Confirmation: A crucial step in copy number variation analysis after exome sequencing in intellectual disabilities

C.A.M. van Heeswijk
Maastricht University & Radboud Nijmegen
kayvanheeswijk@hotmail.com

Prof. Hans van Bokhoven
Radboud Nijmegen
Hans.vanBokhoven@radboudumc.nl

Abstract

Intellectual disability (ID) comprises a group of mental disorders which have underlying genetic causes, among which the monogenic causes are one of the causes for ID. One kind of a monogenic cause is the copy number variations (CNVs). These CNVs can be indicated using exome sequencing (ES) and the CoNVex and CoNIFER algorithms. To confirm the possible causative CNVs quantitative PCR (QPCR) was used. In a Pakistani ID patient a homozygous deletion of *ENTPD3* was indicated and in an Estonian ID patient *CPVL-CHN2* a homozygous duplication was indicated. However the QPCR showed that *ENTPD3* did not segregate and *CPVL-CHN2* was only duplicated heterozygous. Confirmation, like QPCR, is therefore a crucial step in confirming CNVs analysis of ES in ID patients.

Keywords

Intellectual disability, exome sequencing, CNV, QPCR.

Introduction

An intellectual disability (ID) is classified by a significant limitations in adaptive functioning from at least two of the following skill areas: self-direction, work, leisure, health, communication, home living, use of community resources, social or interpersonal skills, functional academic skills or safety (1). On top of that an onset before the 18 years and a low IQ (<70) are classifications of an ID. There is a subdivision of four degrees of severity of ID based on IQ, the mild ID (IQ 50-69), the moderate ID (IQ 35-49), the severe ID (IQ 20-34) and profound ID (IQ <20). The prevalence of total ID is about 1% of the total population, and for severe ID it is 0.6% (1). IDs are among the largest cost factors for healthcare. In 2010 about 5% of the total health costs for brain disorders in Europe were due to ID (2).

IDs can be caused by environmental factors, such as malnutrition during pregnancy, premature birth, pre- and postnatal infections, exposure to neurotoxic compounds and peri- and postnatal asphyxia or other traumas. These environmental factors have a greater contributing factor than monogenic causes to develop mild ID. However, in patients with severe ID genetic causes are observed more frequently than environmental factors as the cause of ID. Mutations in more than 450 different genes can give rise to ID and related cognitive disorders. The discovery of causative ID genes, from which about 14.2% of the monogenic causes are either a duplication or deletion of a gene region, has risen and will be rising due to innovations and implementation of DNA sequencing techniques (3, 4).

Exome sequencing (ES) is a technique to determine the sequence of the DNA sequencing. With ES the protein coding regions, thus the exome, are mapped. These regions together constitute 1-2% of the whole human genome and harbor more than 60% of pathogenic mutations in heritable disorders (5, 6). ES generates reads, which are parts of DNA that are sequenced, of exonic regions. The aligned reads and a genetic reference database are used to check genetic variations. These differences can be mutations like insertions of nucleotides, deletions of nucleotides or substitution of nucleotides.

Exome sequencing makes use of relatively small reads, about <250 base pairs (bp) long. When a region far greater than these 250 bp is duplicated or deleted, it has no use to compare the reads to a reference database, because the system will see this region as a bad sequenced region rather than a big deletion. Small copy number variations (CNVs), which are deletions and duplications, greater than the 250 bp can be calculated by using the read density of ES data. To calculate the CNVs the number of reads of a genetic region of a person is compared to the mean reads of the other genes of the person. These reads should give an average coverage from the reads of the genes. A second way to calculate the CNVs is using a reference database, in which the average reads coverage of controls is used. If there are more or less reads in a region, it can indicate that there are respectively more or less copies of that gene (7, 8).

Another way to measure CNVs is using quantitative polymerase chain reaction (QPCR). With PCR specific regions of DNA can be multiplied. With QPCR the amount of targeted regions is tracked for every amplification, due to a fluorescent label. Every cycle more of the fluorescent is bound to the double stranded DNA and light is emitted. Thus if there are more targeted regions in the whole DNA, more fluorescent is released per cycle. Normally one cycle duplicates the amount of regions. When you set a certain threshold of fluorescent, and measure the PCR cycles needed for reaching this threshold, the threshold cycle (Ct) can be calculated and CNVs can be determined.

The high costs of ID and the great contribution of CNVs to the onset of ID led to the aim of this study to identify causative CNVs in patients with ID from exome data by using algorithms and QPCR confirmation.

Material and methods

CNVs were identified by the tools CoNVex (Copy number variation estimation in exome sequencing) with a score above 5.0 or CoNIFER (Copy number inference from exome reads) with a score above 3.0 for duplications or below -3.0 for deletions. The deletions that were indicated by either CoNVex or CoNIFER were checked in the ES data if the exonic region was really deleted.

For confirmation a genomic quantitative PCR (QPCR) was performed. If indicated necessary, also the family members were analysed. Primers were designed using Primer3Plus **(www.bioinformatics.nl/cgibin/primer3plus/primer3plus.cgi/)**. The FASTA sequence was obtained from UCSC using the GRCh37/hg19. Product size 80-120bp, primer size 18-24bp, GC 40-60%, Melting temperature 58-62°C (ΔTm <1°C) and a GC clamp set to 1. Using the in silico PCR and blat from UCSC and SNPcheck primers were checked for specificity.

Primers to determine the copy number of *ENTPD3* were designed in a way that the whole gene was covered (Exon 2, 4, 6, 8 and 10) and two flanking references genes were included, namely *MYRIP* exon 17 and *RPL14* exon 2 (figure 1). Primers to determine the copy number of *CPVL-CHN2* were designed in a way that one exon before and after the indicated exon were taken into account, thus for *CPVL* exon 1, 2 and 3 and for *CHN2* exon 7, 8 and 9. Also one flanking reference gene was taken into account, namely *PRR15* (Figure 2).

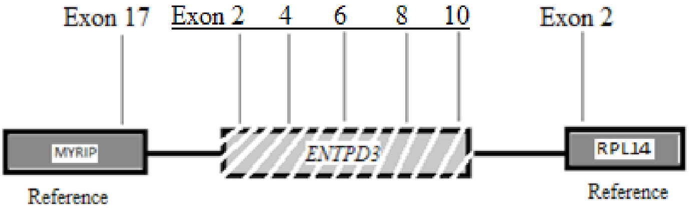

Figure 1. Schematic overview of *ENTPD3* primer design. *MYRIP* and *RPL14* as reference genes. Boxes indicate genes, *ENTPD3* to indicate possible deletion.

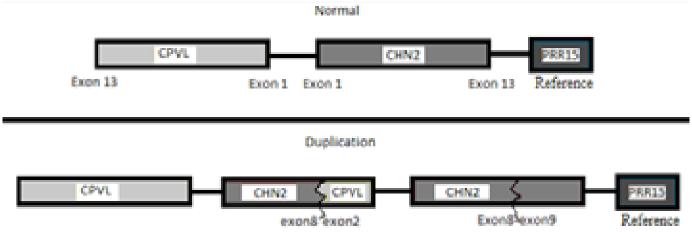

Figure 2. Schematic overview of *CPVL-CHN2* primer design and possible duplication. *PRR15* as reference gene. Boxes indicate genes; Cracks indicate possible breakpoints of the duplication of the genes.

The QPCR mix per reaction consisted of 12.5µL GoTaq2x master mix, 2.5µL 3µM forward primer, 2.5µL 3µM reverse primer, 2.5µL MilliQ and 5µL (1ng/µL) DNA sample were used. Two control samples and one negative were included and everything was performed in duplo. As reference, primers for the *BRWD3* gene on the X-chromosome were used.

The Applied Biosystmes Fast 7500 system with SYBR green was used to perform the QPCR. The stages were

- Stage 1 (1x): 95.0°C for 10:00 minutes
- Stage 2 (40x): 95.0°C for 15 seconds / 60°C for 15 seconds.
- Stage 3 (1x): 95.0°C for 15 seconds / 60°C for 15 seconds.
- Stage 4 (1x): 95.060°C for 15 seconds.

The exact copy number was calculated as follows: Copy number of gene = $2*2^{\Delta\Delta Ct}$ where $\Delta\Delta Ct=\Delta Ct_{patient} - \Delta Ct_{control}$ and $\Delta Ct = Ct_{reference\ gene} - Ct_{gene\ of\ interest}$. The Ct value was obtained by setting the threshold in a way that most of the lines were parallel An independent T-test was used to calculate the significance of the copy numbers using the QPCR $2*2^{\Delta\Delta Ct}$ (9).

Results

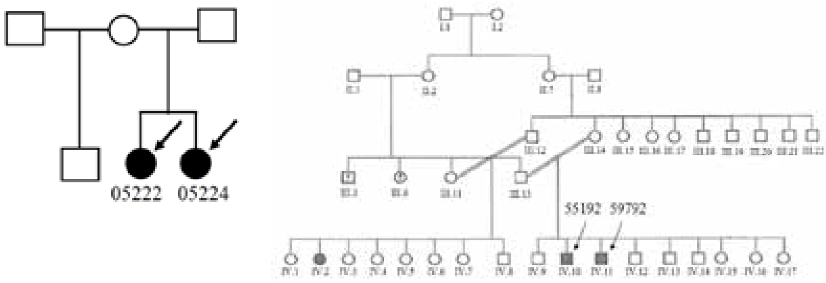

Figure 3. Family pedigrees A) Pakistani family and B) Estonian family. The *filled symbols* indicate the ID affected individuals, the *single horizontal lines* indicate marriage connection and the *double lines* indicate consanguinity and the arrows indicate the CNV analysed patients.

After agreement with and signing of the informed consent DNA samples, blood extractions were performed at the local laboratory, were collected from the ID patients and family members. For one affected Pakistani boy (59792) and two affected Estonian sisters (05222 & 05224) DNA samples were analysed using ES (figure 3).

In the index patient (59792) of the Pakistani family a homozygous (on both chromosomes) deletion of *ENTPD3* was identified covering the whole gene, with a CoNIFER score of -3.65. In the both affected sisters (05222 and 05224) of the Estonian family a homozygous duplication was indicated by ES, with a CoNVex score of 21.34 and 22.97, of the *CPVL*-CHN2. This duplication should reach approximately from exon 2 of *CPVL* until exon 8 of *CHN2*.

For as there was almost no DNA available from 59792 the DNA of his brother (55192), who was also affected, was used. Genomic QPCR analysis, via copy number calculation, showed a copy number of one of the *ENTPD3* gene in the Pakistani index patient's affected brother. The controls and the reference genes all had a copy number of two (figure 4). Hereafter a Genomic QPCR was performed on 59800 (IV:2), 59793 (IV:12) and 59794 (IV:13) (figure 5). These results were that IV:2 and IV:12 had a heterozygous deletion of *ENTPD3* and IV:13 a total deletion. Genomic QPCR analysis, via copy number calculation, showed in the Estonian family that both sisters had a copy number of two of the *CPVL* gene and copy number of three of the *CHN2* gene. The controls and reference had a copy number of two (figure 6).

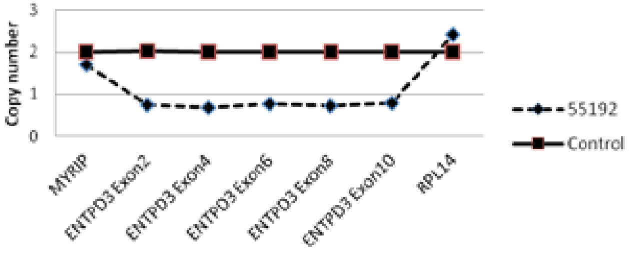

Figure 4. QPCR results of 55192 for *ENTPD3*. *MYRIP* and *RPL14* as reference genes

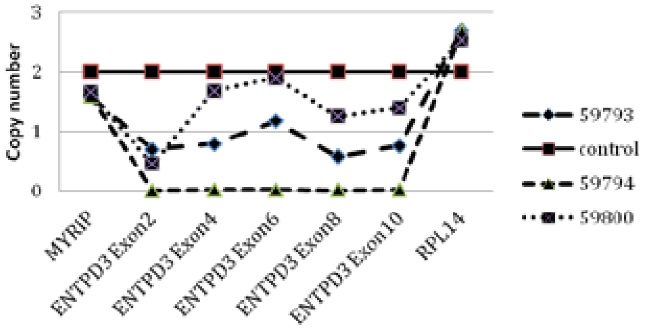

Figure 5. QPCR results of 59793, 59794 and 59800. *MYRIP* and *RPL14* as reference genes

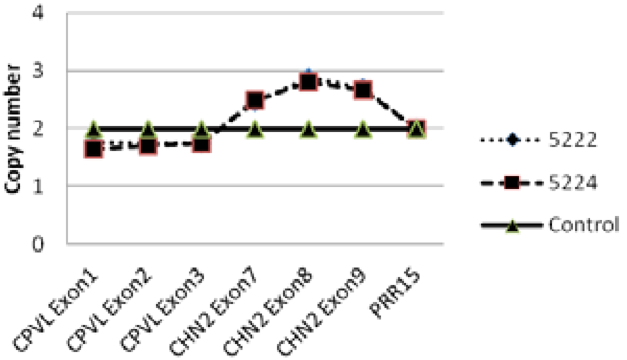

Figure 6. QPCR results of 5222 and 5224 for *CPVL-CHN2*. *PRR15* as reference gene

Discussion/Conclusion

The ES and CNVs analysis by CoNVex and CoNIFER indicated the presence of the potential deletion of *ENTPD3* in the Pakistani ID patient and a duplication of *CPVL-CHN2* in both affected sisters of the Estonian family. However, QPCR did not confirm these CNVs. The brother of the Pakistani index patient only had a heterozygous (on one allele) deletion of *ENTPD3* and for the other tested family members the total deletion of *ENTPD3* did not segregate. The Estonian sisters only had a heterozygous duplication of *CHN2* and no duplication of *CPVL*. The affected brother and the niece of the index patient of the Pakistani family did not have the CNV of *ENTPD3* homozygous and an unaffected brother did have the total deletion, indicating that the *ENTPD3* deletion did not segregate and thus is not the monogenic causative gene contributing to ID development. The same goes for the Estonian sisters with *CPVL-CHN2* double duplication. Because there was only a heterozygous duplication of *CHN2* the *CPVL-CHN2* construct is not the monogenic causative gene.

In this study we used two different algorithms for the calculation of CNVs. This might seem inconsequent, yet the opposite is true. The algorithms were chosen in a way that they served the ES and QPCR methods used in our lab at best. So is CoNVex optimized for UK10K ES files and CoNIFER for non paired exome CNV detection (8, 10). SYBR® Green also shows good results for cross-platform comparison and is a very reliable tool (11). However in essence it differs from some other QPCR mixes. SYBR® Green solution uses Non-specific detection, on the other hand there are techniques, for example the TaqMan probe, that use specific detection with the usage of probes. Where the SYBR® Green releases fluorescent every time there is an amplification reaction because it binds between double stranded DNA, the TaqMan probe binds to a region, directly after region of the primers. This results in fluorescent release after the amplification of the DNA. This technique is more specific than the SYBR Green, although also more expensive.

The *ENTPD3* gene codes for Ectonucleoside triphosphate diphosphohydrolases, which is a class of enzymes that dephosphorylate extracellular ATP and ADP, thus taking phosphate groups of ATP or ADP. In a knockout mouse study it was shown that pair reception did not change, as what they expected. Though they did not test for intelligence of the mice (12). Therefore *ENTPD3* seemed a likely ID gene. For *CPVL-CHN2* it is a different story. Most publication with reference to those two genes are for diabetics, and none about intellectual disorders. For searches alone on the *CPVL*, it has a more important function in macrophage. However the *CHN2* gene has high expression levels in the brain and pancreas, has a role in cell proliferation and migration. Thus a error in the construct of the *CHN2* gene might have made it a possible ID gene (13). What also indicated that these genes might be causative ID genes, is the fact that when looking in online databases of genes. There seems to be almost no loss of function genes (indicating nonsense frameshift or splice site mutations) of the *ENTPD3, CPVL* and *CHN2* genes according to the ExAC database. This database contains the genome of >60,000 unrelated individuals (14).

Exact copy numbers of the genes could not be readily extracted from exome data, as only a duplication of one allele rather than two in the Estonian sisters was found. Adjusting selection criteria, like increasing the CoNVex or CoNIFER score threshold that is used to indicate homozygous CNVs, could result in less false positive results from the ES. On the other hand, possible causative CNVs might be missed. The homozygous deletion was most likely to be present in the affected Pakistani boy, although it did not segregate with the disease since his affected brother had a deletion on only one allele, same for the niece and an unaffected brother had a total deletion. Taken together, ES is a technique that is usefull in identifying potentially causal ID genes through the analysis of CNVs that contain only one gene. However, confirmation with a separate technique, such as genomic QPCR, is essential in CNV analysis of ID patients.

Role of the student

C.A.M. van Heeswijk was an internship student working under the supervision of prof. Hans van Bokhoven and Dr. Arjan de Brouwer when the research in this report was performed. The topic was introduced and the CoNVex and CoNIFER score were calculated by the supervisors. The analysis of the exome sequencing data, primer designing, quantitative PCR, analysis of QPCR results, calculation of copy numbers, statistical analysis and formulating the conclusion was done by the student.

Acknowledgments

I want to thank my supervisors H. van Bokhoven, A.P.M. de Brouwer, E. van Beusekom and D.L. Polla for helping me with my internship and this study. I also want to thank the other colleagues who helped me with various tasks.

References

1. American Psychiatric Association. Diagnostic and statistical manual of metal disorders. Washington, DC: American Psychiatric Association,; 2000.
2. van Bokhoven H. Genetic and epigenetic networks in intellectual disabilities. Annual review of genetics. 2011;45:81-104.
3. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nature genetics. 2011;43(9):838-46.
4. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. European Journal of Human Genetics. 2012;20(5):490-7.
5. Biesecker LG. Exome sequencing makes medical genomics a reality. Nature genetics. 2010;42(1):13-4.
6. 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;22(8):1525-32.
7. Amarasinghe KC, Li J, Halgamuge SK. CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC bioinformatics. 2013;14(Suppl 2):S2.
8. Vijayarangakannan P, Fitzgerald T, Joyce C, McCarthy S, Hurles ME. Detection of Copy Number Variantion from Exomes in the DDD and UK10K Projects. 2012.
9. Zhang, J. D., Ruschhaupt, M., & Biczok, R. (2013). ddCt method for qRT–PCR data analysis.
10. Guo Y, Sheng Q, Samuels DC, Lehmann B, Bauer JA, Pietenpol J, et al. Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. BioMed research international. 2013;2013.
11. Arikawa E, Sun Y, Wang J, Zhou Q, Ning B, Dial SL, et al. Cross-platform comparison of SYBR® Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. BMC genomics. 2008;9(1):328.
12. McCoy E, Street S, Taylor-Blake B, Yi J, Edwards M, Wightman M, et al. Deletion of ENTPD3 does not impair nucleotide hydrolysis in primary somatosensory neurons or spinal cord. F1000Research. 2014;3.
13. Hu C, Zhang R, Yu W, Wang J, Wang C, Pang C, et al. CPVL/CHN2 genetic variant is associated with diabetic retinopathy in Chinese type 2 diabetic patients. Diabetes. 2011;60(11):3085-9.
14. Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org) [Accessed May 2015]