nf-LO: A Scalable, Containerized Workflow for Genome-to-Genome Lift Over

Andrea Talenti * and James Prendergast

The Roslin Institute, University of Edinburgh, Midlothian, United Kingdom

*Corresponding author: E-mail: andrea.talenti@ed.ac.uk.

Accepted: 3 August 2021

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 lift over is computationally and labor intensive and only a limited number are currently publicly available.

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

Key words: liftover, assembly, Nextflow, workflow.

Significance

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 article, we present nf-LO (nextflow-LiftOver), a streamlined, containerized 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.

The advent of third-generation sequencing and ultrafast assemblers (Joseph et al. 2018; Ruan and Li 2020) 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 (Carithers et al. 2013; Amberger et al. 2015; Hu et al. 2019; OMIA 2020).

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 lift over files. Only a few algorithms address the problem in an easy to implement and distributable way, for example, flo for same species lift over (Pracana et al. 2017) and LiftOff for ultrafast lift over (Shumate and Salzberg 2021). In this study, we present nf-LO, a scalable workflow to generate lift over files for any pair of genomes based on the UCSC LiftOver pipeline. nf-

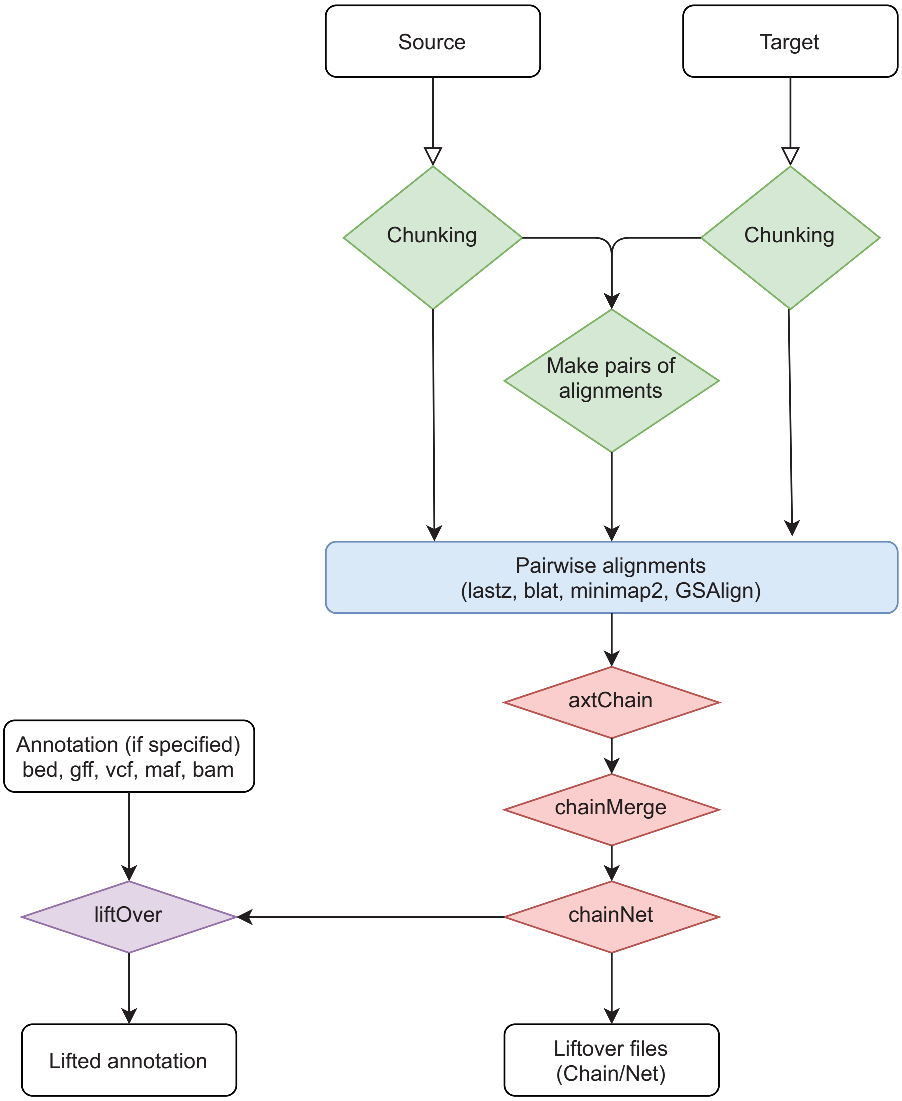

FIG. 1.—Scheme of the workflow of nf-LO with the chunking (step 1, in green), alignment (step 2, in blue), generation of the liftover files (step 3, in red), and optionally lifting of the variants to the target genome (step 4, in purple).

LO can directly pull genomes from public repositories, supports parallelized 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 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 lift over and, if a bed/gff/gtf/vcf/bam/maf file is provided, 4) performing the lift over from source to target. The chunking approach dramatically reduces the runtime of the analysis by parallelizing the alignments.

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 lift over files from the UCSC genome browser, and GSAlign (Lin and Hsu 2020), a new, high speed same-species alignment software. For performing different-species lift overs, 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, Supplementary Material online). nf-LO achieves similar lift over coverage as LiftOver files from UCSC with appropriate tuning of the parameters (supplementary table 2, Supplementary Material online).

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

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgment

This work was supported by BBSRC (Grant Nos. BB/T019468/1 and BBS/E/D/10002070).

Author Contributions

A.T. and J.P. conceived the study. A.T. developed the software. A.T. and J.P. tested the code. A.T. and J.P. contributed to data interpretation and drafted the manuscript. All authors reviewed and approved the final manuscript.

Data Availability

The code described in the article is publicly available on GitHub at the repository https://github.com/evotools/nf-LO, last accessed August 18, 2021. The documentation for the software can be accessed in the wiki page of the website (https://nf-lo.readthedocs.io, last accessed August 18, 2021).

Literature Cited

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 43(Database issue):D789–D798.

Carithers LJ, et al. 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45(6):580–585.

Earl D, et al. 2014. Alignathon: a competitive assessment of whole-genome alignment methods. Genome Res. 24(12):2077–2089.

Harris RS. 2007. Improved pairwise alignment of genomic DNA [PhD thesis]. University Park, PA: College of Engineering, Pennsylvania State University. Available from: https://www.proquest.com/docview/304835295?pq-origsite=gscholar&fromopenview=true. Accessed August 17, 2021.

Hu ZL, Park CA, Reecy JM. 2019. Building a livestock genetic and genomic information knowledgebase through integrative developments of animal QTLdb and CorrDB. Nucleic Acids Res. 47(D1):D701–D710.

Joseph S, et al. 2018. Chromosome level genome assembly and comparative genomics between three falcon species reveals an unusual pattern of genome organisation. Diversity 10(4):113.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12(4):656–664.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 34(7):1812–1819.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100.

Lin HN, Hsu WL. 2020. GSAlign: an efficient sequence alignment tool for intra-species genomes. BMC Genomics. 21(1):182.

Luu P-L, Ong P-T, Dinh T-P, Clark SJ. 2020. Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. NAR Genom Bioinform. 2(3):lqaa054.[33575605]

Navarro Gonzalez J, et al. 2021. The UCSC genome browser database: 2021 update. Nucleic Acids Res. 49(D1):D1046–D1057.

OMIA. 2020, June 10. Online Mendelian Inheritance in Animals. Australia: Sydney School of Veterinary Science, University of Sydney. Available from: https://omia.org/. Accessed August 18, 2021.

Ondov BD, et al. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17(1):132.

Pracana R, Priyam A, Levantis I, Nichols RA, Wurm Y. 2017. The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. Mol Ecol. 26(11):2864–2879.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 17(2):155–158.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. Bioinformatics 37(12):1639–1643.

UCSC. 2018. Minimal steps for liftover. Available from: http://genomewiki.ucsc.edu/index.php/Minimal_Steps_For_LiftOver (accessed June 10, 2020).

Zhao H, et al. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics 30(7):1006–1007.

Zoonomia Consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. Nature 587(7833):240–245.

Associate editor: Bonnie Fraser