Genome Biology

METHOD Open Access

ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data

Egor Dolzhenko[1†], Mark F. Bennett[2,3,4†], Phillip A. Richmond[5†], Brett Trost[6,7], Sai Chen[1], Joke J. F. A. van Vugt[8], Charlotte Nguyen[6,7,9], Giuseppe Narzisi[10], Vladimir G. Gainullin[1], Andrew M. Gross[1], Bryan R. Lajoie[1], Ryan J. Taft[1], Wyeth W. Wasserman[5], Stephen W. Scherer[6,7,9,11], Jan H. Veldink[8], David R. Bentley[12], Ryan K. C. Yuen[6,7,9†], Melanie Bahlo[2,3†] and Michael A. Eberle[1*†]

* Correspondence: meberle@illumina.com
†Egor Dolzhenko, Mark F. Bennett, Phillip A. Richmond, Ryan K. C. Yuen, Melanie Bahlo and Michael A. Eberle contributed equally to this work.
[1]Illumina Inc., 5200 Illumina Way, San Diego, CA 92122, USA
Full list of author information is available at the end of the article

Abstract

Repeat expansions are responsible for over 40 monogenic disorders, and undoubtedly more pathogenic repeat expansions remain to be discovered. Existing methods for detecting repeat expansions in short-read sequencing data require predefined repeat catalogs. Recent discoveries emphasize the need for methods that do not require pre-specified candidate repeats. To address this need, we introduce ExpansionHunter Denovo, an efficient catalog-free method for genome-wide repeat expansion detection. Analysis of real and simulated data shows that our method can identify large expansions of 41 out of 44 pathogenic repeats, including nine recently reported non-reference repeat expansions not discoverable via existing methods.

Keywords: Repeat expansions, Short tandem repeats, Whole-genome sequencing data, Genome-wide analysis, Friedreich ataxia, Myotonic dystrophy type 1, Huntington disease, Fragile X syndrome

Background

High-throughput whole-genome sequencing (WGS) has experienced rapid reductions in per-genome costs over the past 10 years [1] driving population-level sequencing projects and precision medicine initiatives at an unprecedented scale [2–7]. The availability of large sequencing datasets now allows researchers to perform comprehensive genome-wide searches for disease-associated variants. The primary limitations of these studies are the completeness of the reference genome and the ability to identify putative causal variations against the reference background. A wide variety of software tools can identify variations relative to the reference genome such as single nucleotide variants and short (1–50 bp) insertions and deletions [8–13], copy number variants [14, 15], and structural variants [15–17]. A common feature of these variant callers is their reliance on sequence reads that at least partially align to the reference genome. However, because some variants include large amounts of inserted sequence relative to the reference, methods that can analyze reads that do not align to the reference are also needed.

A particularly important category of variants that involve long insertions relative to the reference genome are repeat expansions (REs). An example of which is the expansion in *C9orf72* associated with amyotrophic lateral sclerosis (ALS). This repeat consists of three copies of CCGGGG motif in the reference (18 bp total) whereas the pathogenic mutations are comprised of at least 30 copies of the motif (180 bp total) and may encompass thousands of bases [18, 19]. REs are known to be responsible for dozens of monogenic disorders [20, 21].

Several recently developed tools can detect REs longer than the standard short-read sequencing read length of 150 bp [22–27]. These tools have all been demonstrated to be capable of accurately detecting pathogenic expansions of simple short tandem repeats (STRs). However, recent discoveries have shown that many pathogenic repeats have complex structures and hence require more flexible methods. For instance, (a) REs causing spinocerebellar ataxia types 31 and 37; familial adult myoclonic epilepsy types 1, 2, 3, 4, 6, and 7; and Baratela-Scott syndrome [28–34] occur within an inserted sequence relative to the reference; (b) expanded repeats recently shown to cause spinocerebellar ataxia, familial adult myoclonic epilepsy, and cerebellar ataxia with neuropathy and bilateral vestibular areflexia syndrome have different composition relative to the reference STR [28–33]; (c) Unverricht-Lundborg disease, a type of progressive myoclonus epilepsy, is caused by an expansion of a larger, dodecamer (12-mer) motif repeat [35]. None of the existing variant calling methods is capable of discovering all of these REs.

We have developed ExpansionHunter Denovo (EHdn), a novel method for performing a genome-wide search for expanded repeats, to address the limitations of the existing approaches. EHdn scans the existing alignments of short reads from one or many sequencing libraries, including the unaligned and misaligned reads, to identify approximate locations of long repeats and their nucleotide composition. Unlike other methods designed to identify REs [22–27], EHdn (a) does not require prior knowledge of the genomic coordinates of the REs, (b) can detect nucleotide composition changes within the expanded repeats, and (c) is applicable to both short and long motifs. EHdn is computationally efficient because it does not re-align reads and, depending on the sensitivity settings, can analyze a single 30–40x WGS sample in about 30 min to 2 h using a single CPU thread.

In this study, we demonstrate that EHdn can be used to rediscover the REs associated with fragile X syndrome (FXS), Friedreich ataxia (FRDA), myotonic dystrophy type 1 (DM1), and Huntington disease (HD) using case-control analysis to compare a small number of affected individuals ($N = 14–35$) to control samples ($N = 150$). We also show that REs in individual samples can be identified using outlier analysis. We then characterize large (longer than the read length) repeats in our control cohort to investigate baseline variability of these long repeats. Finally, we demonstrate the capabilities of our method by analyzing simulated expansions of various classes of tandem repeats known to play an important role in human disease. Taken together, our findings demonstrate that EHdn is a robust tool for identifying novel pathogenic repeat expansions in both cohort and single-sample outlier analysis, capable of identifying a new, previously inaccessible class of REs.

Results

ExpansionHunter Denovo

Overview

The length of disease-causing REs tends to exceed the read length of modern short-read sequencing technologies [36]. Thus, pathogenic expansions of many repeats can be detected by locating reads that are completely contained inside the repeats. As in our previous work [24, 26], we call these reads in-repeat reads (IRRs). We implemented a method, ExpansionHunter Denovo (EHdn), for performing a genome-wide search for IRRs in BAM/CRAM files containing read alignments. EHdn computes genome-wide STR profiles containing locations and counts of all identified IRRs. Subsequent comparisons of STR profiles across multiple samples can reveal the locations of the pathogenic repeat expansions.

Genome-wide STR profiles

Genome-wide STR profiles computed by EHdn contain information about two types of IRRs: anchored IRRs and paired IRRs. Anchored IRRs are IRRs whose mates are confidently aligned to the genomic sequence adjacent to the repeat (the "Methods" section). Paired IRRs are read pairs where both mates are IRRs with the same repeat motif. Repeats exceeding the read length generate anchored IRRs (Fig. 1, middle panel). Repeats that are longer than the fragment length of the DNA library produce paired IRRs in addition to anchored IRRs (Fig. 1, right panel). The genomic coordinates where the anchored reads align correspond to the approximate locations of loci harboring REs and the number of IRRs is indicative of the overall RE length.

The information about anchored IRRs is summarized in an STR profile for each repeat motif (e.g., CCG) by listing regions containing anchored IRRs in close proximity to each other together with the total number of anchored IRRs identified (Fig. 2, middle). Note that the mapping positions of anchored IRRs correspond to the positions of anchor reads; mapping positions of IRRs themselves are not used because their alignments are often unreliable. Contrary to anchored IRRs, the origin of paired IRRs cannot be determined if a genome contains multiple long repeats with the same motif. Due to this, STR profiles only contain the overall count of paired IRRs for each repeat motif.

Comparing STR profiles across multiple samples

To compare STR profiles across multiple samples, the profiles must first be merged together across samples. During this process, nearby anchored IRR regions are merged across multiple samples and the associated counts are depth-normalized and tabulated for each sample (Fig. 2, right; the "Methods" section). The total counts of paired IRRs are also normalized and tabulated for each sample. The resulting per-sample counts can be compared in two ways: If the samples can be partitioned into cases and controls where a significant subset of cases is hypothesized to contain expansions of the same repeat, then a case-control analysis can be performed using a Wilcoxon rank-sum test (the "Methods" section). Alternatively, if no enrichment for any specific expansion is expected, an outlier analysis (the "Methods" section) can be used to flag repeats that

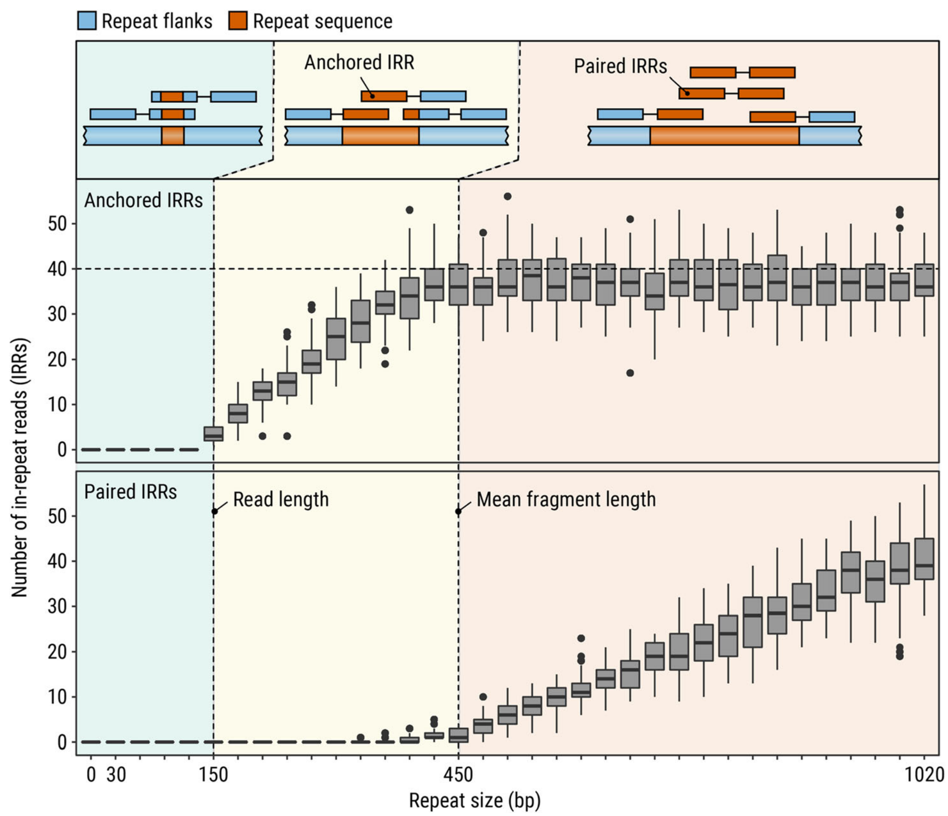
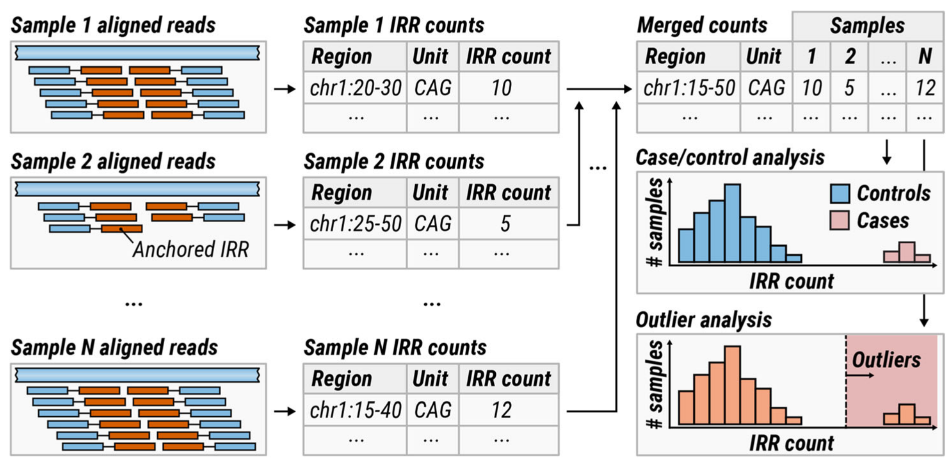


Fig. 1 Diagram illustrating the types and counts of reads generated by simulating repeats of different lengths. When the repeat is shorter than the read length (left panels), there are no IRRs associated with the repeat. When a repeat is longer than the read length but shorter than the fragment length (middle panels), anchored IRRs but no paired IRRs are present. As the repeat length approaches and exceeds the fragment length (right panels), paired IRRs are generated in addition to anchored IRRs



Sample 1 IRR counts

Region	Unit	IRR count
chr1:20-30	CAG	10
...	...	...

Sample 2 IRR counts

Region	Unit	IRR count
chr1:25-50	CAG	5
...	...	...

Sample N IRR counts

Region	Unit	IRR count
chr1:15-40	CAG	12
...	...	...

Merged counts

		Samples			
Region	Unit	1	2	...	N
chr1:15-50	CAG	10	5	...	12
...	...	...	...	...	...

Fig. 2 (Left) A search for anchored IRRs is performed across all aligned reads. (Middle) The IRR counts are summarized into STR profiles. (Right) The resulting STR profiles are merged across all samples. If the dataset can be partitioned into cases and controls, IRR counts in these groups are compared for each locus. Alternatively, if no such partition is possible, an outlier analysis is performed

are expanded in a small subgroup of cases compared to the rest of the dataset. Case-control and outlier analyses can be performed on either anchored IRRs or paired IRRs, which we call locus and motif methods, respectively (the "Methods" section). Thus, the locus method can identify locations of repeat expansions while the motif method can reveal the overall enrichment for long repeats with a given motif.

Baseline simulations

To demonstrate the baseline expectation of how the numbers of anchored and paired IRRs vary with repeat length, we simulated 2 × 150 bp reads at 20x coverage with 450-bp mean fragment length for the repeat associated with Huntington disease and varied the repeat length from 0 to 340 CAG repeats (0 to 1020 bp; Additional file 1). No IRRs occur when the repeat is shorter than the read length (Fig. 1, left panel). When the repeat is longer than the read length, but shorter than the fragment length (Fig. 1, middle panel), the number of anchored IRRs increases proportionally to the length of the repeat. As the length of the repeat approaches and exceeds the mean fragment length (Fig. 1, right panel), the number of paired IRRs increases linearly with the length of the repeat. Because anchored IRRs require one of the reads to "anchor" outside of the repeat region, the number of anchored IRRs is limited by the fragment length and remains constant as the repeat grows beyond the mean fragment length. It is important to note that real sequence data may introduce additional challenges compared to the simulated data. For example, sequence quality in low complexity regions or interruptions in the repeat may impact the ability to identify some IRRs.

Analysis of sequencing data

Detection of expanded repeats in case-control studies

Given a sufficient number of samples with the same phenotype, pathogenic REs may be identified by searching for regions with significantly longer repeats in cases compared to controls (see Fig. 2). To demonstrate the feasibility of such analyses, we analyzed 91 Coriell samples with experimentally confirmed expansions in repeats associated with Friedreich ataxia (FRDA; $N = 25$), myotonic dystrophy type 1 (DM1; $N = 17$), Huntington disease (HD; $N = 14$), and fragile X syndrome (FXS; $N = 35$). This dataset has been previously used to benchmark the performance of existing targeted methods [23, 24, 27].

The pathogenic cutoffs for FRDA, DM1, and FXS repeats are greater than the read length, so our analysis of simulated data suggests that anchored IRRs are likely to be present in each sample with one of these expansions (Fig. 1). The pathogenic cutoff for the HD repeat (120 bp) is less than the read length (150 bp) used in this study, so a subset of samples with Huntington disease may not contain relevant IRRs making this expansion harder to detect de novo even though it is detectable with existing methods [22, 24, 26, 27].

We separately compared samples with expansions in *FXN* (FRDA), *DMPK* (DM1), *HTT* (HD), or *FMR1* (FXS) genes (cases) against a control cohort of 150 unrelated Coriell samples of African, European, and East Asian ancestry [37]. Each case-control comparison revealed a clear enrichment of anchored IRRs at the corresponding repeat region (Fig. 3). This analysis demonstrated that ExpansionHunter Denovo (EHdn) can re-identify known pathogenic repeat expansions without prior knowledge of the

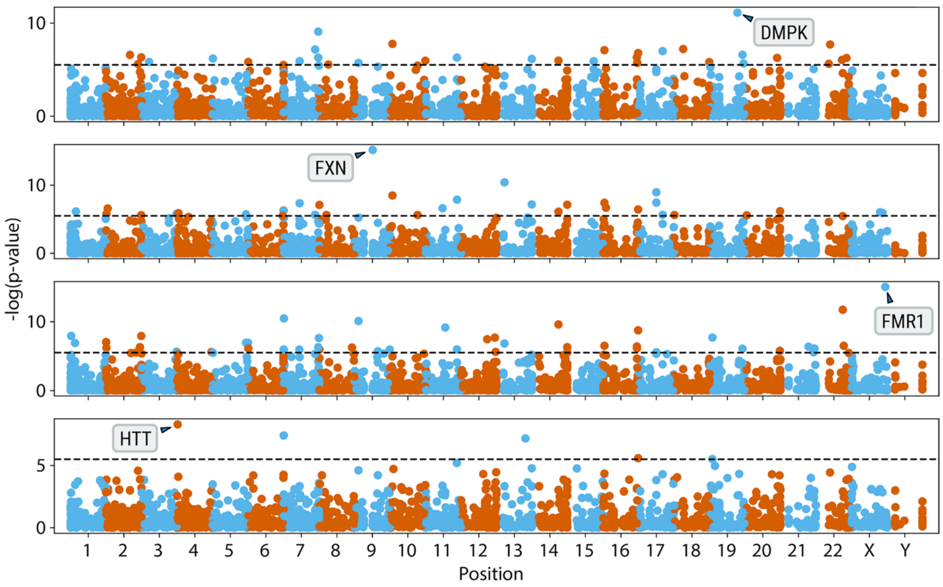


Fig. 3 Genome-wide analysis of anchored IRRs comparing cases with known pathogenic expansions in *DMPK*, *FXN*, *FMR1*, and *HTT* genes (top to bottom) to 150 controls

location or repeat motif when the pathogenic repeat length is equal to or longer than the read length, also assuming that the repeat is highly penetrant.

Detection of expanded repeats in mixed sample cohorts

In many discovery projects, it can be difficult to isolate patients that harbor the same repeat expansion based on the phenotype alone. For instance, the repeat expansion in the *C9orf72* gene is present in fewer than 10% of ALS patients and many ataxias can be caused by expansions of a variety of repeats. Such problems call for analysis methods that are suitable for heterogeneous disease cohorts.

To solve this problem, we follow the approach taken previously by others and compare each case sample against the control cohort to identify outliers [23, 25] (the "Methods" section). To demonstrate the efficacy of this approach, we combined each sample from the pool of samples with expansions in *FXN*, *DMPK*, *HTT*, and *FMR1* genes with 150 controls to generate a total of 91 datasets, each containing 151 samples. We then performed an outlier analysis on the counts of anchored IRRs (the "Methods" section) in each dataset.

In 81% of the datasets, the expanded repeat ranked within the top 10 repeats based on the outlier score (Fig. 4a). This number increased to 84% when the analysis was restricted to short motifs between 2 and 6 bp (Fig. 4b). EHdn performed well for *DMPK* and *FXN* repeats, identifying these REs within the top 10 ranks for 41 out of 42 cases. The *FMR1* expansion was only ranked in the top 10 for 24 out of 35 cases known to have the expansion. This result is consistent with a previous comparison, which found this locus had the poorest performance across all RE detection tools [23]. The performance for the *HTT* repeat is surprisingly good considering that EHdn was not designed to detect REs shorter than the read length. The rankings improved further when the analysis was restricted to repeats located close to exons of brain-expressed genes

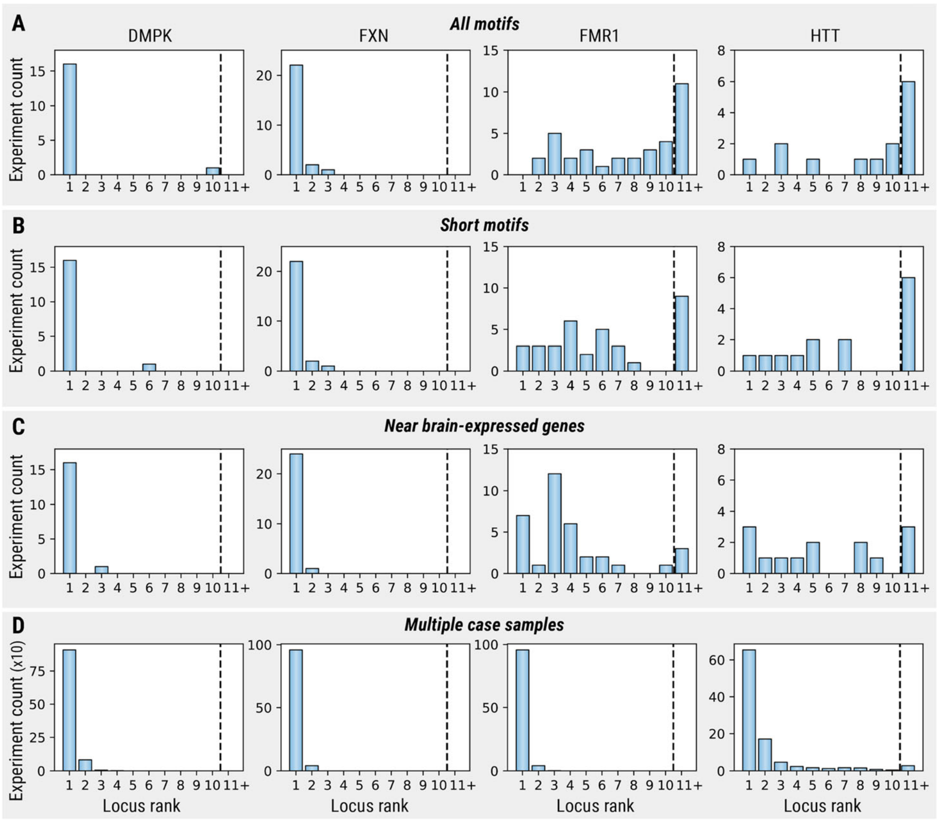


Fig. 4 Ranking of known expansions based on the outlier score computed for anchored IRRs. Each rank originates from a genome-wide analysis of a dataset consisting of one (**a**–**c**) or five (**d**) samples with a known expansion and 150 controls. **a** Ranks for all identified repeats. **b** Ranks for repeats with 2–6-bp motifs. **c** Ranks for repeats located in the 5-kbp region around exons of brain-expressed genes. **d** Ranks for datasets with five case samples

(Fig. 4c, Additional file 1) or when multiple cases (five in this example) were included in the analysis (Fig. 4d, Additional file 1).

The landscape of long repeats within a control population

To explore the landscape of large repeats in the general population, we applied EHdn to 150 unrelated Coriell samples of African, European, and East Asian ancestry [37]. To limit this analysis to higher confidence repeats, we considered loci where EHdn identified at least five anchored IRRs, corresponding to repeats spanning about 150–200 bp and longer, and motifs supported by at least five paired IRRs in a single sample. Altogether, EHdn identified 1574 unique motifs spanning between 2 and 20 bp, 94% of which were longer than 6 bp. Of these, 19% were found in at least half of the samples and 23% were found in just one sample. On average, each person had 660 loci with long repeats. As expected, the telomeric hexamer motif AACCCT is particularly abundant and was found in about ~ 23,000 IRRs per sample. Similarly, the centromeric pentamer motif AATGG was found in ~ 5000 IRRs per sample. To estimate the number of repeats located outside of the telomeric and centromeric regions [38, 39], we stratified the repeats by their distance to the closest telomere/centromere (Additional file 1: Figure S5). We found that, on average, about 170 of the identified repeats are located closer than 2 Mbp from the nearest telomere or centromere and about 200 are located further than 15 Mbp. We also showed that EHdn can accurately detect long repeats in a control sample by validating 77% of repeats supported by two or more reads in Pacific Biosciences long-read data (Additional file 1: Figure S4).

Exploring the limitations of catalog-based RE detection methods

To evaluate the limitations of catalog-based approaches [22–25, 27], we curated a set of 53 pathogenic or potentially pathogenic repeats (Additional file 1 and Additional file 2: Table S1) and checked if they were present in two commonly used catalogs: (a) STRs with up to 6-bp motifs from the UCSC genome browser simple repeats track [38, 40] utilized by STRetch and exSTRa and (b) the GangSTR catalog. Nine of the known pathogenic repeats are not present in the reference genome and hence are absent from both catalogs. Out of the remaining 44 loci, 22 loci are present in both catalogs, 12 are missing from the GangSTR catalog and present in the UCSC catalog, five are missing from the UCSC catalog and present in the GangSTR catalog, and five are missing from both catalogs (Additional file 1: Figure S1). While it is possible to update the catalogs to include these known pathogenic repeats, the number of missing potentially pathogenic REs remains unknown.

To demonstrate that EHdn offers similar performance to catalog-based methods on expansions exceeding the read length, we simulated expansions of the 35 non-degenerate STRs present in the reference (Additional file 1 and Additional file 2: Table S1). We focused our comparisons on STRetch because this method was specifically designed to search for novel expansions using a genome-wide catalog and because it was shown to have similar performance to other existing methods [23]. Our simulations show that EHdn ranks 33 out of 35 pathogenic repeats in the top 10 at sufficiently long lengths (Additional file 1 and Additional files 4, 5, 6, 7: Tables S3-S6). STRetch prioritizes 26 out of 29 repeats in the top 10, and the six remaining repeats are missing from its catalog. One of the REs detected by EHdn and missed by STRetch is the pathogenic *CSTB* repeat with a motif length of 12 bp. This is because STRetch is limited to the detection of motifs with length up to 6 bp. To further highlight this strength of EHdn, we confirmed that it can detect other REs with long motifs (Additional file 1).

Some recently discovered REs are composed of motifs that are not present in the reference genome. One such example is the recently discovered repeat expansion of non-reference motif AAGGG causing cerebellar ataxia with neuropathy and bilateral vestibular areflexia syndrome (CANVAS) [41, 42]. Rafehi et al. [42] demonstrated that EHdn is the only computational method capable of discovering this expansion. To further benchmark EHdn's ability to detect REs with complex structure, we simulated nine complex REs with non-reference motifs known to cause disease (Additional file 1: Figure S3). For eight out of nine REs, including a simulated version of the CANVAS expansion, EHdn was able to detect one or both of the expanded repeats in each locus (Additional file 1 and Additional file 8: Table S7).

Discussion

Here, we introduced a new software tool, ExpansionHunter Denovo (EHdn), that can identify novel REs using high-throughput WGS data. We tested EHdn by comparing samples with known REs against a control group of 150 diverse individuals and performed simulation studies across a range of pathogenic or potentially pathogenic REs. These analyses show that EHdn offers comparable performance to targeted methods on known pathogenic repeats while also being able to detect repeats absent from existing catalogs. In particular, EHdn can be used for discovery of novel repeat expansions not detectable by the current methods because it (a) does not require prior knowledge of the genomic coordinates of the REs, (b) can detect nucleotide composition changes within the expanded repeats, and (c) is applicable to both short and long motifs.

Recent discoveries have highlighted the importance of complex pathogenic repeat expansions involving non-reference insertions [28–33]. EHdn is currently the only method capable of discovering these expansions from BAM or CRAM files without the need for re-alignment of the supporting reads. Additionally, we anticipate that EHdn can replace existing more manual and less computationally efficient discovery pipelines, such as the TRhist-based pipeline [43], where identification of enriched repeat motifs is followed by ad hoc re-alignment of relevant reads to the reference genome and manual evaluation of loci where these reads align.

EHdn has some limitations and areas for further improvement. It is limited to the detection of repetitive sequences longer than the read length and cannot, in general, detect shorter expansions. However, detection of these shorter expansions is feasible with the existing catalog-based methods, or structural variant detection methods. It may be possible to extend the detection limit to shorter repeat expansions; however, increasing the search space will lead to increased runtime and reduced power to detect outlier expansions. It is also important to note that while EHdn can analyze reads produced by a variety of read aligners (Additional file 1: Figure S6), the same aligner should be used for all samples involved in comparative analyses to eliminate false signals due to aligner differences. All parameters of the sequencing assay (sequencing platform and library preparation kits) should also match as closely as possible to avoid coverage biases and other technical artifacts.

In many previous studies, identification of pathogenic REs required years of work and involved linkage studies to isolate the region of interest followed by targeted sequencing to identify the likely causative mutations. EHdn can be used as a front-line tool in such studies to rapidly identify candidate REs. Once identified, these novel REs can be genotyped using targeted methods [22, 24, 26, 27] or molecular assays. The benefits of this approach were demonstrated in a recent study, where EHdn successfully identified a novel complex pathogenic RE [42]. Hundreds of thousands of individuals' genomes have now been sequenced using short-read sequencing from many large disease cohorts, awaiting additional analyses such as RE detection. Additionally, while it is generally easier to analyze the structure of the expanded repeats in long-read data [44, 45], combining short-read sequencing datasets and methods with long-read data can offer a cost-effective way to conduct large-scale repeat expansion discovery projects.

Conclusions

We presented ExpansionHunter Denovo, a new genome-wide and catalog-free method to search for REs in WGS data. We demonstrated that EHdn consistently detects REs in real and simulated data. Given the widespread adoption of WGS for rare disease diagnosis, we expect that EHdn will enable further RE discoveries that will likely resolve the genetic cause of disease in many individuals.

Methods

Identification of IRRs

To determine if a read r is an in-repeat read, we first check the read for periodicity. We define $I_k(i) = 1$ if $r_i = r_{i+k}$ and 0 otherwise, where r_i and r_{i+k} are the ith and $(i+k)$th bases of the read r. We then let $S(k) = \sum_{i=0}^{L-k-1} I_k(i)/(L-k)$ where L is the read length. Note that if a read consists of a perfect stretch of repeat units of length k, then $S(k) = 1$. We search across of range of motif lengths (by default $k \in \{2, 3, \ldots, 20\}$) for the smallest k such that $S(k) \geq t$ where t is a set threshold (we use $t = 0.8$ in all our analyses). If such a value of k is found, we extract the putative repeat unit using the most frequent bases at each offset $0 \leq i \leq k-1$. Since the orientation of the repeat where a given IRR originated is unknown in general, the unit of the repeat is ambiguous. To remove this ambiguity, we select the smallest repeat unit in lexicographical order under circular permutation and reverse complement operations. We then use this putative repeat unit to calculate a weighted-purity (WP) score of a read [24]. We assume that a read is an IRR if it achieves a WP score of at least 0.9. The WP score lowers the penalty for low-quality mismatches in order to account for the possibility of an increased base-call error rate that may occur in highly repetitive regions of the genome.

EHdn searches for IRRs among unaligned reads and reads whose mapping quality (MAPQ) is below a set threshold which in the analysis presented here was set to 40. For this study, we limited our analysis to motif lengths between 2 and 20 bp. Motif lengths equal to 1 were excluded to eliminate the large number of homopolymer repeats from the downstream analyses since we identified over 30 times as many homopolymer IRRs as IRRs with longer repeat motifs.

EHdn designates a read pair as a paired IRR if both mates are IRRs with the same repeat motif. A read is designated as an anchored IRR if it is an IRR and its mate is not an IRR and has MAPQ above a set threshold which was set to 50 for this study. Parameters such as the maximum allowed MAPQ for an IRR, the minimum allowed MAPQ for an anchor, and the range of repeat unit lengths for which to search are all tunable with EHdn. For example, setting the anchor read MAPQ threshold to 0 and the IRR MAPQ threshold to 60 ensures that every read pair in the alignment file is analyzed (assuming that the MAPQ values range from 0 to 60) at the cost of a corresponding increase in runtime.

Merging IRRs

Because an anchored IRR is assigned to the location of the aligned anchor read and not the position of the actual repeat (whose exact location may be unknown), a single repeat may produce anchored IRRs at a variety of locations centered around the repeat. To account for this, anchored IRRs with the same repeat motif are merged if their anchors are aligned within 500 bp of one another. When multiple samples are analyzed, the anchor regions are also merged across all samples and the counts of anchored IRRs (normalized to 40x read depth) are tabulated for each merged region and sample. Additionally, the depth-normalized counts of paired IRRs are tabulated for each repeat motif and sample.

Prioritization of expanded repeats

EHdn supports case-control and outlier analyses of the underlying dataset. The case-control analysis is based on a one-sided Wilcoxon rank-sum test. It is appropriate for situations where a significant subset of cases is expected to contain expansions of the same repeat.

The outlier analysis is appropriate for heterogeneous cohorts where enrichment for any specific expansion is not expected. The outlier analysis bootstraps the sampling distribution of the 95% quantile and then calculates the z-scores for cases that exceed the mean of this distribution. The z-scores are used for ranking the repeat regions. Similar outlier-detection frameworks were also developed for exSTRa [23] and STRetch [25].

Both the case-control and the outlier analyses can be applied either to the counts of anchored IRRs or to the counts of paired IRRs. We refer to these as locus or motif methods, respectively. The high-ranking regions flagged by the analysis of anchored IRRs correspond to approximate locations of putative repeat expansions. The high-ranking motifs flagged by the analysis of paired IRRs correspond to the overall enrichment for repeats with that motif.

Defining relevant repeat expansions

A catalog of pathogenic or potentially pathogenic repeat expansions was collated from the literature. We supplemented this catalog with recently reported STRs linked with gene expression [46], and repeats with longer motifs overlapping with disease genes (Additional file 1).

Simulated repeat expansions

Expanded repeats were simulated using a strategy similar to that taken by BamSurgeon [47]. Briefly, we simulated reads in a 2-kb region around an expanded repeat and then aligned the reads to the reference genome. We then removed reads in the same region from a WGS control sample and merged alignments of real and simulated data together (Additional file 1: Figure S2).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02017-z.

Additional file 1. Supplementary methods, supplementary results, Figure S1-S6, and captions of Tables S1-S8.

Additional file 2: Table S1. Definitions of repeat expansions.

Additional file 3: Table S2. Repeats with long motifs and expression-linked repeats which were used for simulation.

Additional file 4: Table S3. Simulation results for 13 small pathogenic repeat expansions.

Additional file 5: Table S4. Simulation results for 22 large pathogenic repeat expansions.

Additional file 6: Table S5. Simulation results for 27 repeat expansions with long motifs.

Additional file 7: Table S6. Simulation results for expression-linked repeat expansions with short motifs.

Additional file 8: Table S7. Simulation results for nine complex pathogenic repeat expansions.

Additional file 9: Table S8. Overview of existing methods for detecting repeat expansions based on short-read data.

Additional file 10. Review history.

Acknowledgements

We are grateful to Dr. Stacie Taylor and Dr. Kirsten Curnow for their help with preparing the manuscript.

Review history

The review history is available as Additional file 10.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

ED and MAE implemented the algorithm for generating STR profiles. ED, MFB, PAR, MB, and MAE developed methods for comparing STR profiles. PAR performed simulation analyses with guidance from WWW. BT, CN, and RKCY performed confirmatory studies based on long-read data. MFB and PAR curated the catalog of pathogenic repeats. ED, MFB, PAR, BT, SC, JJFAV, CN, GN, VGG, AMG, and BL performed all remaining data analyses. RJT, WWW, SWS, JHV, DRB, RKCY, MB, and MAE supervised and guided the project. ED, MFB, PAR, and MAE wrote the initial draft of the manuscript. The authors read and approved the final manuscript.

Funding

PAR was funded by the BC Children's Hospital Research Institute Graduate Student Scholarship, with computational infrastructure support from the University of British Columbia's Advanced Research Computing. BT was funded by the Canadian Institutes for Health Research Banting Postdoctoral Fellowship and the Canadian Open Neuroscience Platform Research Scholar Award. MB was supported by an Australian National Health and Medical Research Council (NHMRC) Program Grant (GNT1054618) and an NHMRC Senior Research Fellowship (GNT1102971). Additional funding support was provided through the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 772376 - EScORIAL). RKCY was supported by the University of Toronto's McLaughlin Centre Accelerator Grant and the Nancy E.T. Fahrner Award. CN was supported by the Restracomp Award from The Hospital for Sick Children.

Availability of data and materials

ExpansionHunter Denovo is available under the Apache License version 2.0 from GitHub repository [48] and Zenodo repository [49].

The control WGS samples analyzed during the current study are available in the Illumina Polaris repository, https://github.com/Illumina/Polaris [50]. All samples were sequenced on an Illumina HiSeqX instrument using TruSeq DNA PCR-free sample prep. The 91 Coriell samples with experimentally confirmed repeat expansions in *DMPK*, *FMR1*, *FXN*, and *HTT* were introduced in our earlier publication [24] and are available from the European Genome-phenome Archive [51].

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

ED, SC, VGG, AG, BL, RJT, DRB, and MAE are or were employees of Illumina, Inc., a public company that develops and markets systems for genetic analysis.

Author details

[1]Illumina Inc., 5200 Illumina Way, San Diego, CA 92122, USA. [2]Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Australia. [3]Department of Medical Biology, The University of Melbourne, 1G Royal Parade, Parkville, VIC 3052, Australia. [4]Epilepsy Research Centre, Department of Medicine, The University of Melbourne, Austin Health, 245 Burgundy Street, Heidelberg, VIC 3084, Australia. [5]Centre for Molecular Medicine and Therapeutics, BC Children's Hospital, University of British Columbia, Vancouver, BC V5Z 4H4, Canada. [6]Genetics and Genome Biology, The Hospital for Sick Children, 686 Bay Street, Toronto, ON M5G 0A4, Canada. [7]The Centre for Applied Genomics, The Hospital for Sick Children, 686 Bay Street, Toronto, ON M5G 0A4, Canada. [8]Department of Neurology, UMC Utrecht Brain Center, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands. [9]Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 2E5, Canada. [10]New York Genome Center, 101 Avenue of the Americas, New York 10013, USA. [11]The McLaughlin Centre, University of Toronto, 686 Bay Street, Toronto, ON M5G 0A4, Canada. [12]Illumina Cambridge Ltd, Illumina Centre, 19 Granta Park, Great Abington, Cambridge CB21 6DF, UK.

Received: 20 December 2019 Accepted: 14 April 2020
Published online: 28 April 2020

References

1. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 2016;17:53.
2. Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, et al. Whole-genome sequencing of a healthy aging cohort. Cell. 2016;165:1002–11.
3. Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, et al. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A. 2016;113:11901–6.
4. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47:435–44.
5. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun. 2015;6:8018.
6. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
7. Consortium PMAS, Project MinE ALS Sequencing Consortium. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. Eur J Hum Genet. 2018:1537–46. https://doi.org/10.1038/s41431-018-0177-4.

8. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
9. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29:2041–3.
10. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912–8.
11. Poplin R, Newburger D, Dijamco J, Nguyen N, Loy D, Gross SS, et al. Creating a universal SNP and small indel variant caller with deep neural networks. bioRxiv. 2016. p. 092890. Available from: http://biorxiv.org/content/early/2016/12/21/092890.abstract. [cited 2017 Jun 25].
12. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing arXiv [q-bio.GN]. 2012. Available from: http://arxiv.org/abs/1207.3907. 29 Apr 2019.
13. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.
14. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. Bioinformatics. 2016;32:2375–7.
15. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84.
16. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
17. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.
18. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron. 2011;72: 245–56.
19. Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs JR, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011;72:257–68.
20. La Spada AR, Paul TJ. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet. 2010;11: 247–58.
21. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19:286–98.
22. Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, et al. Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. Am J Hum Genet. 2017;101:700–15.
23. Tankard RM, Bennett MF, Degorski P, Delatycki MB, Lockhart PJ, Bahlo M. Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. Am J Hum Genet. 2018;103:858–73.
24. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27:1895–903.
25. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol. 2018;19:121.
26. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35:4754–6.
27. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019;47:e90.
28. Sato N, Amino T, Kobayashi K, Asakawa S, Ishiguro T, Tsunemi T, et al. Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n. Am J Hum Genet. 2009;85:544–57.
29. Seixas AI, Loureiro JR, Costa C, Ordóñez-Ugalde A, Marcelino H, Oliveira CL, et al. A pentanucleotide ATTTC repeat insertion in the non-coding region of DAB1, mapping to SCA37, causes spinocerebellar ataxia. Am J Hum Genet. 2017;101:87–103.
30. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nat Genet. 2018;50:581–90.
31. Corbett MA, Kroes T, Veneziano L, Bennett MF, Florian R, Schneider AL, et al. Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. Nat Commun. 2019;10:4920.
32. Florian RT, Kraft F, Leitão E, Kaya S, Klebe S, Magnin E, et al. Unstable TTTTA/TTTCA expansions in MARCH6 are associated with familial adult myoclonic epilepsy type 3. Nat Commun. 2019;10:4919.
33. Yeetong P, Pongpanich M, Srichomthong C, Assawapitaksakul A, Shotelersuk V, Tantirukdham N, et al. TTTCA repeat insertions in an intron of YEATS2 in benign adult familial myoclonic epilepsy type 4. Brain. 2019;142:3360–6.
34. LaCroix AJ, Stabley D, Sahraoui R, Adam MP, Mehaffey M, Kernan K, et al. GGC repeat expansion and exon 1 methylation of XYLT1 is a common pathogenic variant in Baratela-Scott syndrome. Am J Hum Genet. 2019;104:35–44.
35. Lalioti MD, Scott HS, Buresi C, Rossier C, Bottani A, Morris MA, et al. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. Nature. 1997;386:847–51.
36. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17:507–22.
37. Illumina. Illumina/Polaris. GitHub. Available from: https://github.com/Illumina/Polaris. 30 Apr 2019.
38. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
39. Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004:493D–496. https://doi.org/10.1093/nar/gkh103.
40. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
41. Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yau WY, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019;51:649–58.
42. Rafehi H, Szmulewicz DJ, Bennett MF, Sobreira NLM, Pope K, Smith KR, et al. Bioinformatics-based identification of expanded repeats: a non-reference intronic pentamer expansion in RFC1 causes CANVAS. Am J Hum Genet. 2019;105: 151–65.
43. Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, et al. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. Nat Genet. 2019;51:1222–32.

44. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019;20:58.
45. Roeck AD, De Roeck A, De Coster W, Bossaerts L, Cacace R, De Pooter T, et al. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. Genome Biol. 2019. https://doi.org/10.1186/s13059-019-1856-3.
46. Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, et al. The impact of short tandem repeat variation on gene expression. Nat Genet. 2019;51:1652–9.
47. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. 2015;12:623–30.
48. Dolzhenko, Egor; Bennett, Mark F; Richmond, Phillip A; Trost, Brett; Chen, Sai; van Vugt, Joke J F A; Nguyen, Charlotte; Narzisi, Giuseppe; Gainullin, Vladimir G; Gross, Andrew; Lajoie, Bryan; Taft, Ryan J; Wasserman, Wyeth W; Scherer, Stephen W; Veldink, Jan H; Bentley, David R; Yuen, Ryan K C; Bahlo, Melanie; Eberle, Michael A. ExpansionHunter Denovo. Github; 2019. Available from: https://github.com/Illumina/ExpansionHunterDenovo. 8 Dec 2019.
49. Dolzhenko, Egor; Bennett, Mark F; Richmond, Phillip A; Trost, Brett; Chen, Sai; van Vugt, Joke J F A; Nguyen, Charlotte; Narzisi, Giuseppe; Gainullin, Vladimir G; Gross, Andrew; Lajoie, Bryan; Taft, Ryan J; Wasserman, Wyeth W; Scherer, Stephen W; Veldink, Jan H; Bentley, David R; Yuen, Ryan K C; Bahlo, Melanie; Eberle, Michael A. ExpansionHunter Denovo. 2020. Available from: https://zenodo.org/record/3674022. 18 Feb 2020.
50. Illumina, Inc. Polaris HiSeq X Diversity Cohort. PRJEB20654. The Eur Nucleotide Arch. 2019. Available from: https://www.ebi.ac.uk/ena/data/view/PRJEB20654. 19 Oct 2018.
51. Illumina, Inc. Whole genome sequence data for samples with the validated repeat expansions. EGAS00001002462. Eur Genome-phenome Arch. 2017; Available from: https://www.ebi.ac.uk/ega/studies/EGAS00001002462. 19 Oct 2018.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.