Bioinformatics, 38(5), 2022, 1231–1234
https://doi.org/10.1093/bioinformatics/btab805
Advance Access Publication Date: 2 December 2021
Original Paper

Sequence analysis

trfermikit: a tool to discover VNTR-associated deletions

Peter McHale and Aaron R. Quinlan*

Department of Human Genetics and Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on May 18, 2021; revised on October 25, 2021; editorial decision on November 5, 2021; accepted on November 27, 2021

Abstract

Summary: We present trfermikit, a software tool designed to detect deletions larger than 50 bp occurring in Variable Number Tandem Repeats using Illumina DNA sequencing reads. In such regions, it achieves a better tradeoff between sensitivity and false discovery than a state-of-the-art structural variation caller, Manta and complements it by recovering a significant number of deletions that Manta missed. trfermikit is based upon the fermikit pipeline, which performs read assembly, maps the assembly to the reference genome and calls variants from the alignment.
Availability and implementation: https://github.com/petermchale/trfermikit.
Contact: aquinlan@genetics.utah.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A recent study that comprehensively compared long- and short-read sequencing technologies found that most structural variations (SVs) missed by a suite of standard short-read structural-variant callers lie in tandem repeats (Chaisson *et al.*, 2019). For example, 3895 out of 5031 deletions called from long- but not short-read sequencing data (HG00514) lie in tandem repeats. This may be a simple consequence of the fact that most SVs lie in tandem repeats (Linthorst *et al.*, 2020), but is also likely due to the fact that short reads typically do not span such repetitive sequences, confounding sequence alignment and consequently variant detection.

Thus substantial gains in SV detection sensitivity stand to be realized by optimizing SV calling in tandem repeats. Since callers exist that have been designed specifically to capture SVs in simple tandem repeats with motifs smaller than 6 bp (Dashnow *et al.*, 2018; Dolzhenko *et al.*, 2019; Mousavi *et al.*, 2019), we show here how SV detection (specifically deletions longer than 50 bp) can be optimized in Variable Number Tandem Repeats (VNTRs) composed of repetitive motifs longer than 6 bps.

2 Materials and methods

2.1 Data provenance and caller evaluations

2.1.1 Chaisson *et al.* (2019) study

SVs were discovered in samples HG00514, HG00733 and NA19240 using Illumina Whole Genome Sequencing (WGS) (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/), aligned against GRCh38 and evaluated using truvari (https://github.com/spiralgenetics/truvari) against a set of calls based upon PacBio WGS (https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/genotype/nstd152/) (Chaisson *et al.*, 2019). Alignments of local assemblies of PacBio reads (obtained privately from Mark Chaisson; http://www-rcf.usc.edu/~mchaisso/hgsvg/local_assemblies/) were used in the evaluation stage to filter out VNTRs not covered by Pacbio assemblies.

2.1.2 Ebert *et al.*, (2021) study

SVs were discovered in samples HG02818, HG03125, HG03486 and NA12878 using Illumina WGS obtainable via the following HTTPS hyperlinks:

- https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/additional_698_related/1000G_698_related_high_coverage.sequence.index
- https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/1000G_2504_high_coverage.sequence.index

Reads were aligned to GRCh38. Truvari was used to evaluate the Illumina-based calls against a set of calls based upon PacBio WGS available at the following links:

- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_insdel_alt.vcf.gz
- ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/README_HGSVC_release_v2

The four samples considered here were sequenced using HiFi. SVs in the benchmark that was smaller than 50 bp were filtered out. Evaluation was performed only on VNTRs that overlapped regions covered by PacBio reads; such regions were obtained from:

- https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20200728_Freeze2_PAV_PBSV/pavhifi/${sample}/align/${sample}_hifi_aligned_tig_${haplotype}.bed.gz

where *${haplotype}* was either 'h1' or 'h2' and *${sample}* was one of the four samples: HG02818, HG03125, HG03486 and NA12878. VNTRs that lie on the Y chromosome, or that overlap a set of low-confidence regions, were removed prior to benchmarking; the low-confidence regions are obtainable from:

- http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/filter/20210127_LowConfidenceFilter/LowConfidenceFilter.bed.gz
- http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/filter/20210127_LowConfidenceFilter/README_20210127_LowConfidenceFilter

2.2 trfermikit modules

2.2.1 Make-regions module

Tandem repeats are classified by length (*minRepeatLength* in *config.json*, published at (McHale, 2021)) and period (*minRepeatPeriod*), merged into intervals that contain at least one positive-class tandem repeat, and filtered on length (*maxRegionLength*). If the user provides a set of 'functional' intervals, then only those merged intervals that intersect the functional intervals are retained. Merged intervals are then filtered on short-read coverage.

2.2.2 Make-calls module

Short reads are pulled that originally aligned to the merged intervals, and fermikit is used to assemble those reads into unitigs, map them to the reference genome (using minimap2) and call variants from the alignment.

2.2.3 Filter-calls module

Each SV call is filtered on its size and the mapping quality of its supporting unitigs (i.e. those unitigs comprising a gap that aligns with the call). Each such unitig is composed of 'blocks': aligned regions of the unitig that are free of indels. trfermikit determines the largest block upstream and largest block downstream of the call. If either is smaller than a threshold value (*minUnitigBlockLength*), then trfermikit considers the call to be a FP and filters it out. The remaining calls are ascribed a 'confidence' equal to (*max_block_size_upstream_of_call* + *max_block_size_downstream_of_call*)/len(*blocks*), reflecting our observation that true-positive calls were often supported by 'clean' unitigs, in which blocks were individually long and few in number. In a final filtering step, calls are clustered by genomic position, and the call with the highest 'confidence' (as defined above) in each cluster is retained.

2.3 Exploration of parameter space

By exploring combinations of multiple trfermikit parameter values [see *manta_complementarity__DEL-manta-all_regions.json*, published at McHale (2021)], we generated a region in recall-FDR space, not a curve as would be achieved by varying a single parameter. Part of the boundary of that region is a Pareto front (a term borrowed from economics) and represents a set of operating points that are optimal in that, for each such point, there is no other operating point with better recall and better FDR. Trfermikit's default operating point [as defined in *config.json* and published at McHale (2021)] is indicated by a red circle in Figure 1B and C.

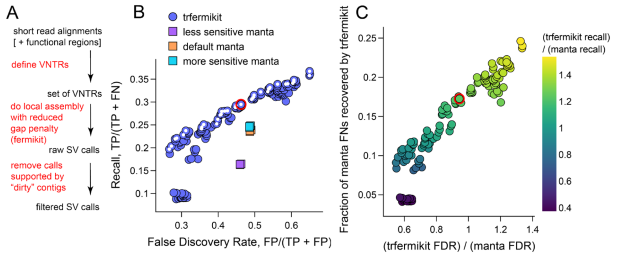

Fig. 1. (**A**) Stages of the trfermikit pipeline. (**B** and **C**) Performance of trfermikit on deletions in sample HG00514. (B) trfermikit is characterized by a better recall-FDR tradeoff than Manta. TP = number of true positives; FP = number of false positives; FN = number of false negatives. The Pareto front of trfermikit operating points is indicated in white (Supplementary Information) and the default operating point is indicated with a red circle (also shown in panel C). (C) trfermikit is complementary to Manta (default mode). Same set of trfermikit parameters as in (B). Throughout, regions were formed by merging tandem repeat regions with motifs >6 bp and tandem-repeat length >100 bp and calls <50 bp in size were filtered out. Functional regions (see panel A) were not used

2.4 Manta operating points

In addition to Manta's standard operating mode (labeled 'default manta' in Fig. 1B), we explored two additional configurations. The configurations labeled 'more sensitive manta' and 'less sensitive manta' are defined by {*minEdgeObservations*: *2*, *minCandidateSpanningCount: 2*} and {*minEdgeObservations: 1*, *minCandidateSpanningCount: 1*}, respectively.

2.5 Figure generation

Two Jupyter notebooks that reproduce the figures in this manuscript are published at (McHale, 2021).

3 Results

We developed a pipeline, which we termed trfermikit, that chooses a set of VNTRs, runs fermikit on those VNTRs (Li, 2015) and filters out false-positive variant predictions (Fig. 1A). We then assessed the performance of trfermikit for a variety of its parameter values on a sample for which both Illumina data and Pacbio data are available (sample HG00514 from (Chaisson *et al.*, 2019)). Trfermikit has a better tradeoff between recall and false discovery than that obtained by running Manta (Chen *et al.*, 2016) on the same set of regions using three different settings of its parameters (Fig. 1B and Supplementary Information). At its default operating point (red circle in Fig. 1B), trfermikit runs in roughly 32 CPU hours per sample (CPU = Central Processing Unit).

The superior performance of trfermikit on VNTRs is due to, primarily, two factors (Fig. 1A). When assembled reads are aligned to the reference genome, the gap penalty is reduced to a value much smaller than is typical of SV callers, enabling trfermikit to detect deletions that other callers would miss. Supplementary Figure S2A, C, D and F illustrate, using a synthetic assembled contig, the efficacy of this approach to uncover deletions accurately, even in tandem repeat regions. The second key ingredient of trfermikit is its call filtering strategy, as reducing the gap penalty ipso facto introduces many false-positive calls. Manual inspection revealed that many of these false positives are characterized by 'noisy' contigs that contain a plethora of gaps and comprising only short 'blocks' of contiguously aligned bases. Experiments revealed that we could selectively remove these false positives via simple heuristics involving the size and number of those blocks [see Section 2.2.3 (filter-calls module) in Section 2].

Given that many tandem-repeat loci that are associated with disease are known to reside in gene bodies (Bakhtiari *et al.*, 2020; Bennett *et al.*, 1997; De Roeck *et al.*, 2018; Hannan, 2018; Kyo *et al.*, 1999; Lalioti *et al.*, 1997; Li *et al.*, 2016; Sulovari *et al.*, 2019; Yang *et al.*, 2000), we ran trfermikit on VNTRs that lie in gene bodies by specifying a corresponding functional region as input to

trfermikit (Fig. 1A), yielding a recall of 0.30 and a false discovery rate of 0.48. These numbers are comparable to those obtained when interrogating all VNTRs (recall = 0.29 and FDR = 0.46). Running trfermikit on VNTRs that not only lie in gene bodies, but also in exons and/or UTRs, yields a recall of 0.23 and a FDR of 0.56. For all three sets of genomic intervals, trfermikit significantly complemented Manta, recovering 17, 18 and 16% of deletions that Manta missed when applied to all VNTRs, those in gene bodies, and those in exons and/or UTRs, respectively (Fig. 1C).

One might expect a tradeoff between VNTR length and ability to assess variation at such loci, e.g. because assembly of Illumina reads is less accurate, or more fragmented, in longer repetitive regions. Consistent with that expectation, we found that the recall of both trfermikit and Manta decreases at larger VNTR loci (Fig. 2A), while their FDRs increase with VNTR length (Fig. 2B). For those VNTR size classes in which deletion events are common, trfermikit's sensitivity exceeds that of Manta (Fig. 2A and C), explaining why trfermikit recall is greater than Manta when pooled over all VNTRs (Fig. 1B).

An oft-employed strategy for combining call sets is to compute their intersection, with a primary objective of reducing FDR. While this strategy does reduce FDR (Fig. 2B), it does not drive it to zero, indicating that some false positives are common to both callers. Moreover, the recall of the intersection call set is lower (Fig. 2A), as many of the true positives are unique to a single caller, e.g. trfermikit recovers 40% of Manta false negatives occurring in VNTRs in the size range of 125–150 bp (Fig. 2D).

In light of the complementarity of trfermikit and Manta, we conjectured that combining their call sets (i.e. creating their union) would increase recall relative to what either caller could achieve alone. Our experiments support this conjecture (Fig. 2A), and, moreover, show that the combined call set has an FDR comparable to the individual callers (Fig. 2B).

We noticed that recall and FDR move in opposite directions, as one varies the VNTR size class in which the metrics are measured (compare Fig. 2A with B). At first glance, this might appear to contradict the well-known tradeoff between recall and FDR in which recall and FDR move in the same direction as one tunes a variant caller. The explanation follows from two facts. First, the

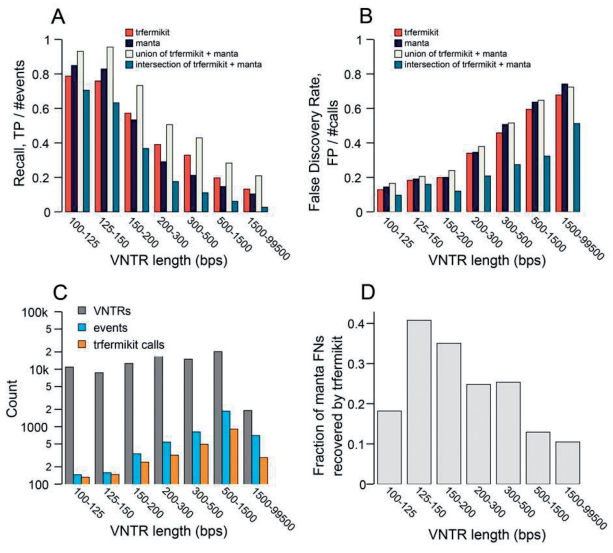

Fig. 2. (**A** and **B**) Performance of trfermikit (red; using default operating point indicated by the red circle in Fig. 1), Manta (dark blue; default mode), their union (ivory; in which variants unique to either callset and variants common to both are combined) and their intersection (teal) on deletions in sample HG00514, stratified by VNTR size. Number of events = TP + FN and number of calls = TP + FP. (**C**) The number of VNTRs in each of the VNTR size classes (gray), together with the distribution of deletion events (blue) and trfermikit deletion calls (orange) across those classes. (**D**) The degree to which trfermikit complements Manta, stratified by VNTR size class

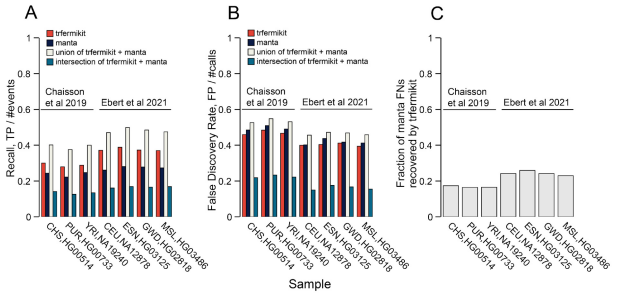

Fig. 3. (**A** and **B**) Performance of trfermikit (red), Manta (dark blue), their union (ivory) and their intersection (teal) across many samples, including three from Chaisson *et al.* (2019), which were sequenced using PacBio Single Molecule, Real-Time (SMRT) sequencing, and four from Ebert *et al.* (2021), which were sequenced using the Circular Consensus Sequencing mode of Pacbio SMRT sequencing (generating HiFi reads). (**C**) The degree to which trfermikit complements Manta across samples. Also see Supplementary Figure S3, where performance is broken down by VNTR size class, for each of the samples indicated here

ratio, *R*, of the number of events to the number of calls per VNTR size class is approximately constant across classes (compare the blue and orange bars, respectively, in Fig. 2C). Second, recall and FDR obey the following formula: recall = (1 − FDR)/R. Together, these facts imply that recall goes down as FDR goes up, as we observe in our data (Fig. 2A and B).

Given that trfermikit was tuned to perform optimally on a single sample (HG00514), and might therefore be overfit, we assessed the degree to which its performance generalizes to other samples. Figure 3 shows that performance of both trfermikit and the combined call sets (the union or intersection of trfermikit and Manta) are qualitatively the same for two other samples from the same study HG00514 was taken from (Chaisson *et al.*, 2019). A more recent study from the same consortium has recently produced PacBio assemblies of higher quality (Ebert *et al.*, 2021), potentially providing a more accurate benchmark on which to assess trfermikit. We therefore ran trfermikit and Manta on four samples from that study. Figure 3 shows that performance measured relative to the more accurate benchmark (Ebert *et al.*, 2021) is higher than that of the less accurate benchmark (Chaisson *et al.*, 2019), further validating the generalizability of trfermikit.

4 Discussion

We describe a new genetic variant detection strategy that discovers a significant number of deletions missed by Manta in VNTRs, suggesting that both trfermikit and Manta should be run in rare disease cases for which compelling single-nucleotide variant or deletion variant candidates have not been identified by tools such as GATK (Genome Analysis ToolKit) (McKenna *et al.*, 2010). Moreover, our results argue that call sets should be combined by taking their union (instead of their intersection), as the intersection strategy significantly reduces recall, whereas the union strategy significantly increases recall without impacting appreciably upon FDR.

We note that trfermikit is less sensitive to insertions than Manta (Supplementary Fig. S1). Two effects explain why. First, reducing the gap penalty to favor alignment gaps exposes deletions but can mask insertions (Supplementary Fig. S2). Second, many insertions harbor a soft-clipped unitig signature (e.g. Supplementary Fig. S2B) that Manta but not trfermikit can detect.

Acknowledgements

The authors acknowledge insightful conversations with Brent Pedersen, Ryan Layer and Harriet Dashnow.

Data Availability

The data underlying this article are available in Zenodo at https://zenodo.org/record/5576328.

Funding

This work was supported by the National Institutes of Health [1S10OD021644-01A1 and 5R01HG010757].

Conflict of Interest: none declared.

References

Bakhtiari, M. *et al.* (2021) Variable number tandem repeats mediate the expression of proximal genes. *Nat Commun* 12, 2075.

Bennett,S.T. *et al.* (1997) Insulin VNTR allele-specific effect in type 1 diabetes depends on identity of untransmitted paternal allele. The IMDIAB Group. *Nat. Genet.*, **17**, 350–352.

Chaisson,M.J.P. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.

Chen,X. *et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**, 1220–1222.

Dashnow,H. *et al.* (2018) STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.*, **19**, 121.

De Roeck,A. *et al.*; On Behalf of the BELNEU Consortium. (2018) An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathol.*, **135**, 827–837.

Dolzhenko,E. *et al.* (2019) ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*, **35**, 4754–4756.

Ebert,P. *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, **372, eabf7117**.

Hannan,A.J. (2018) Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.*, **19**, 286–298.

Kyo,K. *et al.* (1999) Association of ulcerative colitis with rare VNTR alleles of the human intestinal mucin gene, MUC3. *Hum. Mol. Genet.*, **8**, 307–311.

Lalioti,M.D. *et al.* (1997) Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, **386**, 847–851.

Li,H. (2015) FermiKit: assembly-based variant calling for illumina resequencing data. *Bioinformatics*, **31**, 3694–3696.

Li,M. *et al.* (2016) A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 Schizophrenia-Associated Locus. *Nat. Med.*, **22**, 649–656.

Linthorst,J. *et al.* (2020) Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain. *Transl. Psychiatry*, **10**, 369.

McHale,P. (2021) Trfermikit Supporting Software Files. Zenodo. 10.5281/zenodo.5576328.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Mousavi,N. *et al.* (2019) Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.*, **47**, e90.

Sulovari,A. *et al.* (2019) Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl. Acad. Sci. USA*, **116**, 23243–23253.

Yang,F. *et al.* (2000) Variable number tandem repeat in exon/intron border of the cystathionine beta-synthase gene: a single nucleotide substitution in the second repeat prevents multiple alternate splicing. *Am. J. Med. Genet.*, **95**, 385–390.