Identification and correction of genome mis-assemblies due to heterozygosity

Anaïs Gouin[1], Anthony Bretaudeau[2], Claire Lemaitre[1] and Fabrice Legeai[1,2]

[1] Inria/IRISA GenScale, Campus de Beaulieu, 35042 Rennes cedex, France
[2] INRA, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Domaine de la Motte – 35653 Le Rheu, France

Assembly tools are more and more efficient to reconstruct a genome from next-generation sequencing data but some problems remain. One of them corresponds to mis- assemblies due to heterozygosity. Indeed, the assembly of an heterozygous region for which there is a significant divergence between the two haplotypes, could lead to the construction of two different contigs, instead of one consensus sequence. This problem causes an assembly of an heterozygous genome larger than expected, and also a loss of information (heterozygous SNPs or indels cannot be found in the erroneous regions). We propose a strategy to detect and correct false duplications in assemblies based on several metrics.

We identified two specific cases highlighting problems of heterozygosity. The first case involves scaffolds that are completely matching on another one. The second case corresponds to scaffolds matching together by their extremities. The two sequences involved in the match may actually correspond to two distinct alleles of a specific locus instead of two different locations in the genome. Ideally, an erroneous duplication would involve two divergent but similar assembly parts, not containing any heterozygous polymorphisms, and for which the merge of the two would lead to the expected read coverage for the resulting consensus sequence. As a consequence, to distinguish between true genomic duplications and alleles, we used various metrics : sequence similarity, length of the match, average read coverage, presence/absence of SNPs in the two concerned regions, number of mate pairs with expected (or not) insert size... As a result, selected allelic regions are used to construct a single sequence by removal of one of the two alleles or joining of scaffolds by their extremities. This allows to decrease redundancy in the genome assembly, to improve the scaffolding and then to increase the N50 statistic. We applied this method to a 526Mb highly heterozygous wild type insect genome assembly for which we expected a genome size around 400Mb only. A set of user-validated false duplications in this assembly enabled us to validate the method and to fit the set of criteria, in order to distinguish between true and artefactual duplications. We took advantage of this study to compare classical assemblers (Minia, Soap) with more recent tools that handle heterozygosity, such as Platanus. This highlighted the advantages of such new assemblers for diploid genomes. However, for already-built assemblies, we showed that our approach is a fast and easy way to discard as much as possible erroneous duplications, allowing their correction without resorting to a complete new assembly that would be more time-consuming.