Title: MTG-Link: filling gaps in draft genome assemblies with linked read data
Presenter: Anne Guichard

Affiliation: IGEPP, INRAE, Institut Agro, Univ Rennes, 35653, Le Rheu, France

Email: anne.guichard@inrae.fr

Author names: Anne Guichard[1,2], Fabrice Legeai[1,2], Arthur Le Bars[2], Paul Yann Jay[3], Mathieu Joron[3], DenisTagu[1] and Claire Lemaitre[2]
[1] INRAE, Agrocampus Ouest, Univ Rennes, IGEPP, F-35650 Le Rheu, France
[2] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[3] CEFE, CNRS, Univ Montpellier, Univ Paul Valéry Montpellier 3, EPHE, IRD, F-34293 Montpellier, France

Abstract (250 words)

The complete and accurate reconstruction of large genomes remains challenging. The scaffolding step orders and orients contigs but generates undefined sequences, called gaps. Linked read technologies, such as the 10X Genomics Chromium platform, have a great potential for filling the gaps; they provide long-range information while maintaining the power and accuracy of short-read sequencing. Thus, reads that have been sequenced from the same long DNA molecule (30-50 Kb) can be identified by a small barcode sequence. Several tools have been developed for gap-filling, but none uses the long-range information of the linked read data. Here, we present MTG-Link, a novel gap-filling tool dedicated to linked read data generated by 10X Genomics. MTG-Link is a Python pipeline combining the local assembly tool MindTheGap and an efficient read subsampling based on the barcode information. For each gap, it extracts the linked reads whose barcode is observed in the gap flanking sequences, and assembles them into contigs by traversing their de Bruijn graph. MTG-Link tests different parameters values for gap-filling, followed by an automatic qualitative evaluation of the assembly. It returns a GFA file, containing the gap-filled sequences of each gap. Validation was performed on a set of simulated gaps from real datasets with various genome complexities ; it showed that the read subsampling step of MTG-Link enables to get better genome assemblies than using MindTheGap. We applied MTG-Link on individual genomes of a mimetic butterfly (H. numata); it significantly improved the contiguity of a 1.3 Mb locus of biological interest (https://github.com/anne-gcd/MTG-Link).