UPPSALA
UNIVERSITET

Identification and validation of de novo point mutations in patients with intellectual disability

Jin
Zhao

Degree project in biology, Bachelor of science, 2011
Examensarbete i biologi 15 hp till kandidatexamen, 2011
Biology Education Centre and Immunology, Genetics and Pathology, IGP, Uppsala University
Supervisors: Lars Feuk and Ammar Zaghlool

Introduction

Intellectual disability is a common disease that affects more than 1% of the human population[1]. Injury and infection could cause intellectual disability[2], but most cases are caused by genetic mutations[3]. These mutations range from large cytogenetic aberrations to smaller copy number variations and point mutations[4,5,6]. Intellectual disability is usually impossible to cure, as the damage has been done during the development of the child. Nevertheless, there is still great interest in identifying the causes for this disorder. It comforts the whole family to find out the cause of the disease, anticipate the development of the disease, and administer in suitable treatment programs. Identification of the genetic cause also enables family planning and allows other family members to test for carrier status. Finding new causal mutations and investigating their functional effect could generate valuable biological and medical knowledge regarding the processes involved in the normal development of the human brain.

Many genes have been identified to be implicated in intellectual disability by linkage analysis, homozygosity mapping, study of cytogenetic abnormalities, and array-comparative genomic hybridization[3]; however, the causes are still unknown for most patients[7]. Linkage analysis discovers inherited mutations in multiple family members, homozygosity mapping is used to locate recessive alleles only, cytogenetic and array based study are limited to structural variations. One of the possible genetic causes of intellectual disability is de novo dominant point mutation, which is not covered by any of the methods mentioned above.

There are several reasons that de novo point mutations could be causative in many cases of undiagnosed intellectual disability. Firstly, the human genome has a per-generation mutation rate of approximately 12.8×10^{-9} per site per generation, resulting in approximately 0.86 new amino-acid-altering mutation, higher than any other well-studied species[8]. Secondly, the large number of genes associated with intellectual disability makes de novo mutation an important cause of the disease. De novo mutations as the cause of intellectual disability also explain why intellectual disability remains a common disease even though the patients have very low fecundity.

De novo point mutations have been extremely difficult to identify until the development of massively parallel sequencing and the possibility of whole genome sequencing or exome sequencing[9]. Numerous de novo point mutations have been found in intellectual disability patients using exome sequencing[10]. The chance of finding de novo point mutations is shown to be more than 50%[10]. Although functional studies are still needed to investigate which of these mutations are causative, exome sequencing is shown to be an effective, powerful, and relatively unbiased way to identify de novo mutations in unexplainable cases and this method is what we used in this project.

In this study, to increase the chance of finding de novo point mutations, exome sequencing was performed on three samples coming from a trio family, two parents and a child. The child suffers from intellectual disability while both parents are healthy and present with negative family histories. The child could not be genetically diagnosed in the clinic, cytogenetic analysis showed no aberration, and no copy number variation associated with intellectual disability has been found by microarray analysis.

Exome sequencing is performed by target capture of the human exome, followed by massively parallel sequencing. The exome capture is done by using biotinylated probes that are complementary to human exons. The probes are first hybridized to fragmented total DNA, and then bound to magnetic streptavidin beads[11]. Unbound DNA fragments are washed away,

and hybridized DNA are eluted and sequenced by SOLiD system. The sequencing results are mapped to human reference sequence (Hg19), and genetic variants are called.

The exome sequencing initially generates huge amounts of variations; in our case around 60,000 variations were called. It is a challenge for bioinformaticians to filter away the variations that are not causing the disease, and narrow the number down to the few most possible candidate mutations. This is done by first excluding variations known to exist in the general population catalogued in dbSNP (Single Nucleotide Polymorphism database), as well as our in house database. Second, the inherited variants from the parents were filtered away. Next, potential false positives generated by artifacts or variations called with low evidence were removed.

The bioinformatics filtering pipeline identified 13 candidate de novo point mutations. The aim of this project was to experimentally investigate which of the candidate mutations is the potential causative de novo mutation in the affected child.

Material and methods

A total of 13 candidate mutation positions located in 13 respective genes were initially identified (Table 1). Primers were designed to flank those positions (Table 2). DNA sequences that contain the candidates were obtained through the UCSC Genome Browser's human reference sequence (Hg19). The DNA sequences were then used to input to Primer3 (v.0.4.0)[12] to obtain primers that flank the candidate mutation position. Primer pairs with low self-complementarity and with melting temperatures around 60°C are chosen. The chosen primer pairs were then checked with both BLAT and in silico PCR functions of UCSC to make sure single specific DNA sequences would be amplified later by PCR.

The family has three members; their genomic DNA samples are numbered as 27692, 27693, and 27694. The three genomic DNA samples were amplified by Multiple Displacement Amplification using the REPLI-g UltraFast Kit from QIAGEN. The amplification results were checked by gel electrophoresis with 0.4% agarose gel.

PCR was performed using DreamTaq™ DNA polymerase or Platinum® PCR SuperMix. Optimization of PCR protocols was performed to obtain the best results while keeping the amplification cycles as low as possible, in order to avoid mis-incorporation during amplification. DreamTaq™ PCR was performed with initial denaturation at 95 °C for three minutes followed by 30 cycles of 95 °C for 30 s, 60 °C for 30 s and 72 °C for three min. The reaction contained 2.5 u of DreamTaq™ polymerase, 200 μM of each of the deoxyribonucleotide triphosphate (dNTPs), primers (Table 2), and approximately 10 ng of DNA. Platinum® PCR was done with initial denaturation at 95 °C for three min. followed by 35 cycles of 95 °C for 30 s, 60 °C for 30 s and 72 °C for three min. The reaction contained Platinum® PCR Supermix, primers (Table 2), and approximately 10 ng of DNA.

PCR products were checked by gel electrophoresis with 1.5% agarose gel, and purified using QIAquick PCR Purification Kit form QIAGEN. The purified PCR products were checked with NanoDrop™ for purity and concentration. The PCR products were sent to Uppsala Genome Center for Sanger sequencing.

The sequencing results were analyzed using Sequencher 5.0 to check for the presence of the mutations in the patients, the absence of the mutations in the parents, as well as other mutations in the sequenced area.

Table 1. The bioinformatics filtering pipeline identified 13 candidate de novo point mutations located in 13 respective genes.

Gene	Full Name	Function
RYK	Receptor-like tyrosine kinase	A novel member of the family of growth factor receptor protein tyrosine kinases[13]
ANXA6	ANNEXIN A6	Ca(2+)-dependent, and is required for acinar cell membrane trafficking events and digestive enzyme secretion[14]
CYP11A1	Cholesterol side-chain cleavage enzyme	Initiates steroidogenesis by converting cholesterol to pregnenolone[15]
FAM5B	Family with sequence similarity 5, member B	Regulates cell cycle and nervous system development[16]
CTNND1	Catenin, Delta-1	A major component of multiprotein cell-cell adhesion complexes[17]
ATP6AP1	Vacuolar ATPase, subunit 1	A proton-ATPase[18]
SLC9A7	Sodium/hydrogen exchanger 7	An electrogenic vacuolar-type hydrogen ATPase[19]
ATG3	Autophagy 3	A protein-conjugating enzyme essential for autophagy[20]
GMPR3	Guanosine monophosphate reductase 2	Reduce Guanosine monophosphate[21]
TGIF1	Transforming growth factor-beta-induced factor	A transcriptional repressor and co-repressor in retinoid and transforming growth factor[22]
SPATA4	Spermatogenesis associated protein 4	Accelerates cell growth, with more cells traversing S-phase and entering G2-phase[23]
NANOG	Homeobox transcription factor	Affects cell differentiation[24]
ARHGAP22	Rho GTPase activating protein 22	A regulator of Rho GTPase[25]

Results

Exome sequencing was performed for all three members of a trio family, which has two healthy parents and a child with intellectual disability. Bioinformatic analysis identified a large number of mutations, of which 13 candidates were chosen for further validation.

Three genomic DNA samples from the trio family were amplified by whole genome amplification (Figure 1) prior to PCR reactions, because the amount of DNA samples is limited. Primer pairs were designed to amplify products spanning the 13 candidate mutation positions (Table 2). PCR reactions were performed (Figure 2) and the amplicons were sequenced by Sanger sequencing.

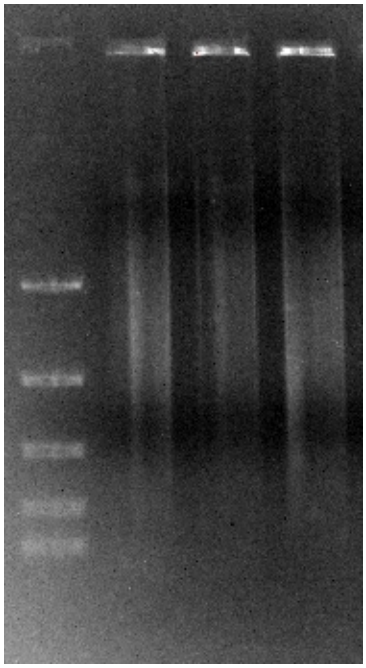

Figure 1. UV photo of 0.4% agarose gel electrophoresis of whole genome amplification for the three samples (the child and the two parents) using the REPLI-g UltraFast Kit from QIAGEN.

Table 2. Results and primers for 13 candidate mutations points.

Gene	Chromosome	Position	Primers	Result
RYK	3	133901869	For:TGATTATCCAGATGTGGGCTTA Rev:GCATTTTCTGCTTTGTTTATGGA	False Positive
ANXA6	5	150515825	For:GCGAGTGTGGAGAGAGGAAG Rev:AAGACAGCCATCCTTGTTGC	False Positive
CYP11A1	15	74659735	For:TGTTGAATTTTGAAATATCCCTGA Rev:GCTGCCAGACCTTTCTGAGT	False Positive
FAM5B	1	177250072	For:TCTTGGGCTGGAGACAGACT Rev:GTCATCTCCAGGGGCTCATA	False Positive
CTNND1	11	57574398	For:GCAGCAGTCCCCATTATCAG Rev:TGCCTTAACAAGCAGTCCTTT	False Positive
ATP6AP1	X	153663672	For:GGTAGGAGCAGAGCTGAGGA Rev:GCAGGGACAGGACTCTCAAC	False Positive
SLC9A7	X	46513124	For:CCTCCATACGGACACTTGCT Rev:CATCATCCGAAACAGAAGCTC	False Positive

ATG3	3	112277230	For:ATGAGAAGTGGCAAACTGACG Rev:TTTTCATGTCTGTTTCCTACATTAGAC	False Positive
GMPR2	14	24707486	For:CAGCAAAAGGAGAAAGCAAAA Rev:CCAGGGACTTTTCCACAGAC	False Positive
TGIF1	18	3457607	For:TGGCTGAGTGAATGAGGAAA Rev:CGTTTGAGTGCAACATCCAC	False Positive
SPATA4	4	177109368	For:CAGCGACTTAGCAGACCACA Rev:AAAACTTACCAAAGACTAACAGGAACA	False Positive
NANOG	12	7945567	For:AATGCATTGGCCACCATTAT Rev:TCCAAGGACAAAACAAGCCTA	False Positive
ARHGAP22	10	49661385	For:GCCTCTTCCATCCACAGAAA Rev:CCCTTGTCCCAGTGACTTTT	False Positive

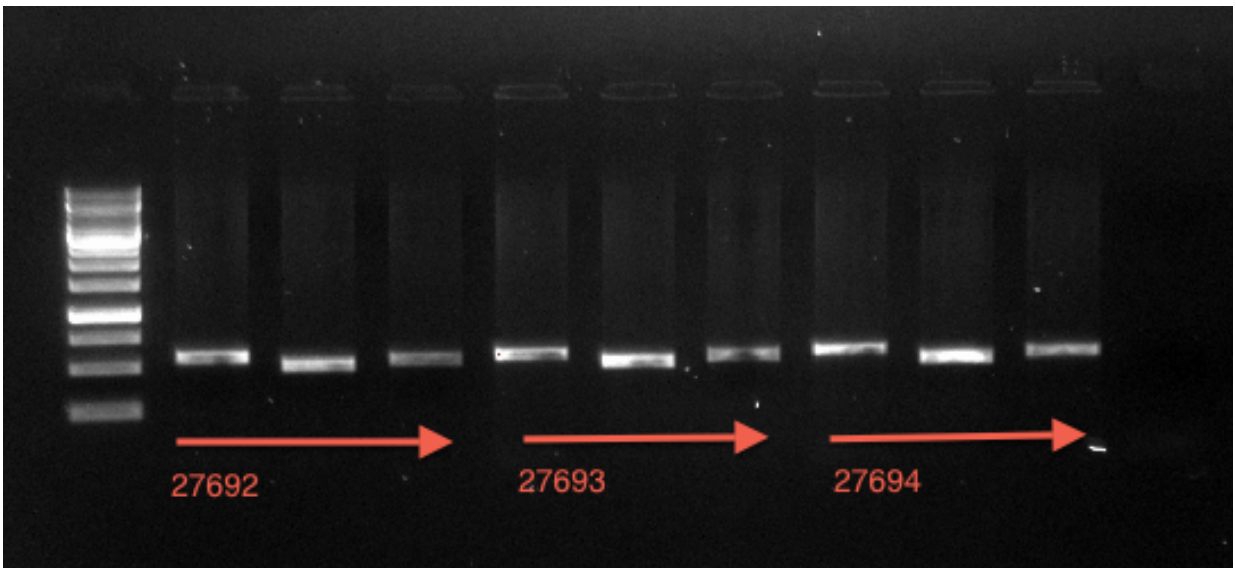

Figure 2. UV photo of 1.5% agarose gel electrophoresis of PCR amplification for three samples, 27692, 27693 and 27694 (the child and the two parents) with three primer pairs using Platinum® PCR SuperMix.

The sequencing results showed all 13 candidate mutations were false positives (Table 2). Besides the 13 positions, there are several mutations found in the sequenced areas but none of them were de novo mutations in the patient. (Figure 3).

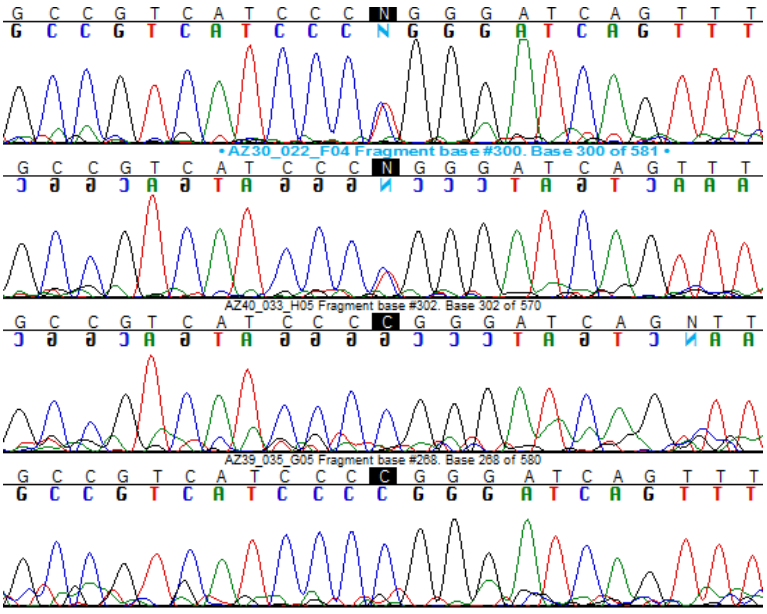

Figure 3. Chromatograms from Sequencher 5.0 showcasing a C to T point mutation form Sanger sequencing results.

Discussion

The overall result is that none of the candidate point mutations could be validated by Sanger sequencing (Table 2), and there are many possible explanations to this. The cause of intellectual disability for this particular patient could be environmental rather than genetic, e.g. by exposure to toxic or harmful agents during early pregnancy. The disorder could also be caused by genetic mutations other than germ line de novo mutation. For example, it could be inherited recessive mutations from both parents, or it could be caused by a somatic mutation in early embryogenesis. A structural de novo mutation could be causing the disease instead of a point mutation. A small insertion/deletion (indel) mutation, for instance, could have escaped detection in both array and sequence-based analysis. A chromosome fragile site could also be causing the disease while being challenging to identify. Even if there is a causative de novo point mutation, it could be positioned outside of the exons, in regulatory sequences or introns. Although the chances are very low, causative de novo point mutations could be filtered out because it already exists in the dbSNP and in house database. Exome sequencing itself does not cover 100% of human exome, and low read coverage from exome sequencing is another pitfall of identification of de novo mutations. The read coverage is usually higher in the middle of an exon and much lower at the edges of an exon, and this creates some biases. The number of probes and length of reads are examples of other limiting factors.

Although de novo point mutations could not be found in this particular family, they have been identified in other family trios. There are also recent publications showing promising result using this approach[9]. Improvements can be made in three steps to increase our chance of identifying causative de novo mutations. The first step is to find better candidate patients for this approach, which involves improvements in both clinical diagnostic and the filtering of causative structural genetic mutations. The second step is to perfect the exome capture technique, for which we need more probes, better coverage and increased read depth. The third step is to follow up with whole genome sequencing when a causative de novo mutations could not be found by exome sequencing. This last step is the most challenging one, because

not only do we need more efficient sequencing techniques, but also a much better understanding of the none coding region of human genome.

With a per-generation mutation rate of approximately 12.8×10^{-9} per site per generation[8], each of us has many de novo point mutations in the non-coding regions of our genome. The impact of those mutations is poorly understood. The sequencing techniques are developing very fast, and we will be able to do whole genome sequencing with much lower cost, but the immense data generated can only be analyzed when we have a much greater knowledge of the human genome.

Acknowledgment

I thank Lars Feuk, Ammar Zaghlool, Jonatan Havardson and Eva Lindholm for their guidance in this project and report writing.

References

1. CDC: Centers for Disease Control and Prevention. State-specific rates of mental retardation – United States, 1993. MMWR Morb Mortal Wkly Rep: 45: 61–65 (1996).
2. CDC: Centers for Disease Control and Prevention. About intellectual disability (2005). from http://www.cdc.gov/ncbddd/dd/ddmr.htm
3. Topper S, Ober C and Das S. Exome sequencing and the genetics of intellectual disability. Clinical Genetics: 80: 117–126. doi: 10.1111/j.1399-0004.2011.01720.x (2011).
4. Leonard H and Wen X. The epidemiology of mental retardation: challenges and opportunities in the new millennium. Ment Retard Dev Disabil Res Rev: 8: 117–134 (2002)
5. de Vries BB, Pfundt R, Leisink M et al. Diagnostic genome profiling in mental retardation. Am J Hum Genet: 77,606–616 (2005).
6. Ropers HH. Genetics of early onset cognitive impairment. Annu Rev Genomics Hum Genet: 11: 161–187 (2010).
7. Rauch A, Hoyer J, Guth S et al. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am J Med Genet A: 140: 2063–2074 (2006).
8. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci USA: 107, 961–968 (2010).
9. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet: 19(R2): R145–R151(2010).
10. Vissers LE, de Ligt J, Gilissen C et al. A de novo paradigm for mental retardation. Nat Genet 2010: 42: 1109–1112.
11. Gnirke A, Melnikov A, Maguire J et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol: 27: 182-189 (2009).
12. Primer3, from http://frodo.wi.mit.edu/primer3/
13. Gough NM, Rakar S, Hovens CM et al. Localization of two mouse genes encoding the protein tyrosine kinase receptor-related protein RYK. Mammalian Genome: 6: 255-256 (1995).
14. Thomas DD, Kaspar KM, Taft WB et al. Identification of annexin VI as a Ca(2+)-

sensitive CRHSP-28-binding protein in pancreatic acinar cells. J. Biol. Chem. 277: 35496-35502 (2002).
15. Katsumata N, Ohtake M, Hojo T et al. Compound heterozygous mutations in the cholesterol side-chain cleavage enzyme gene (CYP11A) cause congenital adrenal insufficiency in humans. J Clin Endocr Metab: 87: 3808-3813 (2002).
16. Takahiro N, Reiko K, Atsushi H et al. Prediction of the Coding Sequences of Unidentified Human Genes. XIX. The complete Sequences of 100 New cDNA Clones from Brain Which Code for Large Proteins in vitro. DNA Res: 7: 347-355 (2000).
17. Matter C, Pribadi M, Liu X et al. Delta-catenin is required for the maintenance of neural structure and function in mature cortex in vivo. Neuron: 64: 320-327 (2009).
18. Supek F, Supekova L, Mandiyan S et al. A novel accessory subunit for vacuolar H(+)-ATPase from chromaffin granules. J Biol Chem: 269: 24102 - 24106 (1994).
19. Numata M, Orlowski J. Molecular cloning and characterization of a novel (Na+,K+)/H+ exchanger localized to the trans-Golgi network. J Biol Chem 276: 17387 - 17394 (2001).
20. Tanida I, Tanida-Miyake E, Komatsu M et al. Human Apg3p/Aut1p homologue is an authentic E2 enzyme for multiple substrates, GATE-16, GABARAP, and MAP-LC3, and facilitates the conjugation of hApg12p to hApg5p. J Biol Chem: 277: 13739 - 13744 (2002).
21. Zhang J, Zhang W, Zou D et al. Cloning and functional characterization of GMPR2, a novel human guanosine monophosphate reductase, which promotes the monocytic differentiation of HL-60 leukemia cells. J Cancer Res Clin Oncol: 129: 76-83 (2003).
22. Aguilella C, Dubourg C, Attia-Sobol J et al. Molecular screening of the TGIF gene in holoprosencephaly: identification of two novel mutations. Hum Genet: 112: 131-134 (2003).
23. Liu, SF, Lu GX, Liu G et al. Cloning of a full-length cDNA of human testis-specific spermatogenic cell apoptosis inhibitor TSARG2 as a candidate oncogene. Biochem Biophys Res Commun: 319: 32-40 (2004).
24. Chambers I, Colby D, Robertson M et al. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell: 113: 643-655 (2003).
25. Katoh M, Katoh M. Identification and characterization of ARHGAP24 and ARHGAP25 genes in silico. Int J Molec Med: 14: 333-338 (2004).