

# Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases

Gaber Bergant, MD<sup>1</sup>, Ales Maver, MD, PhD<sup>1</sup>, Luca Lovrecic, MD, PhD<sup>1</sup>, Goran Čaturilo, MD, PhD<sup>2,3</sup>, Alenka Hodzic, PhD<sup>1</sup> and Borut Peterlin, MD, PhD<sup>1</sup>

**Purpose:** We sought to determine the analytical sensitivity of several extended exome variation analysis approaches in terms of their contribution to diagnostic yield and their clinical feasibility.

**Methods:** We retrospectively analyzed the results of genetic testing in 1,059 distinct cases referred for exome sequencing to our institution. In these, we routinely employed extended exome analysis approaches in addition to basic variant analysis, including (i) copy-number variation (CNV) detection, (ii) nonconsensus splice defect detection, (ii) genomic breakpoint detection, (iv) homozygosity mapping, and (v) mitochondrial variant analysis.

**Results:** Extended exome analysis approaches assisted in identification of causative genetic variant in 44 cases, which represented a 4.2% increase in diagnostic yield. The greatest contribution was

associated with CNV analysis (1.8%) and splice variant prediction (1.2%), and the remaining approaches contributed an additional 1.2%. Analysis of workload has shown that on average nine additional variants per case had to be interpreted in the extended analysis.

**Conclusion:** We show that extended exome analysis approaches improve the diagnostic yield of heterogeneous genetic disorders and result in considerable increase of diagnostic yield of exome sequencing with a minor increase of interpretative workload.

*Genet Med* advance online publication 14 September 2017

**Key Words:** bioinformatics; diagnostic yield; exome sequencing; rare diseases; structural variation

## INTRODUCTION

Next-generation sequencing has significantly facilitated diagnostics and novel gene discovery in patients with rare genetic diseases, especially with the introduction of exome sequencing.<sup>1</sup> Despite this advancement in diagnostics of genetic diseases, a considerable proportion of cases remain undiagnosed even after exome sequencing has been performed. Although undiagnosed cases may be attributed to nongenetic factors or genetic variants outside covered regions, we hypothesized that a proportion of negative results stem from limiting exome analysis to basic inspection of single-nucleotide variations (SNVs) and small insertions or deletions (indels) in coding regions.

In addition to SNVs in coding regions of the nuclear genome, genetic diseases may be caused by pathogenic variants in noncoding regions, structural variants, mitochondrial genome variants, and several other classes of genetic variation. This is also evident from data in the ClinVar database, which is currently the largest resource of disease-associated variation, where such variants amount to almost a fifth of all genetic variants deposited.<sup>2</sup>

Typically, exome sequencing data analysis is limited to detecting and interpreting simple genetic variants in coding

regions of captured genes. Recent reports have shown that the information content of exome sequencing data exceeds the proportion currently harnessed in identification of disease-associated genetic variants.<sup>3–6</sup> Furthermore, several methods for extended analysis of exome sequencing data have been described in recent years. These include various approaches for calling copy-number variants (CNVs) in exome sequencing data.<sup>7</sup> Breakpoint detection in sites of clusters of soft-clipped reads has been shown to facilitate both CNV and chromosomal translocation event detection. Furthermore, analyzing off-target reads offers the possibility to call mitochondrial DNA variants in exome sequencing data.<sup>8</sup> Finally, high-level analysis of simple variants offers an improved detection of variants with cryptic splice defects<sup>9</sup> and analysis of patterns of homozygosity has been shown useful in identification of pathogenic variants associated with recessively inherited diseases.<sup>10</sup>

The significance of these approaches in routine diagnostic use has not yet been well characterized. Although studies have been performed to assess their significance in isolation, there is no analysis of the comprehensive contribution of employing these methods in the clinical setting. A further challenge that

<sup>1</sup>Clinical Institute of Medical Genetics, University Medical Centre Ljubljana, Ljubljana, Slovenia; <sup>2</sup>Faculty of Medicine, University of Belgrade, Belgrade, Serbia; <sup>3</sup>University Children's Hospital, Belgrade, Serbia. Correspondence: Borut Peterlin (borut.peterlin@guest.arnes.si)  
The first two authors contributed equally to this work.

Submitted 10 April 2017; accepted 27 June 2017; advance online publication 14 September 2017. doi:10.1038/gim.2017.142

has not been deservingly addressed in previous studies is the additional interpretive burden introduced by increasing the numbers of variants in extended exome analysis (EEA).

In the present study we thus aimed to show the collective benefit of extended methods of exome analysis in diagnosis of patients with a wide spectrum of genetic diseases affecting various organ systems. We evaluated the contribution of EEA to diagnostic yield in 1,059 consecutive cases of rare diseases, submitted for genetic testing to our institution. Furthermore, we also estimated the relationship between an increased diagnostic rate and increased interpretive burden, which affects the feasibility of these approaches in routine diagnostic use.

## MATERIALS AND METHODS

We analyzed the data for 1,059 consecutive pediatric and adult cases referred for diagnostic whole-exome and Mendeliome sequencing to our institution in the period from July 2014 to November 2016. We included a wide variety of disease categories, with the most common being central nervous system disorders, neuromuscular disorders, neurodevelopmental disorders, movement disorders, hearing loss, hereditary cancer syndromes, hereditary cardiac disorders, myopathies, kidney diseases, and connective tissue disorders. For a fully detailed list see **Supplementary Table 6** online.

Sequencing was performed using a standardized series of procedures, starting with an in-solution capture of exome sequences using various capture kits, including TruSight One, TruSight Exome, and Nextera Coding Exome capture kits (all manufactured by Illumina, San Diego, CA) or Agilent SureSelect Human All Exon v2 or Agilent SureSelect Human All Exon v5 capture kits (both manufactured by Agilent Technologies, Santa Clara, CA). This was followed by sequencing on Illumina MiSeq or Illumina HiSeq 2500 platform. Illumina TruSight One panel (Mendeliome sequencing in further text) was used in a majority (93%) of cases. Since data generated using Mendeliome or whole-exome sequencing was subsequently analyzed using same procedures, results were not stratified on the basis of the next-generation sequencing approach. Basic analysis, including SNV and indel discovery and annotation, was performed according to Genome Analysis Toolkit Best Practices workflow.<sup>2,11-13</sup> The strategy for exome data interpretation was primarily based on the combined disease and phenotype gene target definition approach we previously described.<sup>14</sup> For a detailed description of analytical procedures employed in basic analysis of exome sequencing data, see **Supplementary Materials and Methods**.

Furthermore, we employed the following approaches to expand the spectrum of genetic variation detected and improve identification of causative genetic variants: (i) CNV analysis, (ii) breakpoint detection, (iii) mitochondrial sequence analysis, (iv) detection of noncanonical splice site variants, and (v) identification of long runs of homozygosity.

Only the variants classified as pathogenic or likely pathogenic were considered in estimating positive yield, while

the variants of uncertain significance (VUS) were not considered. As an essential interpretive step of the diagnostic process, all the results were evaluated by a medical doctor who specialized in interpretation of exome sequencing data, followed by evaluation by a clinical geneticist prior to reporting. In all cases, patients were seen by a clinical geneticist prior to exome sequencing diagnostics and the reported variants were clinically evaluated and discussed with them prior to reporting. We only considered and reported variants if they were classified as likely diagnostic findings and if they were compatible with the clinical presentation of referral. The brief methodological details of each utilized method are outlined in the separate sections below; for a fully detailed description of implemented methods see **Supplementary Materials and Methods**.

### EEA methods

Details of the methods employed for EEA are available in the **Supplementary Methods** (sections 2.1 to 2.5). Briefly, we implemented CoNIFER software for detection of CNVs. Secondly, we searched for the presence of breakpoints by detecting clusters of soft-clipped reads in aligned exome sequencing data. Thirdly, to detect mitochondrial variants we reconstructed the mitochondrial sequence from off-target reads mapping to the mitochondrial genome. Next, we used recomputed splice effect predictions in the dbSNV database to identify variants in noncanonical splice regions. Finally, we developed an in-house approach for detecting long runs of homozygosity and validated the regions using the HomozygosityMapper tool ([www.homozygositymapper.com](http://www.homozygositymapper.com)).

### Diagnostic yield of EEA approaches

Firstly, we compared the increase of diagnostic yield for each EEA method in isolation. We then determined the collective contribution of all used methods in increasing the diagnostic yield. We counted as the added diagnostic yield all those cases where the EEA methods detected a pathogenic or likely pathogenic variant that would not have been detected by basic exome analysis approaches.

### Assessment of the interpretation workload in EEA

We evaluated the anticipated increase in interpretation workload resulting from the increased number of variants, generated by EEA. For sake of comparison across different methods, we measured the number of variants that have to be surveyed for each clinically relevant finding for basic and EEA approaches. We expressed this as the ratio between cases solved and the number of variants interpreted. We termed this metric “interpretations per solved case.” The difference of this ratio between basic approach and EEA reflected the additional workload (number of variants to be interpreted) that has to be invested for the increase in diagnostic yield (number of causative variants identified). This metric allowed for comparable assessment of workload increase across different exome analysis approaches used for interpretation of exome data.

## RESULTS

In the present study we surveyed the increase in diagnostic yield when using EEA methods in 1,059 cases referred for exome sequencing at our institution. Analysis using the basic exome analysis for detection of simple genetic variants in coding regions of the exome identified a likely genetic cause in 402 cases, representing a baseline diagnostic yield of 38.0%. Using EEA, we established a diagnosis in a further 44 cases, increasing the diagnostic yield by 4.2% to an overall 42.2%. The breakdown of results for each EEA category is presented in the sections below (Figure 1). Also shown below are four examples of visual representations of diagnostic finds for several EEA approaches (Figure 2). A summary and overview of reported variants found using EEA are shown in Table 1. Detailed figures of variants and clinical data of patients are presented in Supplementary Tables 1–5 and the Supplementary Figures.

### CNV analysis

Using CNV analysis we identified a likely genetic cause in 19 cases (Table 1), which contributed a 1.8% increase to the total diagnostic yield.

On average, we identified 1.6 CNVs in each patient's exome (SD = 3.0), which included common and rare structural variants. The number of detected CNVs was mostly affected by target capture protocol, sequencing coverage, and quality of input DNA samples. As expected, we observed over twice as many duplication events (69.2%) as deletions (30.8%).

On average, the size of called CNVs was within 1 Mb range, with mean size of 442 kb (SD = 955 kb). The majority of observed CNVs were smaller than 1 Mb (88.3%) with 1.3% ultimately reported as causative. The remaining larger CNVs

(>1 Mb) represented 11.7% of detected CNVs and were more often reported as causative (8.3%).

Of the reported CNVs, we identified major structural variants in two cases, specific microdeletions or microduplications in eight cases, and intragenic deletions in nine cases. All the identified CNVs were considered to be diagnostic; the major structural rearrangements, microdeletions, and microduplications identified could be attributed to well-established structural variation syndromes that aligned with the clinical presentation of referral. Furthermore, all the intragenic deletions and duplications were identified in genes with compatible phenotypic consequences and where exonic deletions and duplications were an established pathogenic mechanism.

In all reported cases, we confirmed the presence of the variants using either array-based comparative genomic hybridization, multiplex ligation-dependent probe amplification, or specific polymerase chain reaction-based approaches. Additionally, we identified VUS in an additional three cases.

A detailed list of all reported variants along with their coordinates can be found in Supplementary Table 1.

### Breakpoint analysis

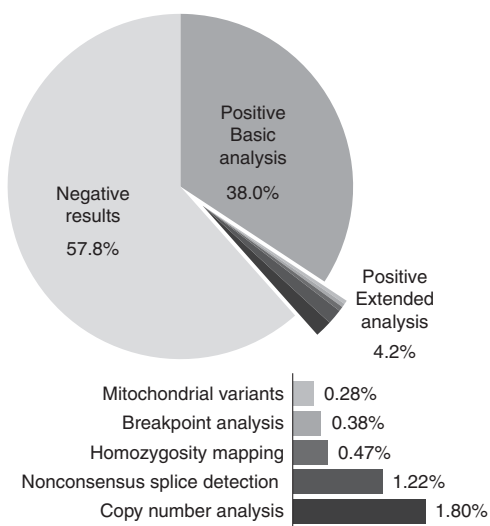
Breakpoint analysis has revealed genetic cause in four cases (Table 1), which represented a 0.4% increase in diagnostic yield. Specifically these cases included an upstream deletion of *VHL* gene in a patient with suspected Von Hippel–Lindau syndrome, a hemizygous intragenic deletion of *UPF3B* gene in a patient with autism, and an *HBB* gene deletion in a patient with porphyria and thalassemia minor. In all cases revealed by breakpoint analysis, the deletion event occurred in the exonic region but was too small to be detected using the CNV detection algorithm employed.

Filtering out the breakpoint variants with frequency over 1% has reduced the average number of suspect breakpoint detections to the average of 1.9 per each case (SD = 4.1).

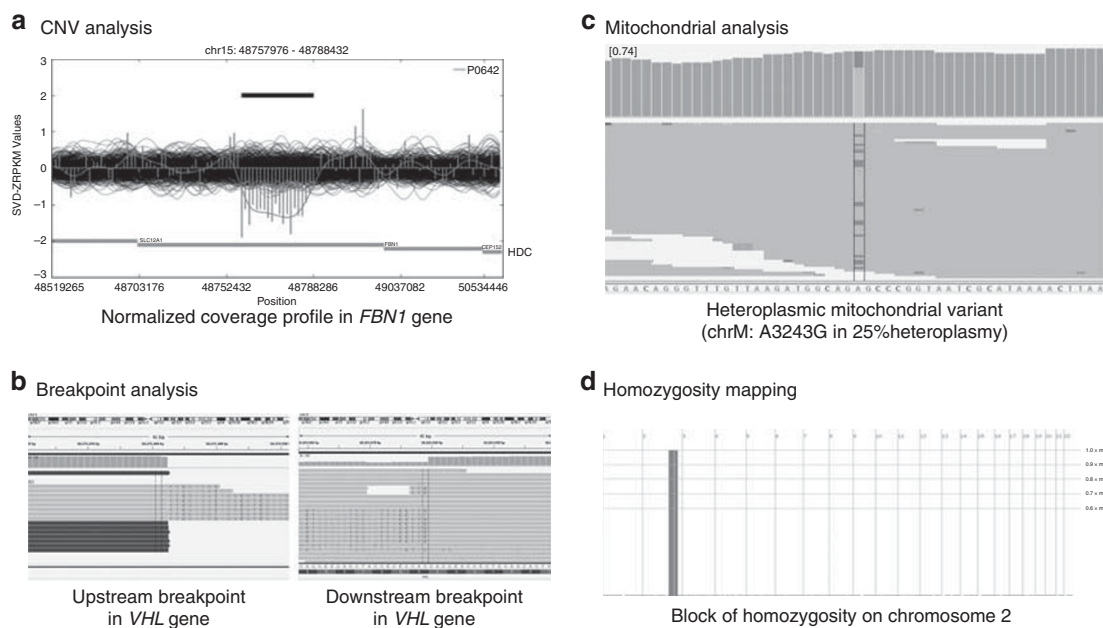
### Mitochondrial sequence analysis

Routine analysis of mitochondrial sequence variation has revealed causative mitochondrial variants in three cases (Table 1), contributing 0.3% to the overall diagnostic yield in our population. Of the identified mitochondrial variants, all were tracked as previously reported pathogenic variants in the MITOMAP database. In one case, causative variant was detected in homoplasmy, while in two cases we detected the variant in heteroplasmy, with minimum of 22% heteroplasmy detected in one case.

We observed the average mitochondrial coverage of 51.7 with SD of 78.3 (Supplementary Figure 35). Only a minor proportion of samples had median mitochondrial coverage below 15 (21.2% of cases) and thus we were able to call mitochondrial sequence variants in a majority of cases. The factor most consistently affecting mitochondrial coverage was source tissue for DNA extraction, where highest coverage was observed for DNA samples isolated from muscle tissue in



**Figure 1 Impact of extended methods of exome analysis on diagnostic yield.** Altogether, methods of extended exome analysis identified the causative variant in 4.2% of cases. The contribution of each method is also presented as a proportion of samples where diagnosis was reached using a specific method.



**Figure 2 Representation of results in extended exome analysis.** (a) Typical output of CoNIFER software, which represents a multiexonic deletion in a patient with neonatal form of Marfan syndrome. (b) Appearance of a typical breakpoint profile with the proximal and distal cluster of soft-clipped reads at the edges of a small deletion in the upstream region of the *VHL* gene, which was not detected using either basic exome analysis or copy-number variant (CNV) analysis. (c) An example of a heteroplasmic mitochondrial variant identified in a patient with stroke and cognitive decline. The ratio between reference and alternative read numbers also allowed for an approximation of the heteroplasmy levels. (d) Profile of homozygosity in a patient with undefined muscular dystrophy and a block on chromosome 3, which was ultimately revealed to harbor homozygous *GMPPB* pathogenic variant. The variant was not previously detected and was only detected after targeted Sanger sequencing of the gene (due to the *GMPPB* gene, which would otherwise not be captured on the Mendeliome panel). The plot was generated using the HomozygosityMapper tool, as described in Materials and Methods.

contrast to lower coverage for samples isolated from peripheral blood (see **Supplementary Figure 35**).

Additionally, VUS was identified in one case. On average, 3.9 variants per case passed the quality and population frequency filters ( $SD = 3.0$ ) and were thus considered in the interpretation.

#### Noncanonical splice variant detection

Predictions of splice effect in the sequences extending consensus  $\pm 1$  and  $\pm 2$  made possible identification of causative genetic variants in further 13 cases (**Table 1**), contributing 1.2% to the final yield. Interestingly, in majority nonconsensus splicing variants were detected at positions +4 and +5 (63.6% splice variants), with other splice defects detected at positions -3, -12, and two synonymous variants were predicted to affect splicing.

VUS were identified in two cases. On average, we reviewed 1.7 variants per case ( $SD = 1.6$ ).

#### Identification of long runs of homozygosity

Homozygosity mapping assisted the discovery of causative variants in five cases in our cohort (**Table 1**), contributing 0.4% to a total diagnostic yield. In most cases homozygosity mapping assisted in narrowing down the set of candidate genes, after no cause was identified using the conventional

approach. An example of these results is the identification of homozygous single-exon deletion in the *DYM* gene in a patient with short stature and developmental delay. Here, we performed a targeted survey of clinically compatible genes after identification of a block of homozygosity on chromosome 18 (**Supplementary Figure 30**) and targeted evaluation of the coverage profile revealed the presence of the deletion, which was not observed with CNV detection algorithm. In one case, we identified causative point variants in a gene that not captured with Mendeliome sequencing (*GMPPB* gene), and in two cases this approach precipitated variants in novel genes that were only later discovered as associated with the referral clinical condition (*GLDN* and *STAG3* genes).

On average, we reviewed 0.25 variants per case ( $SD = 0.7$ ).

#### Assessment of the interpretation workload in EEA

In basic exome analysis, 91 simple coding variants were reviewed per case. Considering the diagnostic yield of 38.0%, this meant that in terms of workload, 239 variants had to be interpreted to reach a diagnostic finding in basic exome analyses (**Figure 3**).

Using the same metric, we assessed the workload incurred by the EEA approaches (raw data available in **Supplementary Table 7**). On average, all the extended exome approaches cumulatively added 9.4 variants to be additionally interpreted

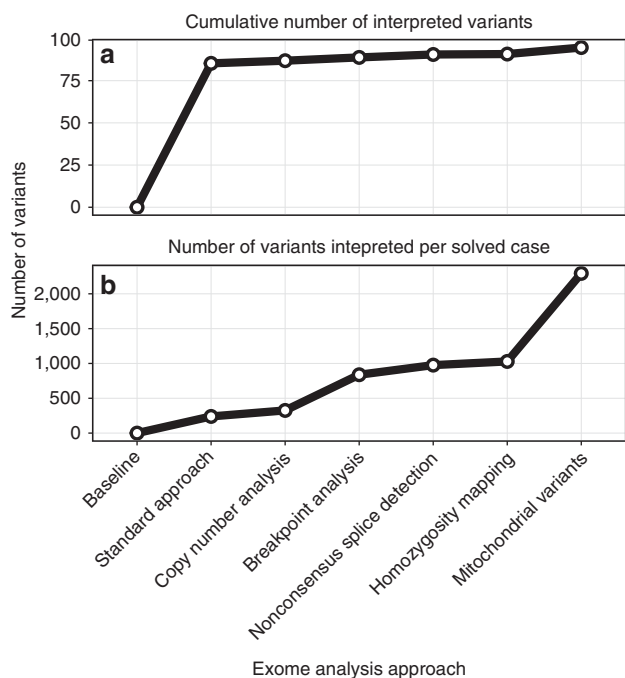
**Table 1** Variants identified using EEA methods

Patient ID	Referral diagnosis	EEA method	Validated causative variant	Genes
P0083	Global developmental delay, epilepsy, inability to speak, sunken nasal bridge	CNV analysis	arr[hg19] Xq28(153,129,811–153,630,671)x2	17 genes including <i>MECP2</i>
P0119	Suspected Lujan–Fryns syndrome	CNV analysis	arr[hg19] 7q11.22(70,232,987–70,257,820)x1	<i>AUTS2</i>
P0175	Undefined neurodevelopmental disorder	CNV analysis	arr[hg19] 19p13.12p13.11(14,802,282–16,652,245)x1	43 genes
P0189	Ocular albinism	CNV analysis	exome[hg19] 15q13.1(28,096,517–28,327,030)	<i>OCA2</i>
P0242	Intellectual disability, polygyria	CNV analysis	arr[hg19] chr2p25.3(30,341–2,908,831)x1	9 genes including <i>MYT1L</i>
P0312	Intellectual delay, motor delay, facial dysmorphism	CNV analysis	exome[hg10] Xq28(153,167,974–153,363,132) exome [hg10] Xq28(153,581,911–153,760,505)	22 genes including <i>MECP2</i>
P0371	Suspected X-linked developmental delay.	CNV analysis	arr[hg19] 1q22q23.1(155,162,566–157,670,289)x1	67 genes
P0377	Sandhoff disease	CNV analysis	arr[hg19] 5q13.3(73,981,082–73,992,976)x0	<i>HEXB</i>
P0405	Myopathy—unspecified	CNV analysis	arr[hg19] 22q11.21(18,706,001–21,561,514)x3	45 genes
P0600	Syndromic progressive cardiomyopathy	CNV analysis	arr[hg19] 3q26.31q26.32(173,525,459–176,767,936) x1, arr[hg19] 3q26.32q27.2(176,771,550–185,226,675)x3, arr[hg19] 3q28q29(188,477,890–195,296,016)x1	3 genes in the proximal deletion, 56 genes in duplication, and 31 genes in the distal deletion (all on chromosome 3)
P0642	Suspected neonatal Marfan syndrome	CNV analysis	arr[hg19] 15q21.21(48,760,124–48,789,598)x1	<i>FBN1</i>
P0815	Global developmental delay, hypotonia, macrocephaly, behavior disorders	CNV analysis	arr[hg19] 1q22(155,957,128–156,126,930)x1	<i>LMNA, RAB25, MEX3A, SEMA4A, LAMTOR2, SSR2, UBQLN4</i>
P0816	Developmental delay, hypotonia	CNV analysis	arr[hg19] 12p13.33(311,657–1,456,832)x1	9 genes
P0880	Suspected Lynch syndrome, endometrial carcinoma, colorectal carcinoma, colorectal carcinoma (daughter), endometrial carcinoma (sister)	CNV analysis	exome[hg19] 2p21p16.3(47,168,700–47,657,090)x1	5 genes including <i>MSH2</i> and <i>EPCAM</i>
P0910	Joint hypermobility, hypotonia	CNV analysis	exome[hg19] 1p36.22(12,016,963–12,027,158)	<i>PLOD1</i>
P0949	Marfan syndrome	CNV analysis	arr[hg19] 15q13.2q13.3(30,954,726–32,509,926)x3	7 genes
P1025	Erythropoietic protoporphyria, thalassemia minor	CNV analysis	arr[hg19] 11p15.4(5,241,108–5,254,510)x1	<i>HBB, HBD</i>
P1050	Polymalformative syndrome	CNV analysis	arr[hg19] 16p13.11p11.2(14,145,698–29,331,350)x3 arr[hg19] Xq21.2q27.3(80,977,681–144,853,235)x3	117 genes 261 genes
P1128	Developmental delay, suspected congenital disorder of glycosylation	CNV analysis	arr[hg19] chr6q25.3(156,922,572–157,159,419)x3 arr[hg19] chr6q25.3(157,341,859–157,870,814)x3	Complex rearrangement involving <i>ARID1B</i>
P0921	Periventricular nodal heterotopia, epilepsy	Breakpoint detection	arr[hg19] Xq28(153,473,833–153,588,359)x1 arr[hg19] Xq28(153,588,359–153,611,490)x3	Complex rearrangement involving <i>FLNA</i> gene
P0929	Suspected Von Hippel–Lindau syndrome	Breakpoint detection	exome[hg19] 3p25.3(10,175,483–10,183,579)	<i>VHL</i>
P0953	Childhood autism	Breakpoint detection	arr[hg19] Xq24(118,985,732–118,986,305)x0	<i>UPF3B</i>
P0955	Porphyria, beta thalassemia minor	Breakpoint detection	exome[hg19] 11p15.4(5,247,858–5,255,271)	<i>HBD, HBB</i>

Table 1 Continued

Patient ID	Referral diagnosis	EEA method	Validated causative variant	Genes
P0597	Suspected Leigh disease	mtDNA analysis	chrM:8993T > G (rs199476133)	<i>MT-ATP6</i>
P0791	Suspected MELAS syndrome	mtDNA analysis	chrM:3243A > G (rs199474657)	<i>MT-TL1</i>
P0918	Leber's hereditary optic neuropathy	mtDNA analysis	chrM:14,598T > C	<i>MT-ND6</i>
P0203	Tuberous sclerosis	NCSplice detection	NM_000548.3:c.5160+4A > C (htz)	<i>TSC2</i>
P0240	Suspected Marfan syndrome	NCSplice detection	NM_000138.4:c.5788+5G > A (htz)	<i>FBN1</i>
P0270	Suspected connective tissue disorder	NCSplice detection	NM_000093.4:c.1389G > A (htz)	<i>COL5A1</i>
P0315	Suspected connective tissue disorder	NCSplice detection	NM_001110556.1:c.7023+4A > T (htz)	<i>FLNA</i>
P0569	Overlapping osteogenesis imperfecta and Ehlers–Danlos syndrome features	NCSplice detection	NM_000089.3:c.432+5G > A (htz)	<i>COL1A2</i>
P0657	Connective tissue disease, cardiac valvular disease	NCSplice detection	NM_016955.3:c.388+5G > A (htz)	<i>SEPSECS</i>
P0843	Congenital insensitivity to pain	NCSplice detection	NM_002977.3:c.377+5C > T (htz)	<i>SCN9A</i>
			NM_002977.3:c.1642C > T (htz)	<i>SCN9A</i>
P0986	Suspected neurofibromatosis type 1	NCSplice detection	NM_001042492.2:c.2252-3T > G (htz)	<i>NF1</i>
P1003	Netherton syndrome	NCSplice detection	NM_001127698.1:c.891C < T (htz)	<i>SPINK5</i>
			NM_001127698.1:c.1431-12G > A (htz)	
P1076	Suspected Bardet–Biedl syndrome	NCSplice detection	NM_144596.2:c.489G > A (hmz)	<i>TTC8</i>
P1080	Progeria	NCSplice detection	NM_170707.3:c.1968+5G > A (htz)	<i>LMNA</i>
P1086	Suspected Myoshi myopathy	NCSplice detection	NM_000070.2:c.632+5G > A (htz)	<i>CAPN3</i>
			NM_000070.2:c.550delA (htz)	
P1087	Suspected Meckel–Gruber syndrome	NCSplice detection	NM_025114.3:c.5710-3C > G (htz)	<i>CEP290</i>
			NM_025114.3:c.4882C > T (htz)	<i>CEP290</i>
P0006	Skeletal dysplasia with mental retardation	Homozygosity mapping	exome[hg19] 18q21.1(46,690,043–46,690,177)	<i>DYM</i>
			(EX14 DEL)	
P0447	Premature ovarian failure	Homozygosity mapping	NM_012447.2:c.2776C > T (hmz)	<i>STAG3</i>
P0567	Severe neurological deterioration, developmental regression, leukodystrophic brain alterations	Homozygosity mapping	NM_001135659.1:c.869del (hmz)	<i>NRXN1</i>
P0811	Congenital contractures, breathing difficulties at birth	Homozygosity mapping	NM_181789.2:c.1305G > A (hmz)	<i>GLDN</i>
P1052	Limb-girdle muscular dystrophy	Homozygosity mapping	NM_013334.3:c.458C > T (hmz)	<i>GMPPB</i>

CNV, copy-number variation; EEA, extended exome analysis; MELAS, mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes; mtDNA, mitochondrial DNA; NCSplice, noncanonical splice site.



**Figure 3** This figure presents an added variant number in each case and the anticipated workload due to additional variants in using extended exome analyses. (a) The upper panel shows the increase in the variant number with the addition of extended exome analysis approaches. The number of these additional variants for evaluation was minor in comparison to the number of variants evaluated in basic exome analysis (on average, nine additional variants had to be considered per case in the full extended exome analysis). (b) The bottom panel shows the assessment of workload increase with introduction of extended exome analysis approaches. To quantify workload, we used the ratio between the number of variants interpreted and cases solved. This ratio effectively reflected the number of variants that have to be interpreted to reach diagnosis in one additional case. While the basic exome analysis had an average ratio of 239 variants per solved case, the combination of CNV analysis, splice analysis, and homozygosity mapping cumulatively had a comparable ratio of 277 interpreted variants per solved case. The ratio was less beneficial for breakpoint analysis as well as mitochondrial variation, likely because several mitochondrial variants were found in each patient, but they were only rarely considered for reporting.

in each case. The best variants interpreted per solved case ratio was obtained for CNV analysis, nonconsensus splice site prediction, and homozygosity mapping approaches in which 50 to 140 additional variants had to be inspected to diagnose an additional patient (Figure 3). The ratios were less favorable for breakpoint detection and mitochondrial sequence analysis, where an additional 513 and 1,264 variants had to be reviewed to diagnose an additional case, respectively (Figure 3).

Expectedly, we observed a favorable impact of aggregating EEA variants in a local database on the number of variants that have to be manually evaluated. Similarly to SNVs (Figure 4), we expected that this beneficial effect would also be observed for other types of genetic variants. As an example, the CoNIFER software used for CNV detection inherently filters out common CNVs as the number of analyzed cases

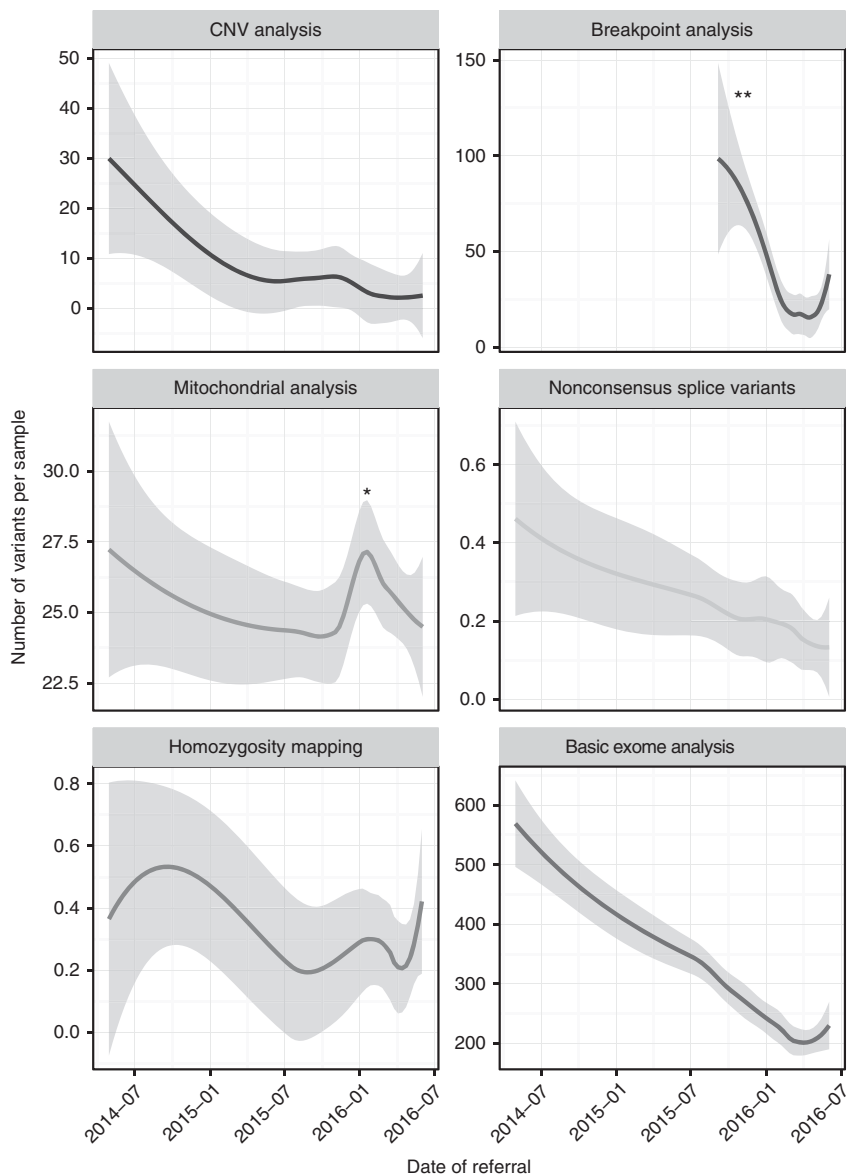
increases. In line with this, we noted that the number of CNVs was reduced from an average of 30 per sample in the beginning of the implementation to an average of three per sample, when all the collected data was included in the CoNIFER analysis. Similarly, we constructed an in-house database of breakpoints and removed those that occurred in over 1% of samples from further analyses. The implementation of the in-house breakpoint database resulted in reducing the number of breakpoints for interpretation from an average of 100 to 30 breakpoints per sample (Figure 3). The integration of databases of locally common variants collected in-house into the EEA algorithm thus appeared to facilitate the interpretation of clinically relevant findings by significantly reducing the number of variants to be interpreted.

## DISCUSSION

In the present study, we have shown that comprehensive utilization of the presented approaches increased diagnostic yield from 38.0% to 42.2% in the sample set of 1,059 cases. Among the EEA approaches used, we found that CNV detection and nonconsensus splice site prediction contributed most towards the increase of diagnostic yield. Our analysis of impact on workload due to additional variants of EEA has shown a favorable workload-to-yield relationship, demonstrating the feasibility of introducing them in daily routine use.

Although exome sequencing has significantly improved diagnosis of rare genetic conditions, a significant proportion of the cases with likely genetic etiology remain undiagnosed even after exome sequencing.<sup>14</sup> While part of these cases likely remain undiagnosed due to having pathogenic variants in noncoding or noncovered regions, we show that a considerable proportion of undiagnosed cases may be attributed to limited analysis in current routine exome diagnostics (4.2%). It has been shown that numerous types of variation in addition to currently analyzed variants can be detected in exome sequencing data, but only a few studies have systematically surveyed the proportion of cases where these extended analyses may be of benefit.<sup>8</sup> In the present study, we characterized the improvement in diagnostic yield due to systematic use of these approaches and delineated their relative contribution. We have shown that CNV analysis and nonconsensus splice site analysis contributed most to the diagnostic yield, while other approaches contributed to a lesser extent.

It is challenging to anticipate the mutational mechanism underlying a suspected genetic condition in patients. This often results in prolonged diagnostic process and use of multiple genetic tests in consecution, before a conclusive diagnosis is reached. This is evidently reflected in our results, where we identified a considerable proportion of causative CNVs (1.8%) and mitochondrial variants (0.3%) in cases where variants in nuclear DNA were initially considered a primary cause of the disease. One notable example was a case with predominant presentation of hypotonia and motor delay, which was suspected to be a monogenic myopathy.



**Figure 4** For the majority of the extended exome analysis approaches, we observed a progressive decrease of the number of interpreted variants, similar to the decrease of variants in basic exome analysis due to the growth of the in-house database of simple coding variants. With the exception of homozygosity mapping, this trend was observed in all extended exome analysis approaches. The plots present the results of analysis in the first 600 samples included in the study. \*An increased number of variants due to update of the pipeline for mitochondrial sequence variant detection. \*\*The introduction of breakpoint analysis into our pipeline in the beginning of 2016. The gray areas represent a 95% standard error interval around the trend line. CNV, copy-number variant.

This patient was thus referred to exome sequencing and no causative sequence variants could be identified. EEA, however, revealed that the patient carried a likely causative 22q11.2 duplication, which may manifest predominantly with severe muscle hypotonia in cases at the extreme spectrum (patient 0405, **Supplementary Figure 10**). This case illustrates an example where the mutational mechanism could not have been anticipated and where the diagnosis has been reclassified after the results of our survey. Without EEA such cases would likely remain undiagnosed or would require further

genetic testing to exclude the possibility of nonsequence genetic variants. Considering the comprehensive nature of exome sequencing, expansion of its scope in terms of detectable mutational mechanisms further strengthens its utility in daily diagnostic practice and contributes to earlier conclusive diagnosis. Furthermore, this expanded scope further reaffirms earlier prioritization of exome sequencing among other diagnostic approaches.

Although some recent studies have evaluated the utility of EEA approaches, most of these surveys were predominantly



focused on the technical aspects of variant detection and most were focused on detection of CNVs. Most recently, Pfundt and colleagues evaluated the impact of CNV detection on the diagnostic yield of exome sequencing in the clinical setting.<sup>15</sup> Interestingly, they could demonstrate a 2.0% increase in diagnostic yield, which is in agreement with our results of 1.8% increase of diagnostic yield with CNV analysis alone. Our results show, however, that the comprehensive use of EEA considerably improves this yield, ultimately reaching an overall 4.2%. To our knowledge, no studies on clinical use of other EEA approaches or their combined use have been published to date.

Interpretation of exome sequencing data involves significant expertise and time input to reach conclusive and clinically useful information. This includes balancing the load of information on one hand, with the possibility of missing or not reporting a medically relevant finding on the other. Expectedly, EEA approaches identified causative findings in considerably fewer cases in comparison to basic exome analysis. One may consider the use of EEA approaches prudent if the amount of work invested in the interpretation would result in a proportional increase in diagnostic yield. For this reason, we also systematically estimated the feasibility of routine use of EEA in terms of its impact on workload increase. We determined that CNV analysis, splice site variant analysis, and homozygosity mapping approaches had a highly favorable workload-to-yield ratio. On average, the cumulative workload-to-yield ratio for these three approaches was comparable (277 interpreted variants per solved case) to basic exome analysis (239 interpreted variants per solved case). A key element in reaching this efficiency was the establishment of databases of EEA findings with the aim to improve separation of pathologic from common genetic variants and technical artifacts. We observed a consistent decrease of the interpreted variants during initial stages of collecting the first few hundreds of cases. This decline reached a plateau after 500 cases were input into our database.

Introduction of additional variant types in exome interpretation may increase the rate at which VUS and incidental findings are identified. In our study population, the rate of VUS findings increased only minimally after the introduction of EEA, with six VUS identified in addition to 93 reported in basic exome analysis (raising the VUS rate from 8.8% to 9.3%). We did not identify any findings that could be classified as incidental findings in our group, despite a relatively large sample taken into consideration. It is likely that introduction of EEA in routine analysis pipelines only marginally increases the rates of VUS and incidental findings.

Although diagnostic whole-genome sequencing (WGS) is anticipated to substitute a majority of diagnostic exome sequencing, it is our opinion that Mendeliome and exome sequencing will nevertheless be used as the cornerstone testing approach in the forthcoming years. First, due to high up-front costs, WGS is primarily available in larger facilities and outside direct reach of the majority of medical

institutions. Furthermore, exome sequencing has been established as the routine testing approach in multiple institutions with a significant body of experience, evidence, and validation data accumulated in the years of routine use. Finally, implementation of WGS brings along an order of magnitude higher load in terms of required processing capacities and data storage, which makes it more difficult for widespread diagnostic use in the current form. In comparison with EEA, WGS is expected to detect all classes of findings reported in the present study, while also enabling the detection of pathogenic small deletions/duplications and nonexonic causative variation. As the interpretation of a large subset of variants identified by WGS is still challenging in the clinical setting, we consider exome sequencing with extended analysis to present a favorable compromise to maximize diagnostic sensitivity, while still retaining the accessibility and economic efficiency of exome sequencing in the clinical setting.

In conclusion, we have shown that routine use of comprehensive EEA approaches improves diagnosis of rare diseases and in our study resulted in 4.2% increase of diagnostic yield of diagnostic exome sequencing. Furthermore, we have shown a minimal increase in interpretation workload per patient, making the introduction feasible in routine diagnostic settings.

#### SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at <http://www.nature.com/gim>

#### ACKNOWLEDGMENTS

This study was supported by grant P3-0326 from the Slovenian Research Agency. We thank our colleagues at the Clinical Institute of Medical Genetics (University Medical Centre Ljubljana) for their clinical and technical assistance in generating and interpreting the data.

#### DISCLOSURE

The authors declare no conflict of interest.

#### REFERENCES

1. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N Engl J Med* 2013;369:1502–11.
2. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;42:D980–D985.
3. Krumm N, Sudmant PH, Ko A, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012;22:1525–32.
4. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat* 2016;37:235–241.
5. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009;37:e67.
6. Imai A, Nakaya A, Fahiminiya S, et al. beyond homozygosity mapping: family-control analysis based on Hamming distance for prioritizing variants in exome sequencing. *Sci Rep* 2015;5:12028.
7. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 2013;14:1.

8. Samuels DC, Han L, Li J, et al. Finding the lost treasures in exome sequencing data. *Trends Genet* 2013;29:593–599.
9. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–423.
10. Bocquet B, Marzouka Na D, Hebrard M, et al. Homozygosity mapping in autosomal recessive retinitis pigmentosa families detects novel mutations. *Mol Vis* 2013;19:2487–2500.
11. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
12. Lucas FAS, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* 2012;28:421–2.
13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
14. Ales M, Luca L, Marija V, et al. Phenotype-driven gene target definition in clinical genome-wide sequencing data interpretation. *Genet Med* 2016;18:1102–1110.
15. Pfundt R, Del rosario M, Vissers LELM, et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet Med* 2017;19:667–675.