Kevlar: a mapping-free framework for accurate discovery of *de novo* variants

Daniel S. Standage[a,*], **C. Titus Brown**[a,b,**], **Fereydoun Hormozdiari**[b,c,d,**]
[a]Population Health and Reproduction, University of California, Davis
[b]Genome Center, University of California, Davis
[c]MIND Institute, University of California, Davis
[d]Biochemistry and Molecular Medicine, University of California, Davis; 1 Shields Ave, Davis, CA 95616

Abstract

Motivation: Discovery of genetic variants by whole genome sequencing has proven a powerful approach to study the etiology of complex genetic disorders. Elucidation of all variants is a necessary step in identifying causative variants and disease genes. In particular, there is an increased interest in detection of *de novo* variation and investigation of its role in various disorders. State-of-the-art methods for variant discovery rely on mapping reads from each individual to a reference genome and predicting variants from difference observed between the mapped reads and the reference genome. This process typically results in millions of variant predictions, most of which are inherited and irrelevant to the phenotype of interest. To distinguish between inherited variation and novel variation resulting from *de novo* germline mutation, whole-genome sequencing of close relatives (especially parents and siblings) is commonly used. However, standard mapping-based approaches tend to have a high false-discovery rate for *de novo* variant prediction, which in many cases arises from problems with read mapping. This is a particular challenge in predicting *de novo* indels and structural variants.

Results: We have developed a mapping-free method, Kevlar, for *de novo* variant discovery based on direct comparison of sequence content between related individuals. Kevlar identifies high-abundance *k*-mers unique to the individual of interest and retrieves the reads containing these *k*-mers. These reads are easily partitioned into disjoint sets by shared *k*-mer content for subsequent locus-by-locus processing and variant calling. Kevlar also utilizes a novel probabilistic approach to score and rank the variant predictions to identify the most likely *de novo* variants. We evaluated Kevlar on simulated and real pedigrees, and demonstrate its ability to detect both *de novo* SNVs and indels with high sensitivity and specificity.

Availability: https://github.com/dib-lab/kevlar

Keywords: *k*-mer, variant discovery, *de novo* variation, indel, autism, human genomics.

***Current affiliation:** National Biodefense Analysis and Countermeasures Center; Fort Detrick, MD 21702

****Contact:** ctbrown@ucdavis.edu and fhormozd@ucdavis.edu

Introduction

It is speculated that genetic variation is a major contributing factor in complex genetic disorders. The genetic heritability of many disorders is estimated to be relatively high. For example, the heritability of autism spectrum disorder (ASD) is over 0.6, and the heritability of Schizophrenia is over 0.5.[1,2] Only a fraction of this heritability is explained by known genetic variants, however, a phenomenon termed "missing heritability".[3] One hypothesis is that *de novo* mutations (especially indels and strucural variants) are a large source of causative variation—and consequently missing heritability—in developmental disorders.[3–5] However, the complexity of *de novo* variant discovery, especially *de novo* indel and structural variant (SV) discovery, has resulted in an incomplete accounting of their contribution to these disorders. The discovery of genetic variants in general, and *de novo* variants in particular, remains a topic of intense research interest. In addition to illuminating the role of genetic variation in the etiology of complex disorders, improved discovery and cataloguing of *de novo* variants across many samples or cohorts will shed additional light on important unresolved questions in human genomics, including rates, biases, and mechanisms of new mutation.

Whole genome sequencing of simplex families (presenting an isolated case of a genetic disorder) is a proven successful approach for discovery of novel genetic variants resulting from *de novo* mutation in the germline.[5–9] Following standard variant calling protocols, mapping-based methods for *de novo* variant prediction begin by aligning reads to the reference genome. Variants are then predicted for each individual based on artifacts observed in the read alignments, such as mismatches, gaps, abrupt shifts in coverage, and discordant read pair distances or orientations.[10–16] This initial process typically results in millions of variant predictions, which *de novo* variant discovery algorithms must examine to discern between inherited variation, true *de novo* variation, and false variant calls.

While reference-based variant discovery methods have proven valuable in the study of complex genetic disorders, we note some of their limitations. Despite consistent improvements in read alignment algorithms, finding the correct mapping for each read is still complicated by sequencing errors, repetitive DNA content, and mis-assemblies in the reference. Also, reads that do not map to the reference genome because they span mutation breakpoints or contain novel sequence are ignored completely by mapping-based variant predictors. Finally, most variant calls determined by analysis of read alignments are not unique to the individual of interest (child, or *proband*) but instead reflect divergence

in ancestry between the family and the reference genome donors. Estimates of human germline mutation rates give an expectation of approximately 80 novel mutations per generation,[17] and distinguishing true *de novo* variation events from millions of inherited or false variants is a substantial challenge.

More generally, accurate and comprehensive *de novo* variant discovery is complicated by several computational and biological factors, and remains an elusive goal. Any algorithm must be confident not only in the *existence* of the variant in the proband but also its *non-existence* in both parents. And while single nucleotide variants (SNVs) are the most common variant type, rarer larger variants affect more nucleotides overall and are hypothesized to have an even greater effect in genetic disorders. Accurate discovery of these larger *de novo* variants is particularly challenging due to the inherent complexity of indel and SV prediction. In a reference-mapping context, calling indels with confidence requires accurate mapping of each read spanning the indel, with all gaps arranged consistently. This is possible only for short indels, and tends to be prone to error and misalignment. Thus, prediction of indels with length >10 bp has proven to be very challenging and accompanied by high false positive and false negative rates. Furthermore, the prediction of SVs is only possible through indirect signatures such as alterations in read depth or read-pair signatures. These signatures can be quite noisy and result in high rate of false negative and false positive prediction. As a result, some basic properties of *de novo* SVs, including their rate of occurrence, remain unknown. It is important to note that there also exists no method for predicting of more complex types of *de novo* SVs, such as inversion-duplication.

Many of the challenges with *de novo* variant prediction can be mitigated by an approach that compares sequence content between related individuals directly, rather than indirectly via a reference genome. Such an approach does not require any read alignments, nor is it sensitive to off-target shared/inherited variation. What a mapping-free approach *does* require is a signature of variation that is not defined in terms of alignment artifacts. In this mapping-free context, we define variants in terms of sequence content and abundance via analysis of k-mers (sequences of a fixed length k). A similar idea was recently utilized for accurate discovery of somatic variants in cancer tumors[18] and indel discovery in whole-exome sequencing data.[19]

Intuitively, a novel germline mutation should result in new sequence content in the proband compared to the parental genomes. Even in the simplest case, a single nucleotide substitution, most if not all of the k-mers spanning the mutation should be unique, given a sufficiently large value of k. Incidentally, this is also true for other classes of variants, such as indels and various types of structural variation. And with sufficiently deep sampling of the proband genome, the expectation is that these novel k-mers are present in high abundance in the read data. Thus, it should be possible to detect both single nucleotide variants and larger variants (indels, structural variants) simultaneously using a single model.

Building on this intuition, we developed Kevlar, a new method based on a mapping-free formulation of the *de novo* variant discovery problem. Kevlar examines k-mer abundances to identify "interesting" k-mers which have significantly high abundance in the proband/child reads while being effectively absent in the reads from both parents. These k-mers are an indicator of the potential existence of a *de novo* variant. We next group the reads into disjoint sets, each reflecting a putative variant, based on the novel k-mers shared between pairs of reads. Kevlar then uses standard algorithms to assemble each set of reads into contigs and align the assembled contigs to a reference genome to make preliminary variant calls. Finally, Kevlar employs a novel probabilistic model to score variant predictions and distinguish miscalled inherited variants and true *de novo* mutations.

We demonstrate the utility of this new method on simulated and real data. We show that Kevlar's achieves similar predictive accuracy to best-in-class tools for SNV and short indel discovery, while at the same time predicting larger events with high accuracy. We demonstrate Kevlar's ability to accurately predict large-scale SV events, defining breakpoints with nucleotide-level precision.

Kevlar is available as an open source software project, and can be invoked via a Python API, a command-line interface, or a standard Snakemake workflow.[20] The stable and actively developed source code is available at `https://github.com/dib-lab/kevlar`, and documentation is available at `https://kevlar.readthedocs.io`.

Results

We present a novel framework for discovery of *de novo* variants based on direct comparisons of sequence content between related individuals, requiring no mapping of short reads to a reference genome. This framework utilizes a single strategy that accurately predicts single nucleotide variants (SNVs), insertions and deletions (indels), and structural variation events simultaneously.

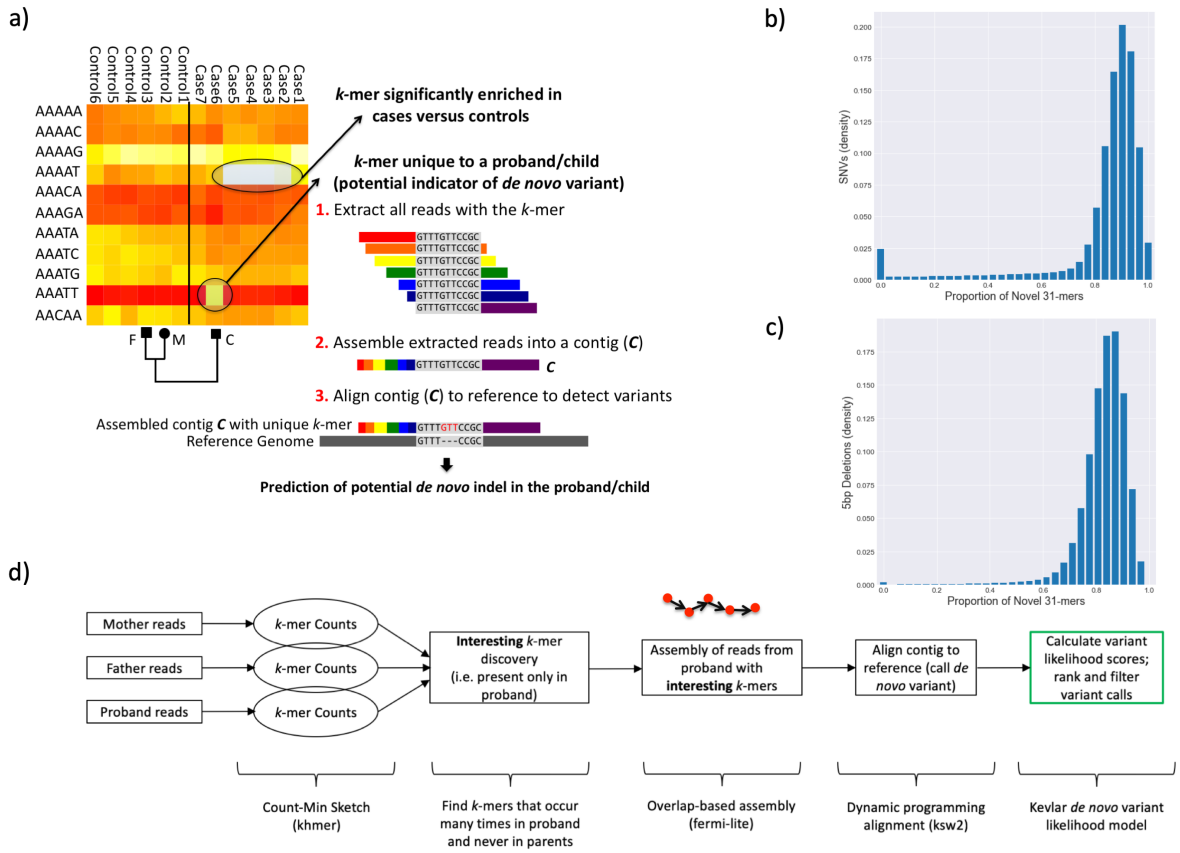

Fig 1 a. The overall mapping-free approach for *de novo* variant discovery b) The proportion N_m of *k*-mers associated with a simulated SNV mutations that are novel, aggregated over all sites in the genome. The trend observed for $k = 31$ holds for a wide range of k values ($\approx 20- \approx 60$) c) The proportion of *k*-mers associated with a simulated 5bp indels that are novel, aggregated over all sites in the genome. d) The Kevlar pipeline..

Overview of Kevlar

Our variant discovery strategy is fundamentally a search for novel DNA content in the sample of interest. It is based on the intuition that a *de novo* mutation will result in novel sequence that can be detected by analysis of *k*-mers, or short subsequences of length *k*. Often the subject is a child affected by a disorder or other trait of interest (referred to as the *focal sample*, *case*, or *proband*), with related individuals (parents and, when applicable, an unaffected sibling) serving as controls. However, the framework is flexible and can be applied in other contexts, such as identifying population-specific variation in an individual using individuals from two or more different populations as controls.

Figure 1d summarizes the Kevlar workflow. In brief, DNA sequence reads from the case and control samples are processed independently. For each sample, the reads are split into overlapping subsequences of length *k*, or *k*-mers, and the abundance of each *k*-mer is stored for subsequent lookup. A second pass over reads from the case sample then identifies all *k*-mers that are unique to the case—that is, *k*-mers that are abundant in the case sample but effectively absent in all control samples. Reads containing any novel *k*-mers are retained for subsequent processing.

After applying filters for contamination and erroneous *k*-mer abundances, the reads containing novel *k*-mers are partitioned such that any two reads sharing at least one novel *k*-mer are grouped together. The reads in each partition are then analyzed independently: they are assembled into a contig, the contig is aligned to the reference genome, and the alignment is used to assess the presence of a variant and make a variant call. Finally, Kevlar employs a novel likelihood score to rank and filter the variant calls.

Each step of the Kevlar workflow is discussed in detail in the Materials and Methods section.

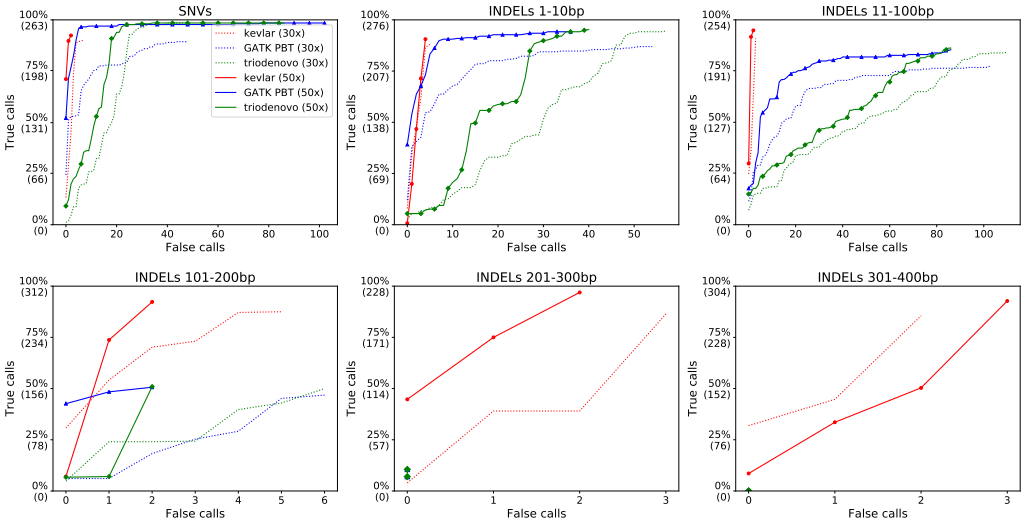

Fig 2 Receiver operating characteristic (ROC) curves comparing variant prediction performance on a simulated data set, shown for 30x coverage (dashed lines) and 50x coverage (solid lines). Each of the six panes shows prediction accuracy for a different variant type: single nucleotide variants (SNVs); insertions or deletions (indels) 1-10bp in length; 11-100bp indels; 101-200bp indels; 201-300bp indels; and 301-400bp indels.

Performance on simulated data

We simulated whole-genome shotgun sequencing of a hypothetical (constructed) family for a fine-grained assessment of Kevlar's accuracy in identifying different variant types at different levels of sequencing depth. Our simulation realistically modeled the inheritance of parental variants, but also included hundreds of "*de novo*" (unique to the proband) SNVs and indels ranging in size from <10bp to 400bp. The sequencing was simulated at 10x, 20x, 30x, and 50x coverage with low error rate. We compared Kevlar's accuracy on this data set to two other widely used *de novo* variant callers: GATK PhaseByTransmission[21] and TrioDenovo.[22]

The accuracy of all three variant callers is poor at low (10x) coverage (see Figure A1). GATK PhaseByTransmission makes very few variant predictions at 10x coverage. Kevlar and TrioDenovo report numerous predictions, but both suffer from low sensitivity (missing many true variants) and poor specificity (including many false calls). At 20x coverage, all three algorithms show marked improvement in SNV detection, in particular TrioDenovo which achieves ≥90% sensitivity with few false calls. Kevlar peforms the best in predicting indels at 20x coverage, but recovers less than 75% of indels in each length range.

At higher levels of coverage (30x and 50x), Kevlar accurately recovers ≥90% of true variants with very few false calls in each variant category (see Figure 2). Both GATK and TrioDenovo recover a slightly higher total number of true SNVs and short (1-10bp) indels than Kevlar at high coverage, but also accumulate a substantial number of false predictions. For indels 11bp or longer, Kevlar exhibits much higher predictive accuracy at high coverage (in both sensitivity and specificity) than GATK or TrioDenovo.

Performance on the SSC 14153 autism trio

To assess Kevlar's performance on real data, we applied Kevlar to predict *de novo* variants in the proband of an autism trio from the Simons Simplex Collection (family 14153). As a reference for comparison, we obtained a potential "truth set" from the denovo-db database (`http://denovo-db.gs.washington.edu/denovo-db/`). This truth set includes 196 *de novo* variant predictions, and represents the union of predictions made for this trio by several recent studies.[23–25] Note that the expected number of *de novo* variants per generation is estimated to be around 100,[17,24] or about half of the number of predictions in the truth set. Annotations in the denovo-db database indicate that experimental validation has confirmed 14 of the 196 calls.

In total, Kevlar predicts 219 *de novo* variants for trio 14153, including 150 SNVs, 68 indels/SVs, and a single 2 bp multinucleotide variant (MNV). We note that Kevlar assigned many of these predicted variants a low likelihood of the

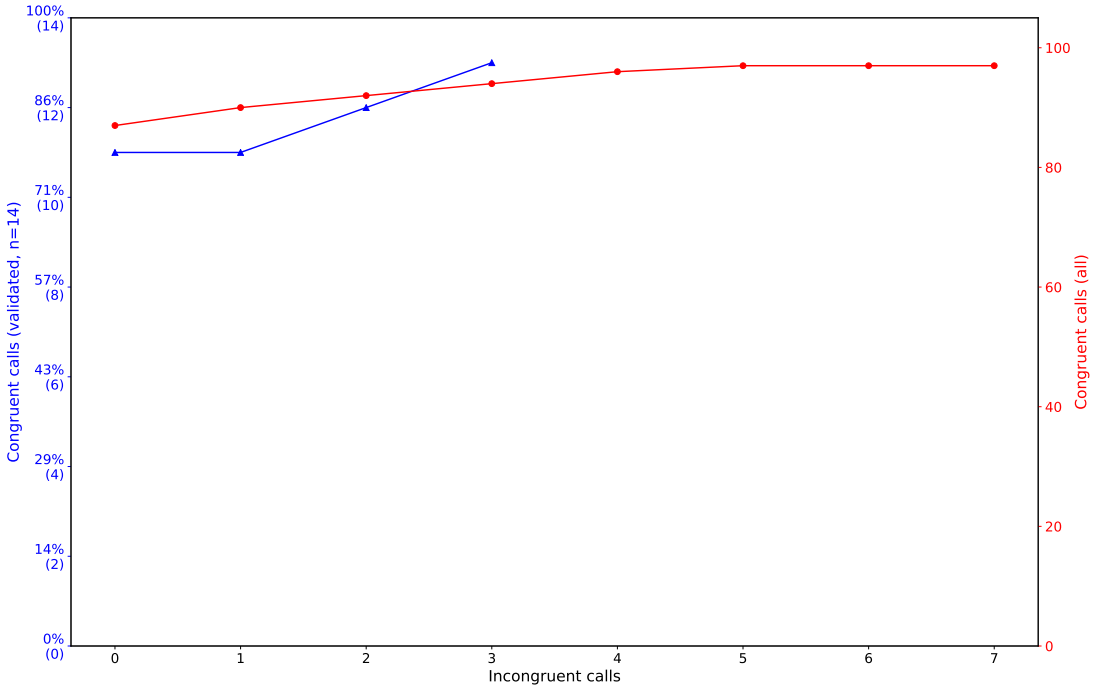

Fig 3 ROC curves showing congruence between *de novo* variant calls made by Kevlar on the SSC 14153 trio and corresponding calls from the denovo-db variant database. The red curve shows Kevlar's performance compared to all denovo-db calls, and the blue curve shows Kevlar's performance compared to denovo-db calls with experimental validation.

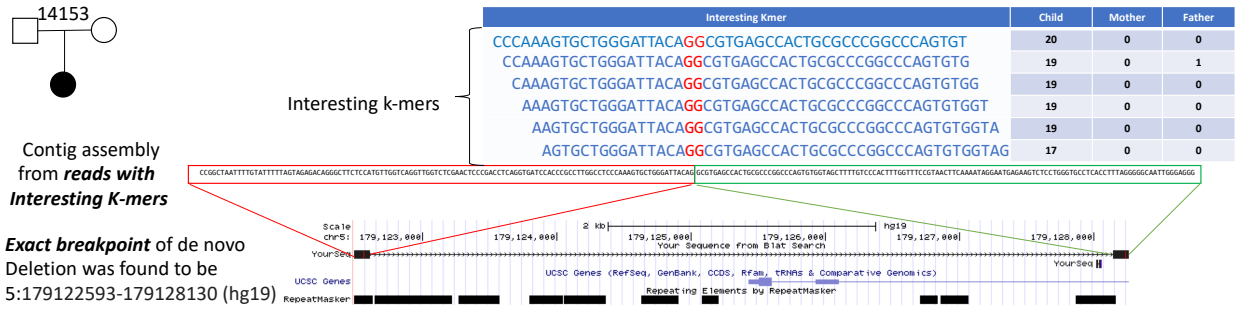

Fig 4 The validated 6kbp *de novo* deletion as predicted by Kevlar. The interesting *k*-mers, their abundances in each sample, the assembled novel sequence, and the breakpoint are depicted.

variant being a true *de novo* event. Figure 3 shows the congruence between the 100 top ranked Kevlar calls and the one reported in denovo-db calls for this trio.

Of the 14 denovo-db calls with experimental validation, 13 were predicted accurately with a very high likelihood score of being true *de novo* assigned by Kevlar. In other words, the predicted *de novo* variants with highest likelihood score assigned by Kevlar cover over 92% (13/14) of validated *de novo* variants. Overall, the 100 Kevlar variant calls ranked highest by the likelihood score include only 4 calls not present in denovo-db (probable false calls). On the other hand, only 5 Kevlar variant calls present in denovo-db (probable true variants) are not among the 100 highest ranked Kelvar calls. Of the 196 denovo-db calls, 95 are absent from the Kevlar predictions. The majority of these calls (75 / 95, $\approx$ 80%) occur in regions of repetitive DNA and have shown to be unreliable in experimental validation (Tychele Turner, personal communication).

Finally, a recent study verified the presence of a *de novo* deletion of approximately 6kbp in the proband of this trio,[24] removing the 5'UTR of the gene *CANX*. Kevlar also predicted this *de novo* deletion successfully, and identified the precise (and previously undetermined) breakpoints at chr5:179,122,593 and chr15:179,128,130 (GRCh37). Inspection of the variant reveals that both of the deletion's breakpoints occur in *Alu* repeat elements abundant throughout the genome (Figure 4). As a result, only 7 of the *k*-mers spanning the variant are unique signatures of mutation not already present elsewhere in the genome. We observe with interest that both breakpoints occur inside a 20 bp identical repeat, indicating this *de novo* deletion is the result of non-allelic homologous recombination (NAHR).

Discussion

De novo variants are a major contributing factor in many disorders (e.g. intellectual disability, autism, and epilepsy). Accurate discovery of these variants has been challenging as prediction methods need to be confident not only in the existence of the event in the proband/child, but also in the absence of the variant in the parents. Current approaches depend on correct alignments of sequence reads to a reference genome. Any complications in computing read alignments due to repeats, gaps in the reference, or variant complexity can result in false predictions or failure to discover a true *de novo* variant. We have introduced a novel method that does not depend on mapping reads to a reference genome, but instead relies on direct comparison of sequence content between related individuals. This enables us to accurately predict several classes of *de novo* mutation (substitutions, insertions, deletions, SVs) simultaneously with a single simple mathematical model. As long as the *de novo* mutation creates a *k*-mer not already present in the reference genome, the proposed algorithm should be able to accurately discover the event. And while we have tuned our model for discovery of novel germline mutations in a simplex scenario, the model is easily adapted to predict population-specific variation or to discovery variants in a large cohort case/control scenario.

We have also developed a novel *k*-mer based likelihood model for scoring and ranking variant calls according to their probability of being true *de novo* events. This likelihood score is effective in discerning *de novo* variants from inherited mutations and false variant calls. We have demonstrated the effectiveness of our discovery method and scoring model using both simulated and real data. Kevlar is competitive with best-in-class tools for discovery of *de novo* SNVs and small ($\leq$10 bp) indels, and substantially outperforms available methods for discovery of larger *de novo* variants. Not only does Kevlar predict indels and SVs with high sensitivity and specificity, but it reports the exact breakpoints of these variants with single base pair precision.

Development of completely reference-free methods are tremendously valuable in scenarios where the availability, quality, or relevance of a reference genome is insufficient. Kevlar's preliminary steps—identifying variant-spanning reads, binning reads into groups corresponding to distinct putative variants, and assembling each read group into a variant-spanning contig—are performed without the use of a reference genome. We note, however, that subsequent steps in the Kevlar workflow to annotate, filter, and score the preliminary variant calls still depend on a reference genome. One promising approach to developing a completely reference-free *de novo* variant discovery method would be to annotate of variants by aligning variant-spanning contigs directly to an assembly/variation graph.

Methods

Assessing diagnostic utility of novel k-mers

We expect that a *de novo* mutation will result in numerous novel *k*-mers, given a sufficiently large value of *k*. We also expect that these novel *k*-mers will be present in high abundance, given sufficiently deep sampling of the proband genome. Intuitively, we can use these novel *k*-mers to identify reads that span the *de novo* variant.

We assessed this intuition by traversing the human reference genome (GRCh38) base by base, simulating variants (SNVs and 5 bp deletions) at each position. For each simulated mutation, we determined the fraction of *k*-mers spanning the mutation that exist nowhere else in the genome, and thus act as a diagnostic signature of that particular variant. We then aggregated over the entire genome the probability that *k*-mer spanning a mutation (in this case 31-mers) will be novel—see Figure 1b and 1c.

Based on the results of this experiment, we formulate the *de novo* variant discovery problem as a search for putatively novel *k*-mers that are abundant in the proband and effectively absent in each parent. For sake of simplicity, we are using the term *proband* to refer generally to the subject or focal individual, and *parent* to refer generally to control individuals.

Here, *abundant* and *effectively absent* are defined in terms of a simple threshold model. Let X be the *absence threshold*, and Y be the presence threshold, and $A = \{A_p, A_m, A_f\}$ be the abundances of a *k*-mer in the proband,

mother, and father. We designate this *k*-mer as "*interesting*" (putatively novel) if and only if $A_p \geq Y$, $A_m \leq X$, and $A_f \leq X$. Based on our experience, the values $Y = 5$ and $X = 1$ produce desirable results for 30x sequencing coverage.

Kevlar workflow

The steps of the Kevlar workflow, summarized at a high level in Figure 1d, are described in detail in the subsequent sections.

Step 0: Compute k-mer counts

Preliminary to identifying novel *k*-mers, the abundance of each *k*-mer in each sample must be counted. Storing exact counts of every *k*-mer requires a substantial amount of space (dozens of gigabytes or more per sample), so Kevlar exploits several strategies to reduce the space required for keeping *k*-mer counts in memory.

First, Kevlar stores approximate *k*-mer counts in a Count-Min sketch, a probabilistic data structure similar to a Bloom filter that operates in a fixed amount of memory, exchanging accuracy for space efficiency.[26] The accuracy of the Count-Min sketch depends on its size and the number of distinct elements (*k*-mers in this case) being tracked. The Count-Min sketch exhibits a one-sided error, meaning that *k*-mer counts are sometimes overestimated but never underestimated. The extent of inaccuracy in the *k*-mer counts is summarized by the *false discovery rate (FDR)* statistic computed from the occupancy of the Count-Min sketch.

Second, Kevlar uses a masked counting strategy in which *k*-mers present in the reference genome and a contaminant database are ignored. This substantially reduces the number of *k*-mers to be stored in the Count-Min sketch, and as a consequence the desired level of accuracy can be maintained using a smaller amount of space.

Third, *k*-mer counts are recomputed with exact precision in subsequent steps of the Kevlar workflow, which means any *k*-mer retained erroneously due to an inflated count can be compensated for at a later stage. As a consequence, it is possible to reduce the size of the Count-Min sketch even further, to a FDR of 0.5 or greater. (Inicidentally, *k*-mers erroneously *discarded* due to inflated counts cannot be compensated for later, putting an upper bound on the imprecision of the initial *k*-mer counts in the control samples.)

Kevlar's *k*-mer counting operations are invoked with the `kevlar count` command, and rely on bulk sequence loading procedures and an implementation of the Count-Min sketch data structure from the khmer library.[26–28] Note that several alternative *k*-mer counting libraries and tools[29,30] have been developed to solve a variety of different biological problems.[31–35]

Step 1: Identifying novel k-mers and reads

To identify sequences spanning *de novo* variants, Kevlar scans each read sequenced from the proband. The per-sample abundances of each *k*-mer are queried from the Count-Min sketches computed in previous steps. If a *k*-mer is present in high abundance in the proband and absent from the parents (that is, it satisfies user-specified abundance thresholds), it is designated as "interesting" or putatively novel. Any read containing one or more interesting *k*-mers is retained for subsequent processing. This step is implemented in the `kevlar novel` command.

Step 2: Contamination, reference, and abundance filters

Reads containing putative novel *k*-mers are filtered prior to subsequent analysis. This filtering step serves two purposes.

First, Kevlar re-computes the abundance of each interesting *k*-mer in the proband sample. The relatively small number of these reads allows Kevlar to compute counts with effectively perfect accuracy using a small amount of memory. Any *k*-mer whose corrected count no longer satisfies the required abundance threshold is discarded. Note that since only proband reads are retained, only the proband *k*-mer abundances can be recomputed. This filtering step will not recover a *k*-mer that is erroneously discarded in the previous step due to an erroneously inflated *k*-mer abundance in one of the control (parent and sibling) samples.

Second, if for any reason *k*-mers from the reference genome and contaminants are not ignored in the initial *k*-mer counting step, this filtering step provides another opportunity to discard these *k*-mers.

After these filters are applied, any read that no longer contains any novel *k*-mers is discarded, and the remainder of the reads are retained for subsequent analysis.

The `kevlar filter` command is used to execute these contamination, reference, and abundance filters.

Step 3: Partitioning reads using shared novel k-mers

Interesting reads spanning the same variant will typically share numerous interesting *k*-mers. These shared novel *k*-mers provide a mechanism for grouping the reads into disjoint sets reflecting distinct variants.

To be precise, we define a *read graph* G as follows: every read containing one or more novel *k*-mers is represented by a node in G, and a pair of nodes is connected by an edge if they have one or more novel *k*-mers in common. With this formulation, if two reads share a novel *k*-mer they must be part of the same connected component in G. Overall G is sparse, but typically each connected component of the graph is highly connected. In subsequent steps, each component or partition $p \in G$ is analyzed independently.

The `kevlar partition` command implements this partitioning strategy.

Step 4: Contig assembly and reference target selection

For each connected component $p \in G$, we assemble the corresponding reads using the overlap-based algorithm implemented in the fermi-lite library.[36] Briefly, fermi-lite performs error correction, trims reads at unique *l*-mers, constructs an FM-index of the trimmed reads, and constructs a transitively reduced overlap graph. The optimal path in the final graph is output as a contig C_p suitable for variant calling.

Next, we select a target reference sequence (or set of candidate targets) for the contig C_p. Briefly, Kevlar decomposes the contig into overlapping subsequences of length l (*seeds*; $l = 51$ by default), and uses BWA MEM[37] to identify locations of exact matches for each seed sequence in the reference genome. The genomic interval that spans all seed exact matches, plus Δ nucleotides in each direction ($\Delta = 50$ by default), is then selected as the target reference sequence for C_p. If any adjacent seed matches are separated by more than D nucleotides ($D = 10,000$ by default), then the seed matches are split at that point and multiple reference targets are selected. The set of reference target sequences corresponding to contig C_p is denoted T_{C_p}.

Read assembly is invoked with the `kevlar assemble` command, and reference target selection is invoked with the `kevlar localize` command.

Step 5: Contig alignment and variant annotation

The contig C_p is aligned to each reference target sequence $t \in T_{C_p}$ using the ksw2 library[38]—specifically its implementation of Green's formulation of dynamic programming global alignment and extension (`ksw2_extz`). If there are multiple candidate targets, only the highest scoring alignment is retained. When a contig aligns to multiple locations with the same optimal score, all optimal alignments are retained for variant calling.

Prior to variant calling, kevlar right-aligns any gaps at the right end of the alignment to minimize the number of alignment blocks/operations. Next, Kevlar inspects the alignment path (represented as a CIGAR string) of each alignment and tests for matches against expected patterns. Alignments matching the pattern `^(\d+[DI])?\d+M(\d+[DI])?$` are classified as SNV events, and the "match" block of the alignment is scanned for mismatches between the contig and the reference target. Any mismatch is reported as a single nucleotide variant. Alignments matching the pattern `^(\d+[DI])?\d+M\d+[ID]\d+M(\d+[DI])?$` are classified as indel events. In addition to reporting the internal gap of this alignment as an indel variant, the flanking "match" blocks are also scanned for mismatches between the contig and the target to be reported as putative SNVs. Any alignment not matching the two patterns described above is designated as an uninterpretable "no-call" and listed in the output along with the corresponding contig sequence.

In some cases, there is a possibility that kevlar will report two or more calls in close proximity. While the probability of two *de novo* variants occurring in close proximity is effectively nil, it is common for an inherited variant to occur proximal to a *de novo* variant. Occasionally one of these inherited variants will not be spanned by any interesting *k*-mers, in which case it can immediately be designated as a "passenger" variant call. However, in cases where an inherited variant is spanned by one or more interesting *k*-mers, we rely on subsequent examination of *k*-mer abundances to distinguish novel variants from inherited variants.

The `kevlar call` command computes the contig alignments and makes preliminary variant calls.

Step 6: Likelihood scoring model for ranking and filtering variant calls

Given the filters already discussed, false *interesting* *k*-mer designations are rare throughout the genome overall. Redundancy from a high depth of sequencing coverage prevents sequencing errors from driving the reported abundance of *k*-mers present in the parents to 0. And for loci where local coverage is low in one parent, it is extremely rare that local coverage will also be low in the other parent by chance. Thus, if a *k*-mer is present in either parent, it is disqualified from the *interesting* or novel designation.

However, false *interesting* *k*-mer designations are enriched around inherited mutations. It is very common for variants present in one parent to be absent from the other parent. If by chance the depth of sequencing coverage is low at such a locus in the donor parent, there may not be enough redundancy to compensate for sequencing errors. As a result, some *k*-mers that are truly present in the donor parent will have a reported abundance of 0. Being truly absent from the other parent, these *k*-mers are erroneously designated as novel.

A related complication occurs when a novel variant is proximal to an inherited variant. Both variants are reflected in the alignment of the associated contig (assembled from proband-derived *interesting* reads) to the reference genome. In both of these cases, distinguishing novel variants from inherited variants benefits from examination of the abundances of all *k*-mers containing each variant, as well as the corresponding reference *k-mers*.

Here, we consider a formulation for calculating the likelihood of a variant. First, for each variant we define a set of alternate *k*-mers A as the *k*-mers indicating existence of the variant (alternate genotype). We consider only *k*-mers that are unique to this variant (that is, they don't appear in any other location in the reference genome). We assume that there are a total of *n* alternate *k*-mers.

Let the random variables v_c, v_f, and v_m indicate the genotype (i.e. $\{0/0, 0/1, 1/1\}$) of the putative variant in the proband/child, father, and mother respectively. The random variable $\mathbf{A}_c = \{A_{c_1}, A_{c_2}, ..., A_{c_n}\}$ denotes the counts of the alternate allele *k*-mers in the proband, $\mathbf{A}_m = \{A_{m_1}, A_{m_2}, ..., A_{m_n}\}$ the alternate allele *k*-mer counts in the mother, and $\mathbf{A}_f = \{A_{f_1}, A_{f_2}, ..., A_{f_n}\}$ the alternate allele *k*-mer counts in the father. The random variable $\mathbf{R}$ denotes the *k*-mer counts of the corresponding reference allele in the reference genome, serving as a proxy for copy number.

The likelihood of that a putative variant is *de novo* can be calculated as follows.

$$\begin{aligned} L(\text{dn} = 1) &= P(\mathbf{A}_c, \mathbf{A}_m, \mathbf{A}_f, \mathbf{R} \mid \text{dn} = 1) \\ &= P(\mathbf{A}_c, \mathbf{A}_m, \mathbf{A}_f, \mathbf{R} \mid v_c = 0/1, v_m = 0/0, v_f = 0/0) \\ &= P(\mathbf{A}_c \mid v_c = 0/1) P(\mathbf{A}_m, \mathbf{R} \mid v_m = 0/0) P(\mathbf{A}_f, \mathbf{R} \mid v_f = 0/0) \end{aligned}$$

To calculate these probabilities we will assume independence of *k*-mer abundances given the genotype. Thus, we have the following.

$$L(\text{dn} = 1) = \prod_{i=1}^{n} P(A_{c_i} \mid v_c = 0/1) P(A_{m_i}, R_i \mid v_m = 0/0) P(A_{f_i}, R_i \mid v_f = 0/0)$$

For calculating the probability of an observed *k*-mer count conditioned on a 0/1 or 1/1 genotype, we assume a normal distribution with mean μ and standard deviation σ. The μ and σ parameters are learned empirically for each sample using only exonic *k*-mers that occur only once in the reference genome. For an observed *k*-mer count conditioned on a 0/0 genotype, we assume it is due to error with probability ϵ. Thus we compute the probability of observing the count y for *k*-mer with a 0/0 genotype as ϵ^y.

We calculate the likelihood that a putative variant is a false prediction as follows (assuming independence of *k*-mer abundances, as before).

$$\begin{aligned} L(\text{fp} = 1) &= P(\mathbf{A}_c, \mathbf{A}_m, \mathbf{A}_f, \mathbf{R} \mid \text{fp} = 1) \\ &= P(\mathbf{A}_c, \mathbf{A}_m, \mathbf{A}_f, \mathbf{R} \mid v_c = 0/0, v_m = 0/0, v_f = 0/0) \\ &= P(\mathbf{A}_c, \mathbf{R} \mid v_c = 0/0) P(\mathbf{A}_m, \mathbf{R} \mid v_m = 0/0) P(\mathbf{A}_f, \mathbf{R} \mid v_f = 0/0) \\ &= \prod_{i=1}^{n} P(A_{c_i}, R_i \mid v_c = 0/0) P(A_{m_i}, R_i \mid v_m = 0/0) P(A_{f_i}, R_i \mid v_f = 0/0) \end{aligned}$$

Finally, we calculate the likelihood of observed *k*-mer counts under the inheritance assumption. Let S denote the set of all trio-genotype combinations representing valid inheritance scenarios ($v_c \in \{0/1, 1/1\}$ and $v_m, v_f \in \{0/0, 0/1, 1/1\}$). We calculate the likelihood that a putative variant is inherited as follows (again assuming independence of *k*-mer abundances as an approximation).

$$
\begin{aligned}
L(\text{ih} = 1) &= P(\mathbf{A}_c, \mathbf{A}_m, \mathbf{A}_f, \mathbf{R} \mid \text{ih} = 1) \\
&= \prod_{i=1}^{n} P(A_{c_i}, A_{m_i}, A_{f_i}, R_i \mid \text{ih} = 1) \\
&= \prod_{i=1}^{n} \frac{P(\text{ih} = 1 \mid A_{c_i}, A_{m_i}, A_{f_i}, R_i) P(A_{c_i}, A_{m_i}, A_{f_i}, R_i)}{P(\text{ih} = 1)} \\
&= \prod_{i=1}^{n} \frac{P(A_{c_i}, A_{m_i}, A_{f_i}, R_i)}{P(\text{ih} = 1)} \times P(\text{ih} = 1 \mid A_{c_i}, A_{m_i}, A_{f_i}, R_i) \\
&= \prod_{i=1}^{n} \frac{P(A_{c_i}, A_{m_i}, A_{f_i}, R_i)}{P(\text{ih} = 1)} \times \sum_{S} \frac{P(A_{c_i}, A_{m_i}, A_{f_i}, R_i \mid v_c, v_m, v_f) P(v_c, v_m, v_f)}{P(A_{c_i}, A_{m_i}, A_{f_i}, R_i)} \\
&= \prod_{i=1}^{n} \frac{1}{P(\text{ih} = 1)} \times \sum_{S} P(A_{c_i}, A_{m_i}, A_{f_i}, R_i \mid v_c, v_m, v_f) P(v_c, v_m, v_f)
\end{aligned}
$$

We can then calculate the value $P(\text{ih} = 1 \mid A_{c_i}, A_{m_i}, A_{f_i})$ as follows

$$
\begin{aligned}
P(\text{ih} = 1 \mid A_{c_i}, A_{m_i}, A_{f_i}) &= \sum_{S} \frac{P(A_{c_i}, A_{m_i}, A_{f_i} \mid v_c, v_m, v_f) P(v_c, v_m, v_f)}{P(A_{c_i}, A_{m_i}, A_{f_i})} \\
&= \frac{1}{P(A_{c_i}, A_{m_i}, A_{f_i})} \sum_{S} P(A_{c_i}, A_{m_i}, A_{f_i} \mid v_c, v_m, v_f) P(v_c, v_m, v_f)
\end{aligned}
$$

Note that all of these probabilities must be computed in log space, and as a result the summation in this final likelihood cannot easily be solved. Thus, instead of calculating the entire summation we can approximate it the most likely inheritance scenario $S_i \in S$ (where $S_i = v_c, v_m, v_f$), assuming one scenario will dominate others. Thus, we can approximate the above summation as follows.

$$
\begin{aligned}
L(\text{ih} = 1) &\approx \frac{1}{P(\text{ih} = 1)} \prod_{i=1}^{n} \max_{S} P(A_{c_i}, A_{m_i}, A_{f_i} \mid v_c, v_m, v_f) P(v_c, v_m, v_f) \\
&= \frac{1}{P(\text{ih} = 1)} \prod_{i=1}^{n} \max_{S} P(A_{c_i} \mid v_c) P(A_{m_i} \mid v_m) P(A_{f_i} \mid v_f) P(v_c, v_m, v_f)
\end{aligned}
$$

For autosomes, there are 15 valid inheritance scenarios, 11 of which result in a proband genotype of 0/1 or 1/1, and therefore $P(\text{ih} = 1) = \frac{11}{15}$. Inheritance scenarios for the sex chromosomes are distinct and depend on the sex of the child. For males, there are 8 valid inheritance scenarios for chromosome X, 4 of which result in a proband genotype of 0/1, and as a result $P(\text{ih} = 1) = \frac{1}{2}$. Alternatively, there are 3 valid inheritance scenarios for chromosome Y, only one of which results in a proband genotype of 0/1, and thus $P(\text{ih} = 1) = \frac{1}{3}$. For females, there are 12 valid inheritance scenarios for chromosome X, 8 of which result in a proband genotype of 0/1 or 1/1, resulting in $P(\text{ih} = 1) = \frac{3}{4}$. When computing the probability of *k*-mer abundances conditional on learned abundance distributions, the abundance distributions are adjusted appropriately for sex chromosomes depending on the sex of the proband and the parent.

The `kevlar simlike` command computes likelihoods for preliminary variant calls, sorts the calls, and filters out low scoring and otherwise problematic calls.

Data simulations

We simulated whole-genome shotgun sequencing for a hypothetical trio (father, mother, and proband) to evaluate the accuracy of our *de novo* variant discovery algorithm. Using the human reference genome (GRCh38) and a catalog of common variants (dbSNP), we constructed two independent diploid genomes representing the two parents. We randomly selected SNPs and indels from dbSNP and assigned the variants to each parental haplotype at a rate of 1 for every 1000 bp.

When then constructed the diploid proband genome through recombination of the parental diploid genomes and simulated germline mutation. SNVs and indels ranging from <10 bp to 400 bp in length were simulated as heterozygous events unique to the proband, representing *de novo* variation.

Finally, we used wgsim[39] to simulate whole-genome shotgun sequencing of each individual. This produced sequences resembling Illumina 2x150bp paired-end reads with low sequencing error rate. The sequencing was repeated at four different average depths of sequencing coverage: 10x, 20x, 30x, and 50x.

Acknowledgements

We would like to acknowledge Dr. Tamer Mansour, Luiz Irber Jr., Camille Scott, and Lisa Johnson for helpful discussions on method development and implementation, and Dr. Tychele Turner for helpful discussions on the method evaluation.

Author Contributions

DSS, CTB, and FH conceived the study. DSS implemented the method. DSS and FH designed the experiments and wrote the manuscript. DSS, CTB, and FH edited and approved the final manuscript.

Declaration of Interests

The authors declare no competing interests.

References

1. J. Hallmayer, S. Cleveland, A. Torres, *et al.*, "Genetic heritability and shared environmental factors among twin pairs with autism," *Archives of General Psychiatry* **68**(11), 1095–1102 (2011).
2. A. G. Cardno, E. J. Marshall, B. Coid, *et al.*, "Heritability estimates for psychotic disorders: the Maudsley twin psychosis series," *Archives of General Psychiatry* **56**(2), 162–168 (1999).
3. T. A. Manolio, F. S. Collins, N. J. Cox, *et al.*, "Finding the missing heritability of complex diseases," *Nature* **461**(7265), 747 (2009).
4. E. E. Eichler, J. Flint, G. Gibson, *et al.*, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics* **11**(6), 446 (2010).
5. J. A. Veltman and H. G. Brunner, "*De novo* mutations in human genetic disease," *Nature Reviews Genetics* **13**(8), 565 (2012).
6. M. Fromer, A. J. Pocklington, D. H. Kavanagh, *et al.*, "*De novo* mutations in schizophrenia implicate synaptic networks," *Nature* **506**(7487), 179 (2014).
7. S. Zaidi, M. Choi, H. Wakimoto, *et al.*, "*De novo* mutations in histone-modifying genes in congenital heart disease," *Nature* **498**(7453), 220 (2013).
8. I. Iossifov, B. J. O'Roak, S. J. Sanders, *et al.*, "The contribution of *de novo* coding mutations to autism spectrum disorder," *Nature* **515**(7526), 216 (2014).
9. B. J. O'Roak, L. Vives, S. Girirajan, *et al.*, "Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations," *Nature* **485**(7397), 246 (2012).
10. A. Soylev, C. Kockan, F. Hormozdiari, *et al.*, "Toolkit for automated and rapid discovery of structural variants," *Methods* **129**, 3–7 (2017).
11. R. M. Layer, C. Chiang, A. R. Quinlan, *et al.*, "Lumpy: a probabilistic framework for structural variant discovery," *Genome biology* **15**(6), R84 (2014).
12. K. Ye, M. H. Schulz, Q. Long, *et al.*, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads," *Bioinformatics* **25**(21), 2865–2871 (2009).

13. T. Rausch, T. Zichner, A. Schlattl, *et al.*, "DELLY: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics* **28**(18), i333–i339 (2012).
14. F. Hormozdiari, C. Alkan, E. E. Eichler, *et al.*, "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes," *Genome Research* (2009).
15. S. S. Sindi, S. Önal, L. C. Peng, *et al.*, "An integrative probabilistic model for identification of structural variation in sequencing data," *Genome Biology* **13**(3), R22 (2012).
16. P. Medvedev, M. Fiume, M. Dzamba, *et al.*, "Detecting copy number variation with mated short reads," *Genome Research* (2010).
17. C. D. Campbell and E. E. Eichler, "Properties and rates of germline mutations in humans," *Trends in Genetics* **29**(10), 575 – 584 (2013). Human Genetics.
18. Z. Chong, J. Ruan, M. Gao, *et al.*, "novoBreak: local assembly for breakpoint detection in cancer genomes," *Nature Methods* **14**(1), 65 (2017).
19. G. Narzisi, J. A. O'Rawe, I. Iossifov, *et al.*, "Accurate *de novo* and transmitted indel detection in exome-capture data using microassembly," *Nature Methods* **11**(10), 1033 (2014).
20. J. Köster and S. Rahmann, "Snakemake: a scalable bioinformatics workflow engine," *Bioinformatics* **28**(19), 2520–2522 (2012).
21. L. C. Francioli, M. Cretu-Stancu, K. V. Garimella, *et al.*, "A framework for the detection of *de novo* mutations in family-based sequencing data," *European Journal Of Human Genetics* **25** (2016).
22. Q. Wei, X. Zhan, X. Zhong, *et al.*, "A Bayesian framework for *de novo* mutation calling in parents-offspring trios," *Bioinformatics* **31**(9), 1375–1381 (2015).
23. D. M. Werling, H. Brand, J.-Y. An, *et al.*, "An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder.," *Nat Genet* **50**, 727–736 (2018).
24. T. N. Turner, F. Hormozdiari, and M. H. D. *et al.*, "Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory dna," *American Journal of Human Genetics* **98**(1), 58–74 (2016).
25. T. N. Turner, B. P. Coe, D. E. Dickel, *et al.*, "Genomic patterns of *de novo* mutation in simplex autism," *Cell* **171**(3), 710–722 (2017).
26. Q. Zhang, J. Pell, R. Canino-Koning, *et al.*, "These are not the k-mers you are looking for: Efficient online k-mer counting using a probabilistic data structure," *PLoS ONE* **9**, 1–13 (2014).
27. M. R. Crusoe, H. F. Alameldin, S. Awad, *et al.*, "The khmer software package: enabling efficient nucleotide sequence analysis," *F1000Research* **4**(900) (2015).
28. D. S. Standage, A. Aliyari, L. J. Cohen, *et al.*, "khmer release v2.1: software for biological sequence analysis," *The Journal of Open Source Software* **2**, 272 (2017).
29. G. Rizk, D. Lavenier, and R. Chikhi, "DSK: k-mer counting with very low memory usage," *Bioinformatics* **29**(5), 652–653 (2013).
30. G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics* **27**(6), 764–770 (2011).
31. N. L. Bray, H. Pimentel, P. Melsted, *et al.*, "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology* **34**(5), 525 (2016).
32. R. Patro, S. M. Mount, and C. Kingsford, "Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms," *Nature Biotechnology* **32**(5), 462 (2014).
33. R. Patro, G. Duggal, M. I. Love, *et al.*, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods* **14**, 417 EP – (2017).
34. A. Rahman, I. Hallgrímsdóttir, M. Eisen, *et al.*, "Association mapping from sequencing reads using k-mers," *eLife* **7**, e32920 (2018).
35. C. Sun and P. Medvedev, "Toward fast and accurate SNP genotyping from whole genome sequencing data for bedside diagnostics," *bioRxiv* , 239871 (2018).
36. H. Li, "fermi-lite: Standalone C library for assembling illumina short reads in small regions." https://github.com/lh3/fermi-lite (2017).
37. H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." arXiv:1303.3997 (2013).
38. H. Li, "KSW2: Global alignment and alignment extension." https://github.com/lh3/ksw2 (2017).
39. H. Li, "wgsim: Read simulator for next generation sequencing." https://github.com/lh3/wgsim (2011).

Appendix

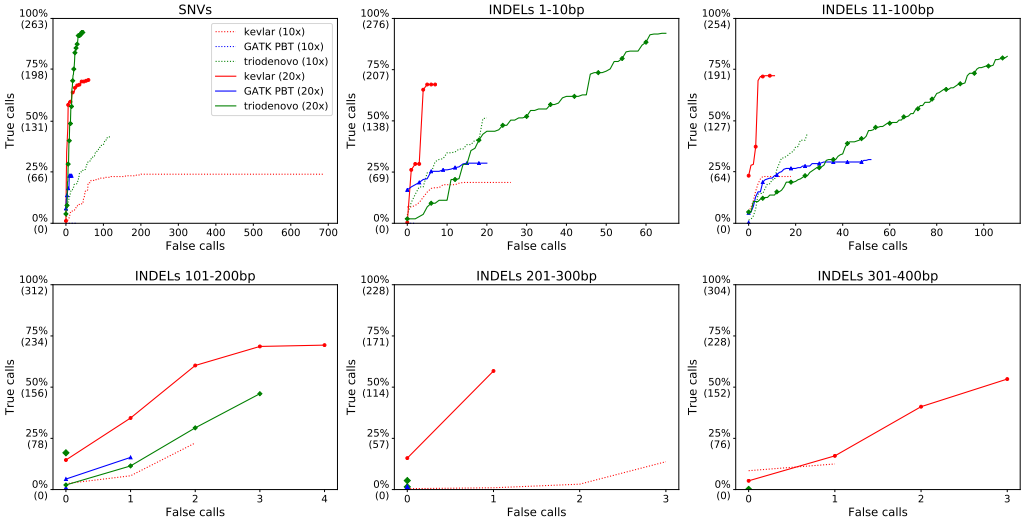

Fig A1 Receiver operating characteristic (ROC) curves comparing variant prediction performance on a simulated data set, shown for 10x coverage (dashed lines) and 20x coverage (solid lines). Each of the six panes shows prediction accuracy for a different variant type: single nucleotide variants (SNVs); insertions or deletions (indels) 1-10bp in length; 11-100bp indels; 101-200bp indels; 201-300bp indels; and 301-400bp indels.