Pseudo-chromosome assembly of large and complex genomes using multiple references

Mikhail Kolmogorov[1], Brian Raney [2], Joel Armstrong[2], Duncan Odom[3,4], Paul Flicek[5], David Thybert [5,6], Benedict Paten[2] and Son Pham[1]

[1]University of California San Diego, USA [2]University of California Santa Cruz, USA, [3]University of Cambridge, UK,
[4]Wellcome Trust Sanger Institute, UK, [5]European Bioinformatic Institute, UK, [6]The Genome Analysis Center, UK

Abstract

• Assembly of mammalian-scale genomes into complete chromosomes is challenging

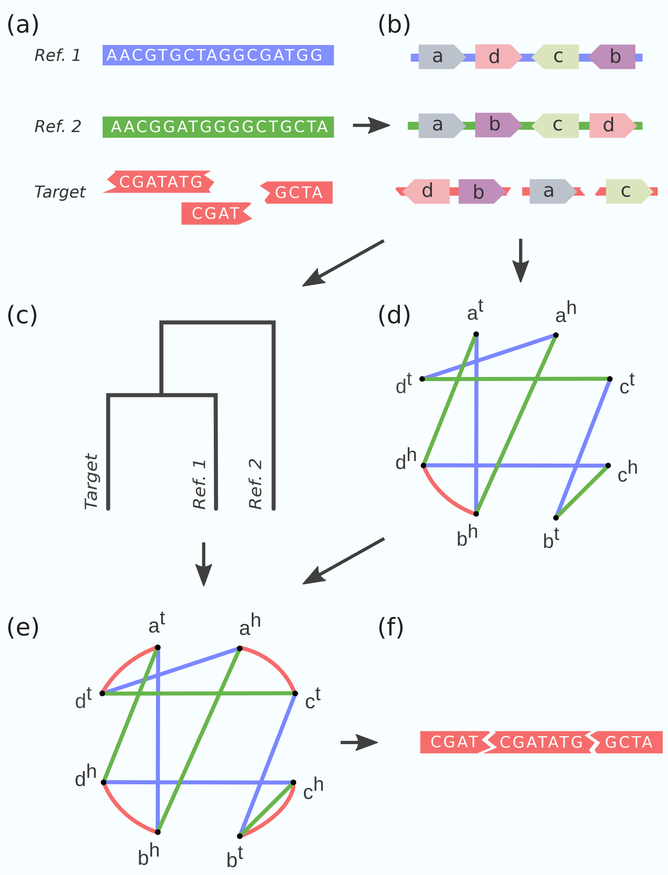
• To address this, we developed Ragout, a reference-assisted assembly tool for large and complex genomes

• Taking as input an NGS assembly and multiple related references, Ragout infers their evolutionary relationship and builds the final assembly of the target genome using a genome rearrangement approach

• Using Ragout we assembled two mice genomes (*M. Caroli* and *M. Pahari*) with comlicated chromosome-scale rearrangements into sets of high-quality pseudo-chromosomes

• Chromosome coloring comfirms most the rearrangements that Ragout has detected

Algorithm Overview



(a) Raw references and target genome sequences

(b) Nucleotide sequences are decomposed into the permutations of synteny blocks

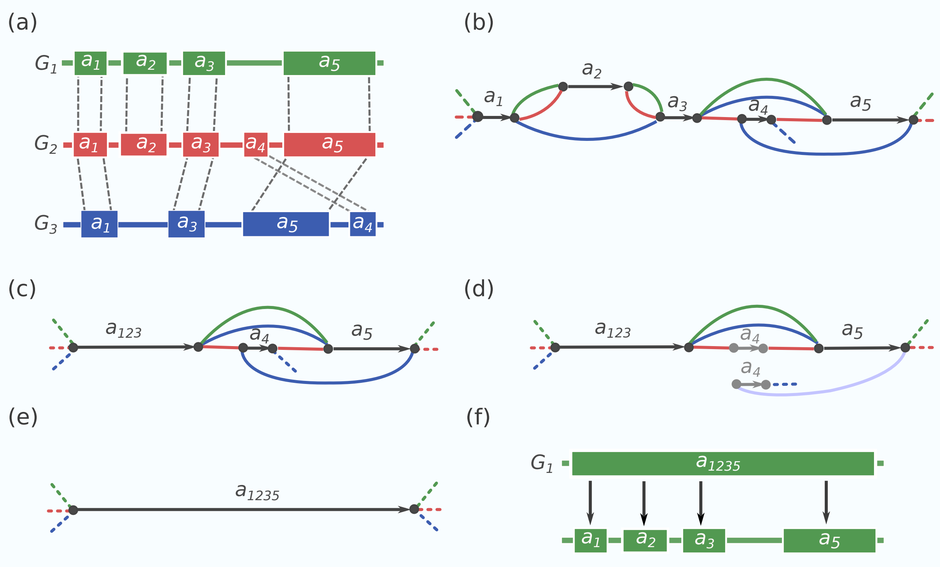
(c) Phylogenetic tree of the input genomes is reconstructed based on the breakpoints data

(d) Incomplete breakpoint graph reflects adjacencies between synteny blocks

(e) Missing adjacencies are reconstructed by analysing rearrangements

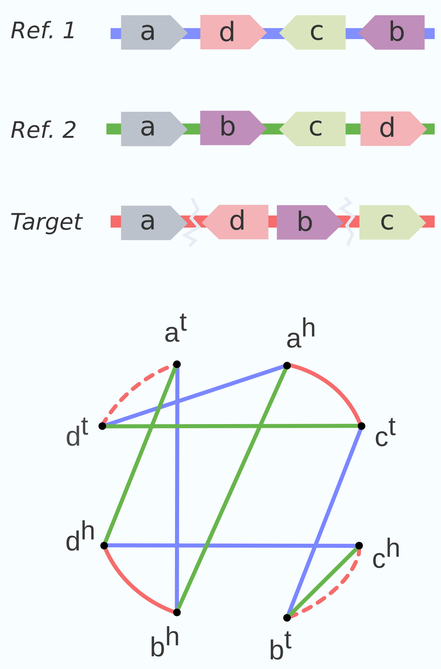
(f) Target fragments are joined into scaffolds with respect to the inferred adjacencies

Synteny Blocks



Synteny blocks help to separate small sequence variations from large-scale rearrangements. (a) An alignment between three genomes with complicated sub-structure. (b) A-Bruijn graph representation of the alignment. Small sequence variations correspond to bubbles, while rearrangements form more complicated structures. (c) A bubble is removed during the graph simplification, forming a larger synteny block a_{123}(d) Masking smaller block a_4 allows to make the graph structure even simplier. (e) After removing another bubble, the whole alignment is represented as a large synteny block. (d) A hierarchical representation of a synteny block.

Incomplete Breakpoint Graph Analysis



• Breakpoint graphs reflect adjacencies between synteny blocks in different genomes

• If all genomes were complete, the edges of each color will define a perfect matching on the graph

• As the target genome is fragmented, some adjacencies of red color are missing

• Ragout recovers the missing adjacencies so as to minimize the weighted number of rearrangements between the genomes

• These adjacencies are then used to merge the target fragments into scaffolds

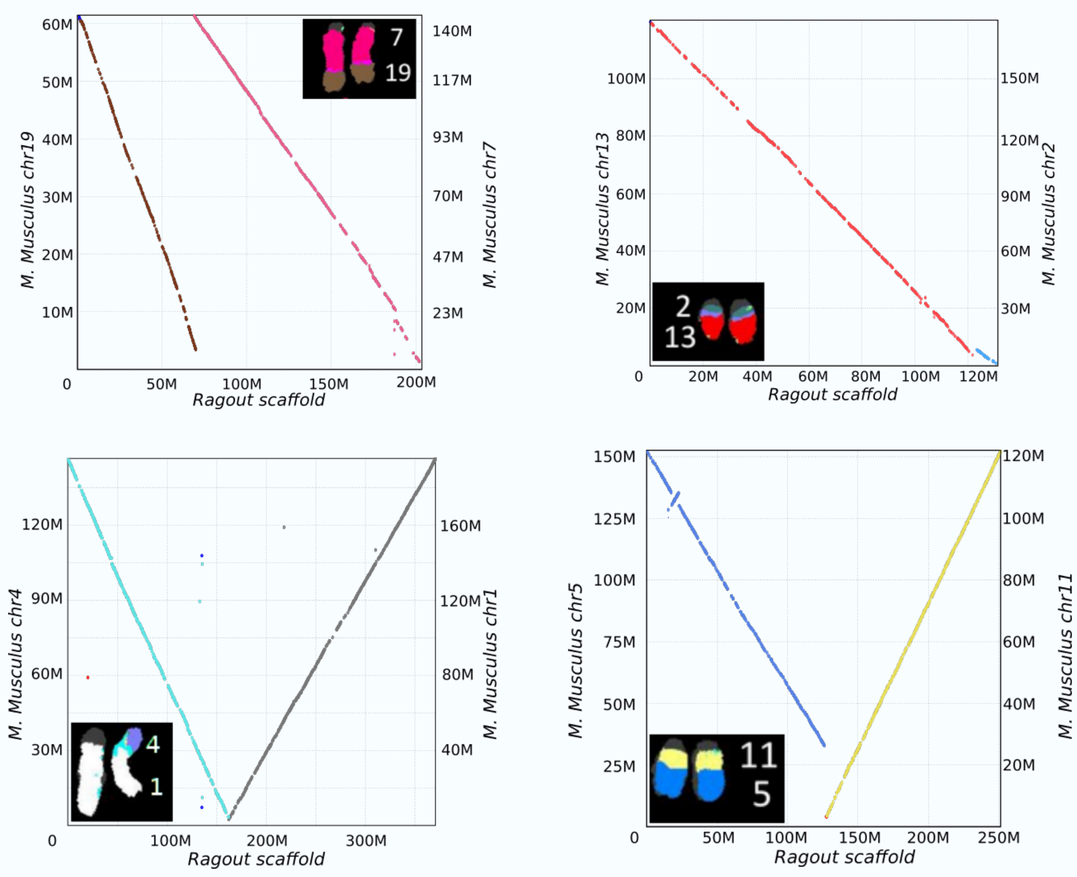
Results

• We assembled two genomes from Murinae family: *Mus. Caroli* and *Mus. Pahari* using *Mus. Musculus* and *Rattus Norvegicus* as references. The assemblies contains 20 and 23 pseudo-chromosomes, respectively with at most 2% of unlocalized sequence.

• *M. Caroli* shows 5% sequence diversity from *Mus. Musculus* and has the same karyotype (which was confirmed by Ragout). However, we have detected a large inversion in chr17.

• *M. Pahari* has 10% sequence diversity and contain many inter-chromosomal rearrangements. Ragout has detected four of them, which were also confirmed by chromosome coloring.

• Some of the rearrangements remain undetected, as the corresponding breakpoints are missing from the NGS data. However, they could be recovered using the aid of chromosomal maps.



Availability & Contacts

• Ragout is an easy to use package, written in Python/C+. It is freely available at http://fenderglass.github.io/Ragout

• Email: fenderglass@gmail.com