Kevlar: A Mapping-Free Framework for Accurate Discovery of De Novo Variants

Rita Rizkallah, Taline El Hany

April 2024

1 Introduction

The genetic heritability, which measures the amount of variation in a phenotypic trait that is due to genetic diversity among individuals, is high for many disorders. However, known genetic variants only account for a portion of this heritability; this phenomenon is known as missing heritability. *De novo* mutations, which are mutations that are only found in the child and absent in both parents, have been thought to be a major cause of missing heritability. Nevertheless, because of the complexity of *de novo* variant discovery, their contribution to these disorders has been partially ignored.

2 Computational and Biological Complications of De Novo Variant Discovery

Accurate *de novo* variant discovery is still a difficult task due to a number of biological and computational reasons. Discovering *de novo* variants calls for methods that can reliably detect variations in the proband, while guaranteeing their absence in both parents. High accuracy and confidence is required, which makes this process difficult. Moreover, larger variants like insertions and deletions can affect more nucleotides and may have a bigger effect on genetic disorders. However, because these are difficult to predict, it is more challenging to identify them. On another note, in a reference mapping context, in order to make confident indels predictions, an accurate and precise mapping of all reads covering the indel with consistent gap alignment is necessary. But because there is a higher chance of mistakes and misalignments in read mapping, this is challenging for longer indels since predicting them tends to have high false-positive and false-negative rates. Complex types of structural variants also completely lack methods for accurate predictions.

3 Mapping-Based Methods

One effective technique for finding new genetic variants from de novo mutations is whole-genome sequencing. First, the sequences are aligned to a reference genome, followed by variant detection through comparison of sequences. Finally, variant discovery algorithms evaluate millions of variant predictions produced by this initial step in order to distinguish between true *de novo* variants and inherited variants. Some limitations of reference-based variant discovery methods are highlighted, despite their proven value in the research of complex genetic disorders.

3.0.1 Challenges In Read Alignment

Even with advancements in read alignment algorithms, it is still challenging to precisely map each read to the reference genome. This is due to the fact that alignment can be hindered by sequencing errors and repetitive DNA sequences which can have numerous possible mapping locations . Moreover, reads that do not map to the reference genome are typically ignored by mapping-based variant predictors. These reads may contain important information about the presence of variations.

3.0.2 Limited Prediction of Multiple Variant Types

Many methods focus exclusively on predicting specific types of variants such as single-nucleotide variants. Few techniques have the ability to actually predict multiple types of variants at once with a single strategy. This limitation can lead to overlooking complex or overlapping variations since different types of variants are analyzed separately.

4 Existing Mapping-Free Methods

Mapping free methods of *de novo* variant prediction do not require read alignments to a reference genome, they compare sequences of related individuals instead. Because of their effectiveness and robustness to the shortcomings of mapping-based methods, mapping-free approaches have gained a lot of prominence.

Cortex was among the first programs to explore a mapping-free method for predicting variants. It presented the idea of a "colored de Bruijn graph", which allows the comparison of sequence content and the prediction of variants. In short, the de Bruijn graph is a data structure used to represent sequence assemblies. It was actually successful in prediciting variants in the 1000 Genomes project. However, it can be computationally demanding to create and analyze colored de Bruijn graphs for large-scale genomic datasets.

Additionally, a unique mapping-free technique for carrying out genome-wide association studies has been introduced by a recent study; HAWK. HAWK quickly finds variant-phenotype connections by detecting k-mers significantly

related with a trait or phenotype, referred to as "significant k-mers." Then, using a local assembly of these important k-mers, it predicts the corresponding significant variants associated with the desired phenotype.

5 Results

The study presents a mapping-free method for de novo variant discovery that is based on k-mer analysis called Kevlar which employs a very similar approach to HAWK's, which was discussed earlier. This approach relies on the idea that short sequences called k-mers, which cover a newly formed genetic mutation, are likely to be unique with high probability. The reads are divided into k-mers for each sample, and the abundance of each k-mer is stored. Following that, a second pass over reads from the case sample identifies every k-mer that is specific to the proband. Reads containing any novel k-mers are retained for subsequent processing. After applying filters for contamination and incorrect k-mer counts, the reads with new k-mers are partitioned so that any two reads sharing at least one new k-mer are grouped together. Then, the reads in each partition are put together to form a contig, which is compared to the reference genome to determine whether a variant is present and to produce a variant call. Lastly, Kevlar ranks and filters the data using a likelihood-based score.

5.1 Performance on Data

Sequencing of an entire genome was modeled in a mock family to assess Kevlar's accuracy in predicting variants at different sequencing depths. Low error rate sequencing was simulated at 10x, 20x, 30x, and 50x coverage. Kevlar's accuracy was compared with two mapping-based de novo variant callers (GATK and TrioDenovo) as well as two mapping-free methods (Scalpel and DiscoSnp++). At low (10x) coverage, the accuracy of all variant callers assessed is poor. However, all five algorithms demonstrate a significant improvement in SNV identification and Kevlar's ability to precisely identify indels ¿100 bp is apparent at at 20x coverage. Kevlar delivers the best results across all variation types at greater coverage levels (30x and 50x). However, TrioDenovo exhibits slightly higher sensitivity than Kevlar, but at the cost of a significant number of incorrect predictions. Additionally, Kevlar matches and even outperforms Scalpel at predicting indels.

In order to evaluate Kevlar's effectiveness on real data, Kevlar was applied to predict de novo variants of the proband of an autistic trio from the Simons Simplex Collection. 14 out of 196 calls have been verified as de novo variants by experimental validation, according to annotations in the denovo-db database. Kevlar predicts 219 de novo variations, many of which were given a poor possibility of being a true de novo event by Kevlar. Actyally, thirteen of the fourteen denovo-db calls that had experimental validation were correctly predicted by Kevlar and given a high likelihood score, indicating a confident de novo variant call. Moreover, Kevlar variant calls include four calls not present in the

database, which are probably false calls.

6 Discussion

Despite its tremendous potential, Kevlar is still facing challenges since it's still undergoing heavy development, which makes its installation more difficult. Furthermore, Kevlar's memory requirements can be quite demanding. Kevlar's memory requirements can be significantly decreased by applying error correction to the input data, but doing so usually results in a decrease in sensitivity. Nevertheless, unlike existing methods, by discovering many variant types at once, Kevlar offers a substantial advancement in de novo variant discovery despite these drawbacks. Moreover, its use of a probabilistic scoring system also improves the accuracy of results. Although Kevlar does not use a reference genome to identify k-mers, it still In the future, Kevlar might enhance its abilities by eliminating the need for a reference genome entirely.