Case report

Loss of function of *PGAP1* as a cause of severe encephalopathy identified by Whole Exome Sequencing: Lessons of the bioinformatics pipeline☆

M. Granzow [a, *, 1], N. Paramasivam [b, c, 1], K. Hinderhofer [a], C. Fischer [a], S. Chotewutmontri [d], L. Kaufmann [a], C. Evers [a], U. Kotzaeridou [e], K. Rohrschneider [f], M. Schlesner [b], M. Sturm [g], S. Pinkert [d], R. Eils [b, h], C.R. Bartram [a], P. Bauer [g], U. Moog [a]

[a] *Institute of Human Genetics, University of Heidelberg, Im Neuenheimer Feld 366, 69120 Heidelberg, Germany*
[b] *Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany*
[c] *Medical Faculty Heidelberg, Heidelberg University, 69120 Heidelberg, Germany*
[d] *Genomics & Proteomics Core Facility, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany*
[e] *Section of Neuropediatrics, Center for Child and Adolescent Medicine, University Hospital Heidelberg, Im Neuenheimer Feld 430, 69120 Heidelberg, Germany*
[f] *Section of Ophthalmologic Rehabilitation, Department of Ophthalmology, University Hospital Heidelberg, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany*
[g] *Institute of Medical Genetics and Applied Genomics, University of Tübingen, Calwerstrasse 7, 72076 Tübingen, Germany*
[h] *Department for Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology (IPMB) and BioQuant, Heidelberg University, 69120 Heidelberg, Germany*

1. Introduction

Whole Exome Sequencing (WES) has been established as a powerful tool for the identification of disease genes in research and has already started to be applied in clinical diagnostics [1]. Sequencing data are processed by a bioinformatics analysis pipeline that filters variants according to a variable minimum base coverage, specific quality scores, presumed inheritance pattern and resulting variant status, allele frequency and listing in public and, if available, in-house databases. In order to assess pathogenicity, they further have to be carefully evaluated for mutation type, their effect

* Corresponding author. Institute of Human Genetics, Heidelberg University, Im Neuenheimer Feld 366, D-69120 Heidelberg, Germany.
E-mail address: martin.granzow@med.uni-heidelberg.de (M. Granzow).
[1] Both authors contributed equally to this work.

predicted by *in silico* tools and (presumed) function in order to avoid annotation errors [2]. The bioinformatics and evaluation pathway thus contains many variables that possibly influence the results of analysis.

We studied a consanguineous Turkish family with two siblings, a boy and a girl, affected by severe encephalopathy and additional features, indicating autosomal recessive inheritance by linkage analysis and WES. Sequencing data were analysed by two similar yet not identical bioinformatics pipelines which resulted in only partially overlapping lists of rare variants including a homozygous splice mutation in *PGAP1* (Post-GPI Attachment To Proteins 1; OMIM 611655) in both children. Recently, homozygous loss-of-function mutations in *PGAP1* have been identified in two Syrian siblings with severe intellectual disability (ID), hypotonia and mild microcephaly [3] (OMIM 615802). *PGAP1* is involved in the biosynthesis and remodeling of GPI (glycosylphosphatidylinositol)-anchor proteins, a conserved post-translational modification of multiple proteins in eukaryotes [4]. More than 30 genes are essential in this process, several of them including *PGAP2* and *PGAP3,* have been implicated in hyperphosphatasia-mental retardation syndrome (HPMRS), an autosomal recessive form of ID with characteristic additional phenotypic features [5,6]. This report not only confirms *PGAP1* as causative for severe syndromic encephalopathy/ID but also illustrates the influence of different variables in the evaluation process of filtering, annotating and prioritizing sequence variants obtained by WES illustrated by analysis with two different bioinformatics pipelines.

2. Material and methods

2.1. Patient reports

Patient 1 and 2 were born at term after an uneventful pregnancy with normal measurements (weight, length and OFC) as the 1st and 2nd child of a consanguineous couple from North Turkey who are both 1st and 2nd cousins (Fig. 1a). The mother was short statured (145 cm), her husband suffered from type 1 diabetes since the age of 20 years. In his family, 4 children of a brother of his father died in infancy for unknown reasons, and a living brother of these children was reported to be severely intellectually and physically disabled.

Patient 1, a girl, was diagnosed with an atrial septal defect II and mild pulmonary stenosis but needed no treatment. She showed congenital nystagmus, failure to thrive, and her psychomotor development was severely delayed from the beginning. She learnt to roll over, could crawl at the age of 6 ½ years, pulled herself up to stand but never learnt to walk and had no speech. She developed microcephaly after 6 months of age and short stature in her 2nd year of life, corresponding at the age of 8 years to − 3.4 and −3.5 SD, respectively. She showed a tonus regulation disorder with hypertonia in the beginning up to age 2 years and later significant hypotonia. At age 2 years and on later examinations, facial features comprised mild synophris, low frontal and posterior hair line, hypertelorism, all compatible with her ethnical and familial background, and in addition a bifid uvula, mild pectus excavatum and broad 1st toes. Retinal dystrophy was suspected at age 3 years by a pathological ERG (electroretinogram) with markedly reduced scotopic and photopic answers. Fundoscopy revealed no anomalies on serial examination at 3, 7, 8 and 9 years of age.

Repeated MRI of the brain, extensive investigations for inborn errors of metabolism (amino acids, organic acids, lactate, ammoniac, congenital disorders of glycosylation, very long fatty acids, homocysteine, lysosomal enzymes, purine/pyrimidine and creatinine deficiency disorders), chromosome analysis, molecular karyotyping by SNP arrray, molecular genetic analyses for Angelman syndrome, Rett syndrome and FragileX syndrome were normal.

Patient 2, a boy, suffered in the neonatal time from congenital pneumonia. He showed severe psychomotor retardation but a better motor development than his sister, and severe, particularly axial hypotonia without further neurologic signs. He learnt to sit unsupported at the age of 1 year and rolled over at 20 months. At age 4 years, his OFC and length were just below the 3rd centile. He showed stereotypic movements, breathing irregularities and frequent laughter. He had no expressive speech and no speech comprehension. Physical examination revealed brachycephaly. ERG at the age of 3 ½ years showed significantly reduced amplitudes, similar to the findings in his sister at the same age indicating retinal dystrophy. Fundoscopy was normal at ages 3 ½ and 5 ½ years. BERA (brainstem evoked response audiometry), MRI of the brain, investigations for inborn errors of metabolism (acyl carnitine profile, amino acids, organic acids, glycosaminoglycan excretion, electrophoresis for mucopolysaccharidoses), chromosome analysis, subtelomere screening by MLPA and molecular karyotyping by SNP array showed normal results.

2.2. Methods

2.2.1. DNA isolation, Sanger sequencing and cDNA analysis

Genomic DNA of the patients and both parents was isolated from peripheral blood using a salting out protocol according to Miller et al. [7]. Written consent for all investigations was given by the parents and the study adhered to the tenets of the declaration of Helsinki and was approved by the local ethics committee.

Validation of the *PGAP1* splice variant was performed by Sanger sequencing. Exon 10 and adjacent intron boundaries of the *PGAP1* gene (RefSeq NM_024989.3; exon numbering according to ENST00000354764) were sequenced from affected individuals and both parents using Big Dye Terminator V1.1 cycle sequencing kit and ABI 3130xl genetic analyzer.

RNA was isolated from peripheral blood of the female patient, the mother and healthy controls via phenol-chlorophorm extraction as described by Chomczynski and Sacchi [8]. cDNA was synthesized through reverse transcription (RevertAid H Minus Reverse Transcriptase, Thermo Fisher Scientific, Inc.) according to manufacturer's instructions.

2.2.2. SNP/CopyNumber array and statistical linkage analysis

Affymetrix® CytoScan HD Oligo/SNP-Array was used to genotype DNA of the affected individuals and both parents (I.1, I.2, II.3 and II.4, Fig. 1a), and to exclude genomic imbalances in the affected patients. Analysis was performed according to the manufacturer's instructions. Analysis for genomic imbalances was done at a resolution of 100 kb. Interpretation was based on human reference sequence GRCh37/hg19, February 2009. Genome-wide parametric linkage analysis with SNP genotypes was performed using ALOHOMORA and MERLIN software [9,10] and the SNP genotype data, assuming affected family members were homozygous at a putative disease locus for an autosomal recessive disease allele inherited from a common ancestor. After performing standard data quality checks, SNP markers with a minor allele frequency (MAF) of 0.15 and a minimum distance of 100 000 bp were selected to ensure low linkage disequilibrium between the markers.

2.2.3. Whole Exome Sequencing (WES)

WES was performed at the German Cancer Research Center (DKFZ), Heidelberg, on both affected children (I.1 and I.2) and their parents (II.3 and II.4). The exons were captured using Agilent SureSelect Human All Exon V4 (without UTR). The downstream data analysis of the raw sequence data was performed in parallel at the DKFZ, Heidelberg (HD), and at the Institute of Medical Genetics and Applied Genomics, Tübingen (TÜ), following different

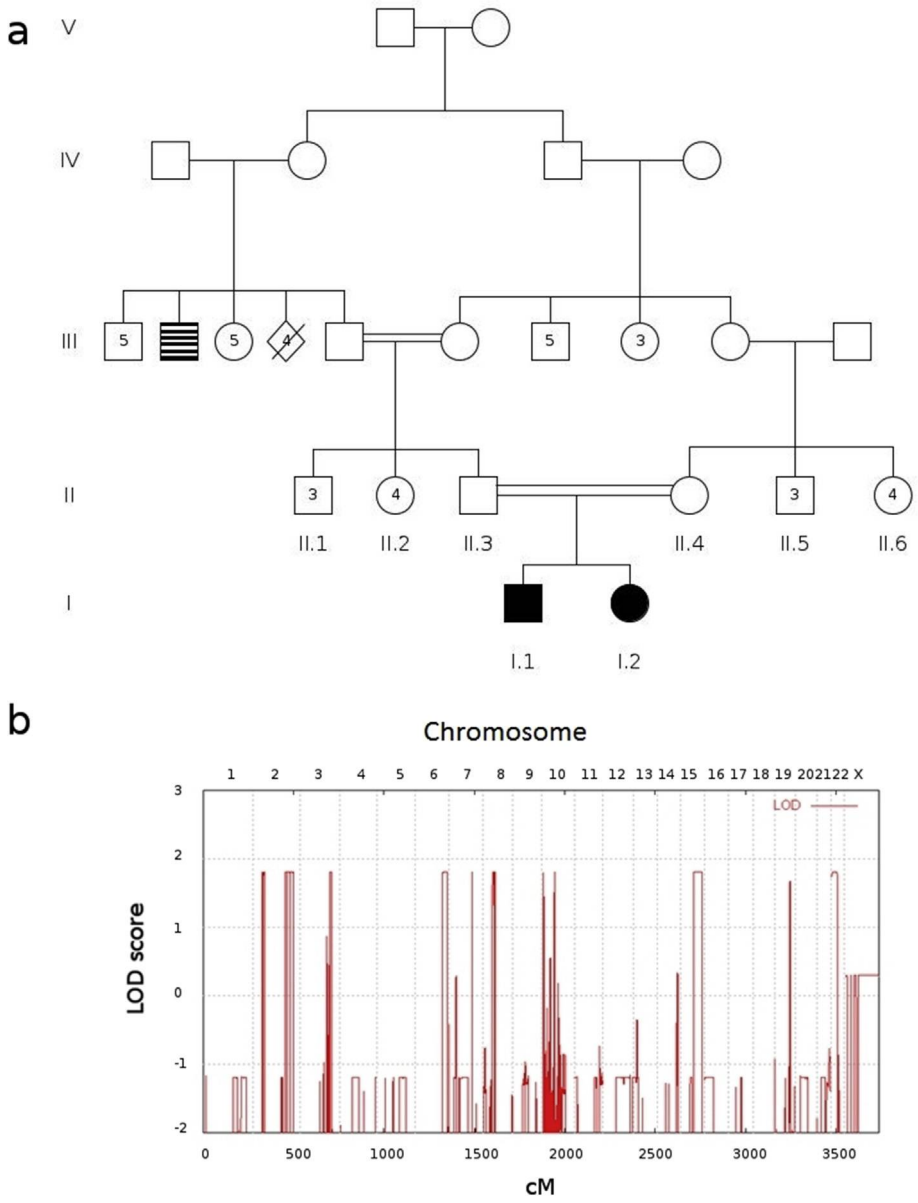

Fig. 1. Pedigree and homozygosity mapping. (a) The pedigree of the family showed double consanguinity of the parents. (b) Linkage analysis: LOD-score distribution relative to chromosomal location indicating several regions as relevant.

bioinformatics pipelines.

2.2.4. Heidelberg and Tübingen exome data analysis pipeline

A detailed overview of the pipelines is presented in Table 1 which lists algorithms, databases, and filter settings. Additionally, the table provides the similarities and differences of each step of the analysis.

Except for differences in software versions, the pipelines from HD and TÜ differed in several aspects: The base coverage was used as a threshold only in the HD pipeline, whereas a lower limit of base quality was set slightly different. The steps of local realignment and indel calling were done by different software tools and different strategies were applied in the pipelines to remove false positives. Referencing minor allele frequency (MAF) in the HD pipeline makes additional use of the Exome Aggregation Consortium (ExAC) database, in contrast to the analysis of TÜ where MAF was matched only to data from the 1000 Genomes Project and ESP6500. Furthermore, HD included variants with a MAF of 1% and lower, whereas TÜ also excluded variants with MAF of exactly 1%. Both pipelines excluded all but variants detected as homozygous alternate found in both affected children due to the suspected inheritance pattern and consanguinity. In the TÜ pipeline, variants that occurred more than 30 times with matching genotypes in an in-house database of 200 exomes were filtered out, unless marked as pathogenic (VUS 3 or more). A similar filter with different sequencing data was used as controls in HD where data from around 80 exomes with different phenotypes were used to remove machine artifacts, and variants with the same genotype as the control samples were removed. Variants without reading frame information were filtered out in the TÜ pipeline. Both pipelines integrated tools to predict functional effects of variants but only HD used this information for filtering.

Table 1
Comparison of two analysis pipelines applied in the WES study. Listed are details of the main analysis steps ranging from alignment and variant calling, quality filters and genetic model to the use of local control exomes and prioritization. EVS6500: Exome Variant Server; ExAC: Exome Aggregation Consortium; HD: Heidelberg; MAF: Minor allele frequency; MT: MutationTaster; PPH2: PolyPhen2; PROVEAN: Protein Variation Effect Analyzer; SIFT: http://sift.jcvi.org/; TÜ: Tübingen.

Task	HD pipeline	TÜ pipeline	Comparison of pipelines
Base quality	Base quality >20	Adaptor contamination Base quality >15	Base quality differences TÜ: Adaptor removal
False positive removal	Coverage ≥10x	VCFutils 'varFilter' (at least 3 observations of alternate allele)	Different strategies
Mapping	BWA 0.6.2 [14] to reference genome hs37d5	BWA 0.7.5 + Stampy 1.0.23 [25] to reference genome hg19	Different versions of BWA TÜ: Refinement with Stampy
PCR duplicates	Picard tools 1.61 [15]	Picard tools 1.122	Different versions
SNV calling	SAMtools 0.1.19 [16]	SAMtools 0.1.19	
Indel realignment	Platypus [17]	In-house tool 'BamLeftAlign'; realigned with GATK2.1-8 [26]	Different tools
Short indel calling	Platypus [17]	SAMtools 0.1.19	Different tools
Functional annotation	ANNOVAR [18] with Gencode v17 [19] as gene model	ANNOVAR + several in-house tools	Both: ANNOVAR HD: Gencode v17 TÜ: in-house tools
MAF references	MAF ≤ 1% in: - 1000 Genomes - EVS6500 - ExAC	MAF < 1% in - 1000 Genomes - EVS6500	Both: 1000 Genomes & EVS6500 HD: ExAC MAF defined differently
Genotype determination	SNVs: VAF > 90% homozygous alternate Indels: Platypus	SAMtools 0.1.19	HD: Manual thresholds TÜ: Automated calculation
Local control	In-house database of ~80 exomes	In-house database of 200 exomes	Different local databases
Prediction tools	PPH2 [20] MT [21] SIFT [22] PROVEAN [23] Intolerance Score [24]	PPH2 [20] MT [21] SIFT [22]	Prioritization of variants by *in silico* predictions only in HD
Functional variant filter	Keep only - non-synonymous - stop gain/loss - splicing variants - Indels	Keep only - non-synonymous SNVs - exonic/splicing variants - 'variant details' annotation present	Filter on 'variant details' annotation only in TÜ

3. Results

3.1. Linkage analysis

Both children (I.1 and I.2) had been screened for CNVs (copy number variants) by SNP array analysis with negative results. Prior filtering of homozygous variants is based on pointwise genotype information. To include also haplotype information, linkage analysis was performed to identify genomic regions linked with the hypothesized gene responsible for the disease in this family. Computational analysis of SNP data showed several regions with maximal lod-scores of 1.8 spread over chromosomes 2, 3, 6, 7, 8, 10, 15, 19 and 22. The longest among them on chromosome 15 covered about 26 Mbp (Fig. 1b).

3.2. Whole Exome Sequencing, filtering and prioritization by two analysis pipelines

Subsequently, exome sequencing using DNA from both affected individuals (I.1 and I.2) and both parents (II.3 and II.4) was performed. The samples were sequenced to an average base coverage of 121x and of a median base coverage of 101.5x on targeted regions, and 98.552% of the targeted bases having at least 10x coverage. After filtering and annotation according to the aforementioned criteria including the suspected autosomal recessive inheritance trait, a total of 15 SNVs and four indels were prioritized by the two analysis pipelines from Heidelberg and Tübingen that were homozygous in both children and heterozygous in the parents (summarized in Table 2). Nine of these were found in both candidate lists, two appeared only in the results from the Heidelberg pipeline, and eight including the four indels were additionally detected by the pipeline applied in Tübingen. Only variants were considered for further analysis which did not show any homozygous allele count in ExAC database.

3.3. Combining linkage results with prioritized variants

We further focused on the variants located in the genomic regions with maximal lod-scores. This narrowed the candidate list to seven putatively causative variants from both pipelines (in the genes *GLB1L*, *KANSL1L*, *MYO9A*, *STAT1*, *TRIM59*, *TRMT61B* and *TTN*) and one additional variant (in *PGAP1*) prioritized only in the Heidelberg analysis pipeline. In the TÜ pipeline, variants without reading frame information were filtered out, preventing the *PGAP1* variant to be listed by the TÜ analysis.

3.4. Variant prioritization

Selected variants were further prioritized by *in silico* prediction algorithms to assess their functional effect and by their reported mRNA and protein expression pattern in mammalian tissue. Results are summarized in Table 3. Three variants (in *MYO9A*, *STAT1* and *TTN*) were classified as benign by functional *in silico* prediction ambiguous results were obtained for the variant in *KANSL1L*. In addition, the variants in *GLB1L*, *KANSL1L* and *STAT1* are listed as rs148825055, rs149500487 and rs140351189, respectively, in the Database of Short Genetic Variations (dbSNP; http://www.ncbi.nlm.nih.gov/SNP; accessed 2015-02-03) and annotated with 'no clinical significance'. The expression pattern of *GLB1L* excluded CNS tissue and no human phenotype has been described for *GLB1L* and *KANSL1L* so far.

Table 2
Homozygous variants present in both children in genes prioritized by the two analysis pipelines from Heidelberg and Tübingen with transcript information, number of homozygous allele counts in ExAC database and indication if genes are located in a linkage region. Genes in which identical variants were found are marked in bold. The last column provides an explanation why the respective pipeline filtered out the variant. HD: Heidelberg; hom.: homozygous; SNV: single nucleotide variant; TÜ: Tübingen.

Affected genes	RefSeq transcript and variant information; genome position	HD	TÜ	ExAC count hom. allele	In linkage region	Explanation of differences
BRE	NM_199194:exon13:c.1102A>G/p.(T368A); chr2:28561330	+	+	3	−	
CSPG4	NM_001897:exon3:c.1094G>A/p.(R365Q); chr15:75982312	−	+	4	+	Not called by SNV pipeline
DFNB59	NM_001042702:exon7:c.793C>T/p.(R265C); chr2:179325735	−	+	273	+	Different MAF references
DSPP	NM_014208:exon5:c.3068_3103del/p.(1023_1035del); chr4:88536882-88536917	−	+	−	−	Putatively false positive
DSPP	NM_014208:exon5:c.3448_3449insAGCAGCGACAGCAGCGAT/p.(S1150delinsKQRQQRC); chr4:88537262	−	+	−	−	Putatively false positive
DSPP	NM_014208:exon5:c.3249A>C/p.(E1083D); chr4:88537063	−	+	197	−	Coverage < 10%
FAM221A	NM_001127365:c.459del/p.(I154fs); chr7:23731209	−	+	2	+	Filtered by genotype likelihood score for indels (Platypus)
FSIP2	NM_173651:exon17:c.20064G>C/p.(K6688N); chr2:186673830	−	+	16	+	Different MAF references
GLB1L	NM_024506:exon7:c.677G>T/p.(C226F); chr2:220104686	+	+	0	+	
KANSL1L	NM_152519:exon2:c.605C>T/p.(P202L); chr2:211018702	+	+	0	+	
MYO9A	NM_006901:exon27:c.5212A>G/p.(T1738A); chr15:72180388	+	+	−	+	
PGAP1	NM_024989.3:c.1090-2A>G:p.?); chr2:197755637	+	−	−	+	No reading frame information
RTL1	NM_001134888:exon1:c.454_455insAGA/p.(E152delinsEK); chr14:101350671	−	+	713	−	MAF > 1% in ExAC
SKIDA1	NM_207371:exon4:c.1241A>C/p.(E414A); chr10:21805511	+	+	−	−	
SLC11A1	NM_000578:exon13:c.1328T>C/p.(V443A); chr2:219258856	+	−	9	+	MAF < 1% in ExAC
STAT1	NM_007315:exon16:c.1341C>A/p.(D447E); chr2:191849042	+	+	0	+	
TRIM59	NM_173084:exon3:c.37T>G/p.(C13G); chr3:160156935	+	+	−	+	
TRMT61B	NM_017910:exon3:c.839G>A/p.(R280Q); chr2:29084138	+	+	0	+	
TTN	NM_001256850:exon173:c.35890G>A/p.(V11964M); chr2:179504489	+	+	−	+	

Therefore, we excluded these six variants from further analysis. The variants in *TRIM59* and *TRMT61B* were classified as damaging. The identified variant in *PGAP1*, NM_024989.3:c.1090-2A>G; IVS9-2A>G/p.?, located in intron 9, was predicted by MutationTaster to be disease causing due to changes of the downstream splice site and therefore assumed not to be recognized as a splice site and to lead to skipping of the respective exon during transcription. Because the variant is intronic, PPH2, PROVEAN and SIFT were not informative for this variant.

The *TRIM59* and *PGAP1* are differentially expressed in adult human and mouse embryonic tissues, while *TRMT61B* is expressed in all tissues. Both *TRIM59* and *PGAP1* have defined expression in brain tissue at adult and embryonic stages. However, analysis of the protein expression data of these three candidates available in the GeneCards database revealed that only PGAP1 protein has been detected in the brain so far, indicating that this may be the most promising candidate variant responsible for the phenotype.

PGAP1 is the only gene of the remaining candidates already described to be associated with a similar human phenotype ([3]; OMIM 615802). Germline mutations in TRIM59 and TRMT61B have not been associated with human phenotypes so far. *TRIM59* is a ubiquitin ligase involved in carcinogenesis and *TRMT61B* is a tRNA methyltransferase that catalyzes the formation of N(1)-methyladenine at position 58 (m1A58) in various tRNAs in the mitochondria.

3.5. Validation of the PGAP1 splice acceptor variant

The *PGAP1* variant was confirmed by Sanger sequencing in the homozygous state in both affected siblings (I.1 and I.2), and in the heterozygous state in their parents (II.3 and II.4). Subsequently, we were able to amplify cDNA in one region 5′ to the splice variant site (exons 3–5) of *PGAP1* in the patient (I.2) and the mother (II.4). The region 3′ to the variant site (exons 8/9 to 12) could not be amplified in the patient's cDNA (Fig. 2). In the mother's cDNA the amount of amplified PCR product was decreased compared to healthy controls. Obviously, the identified splice variant leads to aberrant splicing of the *PGAP1* mRNA resulting in a truncated mRNA product and not in nonsense-mediated decay.

4. Discussion

The application of exome sequencing to identify rare variants causative for Mendelian diseases is a primary focus in research projects and has begun to enter the field of clinical diagnostics. The accomplishment of such a high-level technique is concentrated in centers where typically only one specific analysis pipeline is implemented. This pipeline is a complex multi-step process

Table 3
Prediction of functional effects, mRNA and protein expression tissue pattern of homozygous variants obtained by predictive algorithms MutationTaster (MT; http://www.mutationtaster.org) [21], PolyPhen2 (PPH2; http://genetics.bwh.harvard.edu/pph2) [20], SIFT (http://provean.jcvi.org/index.php) [22] and PROVEAN (http://provean.jcvi.org/index.php) [23]. Information on tissue specific mRNA and protein expression were searched for in integrated databases of human and mammalian genes (GeneCards (http://www.genecards.org), EBI Gene Expression Atlas (http://www.ebi.ac.uk/gxa/help/index.html) and UniGene (http://www.ncbi.nlm.nih.gov/UniGene)) [27–32]. B: benign; dam: damaging; DC: disease causing; del: deleterious; N: neutral; PosD: possibly damaging; ProbD: probably damaging; T: tolerated.

Affected genes	RefSeq transcript and variant information	In silico parameters: MT/PPH2/PROVEAN/SIFT	Tissue expression: mRNA/Protein
GLB1L	NM_024506:exon7:c.677G>T/p.(C226F)	DC/ProbD/del/del	Testis/plasma, platelet, liver, urine
KANSL1L	NM_152519:exon2:c.605C>T/p.(P202L)	DC/ProbD/N/T	Testis, kidney, cerebellum/plasma, platelet, urine
MYO9A	NM_006901:exon27:c.5212A>G/p.(T1738A)	P/B/N/T	Ubiquitous/plasma, platelet, liver
PGAP1	NM_024989.3:c.1090-2A>G:p.?)	DC/−/−/-	Prefrontal cortex, cerebellum/brain, plasma, platelet
STAT1	NM_007315:exon16:c.1341C>A/p.(D447E)	DC/B/N/T	Ubiquitous/ubiquitous
TRIM59	NM_173084:exon3:c.37T>G/p.(C13G)	DC/ProbD/del/dam	Testis, prefrontal cortex, frontal lobe/plasma, platelet, bone
TRMT61B	NM_017910:exon3:c.839G>A/p.(R280Q)	DC/ProbD/del/dam	Ubiquitous/kidney, liver, patelets, heart
TTN	NM_001256850:exon173:c.35890G>A/p.(V11964M)	P/B/N/T	heart/heart, skeletal muscel, plasma, platelet

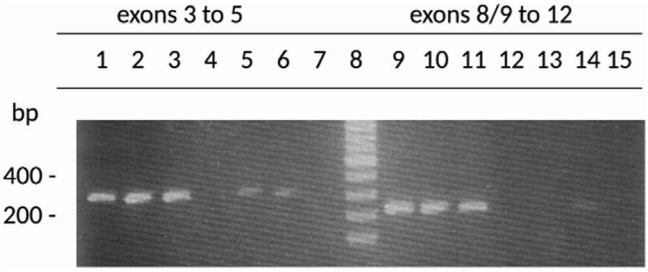

Fig. 2. cDNA analysis of PGAP1: Amplification of exons 3 to 5 (lanes 1–7) and 8/9 to 12 (lanes 9–15) of PGAP1 cDNA, respectively. Lanes 1 – 3 and 9 – 11: cDNA of healthy controls; lanes 4 and 12: genomic DNA of healthy control; lanes 5 and 13: cDNA of patient I.2; lanes 6 and 14: cDNA of her mother (II.2); lanes 7 and 15 negative control; lane 8: marker (100 bp ladder).

deploying dozens of pieces of openly accessible as well as in-house software, online and local databases, reference genomes and specific filtering modalities taking up time and costs. Thus, sequencing data are rarely, if at all, entered into more than one bioinformatics pipeline, which would allow for a comparison of the effect of different settings within the pipelines. We analyzed a multiple consanguineous Turkish family with two children, a boy and a girl, affected by severe encephalopathy, hypotonia, growth retardation, microcephaly and retinal dystrophy by a combination of linkage analysis and WES. The application of two analysis pipelines initially yielded only partially overlapping lists of a total of 19 rare variants. After investigation of the sources of these differences retrospective adaption of filter criteria brought the lists to near congruence. In order to reduce the list of candidate variants, we focused on regions with maximal lod-scores corresponding to homozygous haplotypes shared by both affected children by identity by descent. Within these regions, eight homozygous sequence variants were prioritized in both affected children, seven of them by both pipelines, one by the HD pipeline only. Five of the eight were considered highly unlikely to be pathogenic because of the predicted effects and/or their annotation in dbSNP. Of the remaining, only the variant in *PGAP1* (c.1090-2A>G; IVS9-2A > G; p.?) was both predicted to be disease causing by *in silico* analysis and showed mRNA as well as protein expression in the developing and mature brain. Additionally, *PGAP1* has been described to be associated with a similar human phenotype ([3]; OMIM: 615802) and the homozygous variant affects splicing as shown by cDNA analysis. The analysis revealed a truncated mRNA molecule only containing the exons 5'(at least exons 3–5) to the splice mutation presumably leading to a truncated protein. If the exons 1 to 9 are present in the mRNA of the variant and expressed in the protein, the first transmembrane domain and the complete alpha/beta hydrolase fold domain of PGAP1 protein are present in the truncated protein. This protein might harbor a hydrolase activity with reduced and/or impaired function compared to the wild type PGAP1 protein or be non-functional at all.

So far, inactivating homozygous mutations in *PGAP1* have been described only in one Syrian family with a comparable phenotype in two affected children consisting of severe encephalopathy, hypotonia and mild microcephaly [3]. Retinal dystrophy in the two present children has been detected by ERG only and not by fundoscopy and thus cannot be excluded in the Syrian patients. In addition, Novarino et al. [11] identified a homozygous splice variant in *PGAP1* in a cohort of 55 families with autosomal recessive spastic paraplegia (AR-HSP) complicated by mild ID. The effect on the protein, however, was not analyzed and an association of *PGAP1* with HSP awaits further confirmation.

Thus there is strong evidence that the homozygous splice variant identified in *PGAP1* is responsible for the encephalopathy and associated features in our patients.

In contrast to the different subtypes of HPMRS, *PGAP1* associated disorders are most likely not associated with hyperphosphatasia. In HPMRS, hyperphosphatasia results from increased secretion of non-GPI anchored proteins such as alkaline phosphatase (AP) into the extracellular space (for *PIGV* and *PIGO,* [12]) or of proteins bearing cleaved GPI anchors (for *PGAP2*, [13]). Its mechanism still has to be characterized for PGAP3-deficient cells [6]. In the case of a loss of PGAP1 enzyme activity, however, cells express normal levels of GPI anchors that have an abnormal lipid structure resistant to cleavage and thus do not lead to increased release of AP [3].

5. Conclusion

This report provides evidence that the results of the same sequencing data being processed by two similar analysis pipelines might bring up more differences than commonalities due to the filter settings used. Thus, it should be kept in mind in research to evaluate carefully candidate lists resulting from bioinformatic filtering with respect to selection criteria, e.g., reading frame information and thresholds for the exclusion of variants, and to re-evaluate unsolved cases using different filter settings. The presented case study furthermore confirms homozygous loss-of-function mutations in *PGAP1* as a cause of severe encephalopathy and broadens the phenotypic spectrum by showing possible retinal involvement.

Authors' contributions

Study coordination: MG, UM; Manuscript writing: MG, NP, KH, CF, CE, MS, CRB, PB, UM; Clinical phenotyping: UK, KR, UM; Whole Exome Sequencing and bioinformatic analysis: NP, SC, MSch, MS, SP, RE, PB; Genotype–Phenotype correlation: MG, KH, CF, LK, CE, UM; Molecular karyotyping, linkage analysis and interpretation of data: KH, CF; Sequence analysis, cDNA analysis and interpretation of data: KH, LK; All authors have been involved in the drafting, critical revision and final approval of the manuscript for publication and agree to be accountable for all aspects of the accuracy and integrity of the work.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

We thank the High Throughput Sequencing Unit of the DKFZ Genomics and proteomics Core Facility for the technical support and expertise.

References

[1] Y. Yang, D.M. Muzny, J.G. Reid, M.N. Bainbridge, A. Willis, P.A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, M. Hardison, R. Person, M.R. Bekheirnia, M.S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S.E. Plon, J.R. Lupski, A.L. Beaudet, R.A. Gibbs, C.M. Eng, Clinical whole-exome sequencing for the diagnosis of mendelian disorders, N. Engl. J. Med. 369 (2013) 1502–1511.
[2] D.G. MacArthur, T.A. Manolio, D.P. Dimmock, H.L. Rehm, J. Shendure, G.R. Abecasis, et al., Guidelines for investigating causality of sequence variants in human disease, Nature 508 (2014) 469–476.
[3] Y. Murakami, H. Tawamie, Y. Maeda, C. Büttner, R. Buchert, F. Radwan, et al., Null mutation in PGAP1 impairing GPI-anchor maturation in patients with intellectual disability and encephalopathy, PLoS Genet. 10 (2014) e1004320.
[4] M. Fujita, T. Kinoshita, GPI-anchor remodeling: potential functions of GPI-anchors in intracellular trafficking and membrane dynamics, Biochim. Biophys. Acta 1821 (2012) 1050–1058.
[5] E. Horn, D. Wieczorek, K. Metcalfe, I. Barić, L. Palležac, M. Ćuk, et al.,

Delineation of PIGV mutation spectrum and associated phenotypes in hyperphosphatasia with mental retardation syndrome, Eur. J. Hum. Genet. 22 (2014) 762–767.
[6] M.F. Howard, Y. Murakami, A.T. Pagnamenta, D. Daumer-Haas, B. Fischer, J. Hecht, et al., Mutations in PGAP3 impair GPI-anchor maturation, causing a subtype of hyperphosphatasia with mental retardation, Am. J. Hum. Genet. 94 (2014) 278–287.
[7] S.A. Miller, D.D. Dykes, H.F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells, Nucleic Acids Res 16 (1988) 1215.
[8] P. Chomczynski, N. Sacchi, Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction, Anal. Biochem. 162 (1987) 156–159.
[9] F. Ruschendorf, P. Nurnberg, ALOHOMORA: a tool for linkage analysis using 10K SNP array data, Bioinformatics 21 (2005) 2123–2125.
[10] G.R. Abecasis, S.S. Cherny, W.O. Cookson, L.R. Cardon, Merlin—rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101.
[11] G. Novarino, A.G. Fenstermaker, M.S. Zaki, M. Hofree, J.L. Silhavy, A.D. Heiberg, et al., Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders, Science 343 (2014) 506–511.
[12] Y. Murakami, N. Kanzawa, K. Saito, P.M. Krawitz, S. Mundlos, P.M. Robinson, et al., Mechanism for release of alkaline phosphatase caused by Glycosylphosphatidylinositol deficiency in patients with hyperphosphatasia mental retardation syndrome, J. Biol. Chem. 287 (2012) 6318–6325.
[13] S. Tanaka, Y. Maeda, Y. Tashima, T. Kinoshita, Inositol seacylation of Glycosylphosphatidylinositol-anchored proteins is mediated by mammalian PGAP1 and yeast Bst1p, J. Biol. Chem. 279 (2004) 14256–14263.
[14] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[15] http://broadinstitute.github.io/picard/.
[16] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[17] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S.R. Twigg, WGS500 Consortium, et al., Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications, Nat. Genet. 46 (2014) 912–918.
[18] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010) e164.
[19] http://www.gencodegenes.org/releases/17.html.
[20] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[21] J.M. Schwarz, D.N. Cooper, M. Schuelke, D. Seelow, MutationTaster2: mutation prediction for the deep-sequencing age, Nat. Methods 11 (2014) 361–362.
[22] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081.
[23] Y. Choi, G.E. Sims, S. Murphy, J.R. Miller, A.P. Chan, Predicting the functional effect of amino acid substitutions and indels, PLoS One 7 (2012) e46688.
[24] S. Petrovski, Q. Wang, E.L. Heinzen, A.S. Allen, D.B. Goldstein, Genic intolerance to functional variation and the interpretation of personal genomes, PLoS Genet. 9 (2013) e1003709.
[25] G. Lunter, M. Goodson, Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads, Genome Res. 21 (2011) 936–939.
[26] D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.
[27] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, et al., Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, Bioinformatics 21 (2005) 650–659.
[28] O. Shmueli, S. Horn-Saban, V. Chalifa-Caspi, M. Shmoish, R. Ophir, H. Benjamin-Rodrig, et al., GeneNote: whole genome expression profiles in normal human tissues, Comptes. Rendus Biol. 326 (2003) 1067–1072.
[29] M. Safran, V. Chalifa-Caspi, O. Shmueli, T. Olender, M. Lapidot, N. Rosen, et al., Human gene-Centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE, Nucleic Acids Res. 31 (2003) 142–146.
[30] R.R. Petryszak, T. Burdett, B. Fiorelli, N.A. Fonseca, M. Gonzalez-Porta, E. Hastings, et al., Expression Atlas update – a database of gene and transcript expression from microarray and sequencing-based functional genomics experiments, Nucleic Acids Res. (2013), http://dx.doi.org/10.1093/nar/gkt1270.
[31] M. Kapushesky, T. Adamusiak, T. Burdett, A. Culhane, A. Farne, A. Filippov, et al., Gene Expression Atlas update – a value-added database of microarray and sequencing-based functional genomics experiments, Nucleic Acids Res. 40 (2012) D1077–D1081.
[32] J.U. Pontius, L. Wagner, G.D. Schuler, UniGene: a unified view of the transcriptome, in: The NCBI Handbook, National Center for Biotechnology Information, Bethesda (MD), 2003. Online NCBI resource at, http://www.ncbi.nlm.nih.gov/books/NBK21083/.