Published online 22 May 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
doi: 10.1093/nargab/lqab037

Detection of homozygous and hemizygous complete or partial exon deletions by whole-exome sequencing

Benedetta Bigio[1,2,3], Yoann Seeleuthner[2,3], Gaspard Kerner[2,3], Mélanie Migaud[2,3], Jérémie Rosain[2,3], Bertrand Boisson[1,2,3], Carla Nasca[4], Anne Puel[1,2,3], Jacinta Bustamante[1,2,3,5], Jean-Laurent Casanova[1,2,3,6], Laurent Abel[1,2,3,*,†] and Aurelie Cobat[2,3,*,†]

[1]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065, USA, [2]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, 75015 Paris, France, [3]University of Paris, Imagine Institute, 75015 Paris, France, [4]Laboratory of Neuroendocrinology, The Rockefeller University, New York, NY 10065, USA, [5]Study Center of Immunodeficiencies, Necker Hospital for Sick Children, 75015 Paris, France and [6]Howard Hughes Medical Institute, New York, NY 10065, USA

Received August 20, 2020; Revised March 19, 2021; Editorial Decision April 21, 2021; Accepted May 03, 2021

ABSTRACT

The detection of copy number variations (CNVs) in whole-exome sequencing (WES) data is important, as CNVs may underlie a number of human genetic disorders. The recently developed HMZDelFinder algorithm can detect rare homozygous and hemizygous (HMZ) deletions in WES data more effectively than other widely used tools. Here, we present HMZDelFinder_opt, an approach that outperforms HMZDelFinder for the detection of HMZ deletions, including partial exon deletions in particular, in WES data from laboratory patient collections that were generated over time in different experimental conditions. We show that using an optimized reference control set of WES data, based on a PCA-derived Euclidean distance for coverage, strongly improves the detection of HMZ complete exon deletions both in real patients carrying validated disease-causing deletions and in simulated data. Furthermore, we develop a sliding window approach enabling HMZDelFinder_opt to identify HMZ partial deletions of exons that are undiscovered by HMZDelFinder. HMZDelFinder_opt is a timely and powerful approach for detecting HMZ deletions, particularly partial exon deletions, in WES data from inherently heterogeneous laboratory patient collections.

INTRODUCTION

Copy number variations (CNVs) are unbalanced rearrangements, classically covering >50 base pairs (bp), which increase or decrease the number of copies of specific DNA regions (1,2). There is growing evidence to implicate CNVs in common and rare genetic diseases (1,3–5). CNVs have also been linked to adaptive traits, in environmental contexts for example (3). It has been recently estimated that CNVs affect ∼5–10% of the genome, suggesting that a number of potentially disease-causing CNVs have yet to be discovered (1,6). Next-generation sequencing (NGS) techniques, such as whole-genome and whole-exome sequencing (WGS and WES), provide unprecedent opportunities for studying CNVs. Computational tools using data from WGS have been successfully used to detect CNVs (7–10), but WES-based methods have met with more limited success, mostly due to the nature of targeted enrichment protocols (11–13). Indeed, WES focuses on noncontiguous genomic targets (the exons), and most breakpoints are not sequenced. Hence, current WES-based approaches for detecting CNVs use the read depth (or coverage information) as a proxy for copy number information.

The HMZDelFinder algorithm is a recently developed coverage-based method for detecting rare homozygous and hemizygous (HMZ) deletions (14). This subset of CNVs may result in null alleles and a complete loss of gene function. Their identification may, therefore, lead to the discovery of novel genes or variations underlying Mendelian diseases. HMZDelFinder jointly evaluates the normalized per-interval coverage of all the samples of the entire dataset, making it possible to detect rare exonic HMZ deletions

*To whom correspondence should be addressed. Tel: +33 1 42 75 43 14; Fax: +33 1 42 75 42 24; Email: aurelie.cobat@inserm.fr
Correspondence may also be addressed to Laurent Abel. Email: laurent.abel@inserm.fr
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

while minimizing the number of false-positive calls due to low-coverage regions. HMZDelFinder outperformed other CNV-calling tools, such as CONIFER (15), CoNVex (16), XHMM (17), ExonDel (18), CANOES (19), CLAMMS (20) and CODEX (21), particularly for the detection of single-exon deletions (i.e. deletions spanning only one exon) (14). However, two major limitations remain to be addressed. First, HMZDelFinder has been optimized to detect HMZ deletions from an entire dataset (>500) of homogeneous exome data. Its performance for typical laboratory patient collection, which include exome data generated over time, often under different conditions, is, therefore, not optimal. Second, HMZDelFinder was not designed for the systematic detection of partial exon deletions (i.e. deletions spanning less than one exon). Here, we provide HMZDelFinder_opt, a method that extends the scope of HMZDelFinder by improving the performance of the algorithm for the calling of HMZ deletions in typical laboratory patient collections, which are generated over time, and by allowing the systematic detection of partial exon deletions.

MATERIALS AND METHODS

Patient sample

The 3954 individuals used in this study were recruited in collaborations with clinicians, and most of them (90%) present different severe infectious diseases. Probands' family members account for the remaining 10%. Although these individuals do not form a random sample, they were ascertained through a number of distinct phenotypes and in different countries. Cohort-specific effects are, therefore, not expected to bias patterns of variation. All study participants provided written informed consent for the use of their DNA in studies aiming to identify genetic risk variants for disease. Institutional Review Board (IRB) approval was obtained from The Rockefeller University and Necker Hospital for Sick Children, along with a number of collaborating institutions.

WES and bioinformatic analysis

WES and bioinformatic analysis were performed as previously described (22). Briefly, genomic DNA was extracted and sheared with a Covaris S2 Ultra-sonicator. An adaptor-ligated library (Illumina) was generated, and exome capture was performed with either SureSelect Human All Exon kits (V5-50Mb, V4-50Mb, V4+UTR here referred to as V4-71Mb, or V6-60Mb) from Agilent Technologies, or xGen Exome Research 39 Mb Panel from Integrated DNA Technologies (IDT xGen). Massively parallel WES was performed on a HiSeq 2000 or 2500 machine (Illumina), generating 100- or 125-base paired-end reads. Quality controls were applied at the lane and fastq levels. Specifically, the cutoff used for a successful lane is Pass Filter > 90%, with over 250 M reads for the high-output mode. The fraction of reads in each lane assigned to each sample (no set value) and the fraction of bases with a quality score > Q30 for read 1 and read 2 (above 80% expected for each) were also checked. In addition, the FASTQC tool kit (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to review base quality distribution, representation of the four nucleotides of particular k-mer sequences (adaptor contamination). We used the Genome Analysis Software Kit (GATK) (version 3.2.2 or 3.4–46) best-practice pipeline to analyze our WES data (23). Reads were aligned with the human reference genome (hg19), using the maximum exact matches algorithm in Burrows–Wheeler Aligner (BWA) (24). PCR duplicates were removed with Picard tools (picard.sourceforge.net/). The GATK base quality score recalibrator was applied to correct sequencing artifacts.

Positive controls

The six WES samples used as positive controls carry rare HMZ disease-causing deletions that were confirmed with state-of-the-art molecular approaches (25–27). Specifically, these HMZ deletions comprise one or more exons and have different lengths as follows (Supplementary Table S1). P1 carries a homozygous deletion of exons 21 to 23 in *DOCK8* (10 800 bp) that was validated by multiplex ligation-dependent probe amplification (MLPA). The deletion in *DOCK8* was functionally linked to staphylococcus infection (25). P2 had a homozygous deletion of exon 5 in *NCF2* (134 bp) that was also validated by MLPA and found to be causal in chronic granulomatous disease (CGD) (27). P3's homozygous deletion spanned exons 1–7 in *IL12RB1* (13 000 bp) and was validated by Sanger sequencing. This deletion was demonstrated to be causal for a Mendelian susceptibility to mycobacterial disease (26). P4 has a hemizygous deletion of the entire *CYBB* (3 400 000 bp) validated by MLPA and CGH array that resulted in CGD (27). P5 is a patient with hyper IgE syndrome carrying a homozygous deletion of exons 7–15 in entire *DOCK8* (28 000 bp) that was validated by Sanger sequencing. Finally, P6 has a homozygous deletion of exon 11 in *IFNAR1* gene, validated using Sanger sequencing (28). *CYBB* is on the X chromosome while all other genes are autosomal. The kits used for sequencing were as follows: P1, V4-50Mb; P2 and P5: V6-60Mb; P3 and P4: V5-50Mb; P6: V4-71Mb.

HMZDelFinder_opt

The general workflow used in HMZDelFinder_opt is depicted in Supplementary Figure S1. First, HMZDelFinder_opt computes coverage profiles from the BAM files of the entire dataset. Second, the principal component analysis (PCA) is calculated from a covariance matrix based on standardized coverage profiles and a k nearest neighbors algorithm is used to select the reference control set. Third, the BAM file of a given sample and the BAM files of the reference control set are used as input of HMZDelFinder to detect HMZ deletions. Fourth, when HMZDelFinder_opt is provided with the parameter –sliding_window_size and the related size, it will employ a sliding window approach for identification of partial deletions of exons. Each of these steps is described in the following paragraphs.

Principal component analysis (PCA) and *k* nearest neighbors algorithm

The PCA was performed on the coverage profile of the 3954 WES using per-exon coverage. Specifically, for each sample, the coverage profile was calculated using the mean depth of coverage of the 194 528 exons from the consensus coding sequences (CCDS) annotation of GRCh37 obtained using biomaRt (29). The PCA was then performed using the 'prcomp' function from R 3.5.1 on the scaled coverage profiles. To select the reference control set for a given sample, we computed pairwise weighted Euclidean distances between individuals i and j based on the first 10 principal components from the PCA using the 'dist' function of R 3.5.1, using the formula:

$$dist(i,j) = \sqrt{\sum_{k=1}^{10} \lambda_k \left(PC_{ki} - PC_{kj} \right)^2}$$

where PC is the matrix of principal components (PCs) calculated on common variants and λ_k the eigenvalue corresponding to the *k*-th principal component PC_k.

HMZDelFinder

We used the HMZDelFinder algorithm as described (14). In brief, HMZDelFinder calculates per-exon read depth (reads per thousand base pairs per million reads; RPKM) to detect HMZ deletions. For our purpose of covering all the coding regions, we employed an interval file containing all coding sequences from Gencode. For a given interval, the criteria to call a deletion are as follows: (i) RPKM < 0.65 and (ii) frequency of the deletion within the dataset $\leq 0.5\%$. Filtering criteria at the interval and sample levels include removal of low quality intervals (RPKM median < 7 across all samples) and removal of low quality samples (2% with highest number of calls). When using the optional absence of heterozygosity (AOH) step, HMZDelFinder uses VCF files to filter out deletions not falling in AOH regions, assuming that rare and pathogenic homozygous deletions are likely to be located within larger AOH regions due to the inheritance of a shared haplotype block from both parents. Finally, to prioritize deletions, *z*-scores are computed. The *z*-score of a deletion measures the number of standard deviations between the coverage of the deleted interval in a given sample compared to the mean coverage of the same interval in the rest of the dataset. A very low *z*-score indicates high mean coverage with low variance in the dataset and very low (or no coverage at all) in a given sample. Hence, lower *z*-scores denote higher confidence in a given deletion.

Sliding window approach and simulated data

We simulated deletions of variable size in 200 randomly selected individuals sequenced using the predominant capture kit (SureSelect V4-71Mb) as follows. First, in order to evaluate the impact of the mean exon coverage and the size of the deletion on the sensitivity of the methods, we selected two exons of similar size (~400 bp) but with different coverage profile and deleted a segment of 25%, 50%, 75% or 100% of the exon size. Exon 11 from *LIMCH1* gene was chosen to simulate a favorable case (exon of 409 bp with mean coverage of approximately 85× in our samples, see Supplementary Figure S2) and exon 4 from *RPL15* gene was chosen to simulate an unfavorable case (exon of 406 bp with mean coverage of 15X in our samples, see Supplementary Figure S2). For both exons, we deleted a segment of 25%, 50%, 75% or 100% of the exon size, using the '-v' argument of the 'bedtools intersect' command (bedtools v1.9) on the BAM file to remove all reads overlapping the segment. We then ran HMZDelFinder and HMZDelFinder_opt (with and without the –sliding_windows parameter) on the whole BAM files. Specifically, we applied a sliding window approach, in which each exon was divided into 100 bp windows, with 50 bp overlaps, and BAM files for individual exomes were transformed into per-window read depths. In a separate analysis, we used 50 bp windows with 25 bp overlaps.

Second, we assessed the sensitivity of the methods to detect known real partial exon deletions. Briefly, we identified in the HGMD database (30) deletions >150 bp and spanning less than an exon, and we simulated the three following partial deletions in 200 samples (Supplementary Table S2): (i) a deletion spanning 252 bp in exon 15 of *PKD1* linked to polycystic kidney (31), (ii) a deletion spanning 173 bp in exon 15 of *APC* linked to adenomatous polyposis coli (32) and (iii) a deletion spanning 165 bp in exon 3 of *SERPING1* linked to hereditary angioedema (33). We subsequently ran HMZDelFinder and HMZDelFinder_opt, with and without the –sliding_windows parameter.

Analysis of common deletions

To determine whether some of the called deletions were previously reported as common deletions, we utilized the CNVs from the Gold Standard track (hg19 version dated 15 May 2016) of the Database of Genomic Variants (DGV), a highly curated resource that collects CNVs in the human genome (34). We retained only entries with field 'variant_sub_type' equal to 'Loss' and frequency >1%. We then crossed the retained entries with the deletions called by HMZDelFinder and HMZDelFinder_opt in the positive controls. Deletions were considered common in the DGV database when they overlapped at least 50% with the retained entries from the DGV database.

RESULTS

Optimization of the reference control set in HMZDelFinder_opt

We first aimed to improve the performance of HMZDelFinder for detecting HMZ deletions in typical heterogeneous laboratory patient collections, which were generated over time and in different experimental settings (e.g. capture kit). We reasoned that comparing a given sample with an optimized reference control set would limit the impact of the background variability intrinsic to exome data, thereby improving the performance of HMZDelFinder. We designed the optimized reference control set as a selection of samples with similar coverage profiles (Supplementary Figure S1). We did this by first performing a PCA of the depth of coverage for consensus coding sequences (CCDS) for 3954 exomes from our

in-house cohort, including mostly patients with severe infectious diseases. As expected, given the different sequencing conditions used for WES (Supplementary Table S3), the coverage profiles of the samples were highly variable (Figure 1). The first two principal components (PCs) of the PCA identified six distinct clusters, mostly reflecting the capture kit used (Figure 1). Interestingly, two different clusters (clusters 1 and 2 on Figure 1) corresponded to the V4-71Mb capture kit, the difference between these clusters being associated mostly with a minor change in the sequencing chemistry of the kit, leading to a significant improvement in coverage profile for the more recently generated exome data (Supplementary Figure S3). We then used the first 10 PCs to calculate the pairwise weighted Euclidean distances between all samples (35) (see 'Materials and Methods' section). We used this metric to determine, for each sample of interest, the closest neighbors, for use as the reference control set in HMZDelFinder_opt.

We then compared the performances of HMZDelFinder_opt and HMZDelFinder, using six WES samples carrying validated rare HMZ disease-causing deletions of different lengths as positive controls (Supplementary Table S1, 'Materials and Methods' section). Specifically, we tested the ability of HMZDelFinder_opt and HMZDelFinder to detect the validated deletions, and we also compared the total numbers of deletions called and their *z*-scores (see 'Materials and Methods' section). In HMZDelFinder_opt, we compared reference control sets of different sizes (ranging from 50 to 500, Supplementary Figure S4), selected for each sample as described above. In HMZDelFinder, we used either the entire dataset (consisting of 3954 WES) or a dataset restricted to the samples sequenced with the same capture kit as the tested sample since the type of kit is the major source of coverage variation between samples (Figure 1). For all approaches, the final set of called deletions for each sample was narrowed down to the intervals covered by the capture kit corresponding to the patient WES data.

For positive controls P1 to P5, HMZDelFinder and HMZDelFinder_opt successfully detected the confirmed HMZ deletions, regardless of the reference control set used (Table 1). However, HMZDelFinder using the entire dataset detected a much larger total number of deletions than HMZDelFinder_opt. Specifically, the total number of deletions ranged from 11 for P3 to 2586 for P1. The very large number of potential deletions called in P1, P2 and P5 suggested that HMZDelFinder using the entire dataset called false-positive deletions. Using the optional filtering step based on the absence of heterozygosity (AOH) information for HMZDelFinder (see 'Materials and Methods' section) did not significantly reduce the number of called deletions, in particular for P1 (Table 1). We hypothesized that the large difference between the two methods for P1 reflected the low quality of exome data for this patient. Indeed, the mean coverage and the proportion of bases with coverage above 10× were much lower for P1 than for the other four patients (e.g. only 68.9% of bases had a coverage above 10× for P1, versus >99% for the other patients) (Supplementary Table S1), leading to a large number of likely false positive deletions detected when not using an appropriate reference control set with similar coverage. Consistent with this hypothesis, the total number of detected deletions dropped when using HMZDelFinder with a control dataset restricted to samples sequenced using the same capture kit. However, the total number of deletions detected by HMZDelFinder_opt (from 1 to 11 using 100 or 200 controls) was still slightly smaller than the total number of deletions detected by HMZDelFinder using the same capture kit (from 3 to 17).

For positive control P6, HMZDelFinder, using either the full exome dataset or a dataset restricted to the same capture kit, was unable to detect the known *IFNAR1* deletion (Table 1). Indeed, the read depth at the deletion locus was highly variable in the dataset, even among samples sequenced with the same V4-71Mb kit as P6 (Supplementary Figure S3), probably explaining the absence of detection. Of note, two different clusters corresponded to the V4-71Mb capture kit (clusters 1 and 2 on Figure 1). Interestingly, the deletion was detected when using HMZDelFinder with a control dataset restricted to the samples belonging to the same PCA cluster as P6 (cluster 2). Accordingly, it was also detected by HMZDelFinder_opt that selected the closest neighbors among the exomes of the same cluster. These results further underline the importance of careful control dataset selection as performed by HMZDelFinder_opt since cohorts homogeneous in terms of capture kit could nevertheless show within-kit fluctuations. Overall, the number of deletions detected with HMZDelFinder_opt was consistently larger with the largest reference sample size (500) (Table 1). We therefore performed subsequent HMZDelFinder_opt analyses with a reference sample size of 100, which provided a good trade-off between the algorithm performance and computation time.

We then compared the rankings of the confirmed deletions between HMZDelFinder using the full dataset and HMZDelFinder_opt, using the *z*-score provided by HMZDelFinder (see 'Materials and Methods' section). While the two approaches ranked the confirmed disease-causing deletions for P1 and P5 first, HMZDelFinder_opt ranked higher the confirmed disease-causing deletions for P2, P3 and P4 than HMZDelFinder (Table 1; Figure 2). Moreover, *z*-scores were consistently better with HMZDelFinder_opt (Figure 2) than with HMZDelFinder, leading to a more specific discovery of true HMZ deletions. Again, using the AOH option for HMZDelFinder slightly improved the ranking (Table 1) but did not change the *z*-score ranking. Together, these results suggest that HMZDelFinder_opt gives better *z*-scores for deletions than HMZDelFinder, which should lead to higher sensitivity in the general case. As expected, when using HMZDelFinder with a control data set restricted to the same kit or the same PCA cluster for P6, results in terms of *z*-score distribution were very close to those observed with HMZDelFinder_opt (Figure 2).

Finally, we studied the HMZ deletions called by both approaches, in addition to the validated ones, to determine whether some of the deletions identified were reported as common deletions. We used the CNVs from the gold standard track of the Database of Genomic Variants (DGV), a highly curated resource containing CNVs from the human genome (34). We focused on the positive controls with high data quality (P2, P3, P4, P5 and P6), and found that the

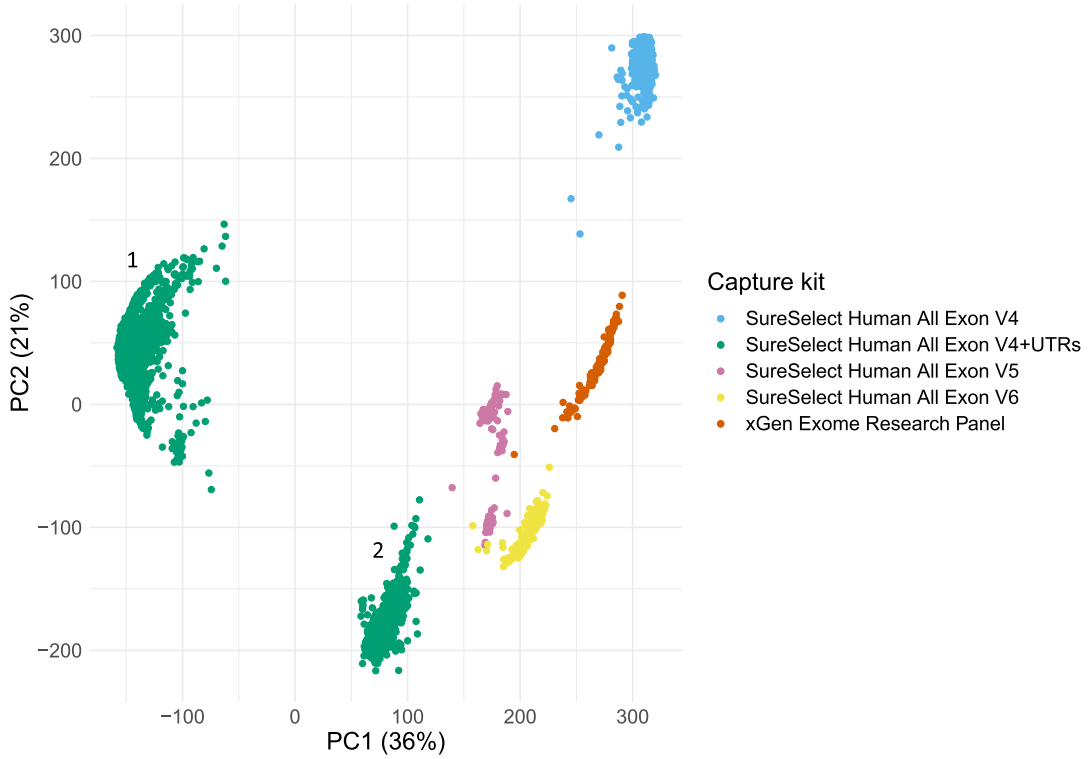

Figure 1. Principal component analysis (PCA) of the WES coverage. The PCA was computed from the coverage profiles of consensus coding sequences (CCDS) from 3954 individuals. Variance explained by each principal component is in parenthesis. Dots are color-coded by the type of the capture kit used for sequencing. Two different clusters (clusters 1 and 2) corresponded to the V4+UTR (here referred as V4-71Mb) capture kit. See also Supplementary Figure S3.

Table 1. Comparison of HMZDelFinder_opt and HMZDelFinder by using six positive controls carrying validated rare HMZ disease-causing deletions. For the first 5 positive controls, both HMZDelFinder_opt and HMZDelFinder (with or without AOH filtering step, using all samples or those restricted to the kit of the tested positive control) detect the confirmed deletions. HMZDelFinder_opt detects a lower number of other deletions and ranks higher the confirmed deletion as compared to HMZDelFinder with or without AOH filtering step, and, to a lower extent, to HMZDelFinder using a control dataset restricted to samples sequenced using the same capture kit. For the last positive control (P6), HMZDelFinder_opt successfully detect the confirmed *IFNAR1* deletion, whereas HMZDelFinder is not able to detect it, except if samples with very similar coverage profile are provided (samples from the same PCA cluster). *N*: Number of samples in the control dataset.

		P1	P2	P3	P4	P5	P6
	KIT	V4-50MB	V6-60MB	V5-50MB	V5-50MB	V6-60MB	V4-71MB
METHOD	N	Confirmed deletion (Rank/Total number of deletions)					
HMZDelFinder_opt	50	*DOCK8* (1/11)	*NCF2* (1/2)	*IL12RB1* (1/1)	*CYBB* (3/5)	*DOCK8* (1/3)	*IFNAR1* (1/1)
	100	*DOCK8* (1/11)	*NCF2* (1/2)	*IL12RB1* (1/1)	*CYBB* (4/5)	*DOCK8* (1/2)	*IFNAR1* (1/1)
	200	*DOCK8* (1/11)	*NCF2* (1/3)	*IL12RB1* (1/1)	*CYBB* (4/5)	*DOCK8* (1/3)	*IFNAR1* (1/1)
	500	*DOCK8* (4/21)	*NCF2* (1/2)	*IL12RB1* (1/3)	*CYBB* (3/5)	*DOCK8* (1/2)	*IFNAR1* (1/1)
HMZDelFinder	All[1]	*DOCK8* (1/2586)	*NCF2* (120/120)	*IL12RB1* (4/11)	*CYBB* (7/13)	*DOCK8* (1/163)	(0/0)
HMZDelFinder w/ AOH	All[1]	*DOCK8* (1/457)	*NCF2* (37/37)	*IL12RB1* (2/5)	*CYBB* (4/7)	*DOCK8* (1/46)	(0/0)
HMZDelFinder (same capture kit)	All[2]	*DOCK8* (1/17)	*NCF2* (1/3)	*IL12RB1* (1/3)	*CYBB* (3/10)	*DOCK8* (1/5)	(0/0)
HMZDelFinder (same PCA cluster)	All[3]	/	/	/	/	/	*IFNAR1* (1/1)

[1]All individuals in the cohort ($N = 3954$).
[2]All individuals of the corresponding kit. Numbers for each kit are provided in Supplementary Table S3.
[3]All individuals of the corresponding cluster ($N = 957$).

HMZ deletions called by HMZDelFinder_opt were more enriched in common deletions (frequency > 1%) than those called by HMZDelFinder using the full dataset (Supplementary Table S4). Among the 6 and 303 additional HMZ deletions called by HMZDelFinder_opt (with the reference control set of 100 exomes) and HMZDelFinder (using the full dataset), 50% and 1%, respectively, were present in the DGV database (Supplementary Table S4), suggesting that the deletions called by HMZDelFinder_opt were enriched in true deletions. As expected, when using HMZDelFinder with a control data set restricted to the same kit, the enrichment of common deletions called by HMZDelFinder improved (Supplementary Table S4). Overall, these findings demonstrate that the use of an appropriate reference control

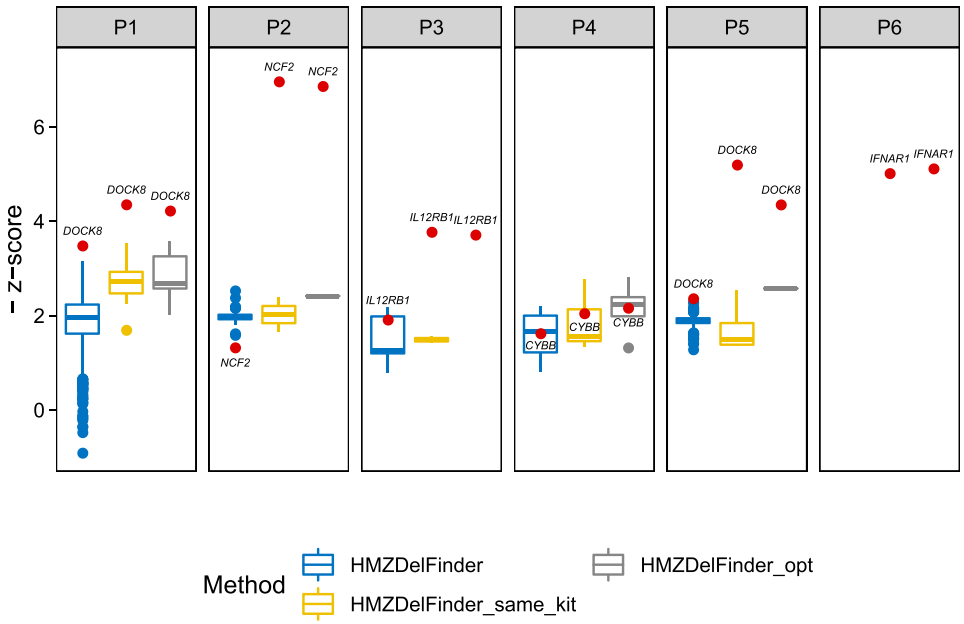

Figure 2. Comparison of the ranking of the deletions called by HMZDelFinder_opt and HMZDelFinder in six positive controls carrying validated rare HMZ disease-causing deletions. Results are shown for HMZDelFinder_opt (grey, right) using 100 as size of the reference control set. For HMZDelFinder either the full dataset (blue, left) or a dataset restricted the same capture kit (yellow, middle) were used as controls. For P6, no homozygous deletion was detected by HMZDelFinder full or same kit and results are shown for control dataset restricted to samples from the same PCA cluster (middle). The ranking is expressed as minus *z*-score. Lower *z*-scores (and higher ranking) indicate more confidence in a given deletion. The confirmed deletions (red dot) ranked first in P1, P2, P3, P5, P6 with HMZDelFinder_opt and HMZDelFinder using only controls sequenced with the same capture kit (HMZDelFinder_same_kit) or same PCA cluster for P6, while they ranked first only in P1 and P5 with HMZDelFinder without any selection of a proper control dataset.

set of WES data based on a PCA-derived coverage distance improves the performance of HMZDelFinder. These results also provided a first validation of HMZDelFinder_opt for six confirmed disease-causing HMZ deletions.

Detection of HMZ partial exon deletions by HMZDelFinder_opt

In HMZDelFinder, individual exome BAM files are transformed into per-exon read depths, facilitating a more efficient detection of single-exon HMZ deletions than can be achieved with other classical CNV-calling algorithms (14). Here, we aimed to address the need for the identification of even smaller HMZ deletions, spanning less than an exon (partial exon deletions). To this end, we used HMZDelFinder_opt with a sliding window approach, in which each exon was divided into 100 bp windows, with 50 bp overlaps, and BAM files for individual exomes were transformed into per-window read depths. We tested this approach by simulating deletions in two exons of similar size (~400 bp) but with different mean coverages in a randomly selected dataset of 200 WES samples from our in-house cohort. The deletions spanned 100%, 75%, 50% or 25% of either exon 11 of *LIMCH1* (409 bp, ~85× mean coverage) or exon 4 of *RPL15* (406 bp, ~15× mean coverage). We used these datasets to compare the performances of HMZDelFinder_opt with sliding windows of 100 bp (HMZDelFinder_opt+sw100) or 50 bp (HMZDelFinder_opt+sw50), HMZDelFinder_opt without sliding windows (HMZDelFinder_opt) and the original HMZDelFinder. For HMZDelFinder_opt+sw100, HMZDelFinder_opt+sw50 and HMZDelFinder_opt, we used reference control sets of size 100. For the original HMZDelFinder, we used all other samples from the same capture kit as controls.

For deletions spanning the full exon (100%), we confirmed that HMZDelFinder_opt had a detection rate (98% and 93% for exons with higher and lower coverage, respectively, Figure 3A) similar to that of HMZDelFinder (98% and 93% for exons with higher and lower coverage, respectively). However, the total number of HMZ deletions called by HMZDelFinder_opt was only one eighth the total number of HMZ deletions called by HMZDelFinder (median number of HMZ deletions: 2 versus 13, Supplementary Figure S5A). The detection rate was slightly higher when sliding windows were used (detection rate for HMZDelFinder_opt+sw100 of 99% and 94% for exons with a higher and lower coverage, respectively) but at the cost of a slightly larger total number of HMZ deletions called than for HMZDelFinder_opt (median number of deletions: 5 versus 2, Supplementary Figure S5A). Nevertheless, the total number of HMZ deletions called by HMZDelFinder_opt+sw100 remained lower than the total number of HMZ deletions called by HMZDelFinder.

For partial exon deletions, the detection rates of HMZDelFinder and HMZDelFinder_opt were much lower, at <10% for deletions spanning 75% of the exon and 0% for deletions spanning 25% or 50% of the exon. Conversely, HMZDelFinder_opt+sw100 succeeded in detecting simulated deletions spanning 50% or 75% (200 or ~300 bp) of both exon 11 of *LIMCH1* and exon 4 of *RPL15* in 99% of the samples, with a median number of called

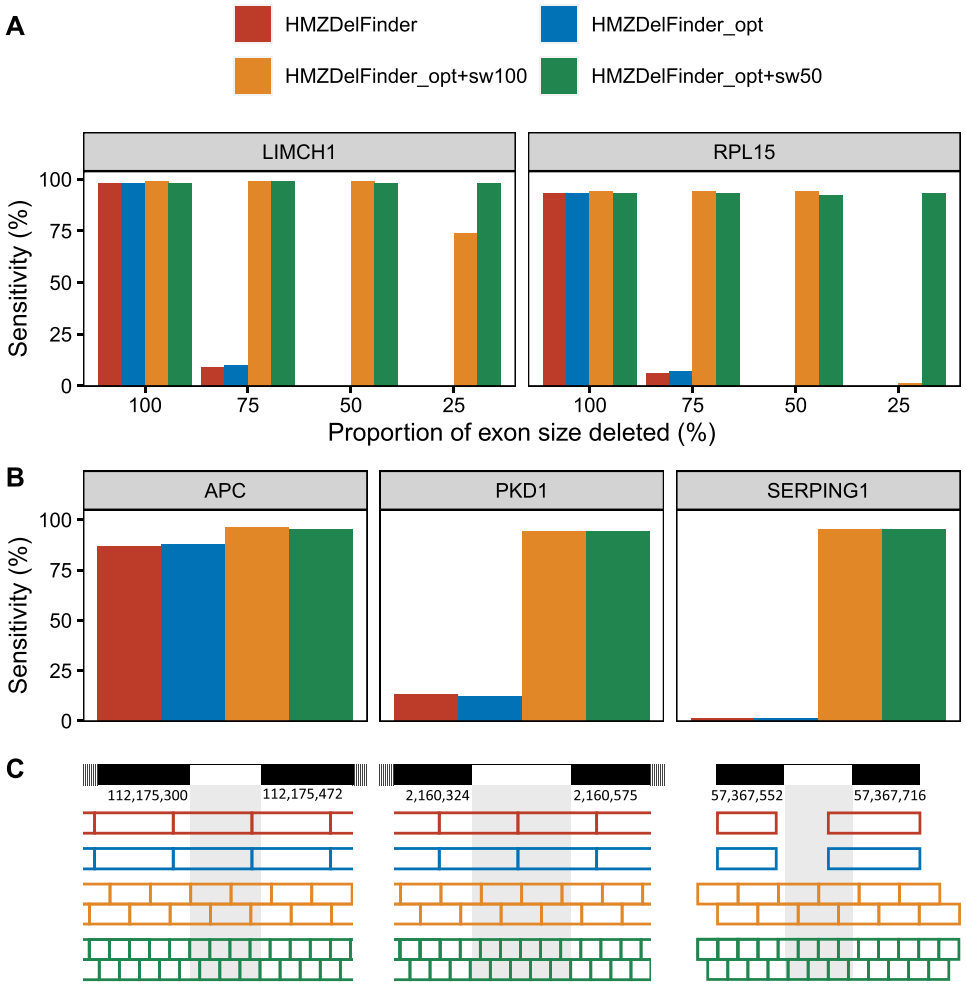

Figure 3. Comparison of HMZDelFinder_opt with or without sliding windows and HMZDelFinder by using simulated partial exon deletions in exome data. (**A**) Proportions of partial exon deletions detected in simulated data ($N = 200$ individuals) in the higher (*LIMCH1*) or lower (*RPL15*) covered exons by using HMZDelfinder (red), HMZDelFinder_opt (blue), HMZDelFinder_opt with 100 bp sliding windows (yellow), HMZDelFinder_opt with 50 bp sliding windows (green). (**B**) Proportions of real-world partial exon deletions detected in simulated data ($N = 200$ individuals). (**C**) Visual depictions of the real-world partial exon deletions (white rectangles) and their surrounding coding sequences (black) in the corresponding exon of *APC*, *PKD1* and *SERPING1* gene. The boxes below represent the positions of the intervals used by the default HMZDelFinder (red), HMZDelFinder_opt (blue), HMZDelFinder_opt with sliding windows of 100 bp (yellow) and HMZDelFinder_opt with sliding windows of 50 bp (green).

HMZ deletions of 5 (Figure 3A and Supplementary Figure S5A). For deletions spanning 25% of the exon (∼100 bp), HMZDelFinder_opt+sw100 had a detection rate of 74% for the exon with the highest coverage in *LIMCH1,* but it failed to detect the deletions in the exon with the lowest coverage in *RPL15*. We assessed the performance of this method further, using a smaller sliding window of 50 bp in size, and a step size of 25 bp, to improve granularity. We found that the use of smaller sliding windows with HMZDelFinder_opt+sw50 greatly increased the detection rate for deletions spanning 25% of the exon with the lowest coverage, exon 4 of *RPL15* (93% for sw50 versus 1% for sw100) and of the exon with the highest coverage in *LIMCH1* (98% for sw50 versus 74% for sw100) (Figure 3A).

We further assessed the sensitivity of the methods to detect known real partial exon deletions by simulating three known partial exon deletions in genes *PKD1, APC* and *SERPING1* reported in the HGMD database (30) (Supplementary Table S2). Again, the use of sliding windows increased the power to detect partial exon deletions (Figure 3B). Specifically, in the *SERPING1* and *PKD1* exons, the partial deletion was detected in only 1% to 13% of the samples using HMZDelFinder or HMZDelFinder_opt but in 94% to 95% of the samples using the sliding windows (sw50 or sw100). In the *APC* exon, HMZDelFinder and HMZDelFinder_opt detected the partial deletion in >87% of the samples. Indeed, the default interval file for HMZDelFinder splits this *APC* exon of 6.5kb in multiple non overlapping intervals of 200 bp. By chance, the partial-deletion of 252 bp almost completely overlaps one interval, allowing the detection (Figure 3C). Nevertheless, the use of sliding windows still improves the sensitivity, as sw50 and sw100 detect the deletion in >95% of the samples (Figure 3B). The median number of detected deletions was lower in HMZDelFinder_opt as compared with HMZDelFinder and slightly higher in HMZDelFinder_opt

using the sliding windows (sw50 or sw100) as compared with HMZDelFinder_opt (Supplementary Figure S5B). Overall, the use of the sliding window strategy makes it possible to detect HMZ partial exon deletions that would otherwise be missed, and the use of simulated data further validated the interest of HMZDelFinder_opt.

DISCUSSION

WES offers unprecedent opportunities for identifying HMZ deletions as novel causal determinants of human diseases, but it poses a number of computational challenges. Most current methods for detecting HMZ deletions compare the depth of coverage between a given exome and the rest of the exomes in the dataset. However, coverage depth is heavily dependent on sequencing conditions, which are continually evolving in typical laboratory settings. Thus, the exome data generated over time are inevitably heterogeneous, complicating the discovery of deletions. Using HMZDelFinder_opt with both validated disease-causing deletions and simulated data, we demonstrated that the *a priori* selection of a reference control set with a coverage profile similar to that of the WES sample studied reduced the number of deletions detected, while improving the ranking of the true HMZ deletion. These results are consistent with a recent report showing that the selection of an appropriate reference control set with multidimensional scaling significantly improves the sensitivity of various CNV callers (36). In further support for our findings, the ranking of the known deletion and the number of additional deletions detected by HMZDelFinder_opt start worsening with increasing numbers of controls in the reference set, including neighbors with a less similar coverage profile, as illustrated, for P1, in Supplementary Figure S4A. In addition, the ability of HMZDelFinder_opt, but not HMZDelFinder, to detect the confirmed *IFNAR1* deletion in positive control P6 further underlines the importance of careful control dataset selection since cohorts homogeneous in terms of capture kit could nevertheless show within-kit fluctuations (Figure 1). A possible limitation of this approach is the presence of the same deletion in several exomes that are used as controls when analyzing a sample of patients with similar medical conditions, for example due to a founder effect. In that case, patients with some degree of relatedness with the tested patient could be removed from the set of controls, while the ideal would be to restrict the controls to subjects without the disease under study if large enough.

In addition to providing an optimized tool for detecting deletions in typical laboratory patient collections, HMZDelFinder_opt also fills the gap in the study of deletions spanning less than an exon, by providing the first tool for the systematic identification of partial exon deletions. Existing CNV callers are optimized for the detection of either large deletions (usually spanning more than three exons), or deletions of full single exons (14,37). Other established callers, such as GATK, are not designed to detect CNVs and can therefore identify deletions of only a few dozen base pairs (typically up to 50 bp, https://gatkforums.broadinstitute.org/gatk/discussion/5938/using-gatk-tool-how-long-insertion-deletion-could-be-detected and (38)). The human genome contains ∼235 000 exons, about 20% of which are >200 bp (39). HMZDelFinder_opt therefore makes possible the systematic discovery of currently unknown HMZ deletions in ∼47 000 exons that are not detectable with other tools. Future extension are also warranted to investigate how HMZDelFinder_opt could be extended for the detection of heterozygous deletions. In sum, we describe HMZDelFinder_opt, a method for improving the detection of HMZ deletions in heterogeneous exome data that can be used to identify partial exon deletions that would otherwise be missed, through an extension of the scope of HMZDelFinder.

DATA AVAILABILITY

The code for the PCA-based selection and sliding window is available in the GitHub repository (https://github.com/casanova-lab/HMZDelFinder_opt/).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank the members of the Human Genetics of Infectious Diseases Laboratory for helpful discussions. We also thank Yelena Nemiroskaya, Dominick Papandrea, Mark Woollett, Dana Liu (St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, New York, USA), and Cécile Patissier, Lazaro Lorenzo-Diaz, Christine Rivalain (Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, Paris, France) for their assistance. J.R. is supported by Inserm PhD program (‘poste d’accueil Inserm’) and the MD-PhD program of Imagine Institute with the support of the Bettencourt-Schueller Foundation.

FUNDING

National Institutes of Health (NIH) [R01AI088364, R37AI095983, U19AI111143, R01AI127564, P01AI061093 to J.-L.C.]; National Center for Research Resources; National Center for Advancing Translational Sciences [8UL1TR001866]; National Human Genome Research Institute (NHGRI) [UM1HG006504, U24HG008956]; High Performance Computing Center [NIH Research Infrastructure Program S10OD018521]; Rockefeller University; St. Giles Foundation; Howard Hughes Medical Institute; Institut National de la Santé et de la Recherche Médicale (INSERM); University of Paris; French National Research Agency (ANR) [ANR-10-IAHU-01], the Integrative Biology of Emerging Infectious Diseases Laboratory of Excellence from the French National Research Agency (ANR) [ANR-10-LABX-62-IBEID], GENMSMD [ANR-16-CE17.0005–01 to J.B.], ANR-LTh-MSMD-CMCD [ANR-18-CE93-0008-01 to A.P.], ProgLegio [ANR-15-CE17-0014], Landscardio [ANR-19-CE15-0010-01 to A.C.]; French Foundation for Medical Research (FRM)

[EQU2019030007798]; SCOR Corporate Foundation for Science; Fonds de Recherche en Santé Respiratoire [SRC2017 to J.B.]; ECOS Nord [C19S01-63407 to J.B.]; the Yale Center for Mendelian Genomics funded by the National Human Genome Research Institute (UM1HG006504).
Conflict of interest statement. None declared.

REFERENCES

1. Zarrei,M., MacDonald,J.R., Merico,D. and Scherer,S.W. (2015) A copy number variation map of the human genome. *Nat. Rev. Genet.*, **16**, 172–183.
2. Collins,R.L., Brand,H., Karczewski,K.J., Zhao,X., Alföldi,J., Francioli,L.C., Khera,A.V., Lowther,C., Gauthier,L.D., Wang,H. *et al.* (2020) A structural variation reference for medical and population genetics. *Nature*, **581**, 444–451.
3. Perry,G.H., Dominy,N.J., Claw,K.G., Lee,A.S., Fiegler,H., Redon,R., Werner,J., Villanea,F.A., Mountain,J.L., Misra,R. *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.*, **39**, 1256–1260.
4. Zhang,F., Gu,W., Hurles,M.E. and Lupski,J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.
5. Lee,C. and Scherer,S.W. (2010) The clinical context of copy number variation in the human genome. *Expert Rev. Mol. Med.*, **12**, e8.
6. Sharp,A.J., Cheng,Z. and Eichler,E.E. (2006) Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 407–442.
7. Handsaker,R.E., Korn,J.M., Nemesh,J. and McCarroll,S.A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.*, **43**, 269–276.
8. Zhou,B., Ho,S.S., Zhang,X., Pattni,R., Haraksingh,R.R. and Urban,A.E. (2018) Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J. Med. Genet.*, **55**, 735–743.
9. Gross,A.M., Ajay,S.S., Rajan,V., Brown,C., Bluske,K., Burns,N.J., Chawla,A., Coffey,A.J., Malhotra,A., Scocchia,A. *et al.* (2019) Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet. Med.*, **21**, 1121–1130.
10. Belkadi,A., Bolze,A., Itan,Y., Cobat,A., Vincent,Q.B., Antipenko,A., Shang,L., Boisson,B., Casanova,J.L. and Abel,L. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 5473–5478.
11. Kadalayil,L., Rafiq,S., Rose-Zerilli,M.J.J., Pengelly,R.J., Parker,H., Oscier,D., Strefford,J.C., Tapper,W.J., Gibson,J., Ennis,S. *et al.* (2015) Exome sequence read depth methods for identifying copy number changes. *Brief. Bioinform.*, **16**, 380–392.
12. Fromer,M., Moran,J.L., Chambert,K., Banks,E., Bergen,S.E., Ruderfer,D.M., Handsaker,R.E., McCarroll,S.A., O'Donovan,M.C., Owen,M.J. *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.*, **91**, 597–607.
13. Tan,R., Wang,Y., Kleinstein,S.E., Liu,Y., Zhu,X., Guo,H., Jiang,Q., Allen,A.S. and Zhu,M. (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat.*, **35**, 899–907.
14. Gambin,T., Akdemir,Z.C., Yuan,B., Gu,S., Chiang,T., Carvalho,C.M.B., Shaw,C., Jhangiani,S., Boone,P.M., Eldomery,M.K. *et al.* (2017) Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res.*, **45**, 1633–1648.
15. Krumm,N., Sudmant,P.H., Ko,A., O'Roak,B.J., Malig,M., Coe,B.P., Quinlan,A.R., Nickerson,D.A. and Eichler,E.E. (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res.*, **22**, 1525–1532.
16. Amarasinghe,K.C., Li,J. and Halgamuge,S.K. (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics*, **14**, S2.
17. Fromer,M. and Purcell,S.M. (2014) Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr. Protoc. Hum. Genet.*, **81**, doi:10.1002/0471142905.hg0723s81.
18. Guo,Y., Zhao,S., Lehmann,B.D., Sheng,Q., Shaver,T.M., Stricker,T.P., Pietenpol,J.A. and Shyr,Y. (2014) Detection of internal exon deletion with exon Del. *BMC Bioinform.*, **15**, 332.
19. Backenroth,D., Homsy,J., Murillo,L.R., Glessner,J., Lin,E., Brueckner,M., Lifton,R., Goldmuntz,E., Chung,W.K. and Shen,Y. (2014) CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res.*, **42**, e97.
20. Packer,J.S., Maxwell,E.K., O'Dushlaine,C., Lopez,A.E., Dewey,F.E., Chernomorsky,R., Baras,A., Overton,J.D., Habegger,L. and Reid,J.G. (2016) CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*, **32**, 133–135.
21. Jiang,Y., Oldridge,D.A., Diskin,S.J. and Zhang,N.R. (2015) CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.*, **43**, e39.
22. Maffucci,P., Bigio,B., Rapaport,F., Cobat,A., Borghesi,A., Lopez,M., Patin,E., Bolze,A., Shang,L., Bendavid,M. *et al.* (2019) Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 950–959.
23. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
24. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
25. Aydin,S.E., Kilic,S.S., Aytekin,C., Kumar,A., Porras,O., Kainulainen,L., Kostyuchenko,L., Genel,F., Kütükcüler,N., Karaca,N. *et al.* (2015) DOCK8 deficiency: clinical and immunological phenotype and treatment options - a review of 136 patients. *J. Clin. Immunol.*, **35**, 189–198.
26. Rosain,J., Oleaga-Quintas,C., Deswarte,C., Verdin,H., Marot,S., Syridou,G., Mansouri,M., Mahdaviani,S.A., Venegas-Montoya,E., Tsolia,M. *et al.* (2018) A Variety of Alu-Mediated Copy Number Variations Can Underlie IL-12Rβ1 Deficiency. *J. Clin. Immunol.*, **38**, 617–627.
27. Blancas-Galicia,L., Santos-Chávez,E., Deswarte,C., Mignac,Q., Medina-Vera,I., León-Lara,X., Roynard,M., Scheffler-Mendoza,S.C., Rioja-Valencia,R., Alvirde-Ayala,A. *et al.* (2020) Genetic, Immunological, and Clinical Features of the First Mexican Cohort of Patients with Chronic Granulomatous Disease. *J. Clin. Immunol.*, **40**, 475–493.
28. Bastard,P., Manry,J., Chen,J., Rosain,J., Seeleuthner,Y., AbuZaitun,O., Lorenzo,L., Khan,T., Hasek,M., Hernandez,N. *et al.* (2021) Herpes simplex encephalitis in a patient with a distinctive form of inherited IFNAR1 deficiency. *J. Clin. Invest.*, **131**, e139980.
29. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
30. Cooper,D.N., Ball,E.V. and Krawczak,M. (1998) The human gene mutation database. *Nucleic Acids Res.*, **26**, 285–287.
31. Rossetti,S., Chauveau,D., Walker,D., Saggar-Malik,A., Winearls,C.G., Torres,V.E. and Harris,P.C. (2002) A complete mutation screen of the ADPKD genes by DHPLC. *Kidney Int.*, **61**, 1588–1599.
32. Nordling,M., Engwall,Y., Wahlström,J., Wiklund,L., Eriksson,M.A., Gustavsson,B., Fasth,S., Larsson,P.A. and Martinsson,T. (1997) Novel mutations in the APC gene and clinical features in Swedish patients with polyposis coli. *Anticancer Res.*, **17**, 4275–4280.
33. Bos,I.G., Lubbers,Y.T., Roem,D., Abrahams,J.P., Hack,C.E. and Eldering,E. (2003) The functional integrity of the serpin domain of C1-inhibitor depends on the unique N-terminal domain, as revealed by a pathological mutant. *J. Biol. Chem.*, **278**, 29463–29470.
34. MacDonald,J.R., Ziman,R., Yuen,R.K.C., Feuk,L. and Scherer,S.W. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
35. Belkadi,A., Pedergnana,V., Cobat,A., Itan,Y., Vincent,Q.B., Abhyankar,A., Shang,L., El Baghdadi,J., Bousfiha,A., Alcais,A. *et al.* (2016) Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *PNAS*, **113**, 6713–6718.
36. Kuśmirek,W., Szmurło,A., Wiewiórka,M., Nowak,R. and Gambin,T. (2019) Comparison of kNN and k-means optimization methods of

reference set selection for improved CNV callers performance. *BMC Bioinform.*, **20**, 266–266.
37. de Ligt,J., Boone,P.M., Pfundt,R., Vissers,L.E.L.M., Richmond,T., Geoghegan,J., O'Moore,K., de Leeuw,N., Shaw,C., Brunner,H.G. *et al.* (2013) Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum. Mutat.*, **34**, 1439–1448.
38. Shigemizu,D., Miya,F., Akiyama,S., Okuda,S., Boroevich,K.A., Fujimoto,A., Nakagawa,H., Ozaki,K., Niida,S., Kanemura,Y. *et al.* (2018) IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis. *Sci. Rep.*, **8**, 5608.
39. Sakharkar,M.K., Chow,V.T.K. and Kangueane,P. (2004) Distributions of exons and introns in the human genome. *In Silico Biol.*, **4**, 387–393.