MUM&Co: accurate detection of all SV types through whole-genome alignment

Samuel O'Donnell, Gilles Fischer

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

MUM&Co: Accurate detection of all SV types through whole genome alignment

Samuel O'Donnell[1] and Gilles Fischer[1,*]

[1] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, Laboratory of Computational and Quantitative Biology, F-75005, Paris, France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: MUM&Co is a single bash script to detect Structural Variations (SVs) utilizing Whole Genome Alignment (WGA). Using MUMmer's nucmer alignment, MUM&Co can detect insertions, deletions, tandem duplications, inversions and translocations greater than 50bp. Its versatility depends upon the WGA and therefore benefits from contiguous *de-novo* assemblies generated by 3rd generation sequencing technologies. Benchmarked against 5 WGA SV-calling tools, MUM&Co outperforms all tools on simulated SVs in yeast, plant and human genomes and performs similarly in two real human datasets. Additionally, MUM&Co is particularly unique in its ability to find inversions in both simulated and real datasets. Lastly, MUM&Co's primary output is an intuitive tabulated file containing a list of SVs with only necessary genomic details

Availability: https://github.com/SAMtoBAM/MUMandCo.

Contact: gilles.fischer@sorbonne-universite.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advent of long read sequencing technologies and increasingly contiguous *de-novo* assemblies provides a better landscape for SV detection. This improves on short read technologies which were unable to either detect complex rearrangements beyond copy number variants or span complex regions during *de-novo* assembly. However, there are a limited number of tools dedicated to discovering SVs through Whole Genome Alignment (WGA), reducing our ability to take advantage of this wealth of information. As long-read sequencing technologies continue to improve in both throughput and read length, alongside improvements in *de-novo* assembly, automated WGA SV detection is becoming increasingly necessary. In response to this, MUM&Co was developed to provide a detailed description of all structural changes between two genomes visible in WGA, namely deletions, insertions, tandem duplication, inversions and translocations.

2 Methods

MUM&Co's initial inputs are two genomes for comparison. First MUM&Co utilizes MUMmer to perform reciprocal WGA, global (g-) and many-to-many (m-) filtering and coordinate parsing using the nucmer, delta-filter and show-coords scripts respectively (Marçais et al. 2018)(Supp figure 1). MUM&Co first associates query contigs to reference chromosomes by selecting a subset of the most accurate, non-overlapping alignments. Alignments outside of these pairings are filtered. Following this, MUM&Co uses alignment characteristics, outlined below, to detect each SV type, to a minimum of 50bp, taking advantage of both the g- and m-alignment filters to provide complementary information that is used to filter out false positives (Supp figure 2). The g-alignment is used to detect translocation fragments based upon multiple contig-chromosome pairings, large inversions using the alignment orientation and possible insertion and deletion events based on alignment gaps (Supp figure 1). Using the reciprocal alignment, gaps within the reference and query are considered deletions and insertions, respectively. The m-alignment is used to find potential inversions and duplications. For inversions, the main contig orientation is calculated using the g-alignment and all alignments with the opposite orientation are labelled as potential. The call is then confirmed as true if there are alignment gaps in the reciprocal g-alignment, represented as both a potential insertion and

deletion at the same loci (Supp figure 2). If used to verify the inversion, these gaps are subsequently removed from the list of potential insertions/deletions. This step greatly reduces the number of false positives for insertions, deletions and inversions. For the resulting insertion and deletion events, an option is provided in which they can be catalogued as involving mobile or novel genetic elements using BLAST. Overlapping regions from the m-alignment are labelled as potential duplications and confirmed if they are in concordance with the g-alignment with respect to any additional events frequently present at their borders. Finally, a tsv file is produced detailing the reference and query chromosome involved, the positions in each and both the type and size of the SV (Supp figure 1).

3 Results

MUM&Co's performance was evaluated and benchmarked against five alternate WGA SV callers, Assemblytics (Nattestad et al. 2016), MUMmer's show-diff script, paftools (Li 2018), SVrefine from SVanalyzer (v0.2) and SVMU (v0.3) (Chakraborty et al. 2017), that all similarly use nucmer alignments except paftools. Each was first tested on randomly simulated SVs created by simuG (Yue and Liti 2019) in the reference genomes of *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and in Human chromosomes 11-20. These included deletions, dispersed duplications, which were considered equivalent to insertions, tandem duplications, inversions and translocations, simulated separately in each background. Deletion and inversion events were simulated 100 times. For tandem and dispersed duplication events, 100 randomly selected sites were chosen to be duplicated 1 to 4 times either in tandem or at random sites within the genome respectively. Lastly, 8, 2 and 5 translocations were introduced into the *S.cerevisiae*, *A.thaliana*, and Human genome respectively, corresponding to the maximum number of translocation events possible with simuG which is limited by events being necessarily reciprocal and only occurring once per chromosome. The original and rearranged genomes were then aligned using nucmer or minimap2 for paftools with recommended settings for each tool if possible. All tools were assessed on all simulated SVs regardless of the limitations in the types of SVs detected by each tool. This highlights false positives potentially introduced by the presence of alternate SVs (Supp table 1). An in-house script was used to compare SV calls available in the MUM&Co github 'comparison_files' folder Following this, it was apparent that SVrefine detected only 1-4% of SVs, in all conditions and therefore results are not shown.

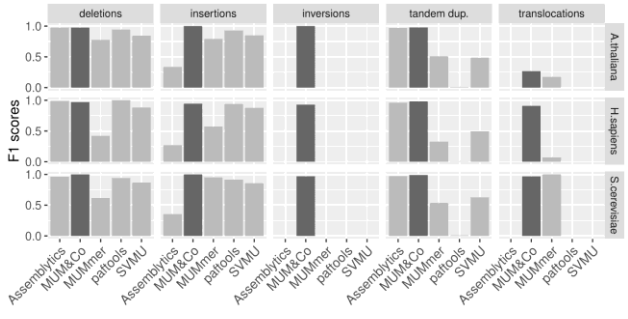

Fig. 1. Benchmarking of MUM&Co against four WGA-based SV-callers. SimuG was used to simulate deletions, tandem duplications, insertions, inversions and translocations in *S. cerevisiae*, *A. thaliana* genomes and *H.sapiens* chromosomes 11-20. Default settings were used for each SV caller. Calls were considered true positives if both the SV type and coordinates (within a 500bp window i.e. 250bp either side) were correct.

Within the benchmarking comparisons, MUM&Co detected the most SVs (97%) and had the highest average F1score (0.92). Additionally inversions were only detected by MUM&Co (average F1-score 0.96) and translocations were detected only by both MUM&Co and MUMmer's show-diff with average F1-scores of 0.9 and 0.41 respectively. Notably, the large number of false positives for both show-diff and SVMU concerning inversions occurs due to a caveat in which only bordering regions of inversions are detected (Supp table 1). The same four tools were then tested on two human genome datasets with validated SV lists, HG002 (Zook et al. 2019) and HG00268 (Audano et al. 2019). Paftools performed better than all tools in detecting deletions and insertions whilst other tools performed similarly (Supp figure 3, Supp table 2 and 3). However, in HG00268, MUM&Co remained the only tool able to detect inversions.

4 Discussion

MUM&Co will complement the increasing number of high quality *de-novo* assemblies by enhancing their use for SV detection. Using WGA and combining both the global and many-to-many delta-filter filters are key to its performance with the latter mainly providing a robust means of removing false positive calls for both duplications and inversions especially in genomes with repetitive elements. MUM&Co and MUMmer's 'show-diff' script are the only two tools able to identify the entire range of simulated SV, including translocations. MUM&Co is unique in its ability to detect inversions in both simulated and real datasets. Additionally, in both datasets it detects deletions and insertions similarly to other tools at the exception of Paftools which performs significantly better on real human datasets.

Acknowledgements

Thank you to Stephane Delmas and Nicolas Agier for advice and troubleshooting

Funding

This work has been supported by the the Agence Nationale de la Recherche [ANR-16-CE12-0019].

Conflict of Interest: none declared.

References

Audano *et al.* (2019) Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*. 176(3), 663-675.

Chakraborty *et al.* (2018) Hiddden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet*. 50, 20-25.

Marçais *et al.* (2018) MUMmer 4: a fast and versatile genome alignment system. *PLoS Comput Biol. 14(1):e1005944*

Li (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34(18), 3094-3100.

Nattestad *et al.* (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*. 32(19), 3021-3023.

Yue and Liti (2019) simuG: a general-purpose gnome simulator. *Bioinformatics*. 35(21), 4442-4444.

Zook *et al.* (2019) A robust benchmark for germline structural variant detection. *bioRxiv*. 664623; doi: https://doi.org/10.1101/664623