome, disease-associated LD blocks or other regions of interest can be targeted for sequencing (5, 11). WES captures only about 1% of human genome to sequence. However it can be used to identify genetic defects underlying single gene disorders. About 85% of the mutations with large effects have been detected in the 1% sequence (5). Massively parallel coupled with targeted capture making whole exome sequencing have the more advantages than whole genome sequencing such as increasing sequence coverage of regions of interest, lower cost and the outcome data adequately handled for functionally interpretable with bioinformatic analysis (5, 11, 12).

Overview of WES is shown in figure 1. Genomic DNA is randomly sheared and coupled by adaptors to construct a shotgun library. Then the targeted library fragments with complementary adaptor are hybridized to biotinylated DNA or RNA baits. The hybridized fragments are captured by biotin-streptavidin-bases. The captured DNA fragments are amplified and massively parallel sequenced on the slide. The identity of the colored fluorescent indicator of each cluster is imaged, the fluorescent indicator is removed, and the cycle is repeated to generate a nucleotide sequence read. The captured DNA is sequenced in single-end or pair-end reads to yield 75-100 per read. After imaged sequencing, the sequence reads are aligned to a reference DNA sequence and a genotype call for each position is made (4, 11, 13).

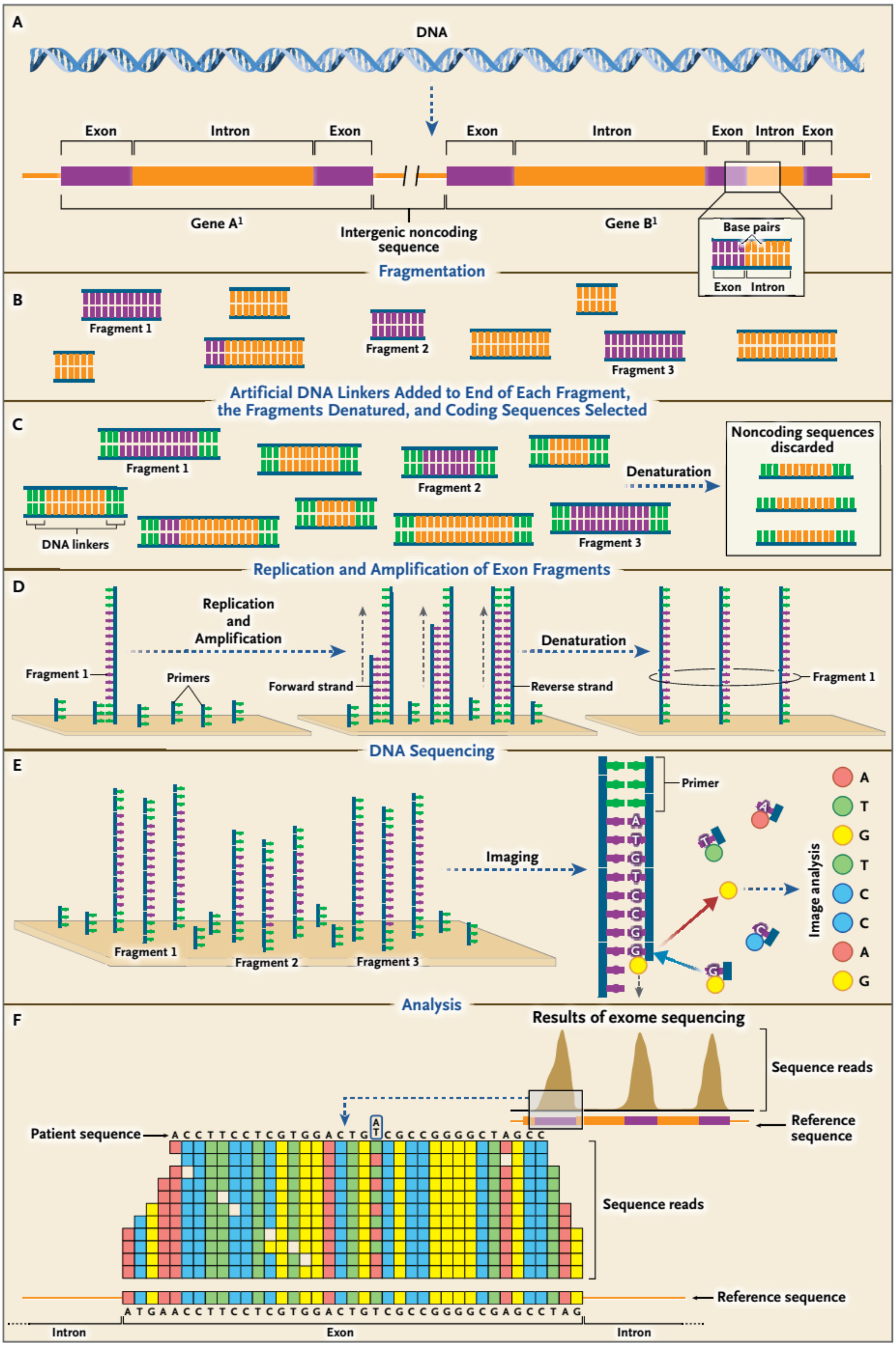


Figure 1: Schematic Overview of Exome Sequencing (13)

WES and identification of genes underlying Mendelian phenotypes

In 2010, Sarah B Ng *et al*. first demonstrated the success of using exome sequencing to identify the cause of a Mendelian disorder inherited in an autosomal recessive manner. They discovered the gene for Miller syndrome by exome sequencing of four affected individuals in three independent kindreds (12). Subsequently exome sequencing was first performed to identify the causative gene for autosomal dominant disorders, Schinzel-Giedion syndrome and Kabuki syndrome (14, 15). NGS technologies have rapidly accelerated the pace of discovery of genes underlying Mendelian phenotype. The pace of gene discovery has increased from an average of ~166 per year between 2005 and 2009 (five years before NGS technology was introduced) to 236 per year between 2010 and 2014. Interestingly, between January 2010 and February 2015, ~613 genes associated with monogenic Mendelian phenotypes were discovered via next-generation sequencing approaches. Since 2003, the number of gene discoveries by WGS/WES has reached almost 3 times comparing with conventional approaches (figure 2) (16).

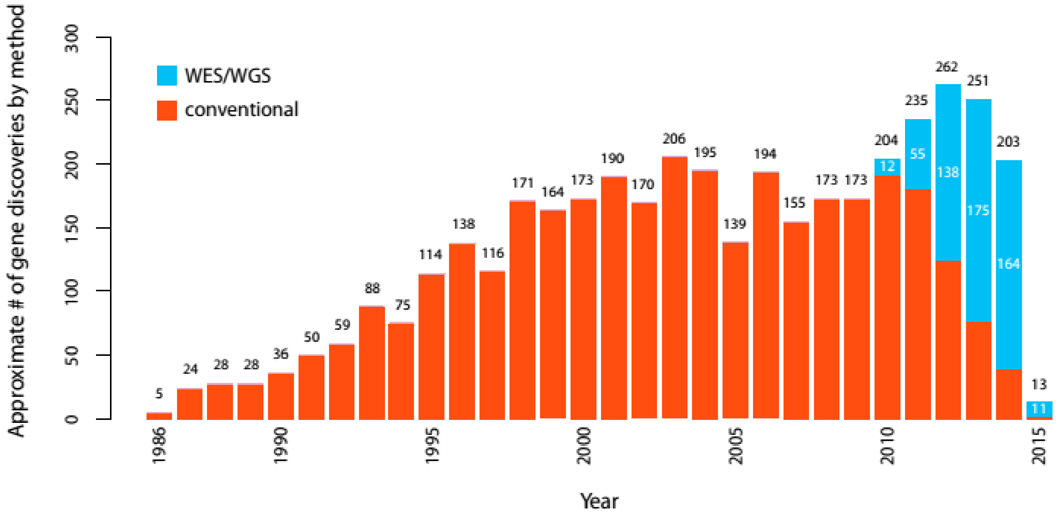
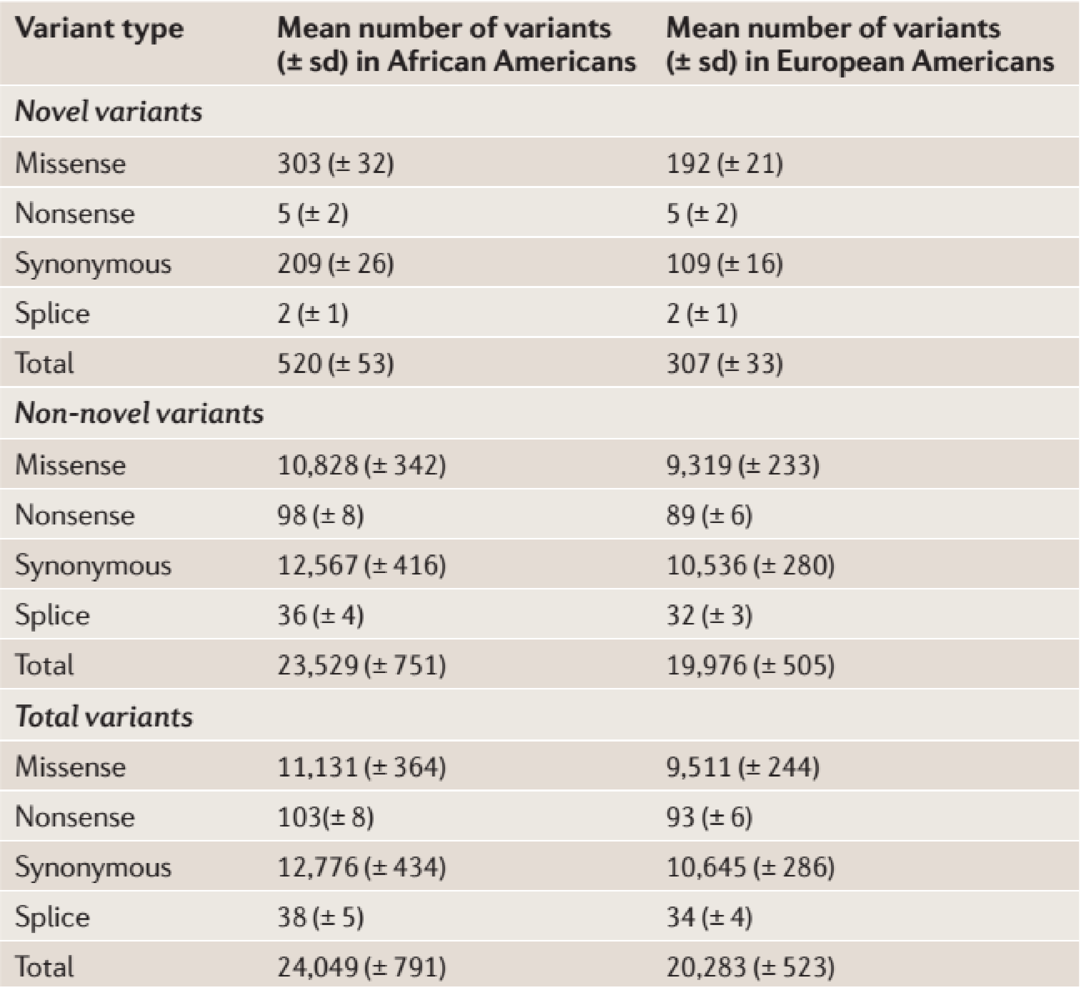


Figure 2: Approximate number of gene discoveries made by WGS and WES versus conventional approaches since 2010 (16)

Disease gene identification strategies for whole exome sequencing

The number of variants detected by exome varies between 20,000-50,000 variants per exome depending on exome enrichment set, sequencing platform and the algorithms used for mapping, and variant calling (17). The mean number of coding variants from 100 sampled African Americans and 100 Europeans are shown in Table 1 revealing an average of single nucleotide variants (SNVs) ~24,000 in African American samples and ~20,000 in European American samples. Novel variants refer to those absent in dbSNPs131 or other 200 controls (11). Identifying the pathogenic mutation amongst these variants is a major challenge, and novel variant prioritization strategies are required.

Table 1: The mean number of coding variants in two populations (11)



Variant type	Mean number of variants (± sd) in African Americans	Mean number of variants (± sd) in European Americans
Novel variants		
Missense	303 (± 32)	192 (± 21)
Nonsense	5 (± 2)	5 (± 2)
Synonymous	209 (± 26)	109 (± 16)
Splice	2 (± 1)	2 (± 1)
Total	520 (± 53)	307 (± 33)
Non-novel variants		
Missense	10,828 (± 342)	9,319 (± 233)
Nonsense	98 (± 8)	89 (± 6)
Synonymous	12,567 (± 416)	10,536 (± 280)
Splice	36 (± 4)	32 (± 3)
Total	23,529 (± 751)	19,976 (± 505)
Total variants		
Missense	11,131 (± 364)	9,511 (± 244)
Nonsense	103(± 8)	93 (± 6)
Synonymous	12,776 (± 434)	10,645 (± 286)
Splice	38 (± 5)	34 (± 4)
Total	24,049 (± 791)	20,283 (± 523)

To discover novel disease genes, different filtering strategies are required to isolate pathogenic mutations. The following assumptions are made about the causal mutations underlying monogenic, Mendelian disease: 1) a single mutation is sufficient to cause the disease 2) which would be rare and 3) they are most likely located in the coding region since they cause large effect and 4) highly penetrance (18). To reduce the number of the variants and prioritize potential pathogenic variants, variants are filtered with initial prioritization including false-positive removal based on quality criteria (eg, at least five independent reads and the percentage of reads; at least 20% for heterozygous variants, at least 80% for homozygous variants), excluding variants outside coding regions as well as synonymous coding variants and excluding known variants from dbSNP, published studies, or in-house databases (17).

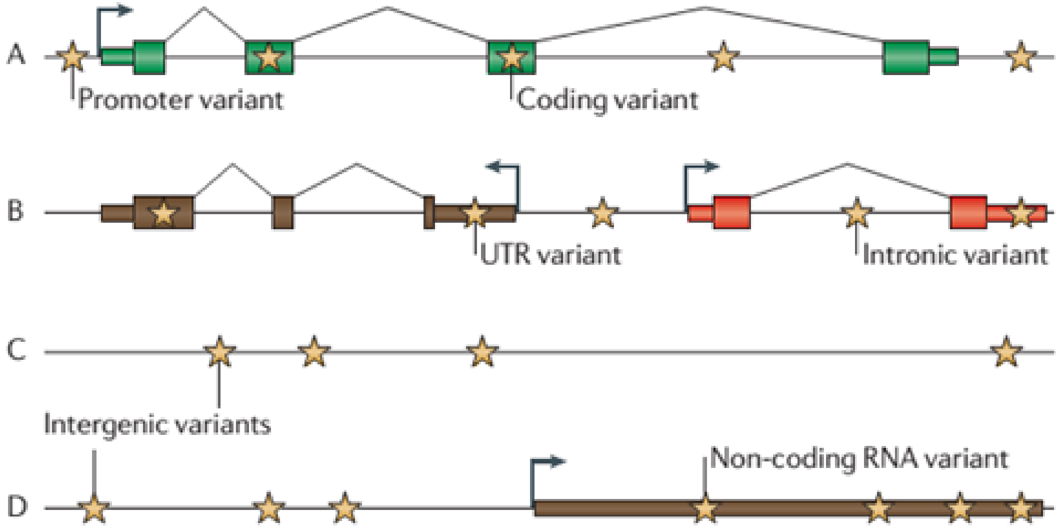


Figure 3: Variants of various functional classes (19)

Additional strategies depending on factors such as the mode of inheritance of a trait, the pedigree or population structure, whether a phenotype arises owing to *de novo* or inherited variants, and the extent of locus heterogeneity for a trait, are needed to find the causative variant among the remaining variants from the initial prioritization step (11, 17). Figure 4 shows 6 strategies to identify disease-causing genes using whole exome sequencing briefly explained by Gilissen *C. et al.*

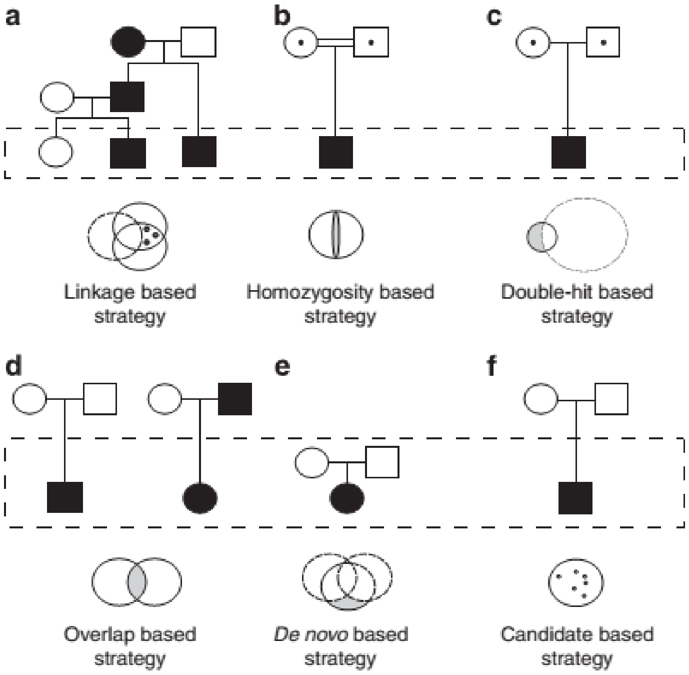


Figure 4: Disease gene identification strategies for exome sequencing (17)

a) Linkage strategy is suitable for a family with a monogenic inherited disorder. Multiple affected family members can be sequenced to identify shared variation together with sequencing of non-affected family members to exclude private benign variation.

b) Homozygosity strategy is good for a rare recessively inherited disorder and suspected consanguinity. The disease is caused by a homozygous variant inherited from both parents and that this variant resides within a large stretch of a homozygous region are assumed. Homozygous variants can therefore be prioritized by their presence in large homozygous regions of the patient's genome.

c) Double-hit strategy works for the disorder which is suspected to be recessively inherited without any indication for consanguinity. If only single patient is available, sequencing only a single patient is possible to identify causative mutations by selecting for genes carrying homozygous as well as compound heterozygous variants.

d) Overlap strategy works for rare diseases that are largely monogenic. Mutations can be identified by sequencing of multiple unrelated patients with a similar phenotype and search for mutations in a single gene.

e) *De novo* strategy is more efficient for common disorders that are genetically highly heterogeneous. Especially when a disorder occurs mostly sporadic and is associated with reduced fitness. *De novo* mutations can be identified by sequencing the patient as well as the parents. All inherited variants will then be filtered out.

f) Candidate strategy is used in case of a single dominantly affected individual, without further availability of family members or other affected individuals. Prioritization is based on the predicted impact of the variant on protein function and structure as well as evolutionary conservation of the variant nucleotide.

In addition, whole exome sequencing combined with linkage analysis or homozygosity mapping is the powerful approach for defining the candidate variants for autosomal dominant and autosomal recessive disorders, respectively (4).

Copy number variation detection from next-generation sequencing data

Copy-number variant (CNV) is a form of structural variations which are defined as genomic rearrangements affecting more than 50 base pairs (20). CNV includes insertions, deletions and duplications (21).

A number of approaches have been proposed to discover structural variants (SVs) from next-generation sequencing data. Four strategies for detection of SVs were summarized in figure 5 (20).

Read-depth (or read count, RC) approaches: With an assumption of a random (Poisson or modified Poisson) distribution in mapping depth, duplications and deletions are highlighted from the divergence of this distribution. Sequencing of duplicated/amplified regions shows higher read depth while deleted regions show reduced read depth when compared to normal regions.

Read-pair (RP) methods: Discordance of the span and/or orientation of the read pairs which are inconsistent with the expected insert size can identify several classes of SVs. Deletions and insertions show read pairs mapping too far apart or closer than expected. Orientation inconsistencies indicate inversions and a specific class of tandem duplications.

Split-read (SR) methods: The SR methods can detect SVs with single base-pair resolution. A split sequence-read signature breaking the alignment to the reference demonstrates the presence of a SV breakpoint. A gap in the read indicates a deletion while insertions show stretches in the reference.

De novo assembly (AS) methods: The *de novo* assembly (AS) methods based on merging and ordering short fragments to reassemble the original sequence from which the short fragments are sampled. These methods can investigate all forms of SVs.

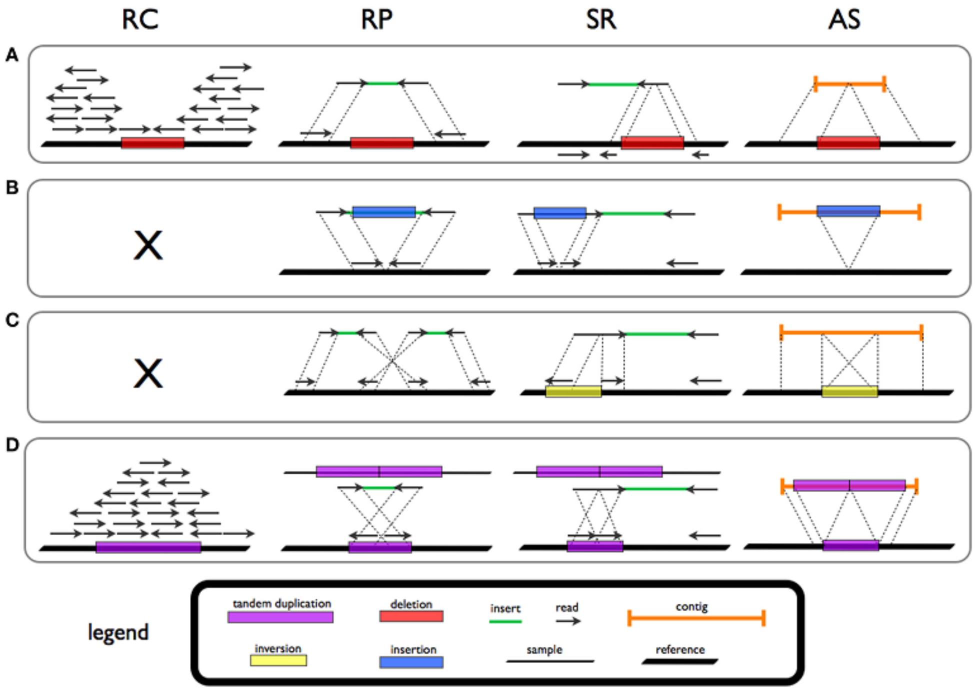


Figure 5: Approaches for detection of structural variations (SVs); RC: read count, RP: read-pair, SR: split-read and AS: *de novo* assembly, signature and pattern of SVs for (A) deletion, (B) novel sequence insertion, (C) inversion and (D) tandem duplication (20).

Besides common single-nucleotide polymorphisms (SNPs), rare sequence variants, and short indels, whole exome sequencing can detect copy number variants (CNVs). Some software has been proposed to identify SVs using WES data mostly based on read count (RC) methods.

With read count method, Krumm *et al.* developed CoNIFER software by combining read-depth data from exome sequencing with singular value decomposition (SVD) methods to discover rare CNVs and genotype known copy number polymorphic (CNP) (Figure 6). CoNIFER first calculates the number of sequencing reads those align to exons and then calculates a normalized RPKM value. Each RPKM value is normalized by median and standard deviation across the analyzed population. A minimum of eight exomes are recommended, but more exomes will improve the quality of the results. The Z-RPKM values are inputted into the SVD transformation, which removes systematic bias. The final SVD-ZRPKM signal is then smoothed and the duplication/deletion breakpoints are found using a threshold algorithm (22).

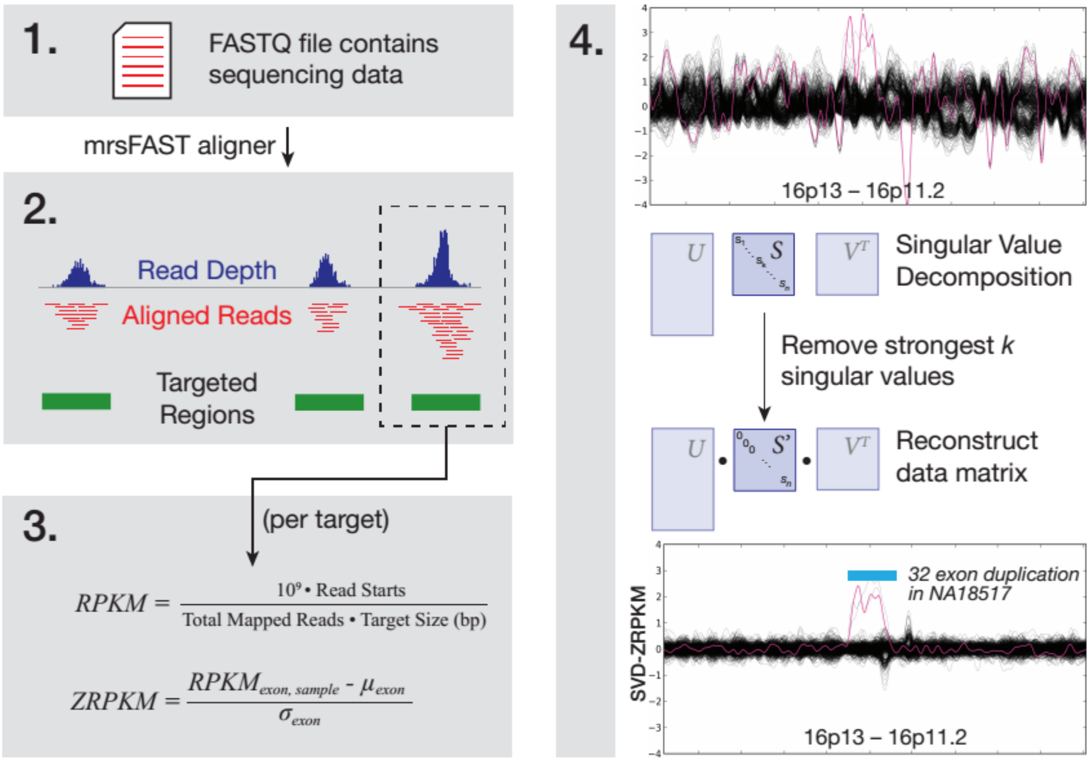

1.
FASTQ file contains sequencing data
mrsFAST aligner
2.
Read Depth
Aligned Reads
Targeted Regions
3.
(per target)
$RPKM = \frac{10^9 \cdot \text{Read Starts}}{\text{Total Mapped Reads} \cdot \text{Target Size (bp)}}$
$ZRPKM = \frac{RPKM_{exon, sample} - \mu_{exon}}{\sigma_{exon}}$
4.
16p13 – 16p11.2
U
S
V^T
Singular Value Decomposition
Remove strongest k singular values
S'
Reconstruct data matrix
SVD-ZRPKM
32 exon duplication in NA18517
16p13 – 16p11.2

Figure 6: Method overview of CNV discovery by CoNIFER software with the read depth approach (22)

Familial comedones

Familial comedones is a skin disorder with autosomal dominant mode of inheritance. The prevalence of familial comedones is very rare. Only three studies have reported 32 cases of familial comedones without dyskeratosis. Since the Mendelian trait is very rare, it doesn't even have an entry in Online Mendelian Inheritance in Man (OMIM). Rodin *et al*. first described the disease in 1967. They identified a 42 year-old Negro-Irish-Indian woman affected with diffuse familial comedones. They noted that most members of her family also had the same condition (23). Cantú *et al.* proposed that familial comedones inherited in an autosomal dominant mode with conclusive evidence. They reported a large family with multiple comedones including 16 affected members. The age of onset, distribution, and severity of the comedones were varied in the family. Greater severity was found in male family members and had direct relationship to age (8). In

2014, the new term “familial disseminated comedones without dyskeratosis” was proposed by Cheng *et al.* to underline its familial and disseminated characteristics. They also described a family of disseminated non-dyskeratotic comedones with 15 affected family members with different degrees of severity (24).

Familial comedones is characterized by the occurrence of numerous comedones. They are distributed all over the body, especially on the face, neck, trunk and forearms. Multiple severe nodulocystic lesions with scar formation may be observed. Some comedones become pustular/acneiform lesions during adolescence and early adulthood. Removal of the comedones can cause pit-like scars (Figure 7). Familial comedones appear to be more common in males. Males are usually more severely affected than females. The histopathological examination of the lesions showed arborescent branching patterns of basaloid cells in the external sheath of the hair follicles without acantholytic dyskeratosis (Figure 8). The genetic defect and pathogenesis of this disease remain unknown. Patients with familial comedones syndrome have not been noted to have squamous cell carcinoma (8, 23-25).

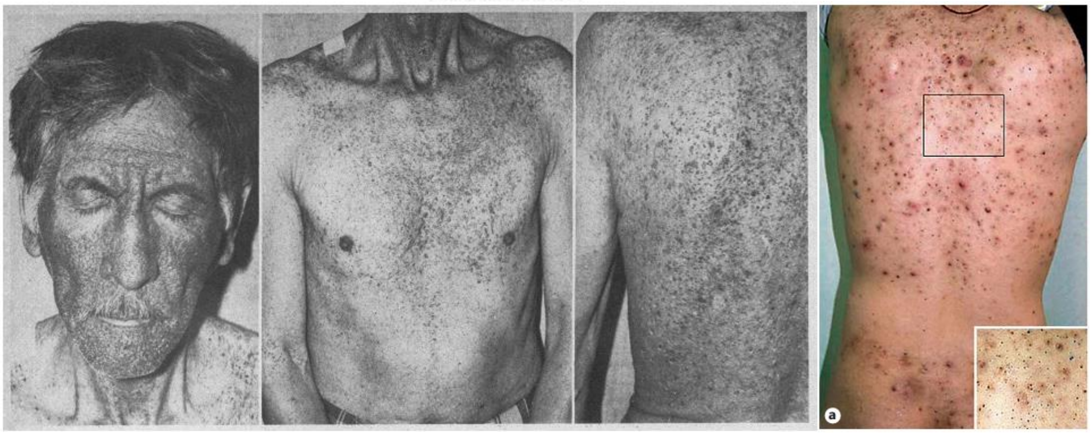


Figure 7: Clinical phenotypes of familial comedones; numerous comedones, comedo-like papules and cystic, nodular acne (8, 24)

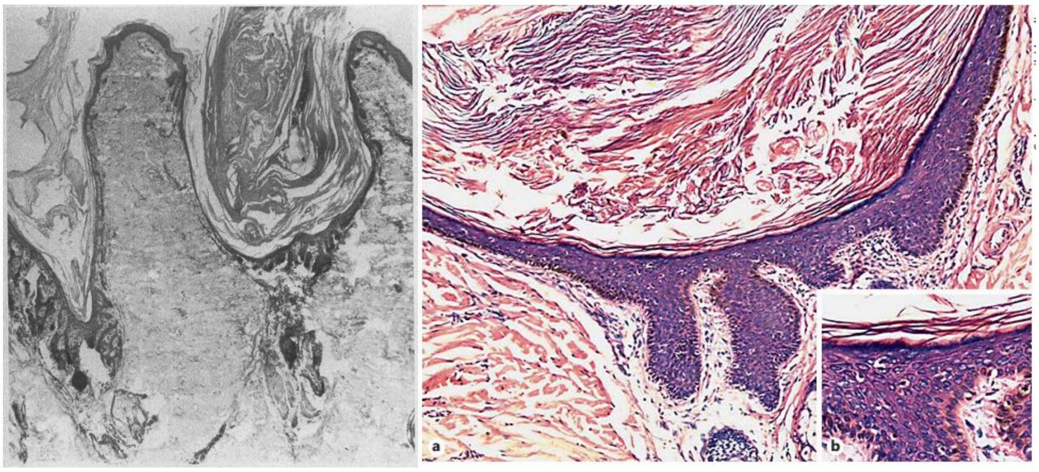


Figure 8: Arborescent branching patterns of basaloid cells in the external sheath of the hair follicles (left panel: hematoxylin-eosin, original magnification x65, right panel: a) hematoxylin-eosin, original magnification ×200, b) hematoxylin-eosin, original magnification x400) (8, 24)

Hydradenitis Suppurativa (HS)

Hidradenitis suppurativa also known as acne inversa (OMIM#142690), is a member of follicular occlusion disease manifested with chronic inflammation of hair follicles. The term hidradenitis suppurativa is noted to be a misnomer because it is considered to be a disease of apocrine glands. HS is characterized by recurrent draining sinuses and abscesses predominantly in apocrine gland-bearing areas leading to disfiguring scars. Symptoms typically present after puberty with an average age of onset 21-23 years old (26-28). HS has been associated with multiple comorbidities e.g. Crohn’s disease, thyroid diseases, acne conglobata and rarely associated with squamous cell carcinoma. Prevalence of HS varies between 1%-4%, the studies mostly conducted in Europe and the United States. HS seems to be more common in woman with range of 2.6:1-3.3:1 female to male ratios (29). Unlike familial comedones, many factors are associated with HS. For example, a hormonal influence is indicated by the female predominance, post puberty onset, frequent premenstrual flares and the improvement often observed during pregnancy and post menopause. Smoking and obesity are also clearly associated with HS (28).

Hidradenitis suppurativa has autosomal dominant inheritance with 100% penetrance (30). In 2006, Gao *et al.* identified a linkage locus on chromosome 1, 1p21.1-1q25.3 (31). Four years later, Wang *et al.* first identified γ *-secretase* gene mutations in six Han Chinese families affected with HS. By combining genome-wide linkage scan and haplotype analysis methods in two large families, they mapped a HS locus on chromosome 19q13. They then performed sequencing analysis of the *PSENEN* gene located in the region. Two frameshift mutations in the *PSENEN* gene were revealed; family 1: c.66delG (p.F23LfsX46) and family 2: c.279delC (p.F94SfsX51). Mutations in other γ *-secretase* genes including *NCSTN* and *PSEN1* were identified in family 3-6 (32). The *NCSTN* gene is located in the previously reported region on chromosome 1(31). To date, 19 mutations in the *NCSTN* gene, 3 mutations in the *PSENEN* gene and 1 mutation in the *PSEN1* gene have been reported (32-41). Most of the identified mutations lead to loss of function of components of the γ -secretase complex.

The γ-secretase complex composed of *PSENEN, PSEN1*, *NCSTN* and *APH1*, is an intramembranous protease complex cleaving transmembrane proteins including amyloid precursor protein, Notch receptors, N-cadherin, and E-cadherin (28). Besides HS, γ*-secretase* genes mutations have been identified in a variety of disorders including familial Alzheimer's disease, dilated cardiomyopathy, breast cancer and leukemia (42).

Intellectual disability (ID)

Intellectual disability (ID) is the current term of mental retardation. The American Association of Intellectual and Developmental Disability (AAIDD) has defined ID as "a disability characterized by significant limitations both in intellectual functioning and in adaptive behavior, which covers many everyday social and practical skills. This disability originates before the age of 18" (43). A limitation in intellectual functioning is commonly characterized by an intelligence quotient (IQ) below 70 (43). Intellectual disability can be found with other clinical findings such as malformations, neurological signs, impairment of the special senses, seizures and behavioral disturbances as part of a syndrome or may occur as an isolated phenomenon. Syndromic and non-syndromic forms of intellectual disability are altogether estimated to affect 1%–3% of the population in Western societies (44).

Although non-genetic factors such as infection, trauma, and teratogens can cause ID, most severe forms of ID with a sizable proportion are caused by genetic abnormalities. The genetic basis of intellectual disability is extremely heterogeneous. Underlying mutations are ranging from large cytogenetic abnormalities to point mutations and also epigenetic alterations. Traditional chromosome analysis revealed about 15% of ID cases with cytogenetically abnormalities (aneuploidies, gross deletions, inversions and rearrangements).. Smaller genomic alterations, submicroscopic copy number variants (CNVs) detected by microarrays cause about 15-20% of cases. Point mutations and small insertions or deletions (indels) explain another ~10% of cases mostly in X-linked forms .However, the etiology of about 60% of ID cases remains unknown (43, 45).

Whole exome sequencing approach has proven to be successful in identifying underlying mutations and strong candidates for ID. The underlying genes of sporadic, syndromic condition of intellectual disability, e.g. Schinzel-Giedion syndrome and Kabuki syndrome were identified by performing whole exome sequencing. They grouped unrelated patients with the same phenotypes. They hypothesized that the causative gene would harbor in all or most of the exome (14, 15, 45). For sporadic, non-syndromic cases of intellectual disability, Vissers *et al.* demonstrated using WES with the approach of trio analysis to identify mutations underlying ID without family history. They performed whole exome sequencing of 10 trios of individuals with moderate to severe ID without family history of ID and their parents. With *de novo* mutation hypothesis, unique non-synonymous *de novo* mutations in nine genes were identified. Two of them were known to cause X-linked ID and autosomal dominant ID. Four of the remaining genes were found to be involved in the disease by some supporting evidence such as gene function, evolutionary conservation and mutation impact (46).

Although many cases of intellectual disability are sporadic, there are families with multiple affected children. Some examples of successful exome sequencing studies have identified familial cases of ID with or without history of consanguinity. Figueiredo *et. al.* recently identified a causal mutation in a large consanguineous family with nine adult members affected by severe ID associated with disruptive behavior. They performed genome-wide human SNP array to determine regions of homozygosity by descent and whole exome sequencing. The identified mutation was a 5-bp duplication in the *IMPA1* gene cosegregated with the disease in 26 genotyped family members (47). For a non-consanguineous family, Krawitz *et. al.* identified a gene responsible for hyperphosphatasia mental retardation (HPMR) syndrome also known as Marby syndrome, an autosomal recessive form of intellectual disability. They performed WES in three siblings of a non-consanguineous family. They hypothesized that all affected children in this family would have inherited identical combination of maternal and paternal haplotypes (identity-by-descent (IBD) = 2 at the disease locus). They have developed an algorithm based on a Hidden Markov Model (HMM) to identify chromosomal regions with IBD = 2. They identified two candidate genes carrying rare variants in all three affected children within these regions. The *PIGV* gene, one of those genes was highly associated with the disease. Subsequently homozygous and compound heterozygous mutations were detected in the *PIGV* gene in individuals from three additional families (48).

Although more than 450 genes contributing to ID have been identified, many more ID genes remain to be identified. Several identified ID genes work together in the same pathway or complex (49). The identification of interacting partners of ID proteins and their targets has enlightened the specific pathways linked to ID. General molecular and cellular mechanisms including neurogenesis, neuronal migration, synaptic functions, and transcription and translation can be recognized as the underlying pathophysiology of ID (50).

Fucosidosis

Fucosidosis (OMIM #230000) is a rare autosomal recessive lysosomal storage disorder first described in 1966 by Durand *et al*. (51) The disease is caused by deficiency of alpha-L-fucosidase (EC3.2.1.51) activity leading to accumulation of fucose-containing glycolipids and glycoproteins in various tissues that affects many areas of the body (52, 53). Willems *et al.* reviewed 77 patients with fucosidosis and revealed that clinical phenotypes mainly consist of progressive intellectual disability (95%), motor deterioration (87%), coarse facies (79%), growth retardation (78%), recurrent infections (78%), dysostosis multiplex (58%), angiokeratoma corporis diffusum (52%), visceromegaly (44%), and seizures (38%). The severe type usually affects patients within the first year of life (54).

Alpha-L-fucosidase is an enzyme which is involved in one of the early steps of glycoprotein and glycolipid breakdown in lysosomes encoded by the *FUCA1* gene. So far, 28 mutations underlying the disease in the *FUCA1* have been reported for fewer than 100 fucosidosis cases in HGMD (the Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac/gene.php?gene=FUCA1). Most mutations are nonsense mutations resulting from either point mutations or deletions that lead to nearly absent enzymatic activity and severely deficient cross-reacting immunomaterial (CRIM) (55).

Shprintzen-Goldberg syndrome (SGS)

Shprintzen-Goldberg syndrome (OMIM #182212) is a rare disorder characterized by intellectual disability, craniosynostosis, distinctive craniofacial features (including hypertelorism, exophthalmos, downslanting palpebral fissures, and maxillary and mandibular hypoplasia), Marfan-like habitus (including dolichostenomelia, arachnodactyly, pectus deformity, scoliosis, and pes planus with foot deformity), camptodactyly, severe skeletal muscle hypotonia and cardiac abnormalities (56-58).

SGS has phenotypic overlapping with Marfan syndrome (OMIM #154700) and Loeys-Dietz syndrome (OMIM #609192/610168). Skeletal and craniofacial phenotypes are similar in all three syndromes. Similar to SGS, Loeys-Dietz syndrome may show craniosynostosis and also have a more severe cardiovascular phenotype. Even though individuals with Loeys-Dietz occasionally have developmental delay, intelligence is normal in both Marfan syndrome and Loeys-Dietz syndrome. Increasing of transforming growth factor beta (TGF-β) signaling has been found to be involved in the pathogenesis of aortic aneurysm in Marfan syndrome and Loeys-Dietz syndrome (58).

Genetic basis of SGS is also associated with TGF-β signaling. In 2012, Carmignac *et al.* and Doyle *et al.* identified mutations in the *SKI* gene which is a known regulator of TGF-β signaling as the underlying gene for the majority of Shprintzen-Goldberg cases. Doyle *et al.* described 10 of 11 Shprintzen-Goldberg cases with confirmed *SKI* mutations, including a recurrent mutation in 2 unrelated probands. They performed whole exome sequencing for an affected SGS child and unaffected parents. With trio analysis for a *de novo* mutation, a heterozygous missense change in exon 1 of the *SKI* gene, c.347G>A (p.Gly116Glu), was identified. The variant was absent in parents, and was a strong functional candidate based upon a described relationship to TGF-β signaling. They subsequently performed sequencing for the *SKI* gene in 11 other sporadic cases and revealed 9 heterozygous mutations including 8 missense mutations and one base pair deletion. These mutations were not present in SNP database, the 1000 genomes project and over 10,000 exomes reported on the NHLBI Exome Variant Server. Carmignac et al. also described 18 of 19 SGS cases with *SKI* mutations. They performed whole exome sequencing of 11 individuals including 3 trios and other 2 cases from 4 families. A 12 bp deletion, c.280_291delTCCGACCGCTCC (p.Ser94_Ser97del) in a highly conserved region of exon1 of *SKI* was found in family 3. They also detected a missense mutation in the *SKI* gene, c.101G>T (p.Gly34Val) in family 4. They then sequenced the *SKI* gene for their entire cohort of SGS individuals. A total of ten *de novo* missense mutations, including somatic mosaicism in a family with recurrence in siblings and two overlapping in-frame deletions (one of them was dominantly inherited in a large family) were identified(57). Au *et al.* also described another two cases with 2 mutations in the *SKI* gene which were previously described by Doyle *et al.(58).* Schepers *et al.* recently identified 8 recurrent and 3 novel *SKI* mutations in 11 SGS cases. Together with their findings, a mutational hotspot was clearly revealed from p.Ser31 to p.Pro35 of SKI residues with 73% (24 out of 33) (59). All identified mutations occurred within exon 1 mostly located in the R-SMAD binding domain and some in the DHD domain. All these findings suggest that the TGF-β signaling pathway is important in the pathogenesis of SGS (57-59).

CHAPTER III
MATERIALS AND METHODS

Familial comedones syndrome

Subjects and clinical descriptions

We described two unrelated Thai families with familial comedones with an interesting expanded phenotypic spectrum. We examined 17 family members from Family I (Figure 9). Of the 11 individuals affected with the disease, only 2 individuals developed into squamous cell carcinoma (IV-11 and IV13).

The proband from Family I (IV-11) was a 33-year-old male. His physical examination showed a large number of open comedones and skin pits covering the majority of his body except the palms and soles. Several recurring abscesses presented on his face, neck and trunk (Figure 10). He was also diagnosed with squamous cell carcinoma. The histological examination from his skin pit revealed a branching pattern of invaginated epidermis. The pit-like lesions first appeared when he was eight years old. The numerous comedones, discharging nodules and large tender abscesses first presented when he was fifteen.

The second proband from Family II (III-4) was a 16-year-old male. His physical examination showed multiple comedones and crater-like pits on his face, upper back and chest. Several purulent lesions, inflammatory nodules, and deep-seated abscesses were detected on his back, intergluteal space and axillary area (Figure 11). The histological examination from his comedonal lesion revealed a dilated hair follicle with a branching pattern projection of basaloid cells. Only the proband and his unaffected mother were recruited for the study (Figure 9, Family II).

After informed consent was obtained, 3 mL EDTA blood samples were collected from the subjects. The genomic DNA and RNA were subsequently extracted following the manufacturer's instructions for the DNA/RNA extraction kits (Qiagen Inc, Valencia, CA).

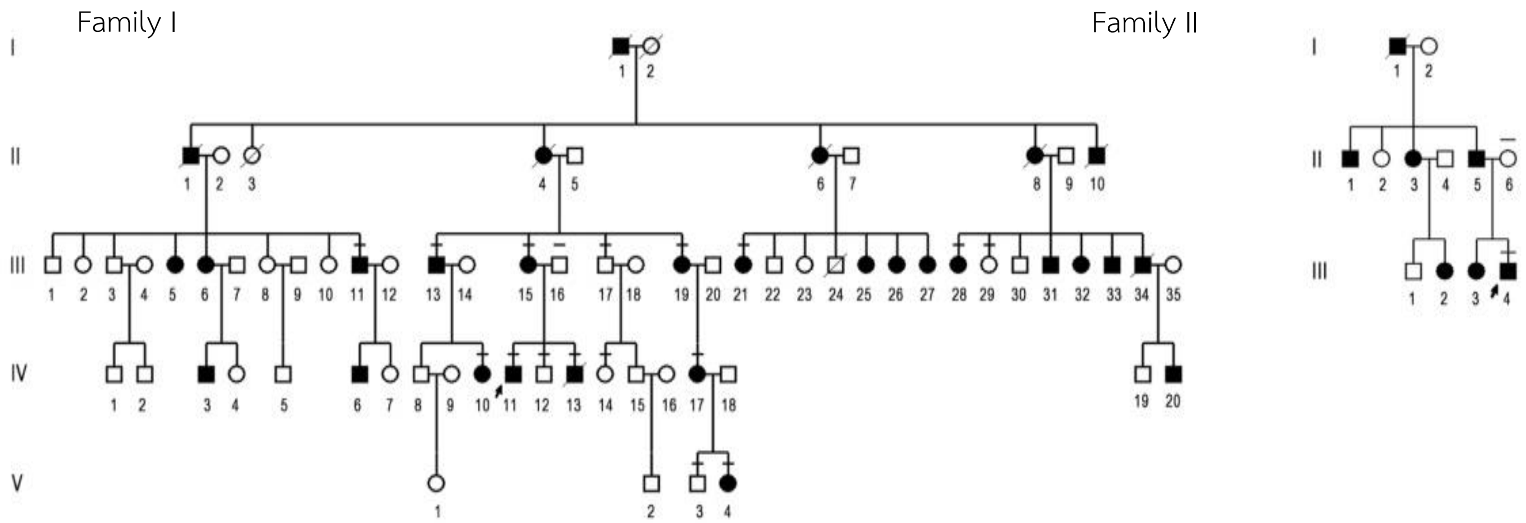


Figure 9: Pedigree of two unrelated Thai families with familial comedones. Probands are indicated by arrows. Squares indicate male subjects, while circles indicate female subjects. The affected individuals are represented by filled symbols. A bar denotes subjects who were clinically examined and had blood collected. A slash through the symbol indicates that the subject is deceased.

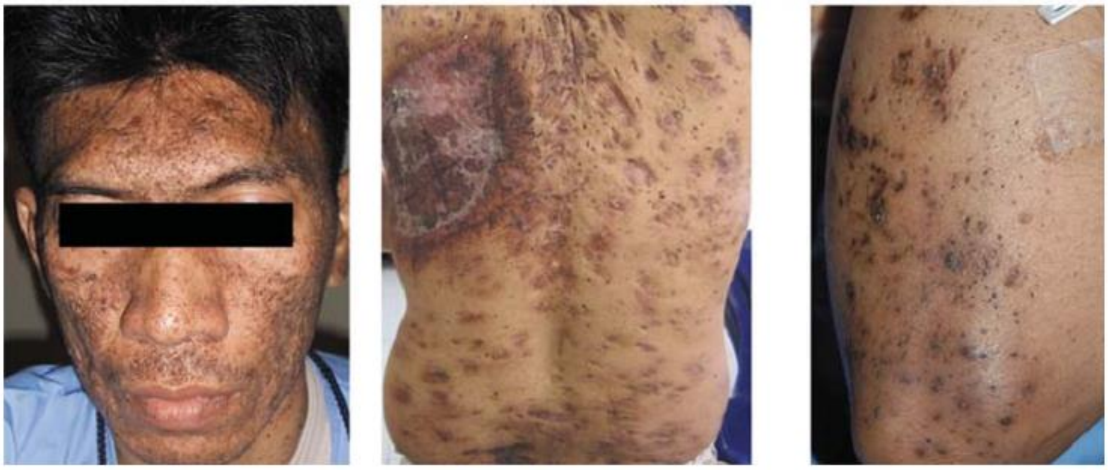


Figure 10: Clinical manifestation of the proband (IV-11) from Family I

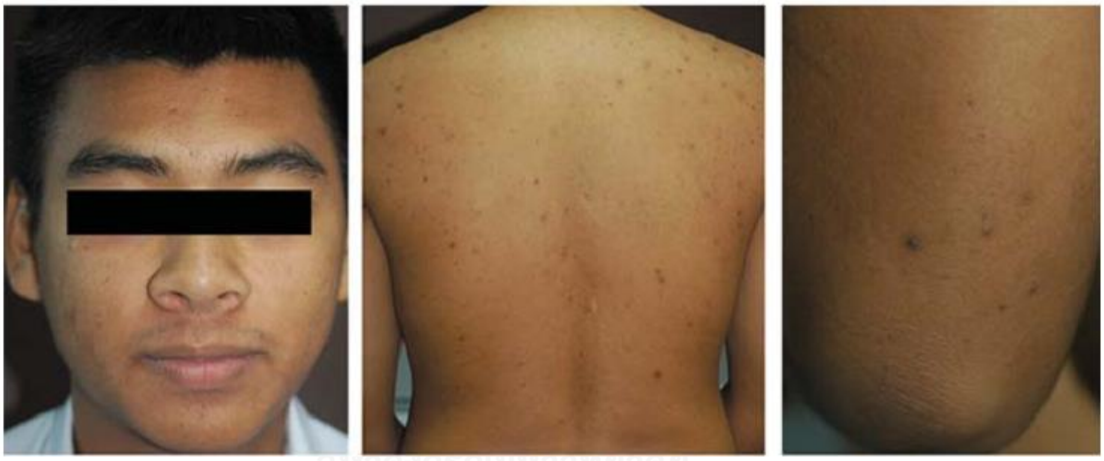


Figure 11: Clinical manifestation of the proband (III-4) from Family II

Whole genome linkage analysis

To identify a linkage region, genotyping of 8 members from Family I (III-13, III-15, III-17, III-19, IV-10, IV-11, IV-13 and IV-17) was performed using a single-nucleotide polymorphism (SNP) array (Human Omni 2.5-4v1 DNA BeadChip, Illumina, San Diego, CA) containing 2,443,177 single-nucleotide polymorphisms.

Genomic DNA was provided for the Illumina Whole Genome SNP service from Macrogen Inc. (Seoul, Korea). The genomic DNA was amplified and hybridized to the Human Omni 2.5-4v1 DNA BeadChip. The BeadChip was imaged by standard Illumina procedures using Illumina iScan scanner. The Intensity files (*.idat) were processed by the GenomeStudio GT module with default analysis settings. Each SNP was analyzed independently to cluster and identify genotypes. Genotype calls were generated by comparing experimental data with those in the supplied cluster file (*.egt). Calls were generally highly accurate and unambiguous for high quality samples.

Merlin 1.1.2 software was used to calculate the LOD score for parametric linkage analysis with an autosomal dominant model assuming a high penetrance. The penetrance values were set at 0.01 and 0.99.

Whole exome sequencing

According to the linkage based strategy for whole exome sequencing (WES), multiple distantly-related affected family members can be sequenced to identify shared variations together with sequencing an unaffected member to exclude private benign variations. Genomic DNA of three affected (III-11, III-21 and IV-11) and one unaffected (III-29) members of Family I were selected to be sequenced. With the WES service from Macrogen Inc. (Seoul, Korea), genomic DNA was captured and enriched by the Agilent SureSelect Human All Exon Capture kit (Agilent Technologies, Santa Clara, CA). The enriched DNA library was subsequently sequenced using a pair-end 100 bp configuration on the Hiseq 2000 platform (Illumina, San Diego, CA).

Sequence reads were aligned to the UCSC hg19 reference genome (http://genome.ucsc.edu/) by BWA software (http://bio-bwa.sourceforge.net/). The SNPs and Indels were detected by SAMTOOLS (http://samtools.sourceforge.net/). The dbSNP & 1000G were used as variant databases.

WES data analysis

To identify a disease-causing variant, all sequencing variants from the three affected and one unaffected members of Family I were analyzed with the following filtering steps:

1. Sequence variants were filtered against public databases (dbSNP135, 1000 Genomes Project) to exclude polymorphisms.

2. To select for pathogenic variants of an autosomal dominant inherited disease, homozygous variants, synonymous variants, and variants located outside of exons and their flanking regions were excluded.

3. The remaining variants were then filtered to select for shared variants present in all three affected members (but absent in the unaffected) and located in the identified linked regions.

PCR and sequencing

To confirm the existence of the identified variants, we performed PCR and sequencing for the identified variant in exon 3 of the *PSENEN* gene (NG_027934.1) for the three affected and the one unaffected who were subjected for WES. We then performed co-segregation analysis for the identified variant using Sanger sequencing in all members from Family I. Moreover, 100 unrelated Thai controls had PCR and sequencing performed for the identified variant. Mutation analysis of the *PSENEN* gene for the proband and his mother from Family II was carried out. PCR amplification of the entire coding sequences and flanking intronic sequences of the *PSENEN* gene was performed using the sets of primers shown in Table 2.

Table 2: List of primers used for amplification of the entire coding sequences and flanking intronic sequences of the *PSENEN* gene

Fragment	Forward Primer (5'-3')	Reverse Primer (5'-3')	PCR size (bp.)	PCR program
PSENEN exon 2	GTTTCGCCCACCCTAGTAAA	TTCGCAGGTCCTTCATCTCT	390	TD 61°C -51°C
PSENEN exon 3	ATCCCAAAGAGGAGCCAGAT	CCTAGACCAGCCTTCCCTTC	398	TD 61°C -51°C
PSENEN exon 4	CTTGGTGGGAAGGGACAAT	GCAGGAAAGTTCCTAGTTCAGAAG	581	TD 63°C -53°C

The PCR reaction was performed in a total reaction volume of 20 μL. The mixture consisted of 0.15 μM dNTPs, 1x Taq buffer with $(NH_4)_2SO_4$, 1.875 mM $MgCl_2$, 0.025 units of Taq DNA polymerase (Thermo Scientific), 0.15 μM of each primer and 50 ng of genomic DNA.

The PCR amplification was performed using the Touchdown PCR (TD-PCR) program described as follows: initial denaturation at 94°C for 5 minutes, followed by 19 cycles of denaturation at 94°C for 45 seconds, annealing at 61°C for PSENEN exons 2 and 3 and 63°C for PSENEN exon 4 for 45 seconds (during which the temperature was reduced by 0.5°C every cycle until the calculated T_m range was reached), extension at 72°C for 1 minute, followed by 16 cycles of denaturation at 94°C for 45 seconds, annealing at 51°C or 53°C for 45 seconds, extension at 72°C for 1 minute and a final extension at 72°C for 2 minutes.

The PCR products were then treated with ExoSAP-IT (USP Corporation, Cleveland, OH) at 37°C for 20 minutes followed by inactivation at 80°C for 15 minutes, and then was sent to Macrogen Inc. (Seoul, Korea) for Sanger sequencing.

Reverse transcription PCR

Total RNA extracted from the peripheral blood was reverse transcribed to generate complementary DNA (cDNA) using the ImProm-II$^{\text{TM}}$ Reverse Transcription System (Promega, Madison, WI). The reverse transcription was performed in a total reaction volume of 20 μl prepared as follows: 10.1 μl of RNA-primer mixture per reaction (which consisted of 1 μl of 50 μg Oligo$(dT)_{15}$Primer, 500 ng of RNA sample and nuclease-free water added to reach a final volume of 10.1 μl). Incubate at 70°C for 5 min and 4°C for 5 min. Prepare the reverse transcription reaction mix by combining the following components: 4 μl of ImProm-IITM 5X Reaction Buffer, 2.4 μl of 25 mM $MgCl_2$, 1 μl of 10 mM dNTPs, 0.5 μl of Recombinant RNasin® Ribonuclease Inhibitor and 1 μl of ImProm-II$^{\text{TM}}$ Reverse Transcriptase. Then add 8.9 μl aliquots of the reverse transcription reaction mix to each reaction tube, mixed with 11.1 μl of RNA/primer mixture. Incubate at 25°C for 5 minutes, 40°C for 60 minutes and 70°C for

15 minutes. The cDNA synthesis reaction could then be stored at -20°C or used for PCR/Real-Time PCR immediately.

cDNA PCR and sequencing

PCR and sequencing of the *PSENEN* (NM_172341.1) complementary DNA from the proband (IV-11) of Family I and a control were performed. The entire coding region of *PSENEN* was amplified using a forward primer (5'-TTCGTGATCCTTGCATCTGT -3') and a reverse primer (5'- GCCCCAGTATGTGCAGAAGT -3') to yield a PCR product size of 413 bp.

The PCR was performed in a total reaction volume of 20 µl, consisting of 0.15 µM dNTPs, 1x Taq buffer with $(NH_4)_2SO_4$, 1.875 mM $MgCl_2$, 0.025 units of Taq DNA polymerase (Thermo Scientific), 0.15 µM of each primer and 150 ng of complementary DNA.

PCR amplification was performed using the Touchdown PCR (65°C -55°C) described as follows: initial denaturation at 94°C for 5 minutes, followed by 19 cycles of denaturation at 94°C for 45 seconds, annealing at 65°C for 45 seconds (in which the temperature was reduced by 0.5°C every cycle until the calculated Tm range was reached), extension at 72°C for 1 minute, followed by 16 cycles of denaturation at 94°C for 45 seconds, annealing at 55°C for 45 seconds, extension at 72°C for 1 minute and a final extension at 72°C for 2 minutes.

The PCR products were then treated with ExoSAP-IT (USP Corporation, Cleveland, OH), according to the manufacturer's instructions, and then sent for direct sequencing at Macrogen Inc., Seoul, Korea.

DNA sequence analysis

All Sanger sequencing results were analyzed using the Sequencher Software (Gene Codes Corporation, Ann Arbor, MI) with a NCBI Reference Sequence.

Quantitative real-time PCR (qRT-PCR)

The *PSENEN* expression level was quantified by TaqMan® Gene Expression Assay on StepOnePlus™ Real-Time PCR Systems. *GAPDH* and *ACTB* were used as reference genes. The Assay-On-Demand products and universal master mix were commercially purchased (Life Technologies, Grand Island, NY, catalog No. Hs01033961_g1, Hs02758991_g1 and Hs01060665_g1 for *PSENEN*, *GAPDH* and *ACTB*, respectively).

The total RNA extracted from the leukocytes of the two affected individuals (III-13 and III-15) from Family I and four unrelated controls were reverse transcribed to cDNA and used for this experiment. The qRT-PCR reaction was performed in a single well of a 96-well plate. The total reaction volume was 20 µl. The components of each reaction are shown in Table 3. The experiment was done twice, each time in triplicate. For the first run, only *GAPDH* was used as a reference gene. The reaction plates were run on an Applied Biosystem real-time quantitative PCR instrument with the default PCR thermal cycling conditions specified in Table 4.

Relative quantitation using comparative C_T method was used. The *PSENEN* expression levels were calculated relative to the internal standard genes *GAPDH* and *ACTB* and also compared with four unaffected controls.

Table 3: Real-Time PCR components for *PSENEN*, *GAPDH* and *ACTB*

Component	Volume/Reaction
TaqMan Gene Expression Assay (20x)	1 μl
TaqMan Universal PCR Master Mix (2x)	10 μl
cDNA template+H_2O	9 μl

Table 4: Thermal cycling conditions

Step	UDG Incubation	AmpliTaqGold, UP Enzyme Activation	PCR	
	HOLD	HOLD	CYCLE (40 Cycles)	
			Denature	Anneal/Extend
Time	2 min	10 min	15 sec	1 min
Temperature	50℃	95℃	95°C	60°C

Statistical analysis

The data obtained from qRT-PCR that followed a normal distribution were analyzed by the unpaired t-test, while those that did not follow a normal distribution were tested by the Mann-Whitney U test.

Haplotype analysis to proof of unrelated status

Haplotype analysis was done with four members from Family I (III-15, III-16, III-21 and IV-11) and two members from Family II (II-6 and III-4). Eight microsatellite markers spanning the linkage region on chromosome 19 (Table 5) were amplified using ABI Prism® True Allele® PCR Premix according to the manufacturer's protocol. The identified variant was located between marker numbers 4 and 5.

Fluorescent-labeled amplified fragments were sent to Macrogen Inc. (Seoul, Korea) to collect raw genotyping data using Macrogen’s fragment analysis service. Peak Scanner™ Software v1.0 (Applied Biosystems) was then used to perform DNA fragment analysis for each marker based on the size of the amplified fragment.

Table 5: Eight microsatellite markers located in the identified linkage region on chromosome 19

Marker No.	Marker name	Dye	Allele Size Range (ASR)		Forward primers (5'-3')	Reverse primers (5'-3')
1	D19S414	NED	163	187	ABI Prism Linkage Mapping Set version 2.5, Panel 25	
2	D19S213	FAM	174	184	CCTCCAATCTGCACCTGACT	TAGGCTTTGTTCTGGGGTTC
3	D19S425	VIC	252	280	CCACAGGTGTGCATAAAAG	GCCATGTGACTGTAGCAGA
4	D19S208	FAM	167	175	CCCAGTGGGCCTTAGAGATA	GGATGCCTGACGGTGTTTAC
5	D19S876	VIC	163	179	GGGTTGCAGTGAGCGG	TGAGGATGCTGGGGGC
6	D19S224	NED	240	262	AACACCATTCCTCATCTTCC	CCCAGGCCCTATCTGA
7	D19S220	FAM	265	283	ABI Prism Linkage Mapping Set version 2.5, Panel 25	
8	D19S420	NED	251	267	ABI Prism Linkage Mapping Set version 2.5, Panel 25	

Intellectual disability syndrome

Subjects and clinical descriptions

Two siblings with undiagnosed intellectual disability (ID) syndrome were recruited for the study. Their parents were non-consanguineous (Figure 12). Besides intellectual disability, they had a variety of other symptoms as listed in Table 6. Some of the symptoms matched those found in fucosidosis and Shprintzen-Goldberg Syndrome (SGS). No other family members were affected. A possible mode of inheritance is an autosomal recessive pattern of inheritance. The karyotyping results were normal.

After written informed consent, DNA from all four family members were extracted from peripheral blood using the QIAamp® DNA blood mini kit according to the manufacturer's instruction (Qiagen, Valencia, CA).

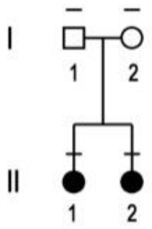


Figure 12: Family ID pedigree

Table 6: Symptoms presented in affected siblings

Symptoms presented in siblings	Fucosidosis	SGS
Intellectual disability	✓	✓
Growth retardation	✓	
Short stature	✓	
Skeletal dysplasia	✓	
Microcephaly		✓
Arachnodactyly		✓
Hyperlaxity		✓

Whole exome sequencing

All four family members of the ID family had WES performed. With WES service provided from Macrogen Inc. (Seoul, Korea), genomic DNA was captured and enriched using the Agilent SureSelect Human All Exon Capture kit (Agilent Technologies, Santa Clara, CA). The enriched DNA library was subsequently sequenced using a pair-end 100 bp configuration on the Hiseq 2000 platform (Illumina, San Diego, CA).

Reads were aligned to NCBI human genome build v37 g1k using BWA software (http://bio-bwa.sourceforge.net/). Duplicate reads were removed using Picard Tools. Variants were called using GATK and annotated using an in-house script.

WES data analysis

As described previously, the siblings had phenotypes similar to fucosidosis and Shprintzen-Goldberg Syndrome (SGS). To ensure that the siblings were not affected with either fucosidosis or Shprintzen-Goldberg Syndrome, we first searched for variants located in known genes related to fucosidosis (*FUCA1*) and Shprintzen-Goldberg Syndrome (*SKI*).

To identify the disease causing gene for the family with siblings affected with intellectual disability (with the disorder suspected to be recessively inherited without any indication for consanguinity), a double-hit strategy was implemented using PERL script showed in the appendix. First we filtered for variants shared between both siblings, and then searched for a single rare homozygous or two rare compound heterozygous mutations. In addition, a *de novo* strategy was also used. The variants shared between both siblings were filtered for *de novo* rare variants. Lastly, a candidate strategy was used.

All variants shared between both siblings were obtained and were then separated into three groups: homozygous, compound heterozygous and *de novo* variants. Each group was sub-grouped into a missense group and a loss of function (LoF) group. Non-synonymous variants and variants found in the dbSNPs database, ExAC database and our in-house exome database were excluded.

With the candidate strategy, we searched for missense variants in genes that shared pathways with *FUCA*1 which included 18 genes in the glycan degradation pathway, 122 genes in the lysosome super pathway and 291 genes (including the *SKI* gene) in the TGF-beta pathway that have been known to cause Shprintzen-Goldberg Syndrome. We also filtered for genes that have been identified as a causative gene for intellectual disability (12 genes for autosomal dominant intellectual disability and 40 genes for autosomal recessive intellectual disability). (Lists of these genes are shown in the appendix.)

CNV detection using WES data

CoNIFER software (Copy Number Inference From Exome Reads http://conifer.sourceforge.net) was used to detect the copy number variation in the ID family from the exome sequencing data. The standard probe file was downloaded from http://sourceforge.net/projects/conifer/files/probes.txt/download. The RPKM files were generated from aligned and indexed BAM files of the four members from the ID family and another 14 samples to meet the minimum requirement of CoNIFER. All the RPKM files were analyzed with default parameters. The SVD value was set at 2 to 6. We then searched for CNVs shared in the siblings.

CHAPTER IV
RESULTS

Familial comedones syndrome

Whole genome linkage analysis

The Merlin 1.1.2 software calculated the LOD score for the parametric linkage analysis with an autosomal dominant model assuming a high penetrance. The penetrance values were set at 0.01 and 0.99. The whole-genome linkage analysis in Family I revealed 8 linkage loci on chromosomes 3, 6, 9, 10, 13 and 19 with a maximum value of linkage with the LOD score of 1.74 (Table 7)

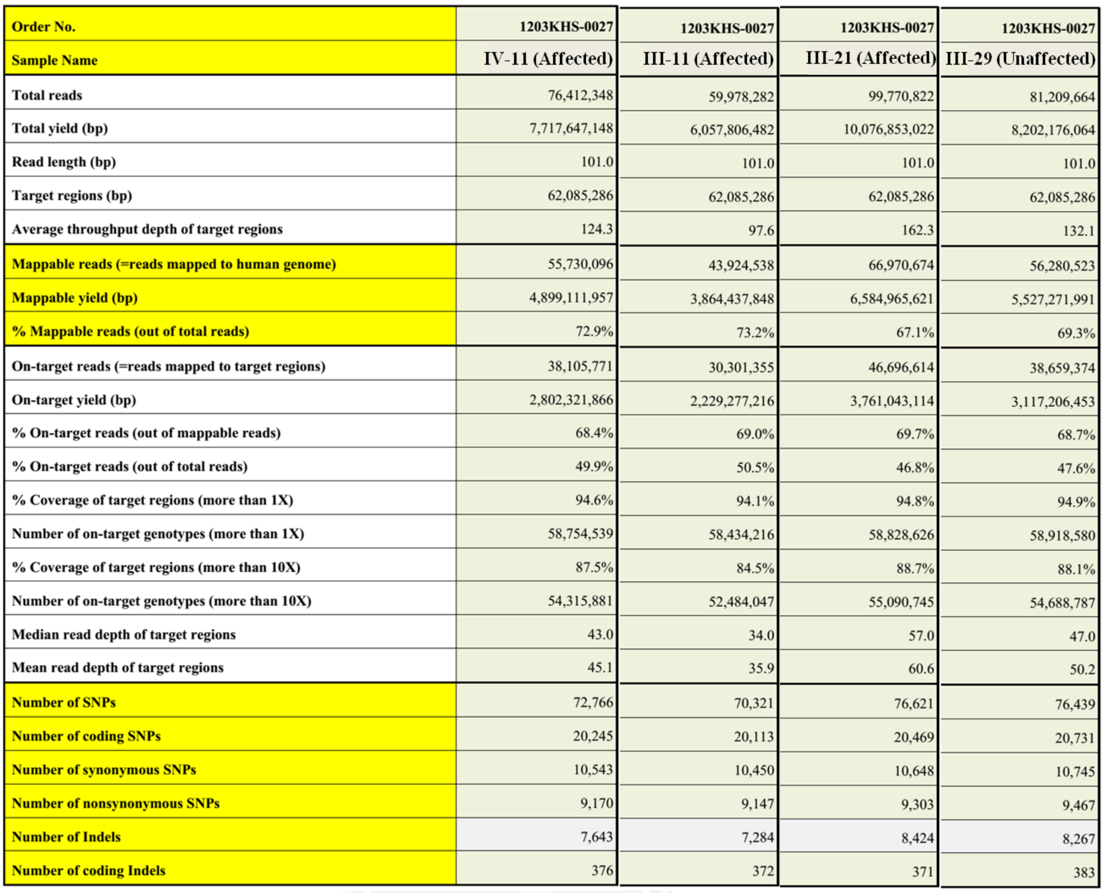
Table 7: Eight linkage loci with the maximum LOD score of 1.74

Locus No.	Chromosome	Position		Size (Mb)
		Start	End	
1	3	22205225	25419008	3.2
2	3	72956648	84038229	11.0
3	3	107152838	122796711	15.6
4	6	33023696	33044638	0.02
5	9	72531140	79979150	7.4
6	10	7642012	15306597	7.6
7	13	95616074	97913181	2.3
8	19	33033939	41668253	8.6

Whole Exome Sequencing

Whole exome sequencing yielded more than 6 gigabases per individual. The capture efficiency varied across the target area with an average of 87.2% of target regions being more than 10X and an average mean read depth of target regions of 48X. Whole exome sequencing results of the three affected and one unaffected members of Family I are summarized in Table 8.

Table 8: Summary of WES results from family I



Order No.	1203KHS-0027	1203KHS-0027	1203KHS-0027	1203KHS-0027
Sample Name	IV-11 (Affected)	III-11 (Affected)	III-21 (Affected)	III-29 (Unaffected)
Total reads	76,412,348	59,978,282	99,770,822	81,209,664
Total yield (bp)	7,717,647,148	6,057,806,482	10,076,853,022	8,202,176,064
Read length (bp)	101.0	101.0	101.0	101.0
Target regions (bp)	62,085,286	62,085,286	62,085,286	62,085,286
Average throughput depth of target regions	124.3	97.6	162.3	132.1
Mappable reads (=reads mapped to human genome)	55,730,096	43,924,538	66,970,674	56,280,523
Mappable yield (bp)	4,899,111,957	3,864,437,848	6,584,965,621	5,527,271,991
% Mappable reads (out of total reads)	72.9%	73.2%	67.1%	69.3%
On-target reads (=reads mapped to target regions)	38,105,771	30,301,355	46,696,614	38,659,374
On-target yield (bp)	2,802,321,866	2,229,277,216	3,761,043,114	3,117,206,453
% On-target reads (out of mappable reads)	68.4%	69.0%	69.7%	68.7%
% On-target reads (out of total reads)	49.9%	50.5%	46.8%	47.6%
% Coverage of target regions (more than 1X)	94.6%	94.1%	94.8%	94.9%
Number of on-target genotypes (more than 1X)	58,754,539	58,434,216	58,828,626	58,918,580
% Coverage of target regions (more than 10X)	87.5%	84.5%	88.7%	88.1%
Number of on-target genotypes (more than 10X)	54,315,881	52,484,047	55,090,745	54,688,787
Median read depth of target regions	43.0	34.0	57.0	47.0
Mean read depth of target regions	45.1	35.9	60.6	50.2
Number of SNPs	72,766	70,321	76,621	76,439
Number of coding SNPs	20,245	20,113	20,469	20,731
Number of synonymous SNPs	10,543	10,450	10,648	10,745
Number of nonsynonymous SNPs	9,170	9,147	9,303	9,467
Number of Indels	7,643	7,284	8,424	8,267
Number of coding Indels	376	372	371	383

WES and data analysis

With the assumption of an autosomal dominant model with full penetrance, the sequencing results were filtered with filtering steps as described. The number of variants reduced after each filtering step is shown in Table 9. Out of 12 variants that were present in the three affected but are absent in the unaffected, there was only one variant located within one of the identified linkage loci. The variant was a heterozygous one base-pair insertion, c.84_85insT (p.L28FfsX93) of *PEN-2*, located within the linked region on chromosome 19 (Table 10).

Table 9: Numbers of variants after each filtering step

Filtering steps \ Samples	IV-11 (Affected)	III-11 (Affected)	III-21 (Affected)	III-29 (unaffected)
All variants[1]	80,409	77,605	85,041	84,706
Private variants[2]	5,158	4,896	5,690	5,778
Heterozygous variants[3]	3,490	3,356	3,814	4,018
Exonic/Splicing variants[4]	776	764	765	825
Co-segregate variants[5]	12*			

*Details are shown in Table 10

[1] Number of SNPs and Indels detected in each individual

[2] Number of SNPs and Indels detected in each individual deducted by polymorphisms

[3] Number of heterozygous SNPs and Indels detected in each individual deducted by polymorphisms

[4] Number of heterozygous SNPs and Indels located in exon or splice site detected in each individual deducted by polymorphisms

[5] Number of heterozygous SNPs and Indels located in exon or splice site detected in each individual deducted by polymorphisms, presence in all three affected individuals and absence in the unaffected individual

Table 10**:** Non-polymorphisms, heterozygous variants located in exonic/splice site regions shared in the three affected but not in the unaffected individuals

(Descriptions of the heterozygous SNPs and Indels located in the exon or splice site detected in each individual deducted by polymorphisms, presence in all three affected individuals and absence in the unaffected individual)

Chr_name	Chr_start	Chr_end	Ref_base	Alt_base	Region	Gene	Change
chr01	211544798	211544798	A	G	exonic	*TRAF5*	nonsynonymous_SNV
chr02	230914604	230914604	C	A	exonic	*SLC16A14*	nonsynonymous_SNV
chr03	75786672	75786672	C	T	exonic	*ZNF717*	nonsynonymous_SNV
chr09	117110115	117110115	T	G	exonic	*AKNA*	nonsynonymous_SNV
chr11	33087551	33087551	A	C	exonic	*TCP11L1*	nonsynonymous_SNV
chr14	24769875	24769875	-	GGA	exonic	*C14orf21*	nonframeshift_insertion
chr17	35581910	35581911	TA	-	splicing	*ACACA*	.
chr17	45234303	45234303	G	C	exonic	*CDC27*	nonsynonymous_SNV
chr19	**36237342**	**36237342**	**-**	**T**	**exonic**	***PSENEN***	**frameshift_insertion**
chr19	56029559	56029559	G	C	exonic	*SSC5D*	nonsynonymous_SNV
chr20	23016692	23016692	G	C	exonic	*SSTR4*	nonsynonymous_SNV
chr22	21384173	21384173	C	T	exonic	*SLC7A4*	nonsynonymous_SNV

PCR and sequencing

Sanger sequencing of the identified variant confirmed a heterozygous one base-pair insertion, c.84_85insT in the three affected individuals who were subjected to WES but not in the unaffected (Figure 13). Co-segregation analysis of Family I demonstrated full co-segregation of the variant with the disease status (Figure 14). The variant was not found in 100 unrelated Thai controls.

Mutation analysis of the *PSENEN* gene of the proband and his mother from Family II revealed the same mutation that was found in Family I (c.84_85insT).

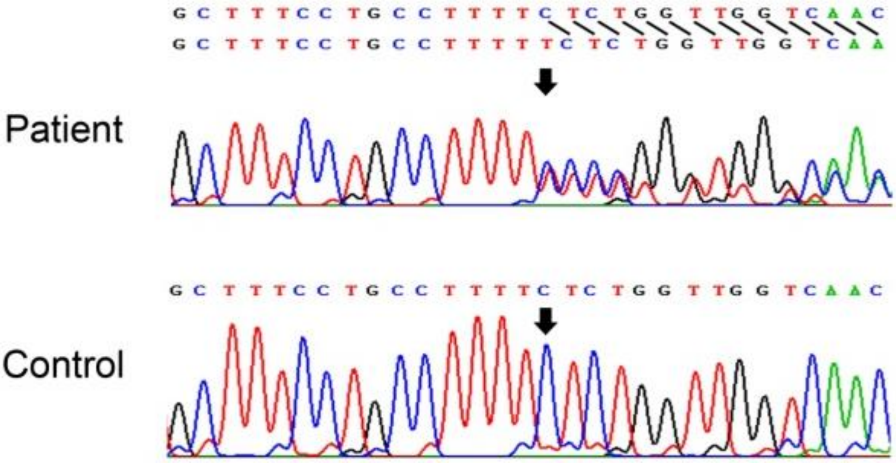


Figure 13: Genomic DNA sequencing result of the identified mutation in *PSENEN*, c.84_85insT

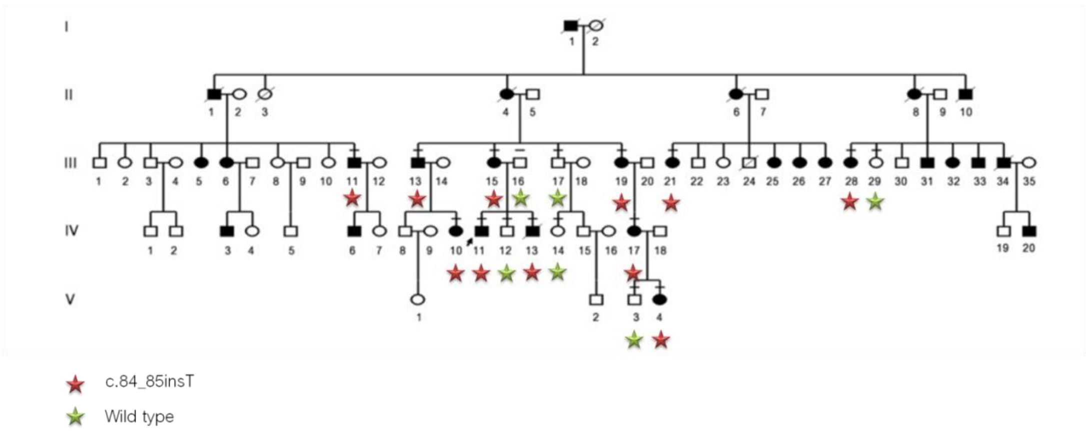


Figure 14: Co-segregation of the *PSENEN*, c.84_85insT with the disease status

cDNA PCR and sequencing

Sanger sequencing of *PSENEN* cDNA revealed that the remaining c.84_85insT had a significantly reduced signal of the mutant allele compared with the wild type (Figure 15). The mutant RNA could possibly be translated into an abnormal protein with an incorrect amino acid sequence containing 120 amino acid residues (19 amino acids longer than the 101 amino acid wild type protein).

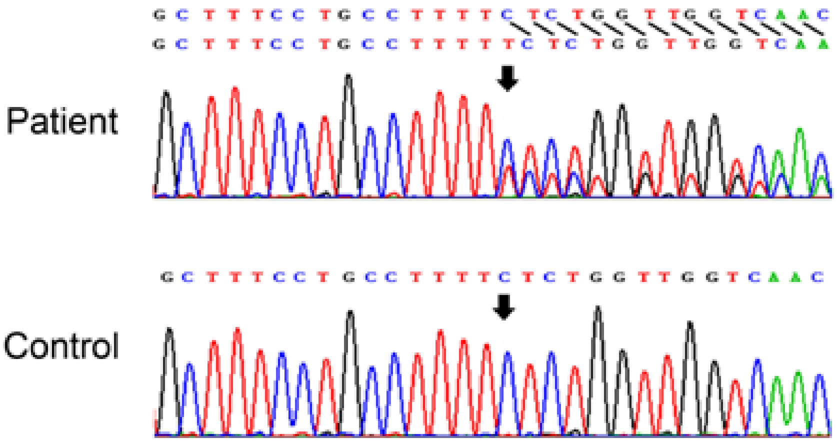
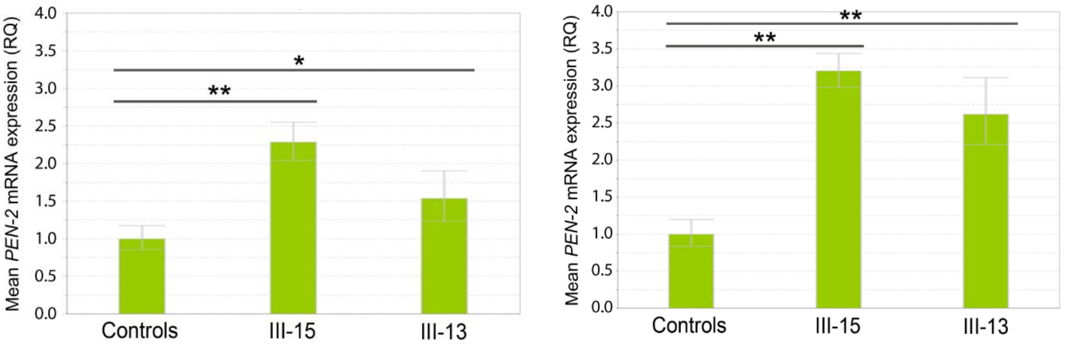


Figure 15: Complementary DNA sequencing result for the identified mutation in *PSENEN*, c.84_85insT

Quantitative real-time PCR (qRT-PCR)

qRT-PCR was performed to study the mRNA expression of *PSENEN* in the patients' leukocytes. The expression levels of *PSENEN* were calculated relative to a set of reference genes, *GAPDH* and *ACTB* (Figure 16). The results revealed an increased mRNA expression in the affected individuals) III-13 and III-15) compared with the four unaffected controls using the unpaired t-test and Mann-Whitney U test.



A) *PSENEN/GAPDH*

B) *PSENEN/ACTB*

Figure 16: Quantitative real-time PCR results showed that the expression levels in patient III-15 and III-13 were significantly increased compared with controls (A) Relative quantification of *PSENEN* with *GAPDH* (B) Relative quantification of *PSENEN* with *ACTB*; The data represents means±SEM. *indicates P<0.05. **indicates P<0.001.

Haplotype analysis

Genotyping results using 8 microsatellite markers spanning the identified linkage region on chromosome 19 are shown in Table 11. The results of the haplotype analysis for these markers revealed that both families had the same haplotype block linked to the identified variant c.84_85insT

Table 11: Genotyping of 8 microsatellite markers located in the identified linkage region on chromosome 19

The orange highlighted cells indicate the same haplotype block in both families linked to the identified variant c.84_85insT

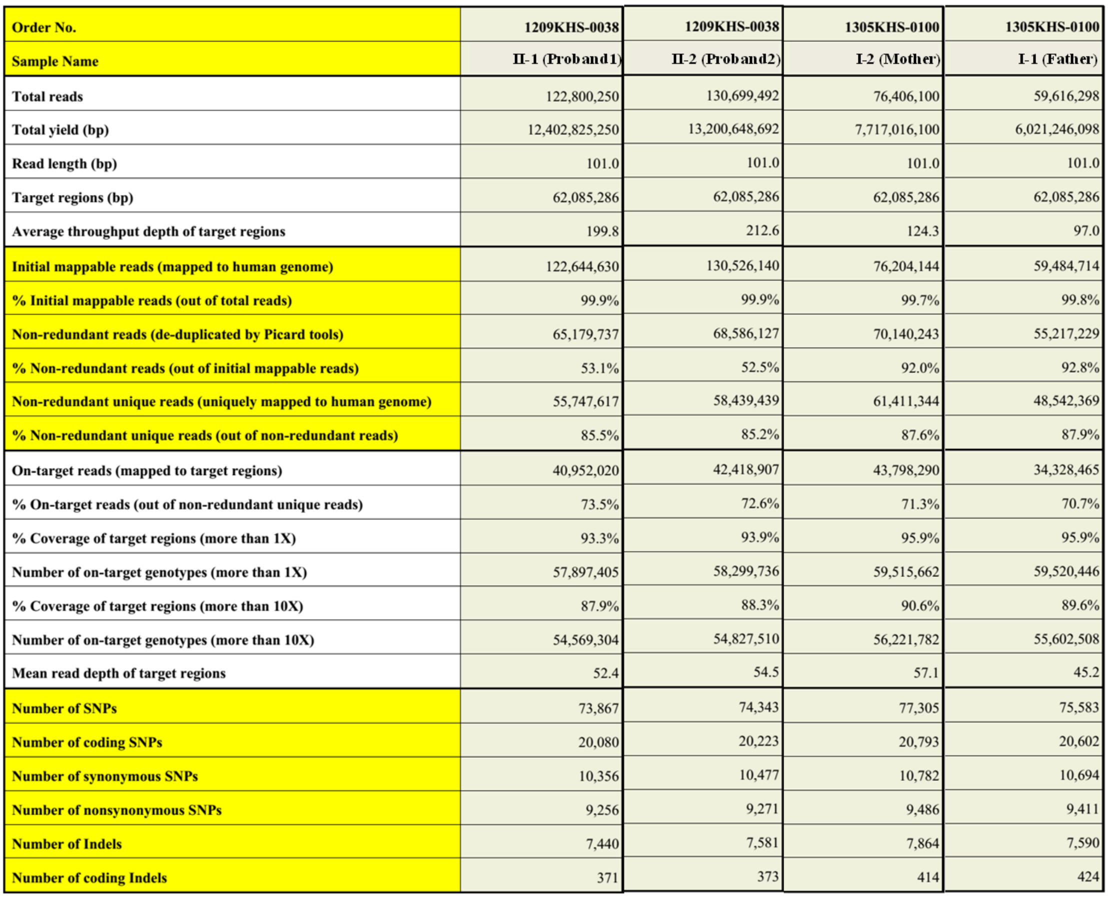
Markers	Family I								Family II			
	III-15		IV-11		III-16		III-21		III-4		II-6	
	Affected mother		Affected proband		Unaffected father		Affected relative		Affected proband		Unaffected mother	
D19S414	-	-	1	2	-	-	1	1	1	1	-	-
D19S213	2	3	1	3	1	1	1	3	3	3	-	-
D19S425	2	2	3	2	2	3	1	2	4	2	2	4
D19S208	1	3	2	3	2	4	2	3	1	3	-	-
c.84_85insT	WT	MT	WT	MT	WT	WT	WT	MT	WT	MT	WT	WT
D19S876	-	-	2	-	-	-	1	-	1	2	-	-
D19S224	-	-	3	1	-	-	4	1	2	1	-	-
D19S220	-	-	3	2	-	-	2	2	3	2	1	3
D19S420	-	-	2	3	-	-	3	3	2	3	1	2

Intellectual disability syndrome

Whole Exome Sequencing

Whole exome sequencing results for all family members are summarized in Table 12. More than 80,000 SNPs/Indels were called in each member with >45 x of mean read depth for target regions.

Table 12: Summary of WES results from the family with intellectual disability



Order No.	1209KHS-0038	1209KHS-0038	1305KHS-0100	1305KHS-0100
Sample Name	II-1 (Proband1)	II-2 (Proband2)	I-2 (Mother)	I-1 (Father)
Total reads	122,800,250	130,699,492	76,406,100	59,616,298
Total yield (bp)	12,402,825,250	13,200,648,692	7,717,016,100	6,021,246,098
Read length (bp)	101.0	101.0	101.0	101.0
Target regions (bp)	62,085,286	62,085,286	62,085,286	62,085,286
Average throughput depth of target regions	199.8	212.6	124.3	97.0
Initial mappable reads (mapped to human genome)	122,644,630	130,526,140	76,204,144	59,484,714
% Initial mappable reads (out of total reads)	99.9%	99.9%	99.7%	99.8%
Non-redundant reads (de-duplicated by Picard tools)	65,179,737	68,586,127	70,140,243	55,217,229
% Non-redundant reads (out of initial mappable reads)	53.1%	52.5%	92.0%	92.8%
Non-redundant unique reads (uniquely mapped to human genome)	55,747,617	58,439,439	61,411,344	48,542,369
% Non-redundant unique reads (out of non-redundant reads)	85.5%	85.2%	87.6%	87.9%
On-target reads (mapped to target regions)	40,952,020	42,418,907	43,798,290	34,328,465
% On-target reads (out of non-redundant unique reads)	73.5%	72.6%	71.3%	70.7%
% Coverage of target regions (more than 1X)	93.3%	93.9%	95.9%	95.9%
Number of on-target genotypes (more than 1X)	57,897,405	58,299,736	59,515,662	59,520,446
% Coverage of target regions (more than 10X)	87.9%	88.3%	90.6%	89.6%
Number of on-target genotypes (more than 10X)	54,569,304	54,827,510	56,221,782	55,602,508
Mean read depth of target regions	52.4	54.5	57.1	45.2
Number of SNPs	73,867	74,343	77,305	75,583
Number of coding SNPs	20,080	20,223	20,793	20,602
Number of synonymous SNPs	10,356	10,477	10,782	10,694
Number of nonsynonymous SNPs	9,256	9,271	9,486	9,411
Number of Indels	7,440	7,581	7,864	7,590
Number of coding Indels	371	373	414	424

WES data analysis

SNPs and indels found in the siblings were first filtered for variants located in the *FUCA1* and *SKI* genes (Table 13-15). No pathogenic variant was found in the known gene for fucosidosis and Shprintzen-Goldberg Syndrome (SGS.). We then moved to the next step to identify a novel causative gene for the undiagnosed disease with Intellectual disability.

Table 13: All variants located in the *FUCA1* gene found in II-1

Chr.	chr_start	chr_end	Ref.	Alt.	hom_het	region	gene	change	dbSNP135 MAF.	1000G MAF.
chr01	24171543	24171543	C	G	hom	downstream	*FUCA1*	.	rs4649119	0.317
chr01	24192200	24192200	A	G	hom	intronic	*FUCA1*	.	rs34902309	0.333
chr01	24194748	24194748	G	C	hom	exonic	*FUCA1*	nonsynonymous_SNV	rs2070956	0.087
chr01	24194773	24194773	G	A	hom	exonic	*FUCA1*	nonsynonymous_SNV	rs2070955	0.1
chr01	24194788	24194788	A	G	hom	UTR5	*FUCA1*	.	rs2070954	0.112
chr01	24194862	24194862	A	C	hom	upstream	*FUCA1*	.	rs2070953	0.133
chr01	24194898	24194898	C	A	hom	upstream	*FUCA1*	.	rs2070952	0.12

Table 14: All variants located in the *SKI* gene found in II-1

Chr.	chr_start	chr_end	Ref.	Alt.	hom_het	region	gene	change	dbSNP135 MAF.	1000G MAF.
chr01	2234903	2234903	C	T	het	intronic	*SKI*	.	rs2256178	0.231
chr01	2240006	2240006	T	C	hom	UTR3	*SKI*	.	rs2173049	0.819
chr01	2241386	2241386	-	TT	hom	UTR3	*SKI*	.	.	.

Table 15: All variants located in the *FUCA1* gene found in II-2

Chr.	chr_start	chr_end	Ref.	Alt.	hom_het	region	gene	change	dbSNP135 MAF.	1000G MAF.
chr01	24171543	24171543	C	G	hom	downstream	*FUCA1*	.	rs4649119	0.317
chr01	24180962	24180962	T	C	het	exonic	*FUCA1*	nonsynonymous_SNV	rs13551	0.217
chr01	24192200	24192200	A	G	hom	intronic	*FUCA1*	.	rs34902309	0.333
chr01	24194748	24194748	G	C	het	exonic	*FUCA1*	nonsynonymous_SNV	rs2070956	0.087
chr01	24194773	24194773	G	A	het	exonic	*FUCA1*	nonsynonymous_SNV	rs2070955	0.1
chr01	24194788	24194788	A	G	het	UTR5	*FUCA1*	.	rs2070954	0.112
chr01	24194862	24194862	A	C	het	upstream	*FUCA1*	.	rs2070953	0.133
chr01	24194898	24194898	C	A	hom	upstream	*FUCA1*	.	rs2070952	0.12

No variants located in the *SKI* gene were found in II-2

To identify a causative variant under the assumption of an autosomal recessive model, sequencing results were filtered and grouped into six groups. The possible causative variant could be homozygous, compound heterozygous and/or *De novo* variants. The number of variants/genes in each group is shown in Table 16.

Table 16: Number of variants/genes shared in the siblings

Group	Loss of function	Missense
Homozygous (variant)	4	300
Compound heterozygous (gene)	5*	37
De novo (variant)	2	43

*Loss of function and missense

The details of the loss of function variants and compound heterozygous genes shared in the siblings are shown in Table 17-19. They were all excluded by their minor allele frequency (MAF) in ExAC or in-house Thai exome with the criteria of MAF < 0.01.

Table 17: List of homozygous, loss of function variants shared in the siblings

Chr_name	chr_start	chr_end	ref_base	alt_base	gene	effect
1	54605319	54605319	G	GC	*CDCP2*	Coding: Frameshift
14	20528448	20528448	TCATAGATTTGCTCACTGAC	T	*OR4L1*	Coding: Frameshift
17	7011225	7011225	T	G	*ASGR2*	3 Splice Site: Canonical_AG_disrupted
17	38858134	38858134	CA	C	*KRT24*	Coding: Frameshift

Table 18: List of *de novo*, loss of function variants shared in the siblings

Chr_name	chr_start	chr_end	ref_base	alt_base	gene	effect
10	99133585	99133585	A	C	*RRP12*	5 Splice Site: Canonical_GT_disrupted
19	46120886	46120886	A	C	EML2	5 Splice Site: Canonical_GT_disrupted

Table 19: List of genes containing compound heterozygous variants shared in the siblings

	Group	Chr.	chr_start	chr_end	ref_base	alt_base	gene	effect
1	Missense	8	142505578	142505578	C	A	*AC100803.1*	Coding: Missense
	LoF	8	142458053	142458053	A	G	*AC100803.1*	Coding: Stop_codon_disrupted
	Missense	8	142446923	142446923	C	T	*AC100803.1*	Coding: Missense
2	Missense	16	81208515	81208515	G	A	*PKD1L2*	Coding: Missense
	Missense	16	81193358	81193358	C	G	*PKD1L2*	Coding: Missense
	Missense	16	81174999	81174999	A	G	*PKD1L2*	Coding: Missense
	LoF	16	81174978	81174978	A	G	*PKD1L2*	Coding: Stop_codon_disrupted
	Missense	16	81173193	81173193	C	T	*PKD1L2*	Coding: Missense
3	Missense	8	68200276	68200276	T	C	*ARFGEF1*	Coding: Missense
	LoF	8	68163686	68163686	C	T	*ARFGEF1*	3 Splice Site: Canonical_AG_disrupted
4	Missense	18	43508856	43508856	C	G	*EPG5*	Coding: Missense
	LoF	18	43483965	43483965	G	A	*EPG5*	Coding: Nonsense
5	LoF	19	20989681	20989681	C	A	*ZNF66P*	Coding: Nonsense
	Missense	19	20988708	20988708	G	A	*ZNF66P*	Coding: Missense

Since there were more than 100 missense variants shared in the siblings, we then used the candidate strategy to look for genes that shared pathways with *FUCA1,* including 18 genes in the glycan degradation pathway, and 122 genes in the lysosome super pathway. A homozygous missense mutation was found in the *AP4E1* gene, but this variant was present at ~43% in the ExAC database (Table 20). We also searched for variants in 291 genes that included the *SKI* gene in the TGF-beta pathway which has been known to cause Shprintzen-Goldberg Syndrome. We found that compound heterozygous variants in the *IL10RA* gene. Both had an ExAC MAF > 0.01 (Table 21). There was no possible causative variant in the list of 12 genes for autosomal dominant intellectual disability and 40 genes for recessive intellectual disability.

Table 20: Homozygous missense variant shared in the siblings in the lysosome super pathway

chr_name	chr_start	chr_end	ref_base	alt_base	gene	effect	MAF.(ExAC)
15	51217361	51217361	T	C	*AP4E1*	Coding: Missense	0.435

Table 21: Compound heterozygous variants shared in siblings the in TGF-beta pathway

chr_name	chr_start	chr_end	ref_base	alt_base	gene	effect	MAF.(ExAC)
11	117857338	117857338	G	C	*IL10RA*	Coding: Missense	0.061
11	117857499	117857499	C	A	*IL10RA*	Coding: Missense	0.012

CNV detection using WES data

We performed CNV detection using whole exome sequencing data with CoNIFER software. We searched for CNVs shared in the siblings. The results did not provide any possible causative copy number variations.

CHAPTER V
DISCUSSION

Familial comedones

Combination of whole genome linkage analysis with whole exome sequencing successfully revealed the gene responsible for familial comedones. With the filtering steps, 12 variants were obtained from WES. There was only one loss of function variant. The rest of them were non-synonymous and non-frameshift insertion variants. In addition, having another family affected with the disease carrying the same mutation provided evidence that the variant was pathogenic. Although the haplotype analysis was not able to confirm unrelated status of the two families, other evidence such as family history and native habitat supported that they were not related.

The identified mutation was a single base pair insertion, c.84_85insT, located in the first transmembrane domain of the *PSENEN* gene (Figure 17). The *PSENEN* or *PEN-2* encodes the presenilin enhancer 2 (PEN-2). The mutation causing frameshift leads to the stop codon at position 121 (p.L28FfsX93). *PEN-2* mRNA expression was evaluated by qRT-PCR showing increased mRNA expression levels in the patients' leukocytes. The reasons and the effects of this increase require further investigations. Protein expression study was not performed because of sample unavailability. However, the predicted amino acid sequences of the mutant PEN-2 were aligned with the wild type (Figure 18) showing that the mutation not only extends 19 amino acid longer but also destroys amino sequences of the two transmembrane domains of PEN-2 and eliminates the evolutionarily conserved DYLSF domain at the C terminus (residues 90–94) necessary for the binding of PEN-2 to other components in the presenilin complex, and the hydrophilic C terminus (residues 90–101) critical for functional γ -secretase activity (60).

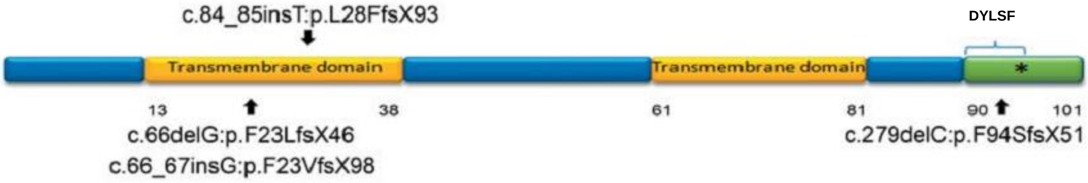


Figure 17: Schematic diagram of the PEN-2 protein. The identified mutation is shown by the arrow above the diagram. The three previously published mutations associated with hidradenitis suppurativa are shown by the arrows under the diagram. The numbers under the diagram indicate amino acid residues. The asterisk denotes the hydrophilic C-terminal domain. The DYLSF domain contains amino acid residues 90–94.

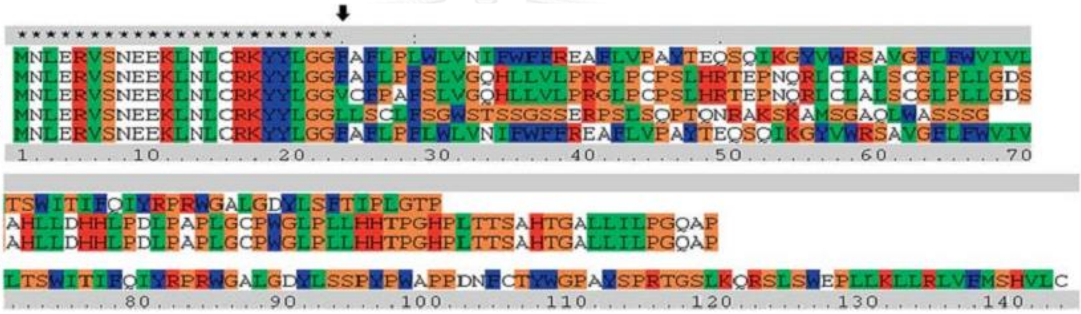

1 MNLERVSNEEKLNLCRKYYLGGFAFLPLWLVNIFWFFREAFLVPAYTEQSQIKGYVWRSAVGFLFWVIVL
2 MNLERVSNEEKLNLCRKYYLGGFAFLPFSLVGQHLLVLPRGLPCPSLHRTEPNQRLCLALSCGLPLLGDS
3 MNLERVSNEEKLNLCRKYYLGGVCFPAFSLVGQHLLVLPRGLPCPSLHRTEPNQRLCLALSCGLPLLGDS
4 MNLERVSNEEKLNLCRKYYLGGLLSCLFSGWSTSSGSSERPSLSQPTQNRAKSKAMSGAQLWASSSG
5 MNLERVSNEEKLNLCRKYYLGGFAFLPFLWLVNIFWFFREAFLVPAYTEQSQIKGYVWRSAVGFLFWVIV
1 . . . . 10 . . . . 20 . . . . 30 . . . . 40 . . . . 50 . . . . 60 70
1 TSWITIFQIYRPRWGALGDYLSFTIPLGTP
2 AHLLDHHLPDLPAPLGCPWGLPLLHHTPGHPLTTSAHTGALLILPGQAP
3 AHLLDHHLPDLPAPLGCPWGLPLLHHTPGHPLTTSAHTGALLILPGQAP
4
5 LTSWITIFQIYRPRWGALGDYLSSPYPWAPPDNFCTYWGPAYSPRTGSLKQRSLSWEPLLKLLRLVFMSHVLC
. . . . 80 . . . . 90 . . . . 100 . . . . 110 . . . . 120 . . . . 130 . . . . 140 . . .

Figure 18: Predicted amino acid change of the PEN-2 protein. Lane 1 shows the amino acid sequence of the wild type. Lane 2 shows that of the novel mutation c.84_85insT (p.L28FfsX92). Lanes 3–5 show those of previously published mutations associated with HS, c.66_67insG (p.F23VfsX98), c.66delG (p.F23LfsX46) and c.279delC (p.F94SfsX51), respectively.

The γ-secretase is a transmembrane protease which cleaves transmembrane proteins such as amyloid precursor protein, Notch receptors, N-cadherin and E-cadherin. The protease complex consists of four integral membrane proteins including presenilin, nicastrin, anterior pharynx defective and PEN-2 (encoded by *PSEN1*or *PSEN2*, *NCSTN*, *APH1A* or *APH1B*, and *PEN-2*, respectively) (28, 61, 62). Mutations in genes encoding subunits of γ-secretase have been described in a variety of disorders including familial Alzheimer's disease, dilated cardiomyopathy, frontotemporal dementia (FTD), breast cancer, leukemia, and hidradenitis suppurativa

(HS) in different roles of γ-secretase (42). HS can be caused by heterozygous mutations in three genes encoding subunits of γ-secretase including PEN-2, PSEN1and NCSTN. There were 3 mutations in PEN-2 that were reported to cause HS shown in figure 17 (32-41). All mutations associated with HS are predicted to reduce γ-secretase activity (28, 32, 37). Haploinsufficiency of the γ-secretase component genes suggests that critical levels of γ-secretase activity are necessary for skin homeostasis.

Familial comedones and HS have some overlapping clinical manifestations including multiple severe purulent nodules and abscesses leaving unsightly scars. They represent, however, two distinct entities. HS lesions are frequently located on the axillae and in the inguinal, perianal, perineal, mammary and submammary regions whereas those in familial comedones are located on the back, abdomen, neck, and legs. Unlike patients with HS, patients with familial comedones have numerous widespread comedones and pits and some only have diffuse comedonal lesions without development of purulent nodules. Histologically, the skin pits of familial comedones show follicular dilatation filled with keratin plugs and branched extraradicular sheaths. HS has either tissue inflammation after rupture of keratin-rich epidermal cysts followed by extensive cutaneous thrombi and infarcts, or has inflammatory destruction of apocrine glands (24, 32).

Different mutations in the *PEN-2* gene could give rise to either familial comedones or HS. This phenomenon has been found in many other genes. For example, mutations in *FGFR2* can result in either Crouzon's or Pfeiffer's syndrome (63). Mutations in the IRF6 gene can result in van der Woude's syndrome or the popliteal pterygium syndrome (64). It has been hypothesized that genetic modifiers and environmental factors may play an important role.

Interestingly, the *PSEN1* mutations causing familial Alzheimer's disease are mostly missense and retain the C-terminal sequence leading to decreased presenilin-dependent neuronal survival, suggesting a gain-of-function mechanism (65, 66).

It is known that different mutations in the same gene can cause distinct disease phenotypes. This study has identified a distinct mutation of *PEN-2* as a cause of yet another skin disorder. How the c.84_85insT leads to familial comedones and whether familial comedones can be caused by other mutations in *PEN*-2 or other genes needs further investigations.

Intellectual disability syndrome

Intellectual disability is a large and heterogeneous group of disorders. The causative factors could be genetic, epigenetic or environmental factors. The genetic causes range from large cytogenetic abnormalities to point mutations.. Since the siblings were affected with the same syndrome, we hypothesized that the undiagnosed syndrome with ID in the family had a genetic defect. We performed WES in order to detect both CNVs and SNVs/indels which caused about 15% and 10% of ID cases, respectively (43, 45). Unfortunately, neither pathogenic CNVs nor SNVs/indels were detected in the family.

We have done data analysis with all possibilities for SNVs/indels detection. First we ensured that there was no mutation in *FUCA1* and *SKI*, the genes responsible for disorders with some clinical features seen in both siblings. We then analyzed for a novel gene with three different models; compound heterozygous, homozygous and *de novo.* We also used the candidate gene strategy to search for variants located in the list of genes causing non-syndromic intellectual disability in either an autosomal dominant (NS-ADID) or autosomal recessive (NS-ARID) manner as well as genes working in the same pathway as *FUCA1* and *SKI*. None of the possible pathogenic variants were detected. However, we might have missed some variants by the limitation of whole exome sequencing since we focused on variants shared in the siblings. It is possible that WES can only capture variants in one sibling. In addition, CNV detection using whole exome sequencing data did not reveal any significant candidates.

In addition, oligogenic heterozygosity, and hypomorphic alterations in multiple genes have recently been noted for both ID and autism spectrum disorder (ASD). Schaaf *et al.* sequenced 21 known autism susceptibility genes in 339 individuals with high-functioning, idiopathic ASD and found that probands were much more likely than controls to carry multiple heterozygous variants in autism susceptibility genes (67). In another recent report, Liu *et al.* performed a large scale of WES showing that there was no single gene significantly associated with the ASD risk. It was speculated that mutations scattered across hundreds of genes could be involved (68).

Finally, it is not surprising that no causative mutations responsible for ID in this family could be identified. The FORGE (Finding of Rare Disease Genes) Canada Consortium studied a total of 264 disorders, 62 disorders were found in non-consanguineous families with two or more affected siblings. They performed WES and used the approach to identify compound heterozygous variants shared between affected siblings. Of 62 disorders, they identified 13 novel genes and 15 genes in which mutations were known to be associated with human diseases. However, the remaining 55% were unsolved (69).

REFERENCES

1. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011 Mar 20;38(3):95-109. PubMed PMID: 21477781. Pubmed Central PMCID: 3076108. Epub 2011/04/12. eng.
2. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46. PubMed PMID: 19997069. Epub 2009/12/10. eng.
3. von Bubnoff A. Next-generation sequencing: the race is on. Cell. 2008 Mar 7;132(5):721-3. PubMed PMID: 18329356. Epub 2008/03/11. eng.
4. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet. 2012 Oct;57(10):621-32. PubMed PMID: 22832387. Epub 2012/07/27. eng.
5. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? Journal of medical genetics. 2011 Sep;48(9):580-9. PubMed PMID: 21730106.
6. Lin X, Tang W, Ahmad S, Lu J, Colby CC, Zhu J, et al. Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. Hear Res. 2012 Jun;288(1-2):67-76. PubMed PMID: 22269275. Epub 2012/01/25. eng.
7. Xuan J, Yu Y, Qing T, Guo L, Shi L. Next-generation sequencing in the clinic: Promises and challenges. Cancer Lett. 2012 Nov 19. PubMed PMID: 23174106. Epub 2012/11/24. Eng.
8. Cantu JM, Gomez-Bustamente MO, Gonzalez-Mendoza A, Sanchez-Corona J. Familial comedones. Evidence for autosomal dominant inheritance. Arch Dermatol. 1978 Dec;114(12):1807-9. PubMed PMID: 153732. Epub 1978/12/01. eng.

9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977 Dec;74(12):5463-7. PubMed PMID: 271968. Pubmed Central PMCID: 431765. Epub 1977/12/01. eng.

10. Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America. 1977 Feb;74(2):560-4. PubMed PMID: 265521. Pubmed Central PMCID: 392330. Epub 1977/02/01. eng.

11. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011 Nov;12(11):745-55. PubMed PMID: 21946919.

12. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010 Jan;42(1):30-5. PubMed PMID: 19915526. Pubmed Central PMCID: 2847889. Epub 2009/11/17. eng.

13. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. The New England journal of medicine. 2014 Jun 19;370(25):2418-25. PubMed PMID: 24941179. Epub 2014/06/19. eng.

14. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010 Sep;42(9):790-3. PubMed PMID: 20711175. Pubmed Central PMCID: 2930028. Epub 2010/08/17. eng.

15. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010 Jun;42(6):483-5. PubMed PMID: 20436468. Epub 2010/05/04. eng.

16. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. American journal of human genetics. 2015 Aug 6;97(2):199-215. PubMed PMID: 26166479.

17. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. European journal of human genetics : EJHG.

2012 May;20(5):490-7. PubMed PMID: 22258526. Pubmed Central PMCID: 3330229.

18. Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. Human molecular genetics. 2010 Oct 15;19(R2):R119-24. PubMed PMID: 20846941. Pubmed Central PMCID: 2953741. Epub 2010/09/18. eng.

19. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011 Sep;12(9):628-40. PubMed PMID: 21850043. Epub 2011/08/19. eng.

20. Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in bioengineering and biotechnology. 2015;3:92. PubMed PMID: 26161383. Pubmed Central PMCID: 4479793. Epub 2015/07/15. eng.

21. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006 Feb;7(2):85-97. PubMed PMID: 16418744. Epub 2006/01/19. eng.

22. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. Genome research. 2012 Aug;22(8):1525-32. PubMed PMID: 22585873. Pubmed Central PMCID: 3409265. Epub 2012/05/16. eng.

23. Rodin HH BM, Bernstein G. Diffuse familial comedones. Arch Dermatol. 1967;95:145 - 6.

24. Cheng MJ, Chen WC, Happle R, Song ZQ. Familial disseminated comedones without dyskeratosis: report of an affected family and review of the literature. Dermatology. 2014;228(4):303-6. PubMed PMID: 24819025. Epub 2014/05/14. eng.

25. Rerknimitr P, Korkij W, Wititsuwannakul J, Panmontha W, Suphapeetiporn K, Shotelersuk V. Expanding phenotypic spectrum of familial comedones. Dermatology. 2014;228(3):215-9. PubMed PMID: 24818872.

26. Wiseman MC. Hidradenitis suppurativa: a review. Dermatol Ther. 2004;17(1):50-4. PubMed PMID: 14756891. Epub 2004/02/06. eng.

27. Jansen I, Altmeyer P, Piewig G. Acne inversa (alias hidradenitis suppurativa). J Eur Acad Dermatol Venereol. 2001 Nov;15(6):532-40. PubMed PMID: 11843212. Epub 2002/02/15. eng.

28. Pink AE, Simpson MA, Desai N, Trembath RC, Barker JN. gamma-Secretase mutations in hidradenitis suppurativa: new insights into disease pathogenesis. The Journal of investigative dermatology. 2013 Mar;133(3):601-7. PubMed PMID: 23096707. Epub 2012/10/26. eng.

29. Martorell A, Garcia-Martinez FJ, Jimenez-Gallo D, Pascual JC, Pereyra-Rodriguez J, Salgado L, et al. An Update on Hidradenitis Suppurativa (Part I): Epidemiology, Clinical Aspects, and Definition of Disease Severity. Actas dermo-sifiliograficas. 2015 Aug 6. PubMed PMID: 26254550. Epub 2015/08/10. Actualizacion en hidradenitis supurativa (I): epidemiologia, aspectos clinicos y definicion de severidad de la enfermedad. Eng

Spa.

30. Nazary M, van der Zee HH, Prens EP, Folkerts G, Boer J. Pathogenesis and pharmacotherapy of Hidradenitis suppurativa. Eur J Pharmacol. 2011 Dec 15;672(1-3):1-8. PubMed PMID: 21930119. Epub 2011/09/21. eng.

31. Gao M, Wang PG, Cui Y, Yang S, Zhang YH, Lin D, et al. Inversa acne (hidradenitis suppurativa): a case report and identification of the locus at chromosome 1p21.1-1q25.3. The Journal of investigative dermatology. 2006 Jun;126(6):1302-6. PubMed PMID: 16543891. Epub 2006/03/18. eng.

32. Wang B, Yang W, Wen W, Sun J, Su B, Liu B, et al. Gamma-secretase gene mutations in familial acne inversa. Science. 2010 Nov 19;330(6007):1065. PubMed PMID: 20929727. Epub 2010/10/12. eng.

33. Miskinyte S, Nassif A, Merabtene F, Ungeheuer MN, Join-Lambert O, Jais JP, et al. Nicastrin mutations in French families with hidradenitis suppurativa. The Journal of investigative dermatology. 2012 Jun;132(6):1728-30. PubMed PMID: 22358060. Epub 2012/02/24. eng.

34. Liu Y, Gao M, Lv YM, Yang X, Ren YQ, Jiang T, et al. Confirmation by exome sequencing of the pathogenic role of NCSTN mutations in acne inversa

(hidradenitis suppurativa). The Journal of investigative dermatology. 2011 Jul;131(7):1570-2. PubMed PMID: 21430701. Epub 2011/03/25. eng.

35. Zhang C, Wang L, Chen L, Ren W, Mei A, Chen X, et al. Two novel mutations of the NCSTN gene in Chinese familial acne inverse. J Eur Acad Dermatol Venereol. 2013 Dec;27(12):1571-4. PubMed PMID: 22759192. Epub 2012/07/05. eng.

36. Pink AE, Simpson MA, Desai N, Dafou D, Hills A, Mortimer P, et al. Mutations in the gamma-secretase genes NCSTN, PSENEN, and PSEN1 underlie rare forms of hidradenitis suppurativa (acne inversa). The Journal of investigative dermatology. 2012 Oct;132(10):2459-61. PubMed PMID: 22622421. Epub 2012/05/25. eng.

37. Pink AE, Simpson MA, Brice GW, Smith CH, Desai N, Mortimer PS, et al. PSENEN and NCSTN mutations in familial hidradenitis suppurativa (Acne Inversa). The Journal of investigative dermatology. 2011 Jul;131(7):1568-70. PubMed PMID: 21412258. Epub 2011/03/18. eng.

38. Nomura Y, Nomura T, Sakai K, Sasaki K, Ohguchi Y, Mizuno O, et al. A novel splice site mutation in NCSTN underlies a Japanese family with hidradenitis suppurativa. The British journal of dermatology. 2013 Jan;168(1):206-9. PubMed PMID: 22834455. Epub 2012/07/28. eng.

39. Yang JQ, Wu XJ, Dou TT, Jiao T, Chen XB, Min M, et al. Haploinsufficiency caused by a nonsense mutation in NCSTN underlying hidradenitis suppurativa in a Chinese family. Clinical and experimental dermatology. 2015 Jul 30. PubMed PMID: 26224166. Epub 2015/08/01. Eng.

40. Jiao T, Dong H, Jin L, Wang S, Wang J. A novel nicastrin mutation in a large Chinese family with hidradenitis suppurativa. The British journal of dermatology. 2013 May;168(5):1141-3. PubMed PMID: 23517242. Epub 2013/03/23. eng.

41. Nomura Y, Nomura T, Suzuki S, Takeda M, Mizuno O, Ohguchi Y, et al. A novel NCSTN mutation alone may be insufficient for the development of familial hidradenitis suppurativa. Journal of dermatological science. 2014 May;74(2):180-2. PubMed PMID: 24581508. Epub 2014/03/04. eng.

42. Jurisch-Yaksi N, Sannerud R, Annaert W. A fast growing spectrum of biological functions of gamma-secretase in development and disease. Biochimica et biophysica acta. 2013 Dec;1828(12):2815-27. PubMed PMID: 24099003. Epub 2013/10/09. eng.

43. Ellison JW, Rosenfeld JA, Shaffer LG. Genetic basis of intellectual disability. Annual review of medicine. 2013;64:441-50. PubMed PMID: 23020879. Epub 2012/10/02. eng.

44. Srivastava AK, Schwartz CE. Intellectual disability and autism spectrum disorders: causal genes and molecular mechanisms. Neuroscience and biobehavioral reviews. 2014 Oct;46 Pt 2:161-74. PubMed PMID: 24709068. Pubmed Central PMCID: 4185273. Epub 2014/04/09. eng.

45. Topper S, Ober C, Das S. Exome sequencing and the genetics of intellectual disability. Clin Genet. 2011 Aug;80(2):117-26. PubMed PMID: 21627642. Epub 2011/06/02. eng.

46. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, et al. A de novo paradigm for mental retardation. Nat Genet. 2010 Dec;42(12):1109-12. PubMed PMID: 21076407. Epub 2010/11/16. eng.

47. Figueiredo T, Melo US, Pessoa AL, Nobrega PR, Kitajima JP, Rusch H, et al. A homozygous loss-of-function mutation in inositol monophosphatase 1 (IMPA1) causes severe intellectual disability. Molecular psychiatry. 2015 Sep 29. PubMed PMID: 26416544. Epub 2015/09/30. Eng.

48. Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet. 2010 Oct;42(10):827-9. PubMed PMID: 20802478. Epub 2010/08/31. eng.

49. van Bokhoven H. Genetic and epigenetic networks in intellectual disabilities. Annual review of genetics. 2011;45:81-104. PubMed PMID: 21910631. Epub 2011/09/14. eng.

50. Chelly J, Khelfaoui M, Francis F, Cherif B, Bienvenu T. Genetics and pathophysiology of mental retardation. European journal of human genetics : EJHG. 2006 Jun;14(6):701-13. PubMed PMID: 16721406. Epub 2006/05/25. eng.

51. Durand P, Borrone C, Della Cella G. A NEW MUCOPOLYSACCHARIDE LIPIDSTORAGE DISEASE? The Lancet. 1966;288(7476):1313-4.

52. Durand P, Borrone C, Della Cella G. Fucosidosis. The Journal of pediatrics. 1969 Oct;75(4):665-74. PubMed PMID: 4241464. Epub 1969/10/01. eng.

53. Van Hoof F, Hers HG. Mucopolysaccharidosis by absence of alpha-fucosidase. Lancet. 1968 Jun 1;1(7553):1198. PubMed PMID: 4172303. Epub 1968/06/01. eng.

54. Willems PJ, Gatti R, Darby JK, Romeo G, Durand P, Dumon JE, et al. Fucosidosis revisited: a review of 77 patients. American journal of medical genetics. 1991 Jan;38(1):111-31. PubMed PMID: 2012122. Epub 1991/01/01. eng.

55. Willems PJ, Seo HC, Coucke P, Tonlorenzi R, O'Brien JS. Spectrum of mutations in fucosidosis. European journal of human genetics : EJHG. 1999 Jan;7(1):60-7. PubMed PMID: 10094192. Epub 1999/03/27. eng.

56. van Steensel MA, van Geel M, Parren LJ, Schrander-Stumpel CT, Marcus-Soekarman D. Shprintzen-Goldberg syndrome associated with a novel missense mutation in TGFBR2. Experimental dermatology. 2008 Apr;17(4):362-5. PubMed PMID: 17979970. Epub 2007/11/06. eng.

57. Carmignac V, Thevenon J, Ades L, Callewaert B, Julia S, Thauvin-Robinet C, et al. In-frame mutations in exon 1 of SKI cause dominant Shprintzen-Goldberg syndrome. American journal of human genetics. 2012 Nov 2;91(5):950-7. PubMed PMID: 23103230. Pubmed Central PMCID: 3487125. Epub 2012/10/30. eng.

58. Au PY, Racher HE, Graham JM, Jr., Kramer N, Lowry RB, Parboosingh JS, et al. De novo exon 1 missense mutations of SKI and Shprintzen-Goldberg syndrome: two new cases and a clinical review. American journal of medical genetics Part A. 2014 Mar;164A(3):676-84. PubMed PMID: 24357594. Epub 2013/12/21. eng.

59. Schepers D, Doyle AJ, Oswald G, Sparks E, Myers L, Willems PJ, et al. The SMAD-binding domain of SKI: a hotspot for de novo mutations causing Shprintzen-Goldberg syndrome. European journal of human genetics : EJHG.

2015 Feb;23(2):224-8. PubMed PMID: 24736733. Pubmed Central PMCID: 4297897. Epub 2014/04/17. eng.

60. Hasegawa H, Sanjo N, Chen F, Gu YJ, Shier C, Petit A, et al. Both the sequence and length of the C terminus of PEN-2 are critical for intermolecular interactions and function of presenilin complexes. The Journal of biological chemistry. 2004 Nov 5;279(45):46455-63. PubMed PMID: 15322109. Epub 2004/08/24. eng.

61. Wolfe MS. gamma-Secretase in biology and medicine. Seminars in cell & developmental biology. 2009 Apr;20(2):219-24. PubMed PMID: 19162210. Epub 2009/01/24. eng.

62. Bergmans BA, De Strooper B. gamma-secretases: from cell biology to therapeutic strategies. The Lancet Neurology. 2010 Feb;9(2):215-26. PubMed PMID: 20129170. Epub 2010/02/05. eng.

63. Rutland P, Pulleyn LJ, Reardon W, Baraitser M, Hayward R, Jones B, et al. Identical mutations in the FGFR2 gene cause both Pfeiffer and Crouzon syndrome phenotypes. Nat Genet. 1995 Feb;9(2):173-6. PubMed PMID: 7719345. Epub 1995/02/01. eng.

64. Little HJ, Rorick NK, Su LI, Baldock C, Malhotra S, Jowitt T, et al. Missense mutations that cause Van der Woude syndrome and popliteal pterygium syndrome affect the DNA-binding and transcriptional activation functions of IRF6. Human molecular genetics. 2009 Feb 1;18(3):535-45. PubMed PMID: 19036739. Pubmed Central PMCID: 2638798. Epub 2008/11/28. eng.

65. Newman M, Wilson L, Verdile G, Lim A, Khan I, Moussavi Nik SH, et al. Differential, dominant activation and inhibition of Notch signalling and APP cleavage by truncations of PSEN1 in human disease. Human molecular genetics. 2014 Feb 1;23(3):602-17. PubMed PMID: 24101600. Epub 2013/10/09. eng.

66. Kelleher RJ, 3rd, Shen J. Genetics. Gamma-secretase and human disease. Science. 2010 Nov 19;330(6007):1055-6. PubMed PMID: 21097925. Pubmed Central PMCID: 4556372. Epub 2010/11/26. eng.

67. Schaaf CP, Sabo A, Sakai Y, Crosby J, Muzny D, Hawes A, et al. Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. Human molecular genetics. 2011 Sep 1;20(17):3366-75. PubMed PMID: 21624971. Pubmed Central PMCID: 3153303. Epub 2011/06/01. eng.

68. Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, Lim E, et al. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS genetics. 2013 Apr;9(4):e1003443. PubMed PMID: 23593035. Pubmed Central PMCID: 3623759. Epub 2013/04/18. eng.

69. Beaulieu CL, Majewski J, Schwartzentruber J, Samuels ME, Fernandez BA, Bernier FP, et al. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. American journal of human genetics. 2014 Jun 5;94(6):809-17. PubMed PMID: 24906018. Pubmed Central PMCID: 4121481. Epub 2014/06/07. eng.

APPENDICES

APPENDIX A

GENE LISTS

Table 22: 122 Genes in the Lysosome SuperPath

ABCA2	ABCB9	ACP2	ACP5	AGA	AP1B1	AP1G1	AP1G2
AP1M1	AP1M2	AP1S1	AP1S2	AP1S3	AP3B1	AP3B2	AP3D1
AP3M1	AP3M2	AP3S1	AP3S2	AP4B1	**AP4E1**	AP4M1	AP4S1
ARSA	ARSB	ARSG	ASAH1	ATP6AP1	ATP6V0A1	ATP6V0A2	ATP6V0A4
ATP6V0B	ATP6V0C	ATP6V0D1	ATP6V0D2	ATP6V1H	CD164	CD63	CD68
CLN3	CLN5	CLTA	CLTB	CLTC	CLTCL1	CTNS	CTSA
CTSB	CTSC	CTSD	CTSE	CTSF	CTSG	CTSH	CTSK
CTSL	CTSO	CTSS	CTSV	CTSW	CTSZ	DNASE2	DNASE2B
ENTPD4	**FUCA1**	GAA	GALC	GALNS	GBA	GGA1	GGA2
GGA3	GLA	GLB1	GM2A	GNPTAB	GNPTG	GNS	GUSB
HEXA	HEXB	HGSNAT	HYAL1	IDS	IDUA	IGF2R	LAMP1
LAMP2	LAMP3	LAPTM4A	LAPTM4B	LAPTM5	LGMN	LIPA	M6PR
MAN2B1	MANBA	MCOLN1	MFSD8	NAGA	NAGLU	NAGPA	NAPSA
NEU1	NPC1	NPC2	PLA2G15	PPT1	PPT2	PSAP	PSAPL1
SCARB2	SGSH	SLC11A1	SLC11A2	SLC17A5	SMPD1	SORT1	SUMF1
TCIRG1	TPP1						

Table 23: 18 Genes in the other glycan degradation SuperPath

AGA	ENGASE	**FUCA1**	FUCA2	GBA	GBA2	GLB1	HEXA
HEXB	HEXDC	MAN2B1	MAN2B2	MAN2C1	MANBA	NEU1	NEU2
NEU3	NEU4						

Table 24: 291 Genes in the TGF-Beta Pathway

AKT1	ANGPT1	ANGPT2	ANGPT4	ANGPTL1	ANGPTL6	AREG	ATF2
BAX	BCL2	BDNF	BDNF-AS	BMP1	BMP10	BMP15	BMP2
BMP3	BMP4	BMP5	BMP6	BMP7	BMP8A	BMP8B	BTC
BTK	CCL1	CCL11	CCL13	CCL14	CCL15	CCL16	CCL17
CCL18	CCL19	CCL2	CCL20	CCL21	CCL22	CCL23	CCL24
CCL25	CCL26	CCL27	CCL28	CCL3	CCL3L1	CCL4	CCL5
CCL7	CCL8	CD2	CD4	CD80	CDH1	CLEC11A	CREBBP
CSF2RB	CTGF	CXCL12	EGF	EGFR	EP300	FGF1	FGF10
FGF11	FGF12	FGF13	FGF14	FGF16	FGF17	FGF18	FGF19
FGF2	FGF20	FGF21	FGF22	FGF23	FGF4	FGF5	FGF6
FGF7	FGF8	FGFR1	FGFR2	FGFR3	FGFR4	FIGF	FLT1
FLT4	FOS	FOXH1	GAS1	GAS2	GAS6	GAS8	GDF10
GDF15	GDF2	GDF5	GDF6	GDF9	GFER	GH1	GH2
GHRH	GMFG	GNRH1	GNRH2	GPC6	GRB2	HDGF	HGF
HRAS	IGF1	IGF2	IGF2-AS	IL10	**IL10RA**	IL10RB	IL11
IL12A	IL12B	IL13	IL15	IL16	IL17A	IL17B	IL17RC
IL18	IL18R1	IL19	IL1A	IL1B	IL1R1	IL2	IL20
IL22	IL23A	IL24	IL26	IL2RA	IL2RB	IL2RG	IL3
IL3RA	IL4	IL4R	IL5	IL6	IL6R	IL7	IL7R
IL8	IL9	IL9R	JUN	KDR	KITLG	KLF10	KRAS
LEFTY1	LEFTY2	LHB	LTBP1	LTBP2	LTBP3	LTBP4	MAP2K1
MAP2K2	MAP2K3	MAP2K4	MAP2K5	MAP2K6	MAP3K7	MAP4K1	MAPK1
MAPK10	MAPK11	MAPK12	MAPK13	MAPK14	MAPK3	MAPK8	MAPK9
MDK	MET	MMDK	MRAS	NGF	NGFR	NRAS	NRG1
NRG2	NRG3	NTF3	NTF4	NUDT6	OGFR	OGN	PDF
PDGFA	PDGFB	PDGFC	PDGFD	PENK	PGF	PIK3R1	PIK3R2
PIK3R3	PIK3R4	PIK3R5	PPAP2B	PPBP	PPH2	PPP2CA	PPP2CB
PPP2R1A	PPP2R1B	PPP2R2A	PPP2R2B	PPP2R2C	PPP2R3A	PPP2R3B	PPP2R4
PPP2R5A	PPP2R5B	PPP2R5C	PPP2R5D	PPP2R5E	PROK1	PSIP1	PTN
RA1	RA2	RHOA	RNMT	RRAS	RRAS2	S1PR1	SAR1A
SAR1B	SEMA3A	SEMA4B	SEMA4G	SEMA5A	**SKI**	SKIL	SLTM
SMAD1	SMAD2	SMAD3	SMAD4	SMAD5	SMAD6	SMAD7	SMAD9

SOS1	SOS2	SPON1	TAB1	TBRG1	TDGF1	TDGF1P3	TGFA
TGFB1	TGFB2	TGFB3	TGFBR1	TGFBR2	THPO	TIMP1	TIMP2
TIMP3	TIMP4	TNFRSF10A	TNFRSF10B	TNFRSF10C	TNFRSF10D	TNFRSF13C	TNFRSF17
TNFRSF18	TNFRSF1A	TNFRSF1B	TNFRSF25	TNFSF10	TNFSF13B	TYMP	UTS2
VEGFB	VEGFC	XIAP					

Table 25: 40 Autosomal Recessive Intellectual Disability Genes (ARID)

PRSS12	CRBN	CC2D1A	GRIK2	TUSC3	TRAPPC9	ZC3H14	MED23
ADK	ADRA2B	ASCC3	ASCL1	C11orf46	TTI2	RABL6	CASP2
CCNA2	COQ5	CRADD	EEF1B2	ELP2	ENTPD1	FASN	HIST3H3
INPP4A	KIAA1033	MAN1B1	NDST1	PECR	PRMT10	PRRT2	RALGDS
RGS7	SCAPER	ST3GAL3	TECR	TRMT1	UBR7	ZCCHC8	ZNF526

Table 26: 12 Autosomal Dominant Intellectual Disability Genes (ADID)

CDH15	KIRREL3	SYNGAP1	GRIN2B	DYRK1A	DOCK8	GRIN1	KIF1A
CACNG2	EPB41L1	DYNC1H1	TSPAN7				

APPENDIX B
THE PERL SCRIPT USED IN FILTERING STEPS

```
#!/usr/bin/perl
use warnings;
use strict;
use List::Util 'sum';

#check that user-supplied arguments are correct
my $arguments = @ARGV;
if ($arguments != 1){
        print "Error: Incorrect arguments. Example usage\n";
        print "\nperl all.pl file1.txt\n";
        die;
}

#declare universal variables
##files
our ($fileA, $fileB)=($ARGV[0], $ARGV[1]);
##cols #note this should be flexible once column names stabilize
my ($pQual, $sQual, $fQual, $mQual) = (27, 52, 46, 49);
my ($pUA, $sUA, $fUA, $mUA) = (56, 68, 60, 64);
my ($pUV, $sUV, $fUV, $mUV) = (57, 69, 61, 65);
my ($pUP, $sUP, $fUP, $mUP) = (59, 71, 63, 67);
my ($gene, $effect, $dn, $accession) = (77, 79, 55, 76);
my @toGrab = (0..4, $gene, $accession, $effect, $dn, $pQual, $sQual, $fQual, $mQual,
$pUA, $sUA, $fUA, $mUA, $pUV, $sUV, $fUV, $mUV, $pUP, $sUP, $fUP, $mUP);

#trim to high quality variants
my ($lofA, $missenseA, $lofHomoA, $missenseHomoA, $lofDNA, $missenseDNA,
$lofHet, $missenseHet, $countA) = trim($fileA);
```

```
print "\n";
print "There are ".(scalar(@$lofA)-1)." high quality LoF mutations and
".(scalar(@$missenseA)-1)." high quality missense mutations out of ".$countA." in
$fileA\n";
print ((scalar(@$lofHomoA)-1)." of the high quality LoF mutations are homozygous
and ".(scalar(@$missenseHomoA)-1). " of the high quality missense mutations are
homozgyous\n");
print ((scalar(@$lofDNA)-1)." of the high quality LoF mutations are de novo and
".(scalar(@$missenseDNA)-1). " of the high quality missense mutations are de
novo\n\n");

print ((scalar(@$lofHet) -1)." of the high quality LoF mutations are heterozygous and
".(scalar(@$missenseHet) -1)." of the high quality missense mutations are
heterozygous\n");


printToFile($lofHomoA, "sharedHomozygousLoF.txt");
printToFile($missenseHomoA, "sharedHomozygousMissense.txt");
printToFile($lofDNA, "sharedDNLoF.txt");
printToFile($missenseDNA, "sharedDNMissense.txt");
printToFile($lofHet, "hetLoF.txt");
printToFile($missenseHet, "hetMissense.txt");

my @compoundHetLoF = compoundHet($lofHet);
my @compoundHetMissense = compoundHet($missenseHet);
printToFile(\@compoundHetLoF, "compoundHetLoF.txt");
printToFile(\@compoundHetMissense, "compoundHetMissense.txt");
```

```
print ((scalar(@compoundHetLoF) -1)." of the high quality LoF mutations are
compound heterozygous and ".(scalar(@compoundHetMissense) -1)." of the high
quality missense mutations are compound heterozygous\n");

print "\n";

#subroutines

#this subroutine trims to high quality variants and separates by LoF versus missense
sub trim {
        open(IN, "<", @_) or die "Could not open @_\nError:\n$!";
        my (@lof, @missense, @lofHomo, @missenseHomo, @lofDN, @missenseDN,
@lofHet, @missenseHet);
        my $counter=0;
        while (my $line = <IN>) {
                $line =~ s/\n|\r//g;
                my @split_line = split /\t/, $line;
                if ($counter == 0){
                        push @lof, $line;
                        push @lofHomo, $line;
                        push @missense, $line;
                        push @missenseHomo, $line;
                        push @lofDN, $line;
                        push @missenseDN, $line;
                        push @lofHet, $line;
                        push @missenseHet, $line;
                        $counter++;
                } else {
                        #skip any lines where any entry is "."
                        my @params=@split_line[$pQual, $sQual, $pUA, $sUA, $fUA,
$mUA, $pUV, $sUV, $fUV, $mUV, $pUP, $sUP, $fUP, $mUP, $gene, $effect, $dn];
```

```
            $counter++;
            next if grep /^\.$/, @params;
            #split into LoF mutations versus missense mutations
            if ( $split_line[$pQual] >=100 && $split_line[$sQual] >=100 &&
$split_line[$pUA] >= 20 && $split_line[$sUA] >=20 && $split_line[$fUA] >=20 &&
$split_line[$mUA] >=20 && $split_line[$pUV] >= 8 && $split_line[$sUV] >=8 ){
                if ($split_line[$effect] =~
/Start_codon|Stop_codon|Nonsense|Canonical|Frameshift/){
                    push @lof, $line;
                    #capture homozygous LoF
                    if ( $split_line[$dn] !~ /Y/ &&
$split_line[$pUP]>=70 && $split_line[$sUP]>=70 && $split_line[$fUP]>=30 &&
$split_line[$fUP]<=70 && $split_line[$mUP]>=30 && $split_line[$mUP]<=70 &&
$split_line[$fUV] >=8 && $split_line[$mUV] >=8 ){
                        push @lofHomo, $line;
                    }
                    #capture de novo LoF
                    if ( $split_line[$dn] =~ /Y/ &&
$split_line[$pUP]>=30 && $split_line[$pUP]<=70 && $split_line[$sUP]>=30 &&
$split_line[$sUP]<=70 && $split_line[$fUP]<=30 && $split_line[$mUP]<=30 ) {
                        push @lofDN, $line;
                    }
                    #capture heterozygous LoF
                    if ( $split_line[$pUP]>=30 &&
$split_line[$pUP]<=70 && $split_line[$sUP]>=30 && $split_line[$sUP]<=70 &&
$split_line[$fUP]<=70 && $split_line[$mUP]<=70 ) {
                        push @lofHet, $line;
                    }
                } elsif ($split_line[$effect] =~ /Missense/){
                    push @missense, $line;
                    #capture homozygous missense
                    if ( $split_line[$dn] !~ /Y/ &&
$split_line[$pUP]>=70 && $split_line[$sUP]>=70 && $split_line[$fUP]>=30 &&
$split_line[$fUP]<=70 && $split_line[$mUP]>=30 && $split_line[$mUP]<=70 &&
$split_line[$fUV] >=8 && $split_line[$mUV] >=8 ){
                        push @missenseHomo, $line;
                    }
                    #capture de novo missense
                    if ( $split_line[$dn] =~ /Y/ &&
$split_line[$pUP]>=30 && $split_line[$pUP]<=70 && $split_line[$sUP]>=30 &&
$split_line[$sUP]<=70 && $split_line[$fUP]<=30 && $split_line[$mUP]<=30) {
                        push @missenseDN, $line;
                    }
                    #capture heterozygous missense
                    if ( $split_line[$pUP]>=30 &&
$split_line[$pUP]<=70 && $split_line[$sUP]>=30 && $split_line[$sUP]<=70 &&
$split_line[$fUP]<=70 && $split_line[$mUP]<=70 ) {
                        push @missenseHet, $line;
                    }
                }
            }
        }
    }
    close IN;
    return (\@lof, \@missense, \@lofHomo, \@missenseHomo, \@lofDN,
\@missenseDN, \@lofHet, \@missenseHet, $counter-1);
}

#this subroutine prints an array reference to a file
sub printToFile{
    my($arrayRef, $fileName) = @_;
    if (scalar(@$arrayRef)>1) {
        open(OUT, ">", $fileName) or die "Could not open $_[1]\nError:\n$!";
        foreach my $item (@$arrayRef){
            my @toPrint = split /\t/, $item;
            print OUT join("\t", @toPrint[@toGrab]), "\n";
        }
    }
}

#this subroutine finds compound heterozygous mutations
sub compoundHet{
    my @array = @{$_[0]};
    #make hash with unique ID --> line #
    my ($gene_id_ref, $gene_location_ref, $geneList_ref) = mergeline(@array);
    my %gene_id = %$gene_id_ref;
    my %gene_location = %$gene_location_ref;
    my @geneList = @$geneList_ref;
    my $counter = 1; #no header after mergeline subroutine
    my @finalArray;
    my @hetGenes;
    push @finalArray, $array[0];
    foreach my $tmpGene (@geneList){
        my %properHetFather;
        my %properHetMother;
        my @split_value = split /,/, $gene_id{$tmpGene};
        if ( scalar(@split_value) > 1 ) {
            if (scalar uniq(@split_value) == 2 ){
                my @geneEntries=@array[split(/,/,
$gene_location{$tmpGene})];
                foreach my $entry (@geneEntries){
                    my $counter2 = 0;
                    my @split_entry=split(/\t/, $entry);
```

```
                                        if ( ($split_entry[$fUP] >=30 &&
$split_entry[$mUP] <= 30) ) {
                                                if ( !exists
$properHetFather{$split_value[$counter2]} ){

        $properHetFather{$split_value[$counter2]}=1;
                                                }
                                        }
                                        if ( ($split_entry[$fUP] <=30 &&
$split_entry[$mUP] >= 30) ) {
                                                if ( !exists
$properHetMother{$split_value[$counter2]} ){

        $properHetMother{$split_value[$counter2]}=1;
                                                }
                                        }
                                        $counter2++;
                                }
                                if ( sum values %properHetFather > 0 && sum values
%properHetMother > 0 ){
                                        push @finalArray, @array[split(/,/,
$gene_location{$tmpGene})];
                                        push @hetGenes, $tmpGene;
                                }
                        }
                }
        $counter++;


        }
        print "There are ".scalar(@hetGenes)." compound heterozgous mutations\n";
        return(@finalArray)
```

```
}

#merges first 5 columns (chr-start-stop-ref-alt) to create a unique ID for each variant.
sub mergeline {
        my $counter=0;
        my (%gene_lines, %gene_uniqPos, @genes);
        foreach my $line (@_){
                unless ( $counter == 0 ) {
                        $line =~ s/\n|\r//gi;
                        my @split_line = split(/\t/, $line);
                        my $tmp = join ("_", @split_line[0..4]);
                        if ( exists ($gene_uniqPos{$split_line[$gene]}) ){
                                $gene_uniqPos{$split_line[$gene]}=join(",",
($gene_uniqPos{$split_line[$gene]}, $tmp));
                        } else {
                                $gene_uniqPos{$split_line[$gene]}=$tmp;
                        }
                        if ( exists ($gene_lines{$split_line[$gene]}) ) {
                                $gene_lines{$split_line[$gene]}=join(",",
($gene_lines{$split_line[$gene]}, $counter));
                        } else {
                                $gene_lines{$split_line[$gene]}=$counter;
                        }
                        push @genes, $split_line[$gene];

                }
                $counter++;
        }
        @genes=uniq(@genes);
        return (\%gene_uniqPos, \%gene_lines, \@genes);
}
```

```
#returns uniq elements from an array
sub uniq {
  my %seen;
  return grep { !$seen{$_}++ } @_;
}
```

APPENDIX C
ADDITIONAL PROJECTS

1. Fuchs' Endothelial Corneal Dystrophy (FECD)

Subjects and clinical descriptions

Four families with Fuchs' Endothelial Corneal Dystrophy (FECD) were studied. All study subjects underwent a complete ophthalmic examination, including confocal microscopy and contact ultrasound pachymetry by a cornea fellowship-trained ophthalmologist. After informed consents were acquired, 3-mL EDTA blood samples were collected from the subjects. The genomic DNA was subsequently extracted following the manufacturer's instructions (Qiagen Inc, Valencia, CA).

Excluding known genes by Sanger sequencing and linkage analysis

Mutation screening for *COL8A2* was performed using PCR and Sanger sequencing of the entire coding region of the gene. No mutation was found in any subjects. We then performed linkage analysis with 3 loci in all families using microsatellite markers. A total of 12 microsatellite markers spanning around the *COL8A2* gene (D1S496, D1S2729, D1S186, D1S2892), the *SLC4A11* gene (D20S117, D20S842, D20S889, D20S482) and the *TCF8* gene (D10S1214, D10S1426, D10S208, D10S1208) were analyzed and revealed that they were unsegregated with the disease. The informative markers were showed in figures 19 to 22.

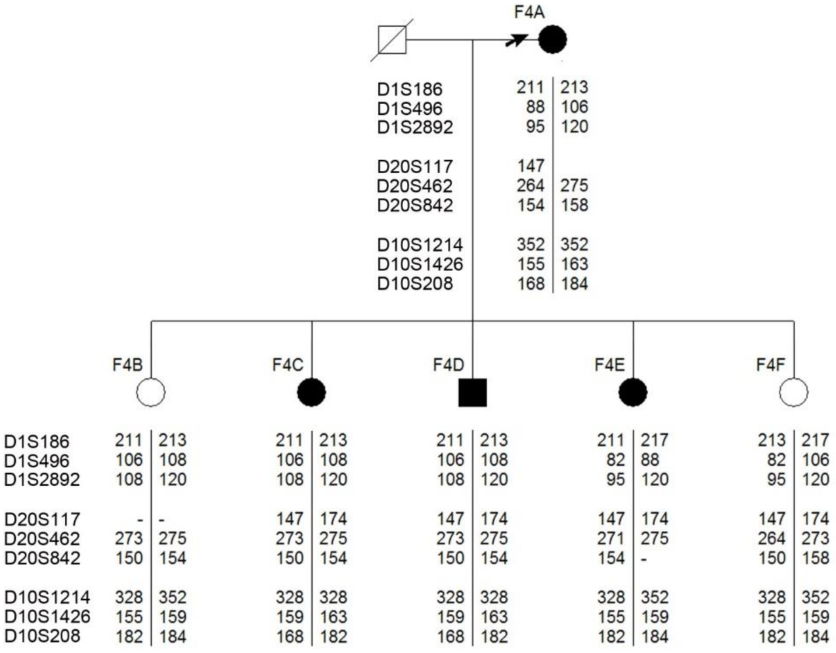

F4A
D1S186 211 | 213
D1S496 88 | 106
D1S2892 95 | 120
D20S117 147
D20S462 264 | 275
D20S842 154 | 158
D10S1214 352 | 352
D10S1426 155 | 163
D10S208 168 | 184
F4B
F4C
F4D
F4E
F4F
D1S186
D1S496
D1S2892
D20S117
D20S462
D20S842
D10S1214
D10S1426
D10S208
F4B: 211 | 213, 106 | 108, 108 | 120, - | -, 273 | 275, 150 | 154, 328 | 352, 155 | 159, 182 | 184
F4C: 211 | 213, 106 | 108, 108 | 120, 147 | 174, 273 | 275, 150 | 154, 328 | 328, 159 | 163, 168 | 182
F4D: 211 | 213, 106 | 108, 108 | 120, 147 | 174, 273 | 275, 150 | 154, 328 | 328, 159 | 163, 168 | 182
F4E: 211 | 217, 82 | 88, 95 | 120, 147 | 174, 271 | 275, 154 | -, 328 | 352, 155 | 159, 182 | 184
F4F: 213 | 217, 82 | 106, 95 | 120, 147 | 174, 264 | 273, 150 | 158, 328 | 352, 155 | 159, 182 | 184

Figure 19: Pedigree of FECD family 1 with genotypes of the nine informative markers of the three loci

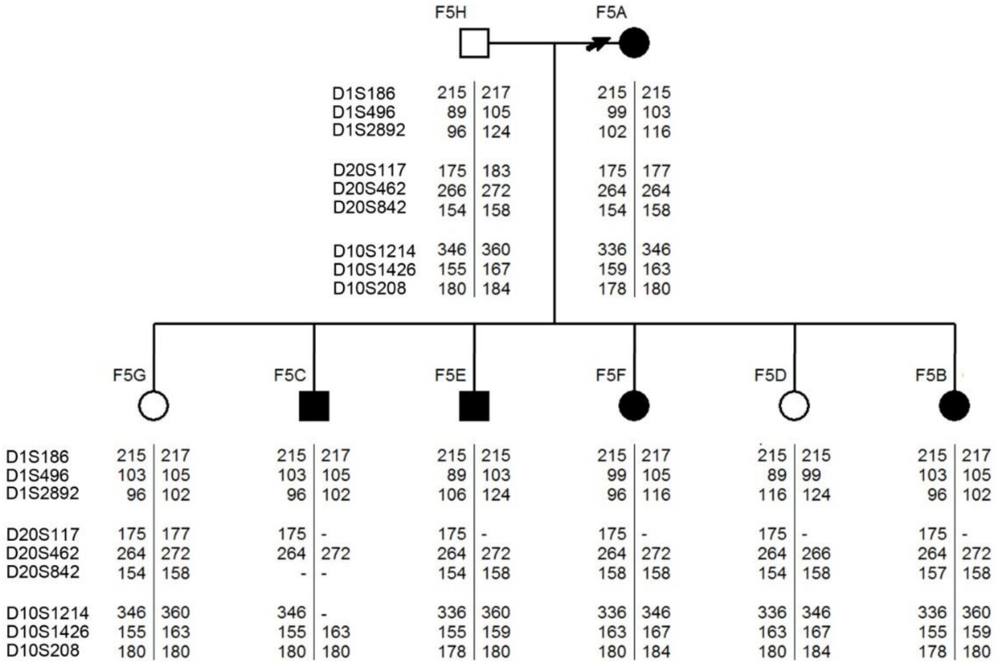

F5H
F5A
D1S186
D1S496
D1S2892
D20S117
D20S462
D20S842
D10S1214
D10S1426
D10S208
F5H: 215 | 217, 89 | 105, 96 | 124, 175 | 183, 266 | 272, 154 | 158, 346 | 360, 155 | 167, 180 | 184
F5A: 215 | 215, 99 | 103, 102 | 116, 175 | 177, 264 | 264, 154 | 158, 336 | 346, 159 | 163, 178 | 180
F5G
F5C
F5E
F5F
F5D
F5B
F5G: 215 | 217, 103 | 105, 96 | 102, 175 | 177, 264 | 272, 154 | 158, 346 | 360, 155 | 163, 180 | 180
F5C: 215 | 217, 103 | 105, 96 | 102, 175 | -, 264 | 272, - | -, 346 | -, 155 | 163, 180 | 180
F5E: 215 | 215, 89 | 103, 106 | 124, 175 | -, 264 | 272, 154 | 158, 336 | 360, 155 | 159, 178 | 180
F5F: 215 | 217, 99 | 105, 96 | 116, 175 | -, 264 | 272, 158 | 158, 336 | 346, 163 | 167, 180 | 184
F5D: 215 | 215, 89 | 99, 116 | 124, 175 | -, 264 | 266, 154 | 158, 336 | 346, 163 | 167, 180 | 184
F5B: 215 | 217, 103 | 105, 96 | 102, 175 | -, 264 | 272, 157 | 158, 336 | 360, 155 | 159, 178 | 180

Figure 20: Pedigree of FECD family 2 with genotypes of the nine informative markers of the three loci

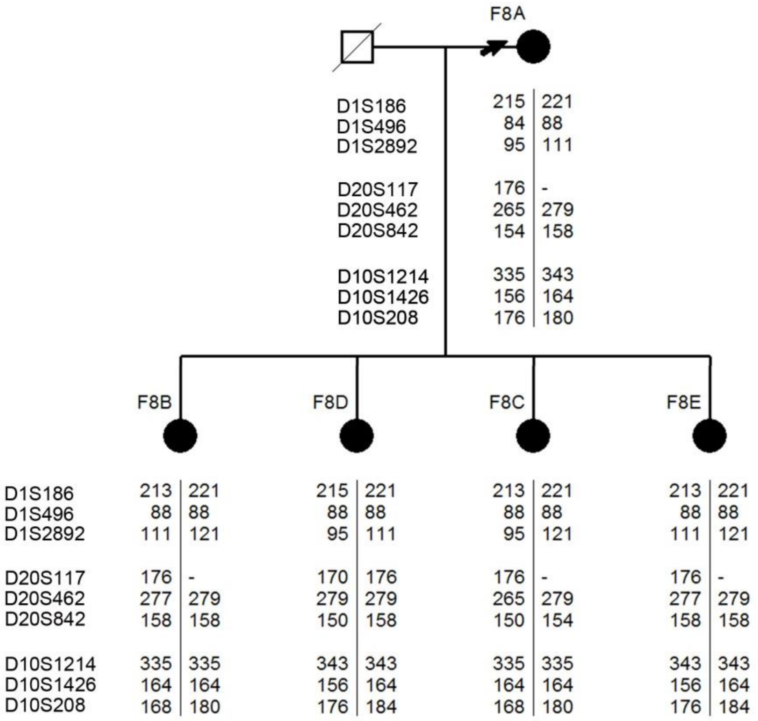

F8A
D1S186 215 | 221
D1S496 84 | 88
D1S2892 95 | 111
D20S117 176 | -
D20S462 265 | 279
D20S842 154 | 158
D10S1214 335 | 343
D10S1426 156 | 164
D10S208 176 | 180
F8B
F8D
F8C
F8E
D1S186 213 | 221 215 | 221 213 | 221 213 | 221
D1S496 88 | 88 88 | 88 88 | 88 88 | 88
D1S2892 111 | 121 95 | 111 95 | 121 111 | 121
D20S117 176 | - 170 | 176 176 | - 176 | -
D20S462 277 | 279 279 | 279 265 | 279 277 | 279
D20S842 158 | 158 150 | 158 150 | 154 158 | 158
D10S1214 335 | 335 343 | 343 335 | 335 343 | 343
D10S1426 164 | 164 156 | 164 164 | 164 156 | 164
D10S208 168 | 180 176 | 184 168 | 180 176 | 184

Figure 21: Pedigree of FECD family 3 with genotypes of the nine informative markers of the three loci

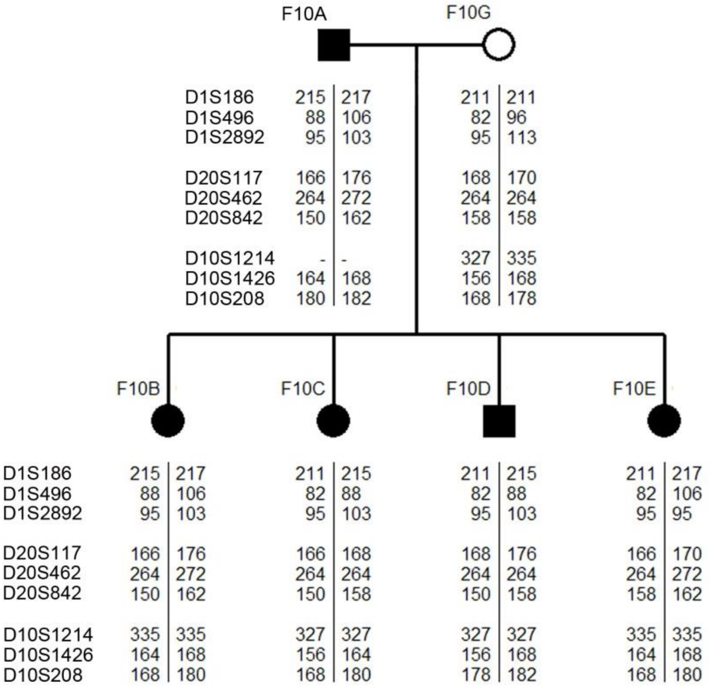

F10A
F10G
D1S186 215 | 217 211 | 211
D1S496 88 | 106 82 | 96
D1S2892 95 | 103 95 | 113
D20S117 166 | 176 168 | 170
D20S462 264 | 272 264 | 264
D20S842 150 | 162 158 | 158
D10S1214 - | - 327 | 335
D10S1426 164 | 168 156 | 168
D10S208 180 | 182 168 | 178
F10B
F10C
F10D
F10E
D1S186 215 | 217 211 | 215 211 | 215 211 | 217
D1S496 88 | 106 82 | 88 82 | 88 82 | 106
D1S2892 95 | 103 95 | 103 95 | 103 95 | 95
D20S117 166 | 176 166 | 168 168 | 176 166 | 170
D20S462 264 | 272 264 | 264 264 | 264 264 | 272
D20S842 150 | 162 150 | 158 150 | 158 158 | 162
D10S1214 335 | 335 327 | 327 327 | 327 335 | 335
D10S1426 164 | 168 156 | 164 156 | 168 164 | 168
D10S208 168 | 180 168 | 180 178 | 182 168 | 180

Figure 22: Pedigree of FECD family 4 with genotypes of the nine informative markers of the three loci

Whole Exome Sequencing

We performed whole exome sequencing in 14 individuals including 3 affected individuals and 1 unaffected individual from family 1 (F4A, F4C, F4D and F4B), 3 affected individuals and 1 unaffected individual from family 2 (F5A, F5C, F5E and F5G), 3 affected individuals from family 3 (F8A, F8B, F8D) and 3 affected individuals from family 4 (F10A, F10B, F10C). The genomic DNA was captured and enriched by the Agilent SureSelect Human All Exon Capture kit (Agilent Technologies, Santa Clara, CA). The enriched DNA library was subsequently sequenced using a pair-end 100 bp configuration on the Hiseq 2000 platform (Illumina, San Diego, CA).

Reads were aligned to NCBI human genome build v37 g1k using BWA software (http://bio-bwa.sourceforge.net/). Duplicate reads were removed using Picard Tools. Variants were called using GATK and annotated using an in-house script.

WES data analysis

To identify disease-causing variants under the assumption of an autosomal dominant pattern of inheritance, sequencing results were analyzed with 3 filtering steps. First, homozygous variants, synonymous variants, and variants locating outside exons and their flanking regions were excluded. Secondly, all inherited variants, variants found in all three affected but absent in the unaffected individuals (available for families 1 and 2) from each family were selected. The remaining variants were filtered for SNPs quality higher than 100 and grouped into loss of function and missense groups in each family. The number of variants during steps of filtering is shown in Table 27.

Table 27: The number of variants during steps of filtering

Family	Individual	All detected variant	Heterozygous variant located in exonic or splice site	Inherited and Quality >100	
				LoF	Missense
1	F4A	211,437	8,166	25	405
	F4B	199,962	8,258		
	F4C	196,088	8,222		
	F4D	164,630	6,207		
2	F5A	249,574	10,074	31	461
	F5C	180,688	7,899		
	F5E	183,506	7,865		
	F5G	170,604	7,772		
3	F8A	201,530	8,253	166	2,023
	F8B	200,904	8,286		
	F8D	185,461	7,983		
4	F10A	172,928	7,764	137	2,070
	F10B	167,833	7,829		
	F10C	189,073	7,837		

We then prioritized the remaining variants in each group to be candidate variants. The remaining variants in each group were searched for 1) variants locating in any known genes; 2) variants in genes encoded known or predicted protein interacting with known genes (http://string-db.org/); 3) variants located in known loci and 4) variants shared between families. All variants with minor allele frequency < 0.3 in dbSNP, NHLBI Exome Sequencing Project (ESP), The Exome Aggregation Consortium (ExAC) and our in-house exome database were excluded.

Confirmation of candidate variants by PCR and sequencing

All candidate variants were PCR and sequenced to confirm the existence of the variant. Cosegregation analysis was subsequently performed. Only two variants located in a known locus (FCD3; 5q33.1-q35.2) found in family 1 were confirmed and showed co-segregation with the disease (Table 28).

Table 28: Two candidate variants identified in FECD family 1

CHROM	POS	REF.	ALT.	Gene	Effect	AA_Change	BP_Change	120 Thai MAF.	ExAC MAF.
5	150911154	G	A	FAT2	Coding:Missense	R3269C	C9805T	0	0.0001931
5	154173395	C	A	LARP1	Coding:Missense	P302T	C904A	0	0.000008264

Confirm linkage using microsatellite markers

To confirm linkage of the FCD3 locus, 5 microsatellite markers from ABI prism mapping set located in 5q33.1-q35.2 were genotyped in family F1. The results are shown in Table 29.

Table 29: Genotyping results of 5 microsatellite markers and the 2 identified variants in FECD family 1

Chr	Start	Stop	Marker, Variant	Genotyping result
5	134719248	134719548	D5S2115	cosegregated
5	145203918	145204281	D5S436	cosegregated
5	150911154	150911154	FAT2, G>A	cosegregated
5	152774975	152775361	D5S410	uninformative
5	154173395	154173395	LARP1, C>A	cosegregated
5	162153859	162154179	D5S422	uninformative
5	168442870	168443153	D5S400	uninformative

Genotyping for the two candidate variants by pyrosequencing

We then performed genotyping for the identified variants in *FAT2* and *LARP2* in our 89 cohorts. The *FAT2*, G>A variant was found in another family with two affected members.

The *FAT2* gene (FAT Atypical Cadherin 2) is a human homolog of Drosophila that encodes 4,349 amino acid of a tumor suppressor which is involved in tumor suppression and planar cell polarity (PCP) and essential for controlling cell proliferation during Drosophila development.

2. Fucosidosis

Subjects and clinical descriptions

The subject was a Thai male, first born of a non-consanguineous marriage with uneventful perinatal history. His milestones were reportedly normal until age 2. He started to have unsteady gait with toe walking. The clinical features of neurodegeneration with coarse facies, skeletal change, angiokeratoma and neuroimaging are suggestive of lysosomal storage disease including fucosidosis, mucopolysacharidosis type IH or Hurler syndrome, mucopolysacharidosis type II: α and β Mannosidosis caused by mutations in *FUCA, IDUA, IDS, MAN2B1* and *MANBA* respectively.

Whole exome sequencing:

Patient's genomic DNA was captured and enriched by Agilent SureSelect Human All Exon Capture kit (Agilent Technologies, Santa Clara, CA). The enriched DNA library was subsequently sequenced using a pair-end 100 bp configuration on Hiseq 2000 platform (Illumina, San Diego, CA). Sequence reads were aligned to the reference genome UCSC hg19 (http://genome.ucsc.edu/) by BWA software (http://bio-bwa.sourceforge.net/). The SNPs and Indels were detected by SAMTOOLS (http://samtools.sourceforge.net/). dbSNP & 1000G were used as variant databases.

To identify a causative variant for the disease, we filtered for homozygous and heterozygous variants located in exonic region of FUCA1gene. The variants found in dbSNPs database, ExAC database and our in-house exome database were excluded.

There were 5 variants in *FUCA1* gene showing in Table 30. The first one is most likely be a pathogenic variant since it caused frameshift.

Table 30: Variants detected in *FUCA1* gene

chr.	chr_start	chr_end	Ref.	Alt.	region	change
chr01	**24186386**	**24186386**	**G**	**-**	**exonic**	**frameshift_deletion**
chr01	24192200	24192200	A	G	intronic	.
chr01	24194748	24194748	G	C	exonic	nonsynonymous_SNV
chr01	24194773	24194773	G	A	exonic	nonsynonymous_SNV
chr01	24194788	24194788	A	G	UTR5	

PCR and Sanger sequencing:

To confirm the identified variant, we performed PCR and Sanger sequencing for the entire sequence of exon4 and flanking sequence using forward primer (5’-AAGGGAGCCAGGGAAGATTA-3’) and reverse primer (5’-GGAGGCTTAGGCAAGAGGAT-3’) with touchdown PCR program. The results show in figure 23. The patient has a hemizygous variant inherited from his father. Mother’s chromatogram was normal.

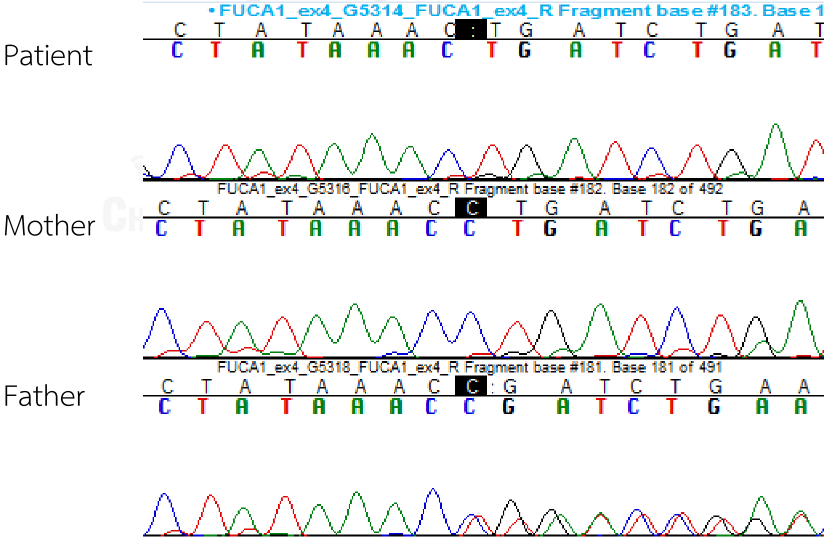


Figure 23: Sanger sequencing results

CGH array:

As only one mutant allele was detected, the results suggested the possibility of a large deletion occurring in the other allele. Further experiments using array comparative genomic hybridization (CGH) covering chromosome 1 using the Agilent Human Genome CGH Microarray.

CGH array results revealed a deletion in *FUCA1* gene 5 containing probes with mean log2ratio of -0.696971 at chr1:24187450-24189418.

VITA

Name	Ms. Wipa Panmontha
Date of birth	December 20th, 1988
Place of birth	Bangkok, Thailand

Education

She got her bachelor degree with a second class honor in Medical Technology from Faculty of Allied Health Science, Chulalongkorn University in 2010. She then got a Royal Golden Jubilee (RGJ) Ph.D. Scholarship from the Thailand Research Fund (TRF) and participated in Medical Sciences Ph.D. program, Faculty Medicine, Chulalongkorn University since 2011.

Research Grants

1. The Royal Golden Jubilee Ph.D. Program, Thailand Research Fund (TRF)

Publications

1. Panmontha W, Rerknimitr P, Yeetong P, Srichomthong C, Suphapeetiporn K, Shotelersuk V. A Frameshift Mutation in PEN-2 Causes Familial Comedones Syndrome. Dermatology. 2015:77-81.

2. Rerknimitr P, Korkij W, Wititsuwannakul J, Panmontha W, Suphapeetiporn K, Shotelersuk V. Expanding phenotypic spectrum of familial comedones. Dermatology. 2014;228(3):215-9.

3. Utokpat P, Panmontha W, Tongkobpetch S, Suphapeetiporn K, Shotelersuk V. Novel CTSK mutation resulting in an entire exon 2 skipping in a Thai girl with pycnodysostosis. Pediatrics international: official journal of the Japan Pediatric Society. 2013;55(5):651-5.