segment_liftover :

a Python tool to convert segments between genome assemblies

Introduction

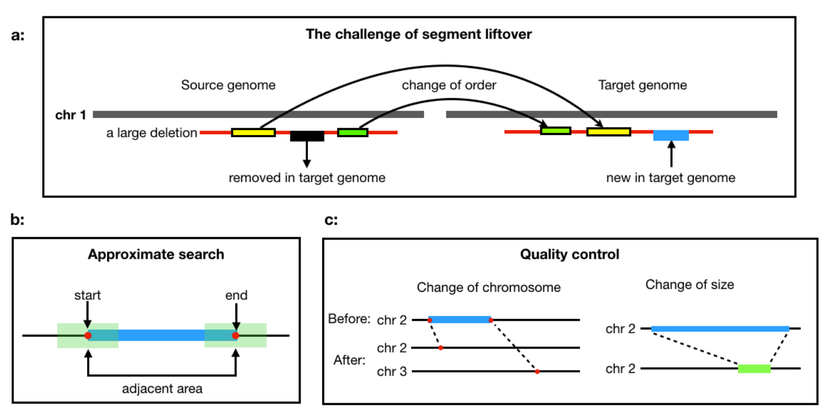
The process of assembling a species' reference genome may be performed in a number of iterations, with subsequent genome assemblies differing in the coordinates of mapped elements. The conversion of genome coordinates between different assemblies is required for many integrative and comparative studies. While currently a number of bioinformatics tools are available to accomplish this task, most of them are tailored towards the conversion of single genome coordinates. When converting the boundary positions of segments spanning larger genome regions, segments may be mapped into smaller sub-segments if the original segment's continuity is disrupted in the target assembly. Such a conversion may lead to a relevant degree of data loss in some circumstances such as copy number variation (CNV) analysis, where the quantitative representation of a genomic region takes precedence over base-specific accuracy. segment_liftover aims at continuity-preserving remapping of genome segments between assemblies and provides features such as approximate locus conversion, automated batch processing and comprehensive logging to facilitate processing of datasets containing large numbers of structural genome variation data.

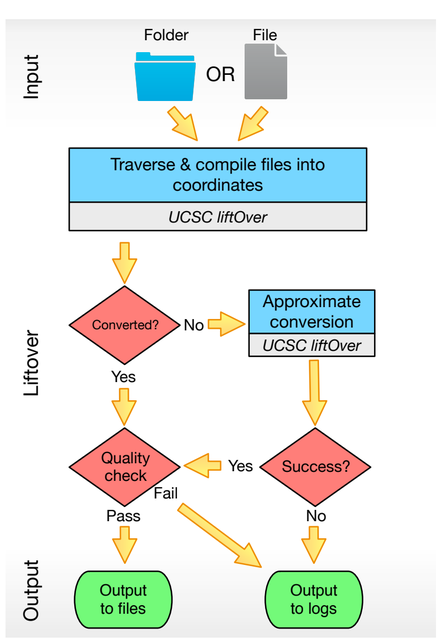
The challenge

(a) When lifting a segment to another assembly, the landscape of the segment may be affected by indels and copy number variations, but the overall span of the segment does not change significantly.

(b) When the end positions cannot be converted by the UCSC liftOver, the nearby regions will be searched for convertible positions as approximation.

(c) Quality control checks for changes of chromosome or size to make sure the segment is converted properly.





The workflow

(1) It can take either a folder or a file containing the list of files as the input.

(2) It will try to convert by approximation when UCSC liftOver fails to convert a coordinate.

(3) The directory structure will be kept in the output folder and detailed log files are also available.

Availability

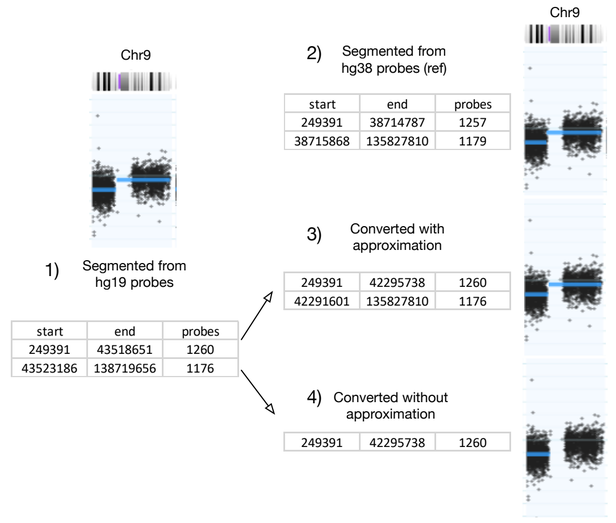
1. pip version: https://pypi.python.org/pypi/segment-liftover
2. github: https://github.com/baudisgroup/segment-liftover
3. Software license: MIT

Use cases

We provide two examples of using segment_liftover to convert probes and segments, respectively. The two examples are part of the pipeline which updates the arrayMap database, a reference resource of somatic genome copy number variations in cancer, from human genome assembly hg19 to hg38. In the first example, we converted 44,632 probe files and 44,471 segment files. In the second example, we compared the performance of different conversion strategies using 1,000 samples randomly drawn from the first example.

Number of samples from nine platforms in use case examples

	all	1000 samples
CytoScanHD_Array	2963	73
CytoScan750K_Array	173	2
Mapping50K_Hind240	2699	58
Mapping50K_Xba240	3303	72
Mapping10K_Xba142	912	19
Mapping250K_Nsp	9738	223
Mapping250K_Sty	7561	184
GenomeWideSNP_6	16570	359
GenomeWideSNP_5	552	10



Number of segments with or without approximation on average

	perfect	minor difference	significant difference	sum
reference hg19	*na*	*na*	*na*	218.42
reference hg38	*na*	*na*	*na*	215.04
approximation	198.18	6.24	10.25	214.67
no approximation	198.18	5.23	10.01	213.42

A demonstration of the effectiveness of approximate conversion

Summary

With the functionalities of automated batching, approximate conversion and segment conversion, segment_liftover can dramatically reduce the complexity and workload of such data processing. Furthermore, segment_liftover's detailed logs of execution result provide an easy and clear foundation for follow up analysis.

Contact:

Bo Gao, Qingyao Huang, Michael Baudis

Institute of Molecular Life Sciences,
Swiss Institute of Bioinformatics,
University of Zürich, Switzerland
Email: bo.gao@imls.uzh.ch