nf-LO

Brief Communication

nf-LO: A scalable, containerised workflow for genome-to-genome lift over

Andrea Talenti[1] and James Prendergast[1]

[1] The Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, UK

Correspondence to: Andrea Talenti – andrea.talenti@ed.ac.uk

Keywords: liftover, assembly, Nextflow, workflow

Significance Statement

Studies such as the vertebrate genomes project (VGP) aim to produce high quality genome assemblies for tens of thousands of species. However, these new genomes most often come with limited annotations, reducing their utility. One solution is to “lift over” annotations from better annotated genomes. This process is though complex, requiring multiple steps which differ depending on the distance between the species. In this paper we present nf-LO, a streamlined, containerised Nextflow workflow that can enable rapid genome lift over between any pair of species and which can be easily implemented on any system. We believe that its ease of implementation, scalability and flexibility will allow for widespread use and rapid adoption by the scientific community.

Abstract

The increasing availability of new genome assemblies often comes with a paucity of associated genomic annotations, limiting the range of studies that can be performed. A common workaround is to lift over annotations from better annotated genomes. However, generating the files required to perform a liftover is computationally and labour intensive and only a limited number are currently publicly available.

Here we present nf-LO (nextflow-LiftOver), a containerised and scalable Nextflow pipeline that enables liftovers within and between any species for which assemblies are available. nf-LO will consequently facilitates data interpretation across a broad range of genomic studies.

Main body

The advent of third generation sequencing and ultra-fast assemblers (Ruan & Li 2020; Joseph et al. 2018) allows for the generation of high quality *de novo* assemblies in a fraction of the previous time. As a result increasingly large numbers of new genomes for several species are being generated (Zoonomia consortium 2020).

Despite this increased availability, novel assemblies most often lack the extensive annotation data required to perform downstream analyses. Not only simple annotations such as gene models, but also supplementary resources for researcher to understand the biological significance of their studies. Unfortunately, such resources are generally only available for a small number of model organisms (OMIA; Amberger et al. 2015; Carithers & Moore 2015; Hu et al. 2019).

A solution to the problem is to lift over positions and annotations (i.e. cross-mapping of the loci) to the new genome from well-annotated assemblies, using tools such as LiftOver (Navarro Gonzalez et al. 2021) and NCBI Remap (Luu et al. 2020). However, the alignment files required to perform these analyses are not simple to generate, and are therefore limited to a few popular reference genomes. For all other pairs of genomes researchers have to generate their own liftover files. Only a few algorithms address the problem in an easy to implement and distributable way, e.g. flo for same species liftovers (Pracana et al. 2017) and LiftOff for ultra-fast liftovers (Shumate & Salzberg 2020). In this study we present nf-LO, a scalable workflow to generate liftover files for any pair of genomes based on the UCSC liftover pipeline. nf-LO can directly pull genomes from public repositories, supports parallelised alignment using a range

of alignment tools and can be finely tuned to achieve the desired sensitivity, speed of process and repeatability of analyses.

nf-LO is a workflow to facilitate the generation of genome alignment chain files compatible with the LiftOver utility. It is written in Nextflow, a domain specific language (DSL) and workflow manager, that allows easy implementation, redistribution and scalability of complex workflows across every Unix-based operating system; ranging from a desktop machine to cloud computing and HPC clusters. The dependencies are shipped alongside the workflow as docker containers or as an anaconda environment, facilitating the diffusion and adoption of the workflow across different systems.

The software accepts any two input genomes in fasta format, or alternatively can download a resource by providing a web address, an iGenome identifier or an NCBI GenBank or RefSeq accession. The workflow is shown in Figure 1, and in brief consists of three core steps, and one optional one: 1) chunking the two genomes, 2) pairwise alignment of the blocks, 3) generating the chain-net file that can be used to perform the liftover and, if a bed/gff/gtf/vcf/bam/maf file is provided, 4) performing the liftover from source to target. The chunking approach dramatically reduces the runtime of the analysis by parallelizing the alignments.

Figure 1

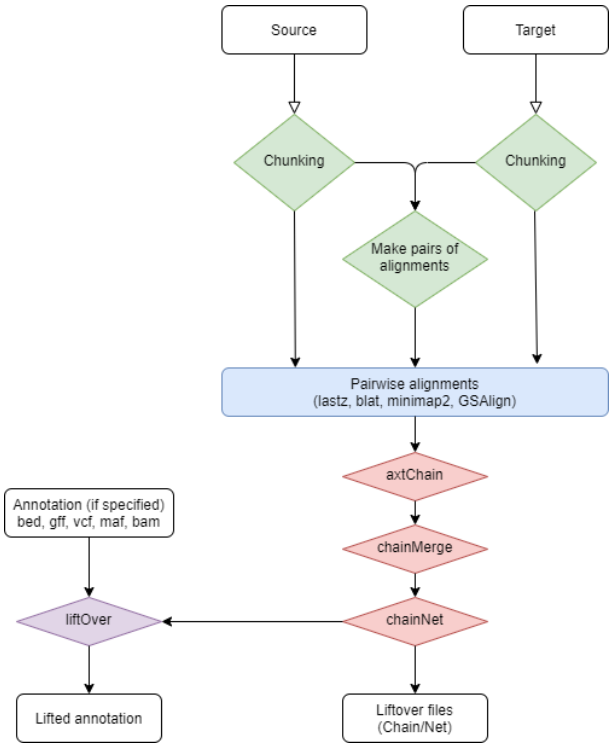

Figure 1 - Scheme of the workflow of nf-LO with the chunking (step 1, in green), alignment (step 2, in blue), generation of the liftover files (step 3, in red) and optionally lifting of the variants to the target genome (step 4, in purple).

The alignment phase can be performed in different ways, depending on the type and sensitivity required by the user. For same-species alignments, we provide native support for both blat (Kent 2002), the aligner of choice for same species liftover files from the UCSC genome browser, and GSAlign (Lin & Hsu 2020), a new, high speed same-species alignment software. For performing different-species liftovers, nf-LO also incorporates lastz (Harris 2007), used by the UCSC genome browser to generate between species liftover files, and minimap2 (Li 2018), one of the fastest genome-to-genome aligners. All these aligners are integrated within the workflow, keeping unchanged the UCSC backbone for downstream stages (UCSC 2018). We provide canned configurations for each aligner based on how distant the two genomes are (e.g. near or far), with the possibility to provide sets of custom parameters to achieve the desired balance between speed and sensitivity (Supplementary table 1). nf-LO achieves similar liftover coverage as liftover files from UCSC with appropriate tuning of the parameters (Supplementary table 2).

The third stage processes the alignments analogously to the UCSC processing pipeline, obtaining the chain-net files to perform the actual liftover. Finally, the fourth

step supports both the standard bed format with the LiftOver software, or several additional formats using CrossMap (Zhao et al. 2014), including popular formats such as VCF, BAM and GFF. Optionally, the workflow can collect metrics on the lifted annotation when provided, as well as take advantage of mafTools (Earl et al. 2014) to report metrics for the chain file generated by the workflow. These metrics are then provided in HTML format to facilitate the interpretation and collection across multiple runs.

In conclusion, we provide a transposition of the UCSC liftover pipeline within the Nextflow language, together with the necessary containers to run the analyses, allowing an easy, streamlined implementation in any Unix-based system. We believe that this workflow will be of use across genomics studies, facilitating research work and enabling data interpretation.

Code availability

The code described in the paper is publicly available on GitHub at the repository https://github.com/evotools/nf-LO. The documentation for the software can be accessed in the wiki page of the website (https://nf-lo.readthedocs.io).

Authors' contributions

AT and JP conceived the study. AT developed the software. AT and JP tested the code. AT and JP contributed to data interpretation and drafted the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

This work was supported by BBSRC grants BB/T019468/1 and BBS/E/D/10002070.

Captions

Figure 1 - Scheme of the workflow of nf-LO with the chunking (step 1, in green), alignment (step 2, in blue), generation of the liftover files (step 3, in red) and optionally lifting of the variants to the target genome (step 4, in purple).

Supplementary Table 1 – Comparison of the run times of different aligners and configurations using the human genome GRCh38 as the source and four other large genomes (>1Gbp) as targets on a Scientific Linux 6.9 system with AMD Opteron 6376 2.3GHz 64-cores and 500 GB of RAM. The genomic distances are represented as

MASH v2.2(Ondov et al. 2016) distances (-k32 -s5000) and TimeTree divergence times (http://www.timetree.org/; (Kumar et al. 2017)).

Supplementary Table 2 – Coverage for the liftover chain files both generated by us and those available from the UCSC genome database, calculated by converting the chain files to maf (chainToAxt > axtToMaf) and then using mafCoverage (Earl et al. 2014).

140 References

141 Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org:
142 Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes
143 and genetic disorders. Nucleic Acids Res. 43:D789–D798. doi:
144 10.1093/nar/gku1205.

145 Carithers LJ, Moore HM. 2015. The Genotype-Tissue Expression (GTEx) Project.
146 Biopreservation Biobanking. doi: 10.1038/ng.2653.

147 Earl D et al. 2014. Alignathon: A competitive assessment of whole-genome
148 alignment methods. Genome Res. 24:2077–2089. doi: 10.1101/gr.174920.114.

149 Harris RS. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania
150 State University.

151 Hu ZL, Park CA, Reecy JM. 2019. Building a livestock genetic and genomic
152 information knowledgebase through integrative developments of Animal QTLdb and
153 CorrDB. Nucleic Acids Res. 47:D701–D710. doi: 10.1093/nar/gky1084.

154 Joseph S et al. 2018. Chromosome level genome assembly and comparative
155 genomics between three falcon species reveals an unusual pattern of genome
156 organisation. Diversity. 10. doi: 10.3390/d10040113.

157 Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Res. 12:656–664.
158 doi: 10.1101/gr.229202.

159 Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for
160 Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34. doi:
161 10.1093/molbev/msx116.

162 Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics.
163 34:3094–3100. doi: 10.1093/bioinformatics/bty191.

164 Lin HN, Hsu WL. 2020. GSAlign: An efficient sequence alignment tool for intra-
165 species genomes. BMC Genomics. 21. doi: 10.1186/s12864-020-6569-1.

166 Luu P-L, Ong P-T, Dinh T-P, Clark SJ. 2020. Benchmark study comparing liftover
167 tools for genome conversion of epigenome sequencing data. NAR Genomics
168 Bioinforma. 2. doi: 10.1093/nargab/lqaa054.

169 Navarro Gonzalez J et al. 2021. The UCSC genome browser database: 2021
170 update. Nucleic Acids Res. 49. doi: 10.1093/nar/gkaa1070.

171 OMIA. Online Mendelian Inheritance in Animals. Syd. Sch. Vet. Sci. https://omia.org/
172 (Accessed June 10, 2020).

173 Ondov BD et al. 2016. Mash: Fast genome and metagenome distance estimation
174 using MinHash. Genome Biol. 17. doi: 10.1186/s13059-016-0997-x.

175 Pracana R, Priyam A, Levantis I, Nichols RA, Wurm Y. 2017. The fire ant social
176 chromosome supergene variant Sb shows low diversity but high divergence from SB.
177 Mol. Ecol. 26:2864–2879. doi: 10.1111/mec.14054.

178 Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat.
179 Methods. 17:155–158. doi: 10.1038/s41592-019-0669-3.

180 Shumate A, Salzberg SL. 2020. Liftoff: accurate mapping of gene annotations.
181 Bioinformatics. doi: 10.1093/bioinformatics/btaa1016.

182 UCSC. 2018. Minimal steps for liftover.
183 http://genomewiki.ucsc.edu/index.php/Minimal_Steps_For_LiftOver (Accessed June
184 10, 2020).

185 Zhao H et al. 2014. CrossMap: A versatile tool for coordinate conversion between
186 genome assemblies. Bioinformatics. 30. doi: 10.1093/bioinformatics/btt730.

187 Zoonomia consortium. 2020. A comparative genomics multitool for scientific
188 discovery and conservation. Nature. 587:240–245. doi: 10.1038/s41586-020-2876-6.

189