Redundans: an assembly pipeline for highly heterozygous genomes

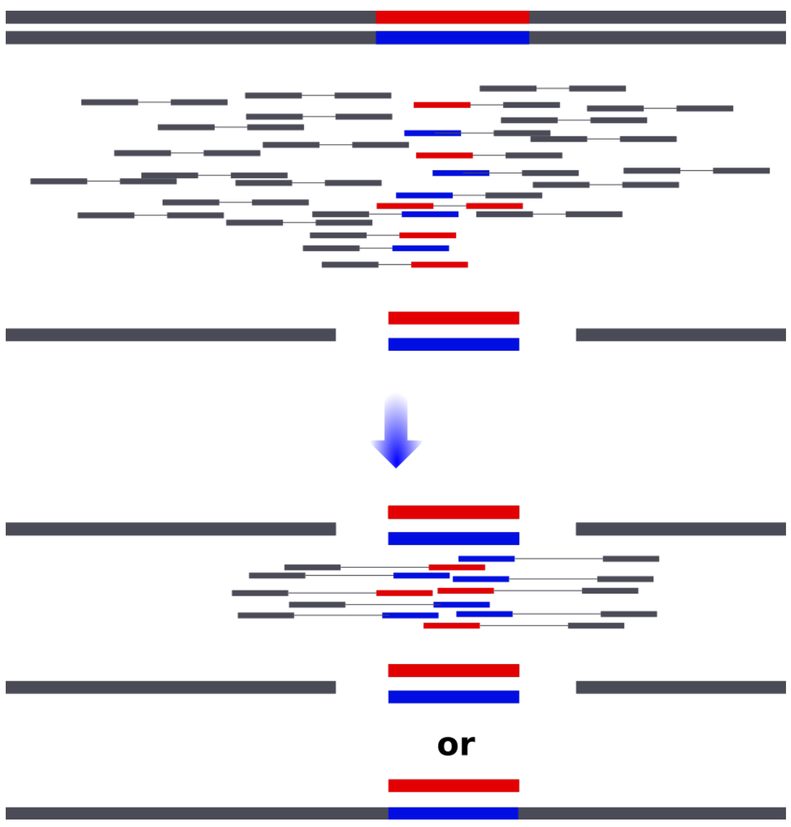
Leszek P. Pryszcz[1,2] and Toni Gabaldón[1,3]

l.p.pryszcz@gmail.com

1) Centre for Genomic Regulation (CRG) and UPF, Dr. Aiguader 88, 08003 Barcelona, Spain
2) International Institute of Molecular and Cellular Biology (IIMCB), Ks. Trojdena 4, 02-109 Warsaw, Poland
3) Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain

Heterozygous genome assembly

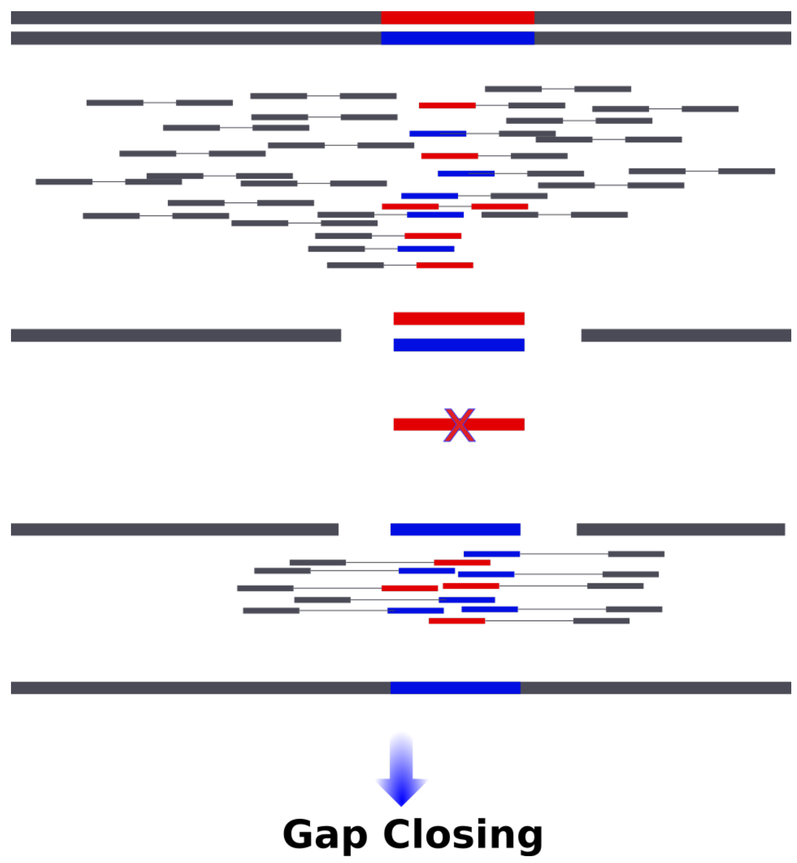
The assembly of highly heterozygous genomes from short sequencing reads is a **challenging task** because it is difficult to accurately recover the different haplotypes.



Standard assembly process tends to **collapse homozygous regions** and reports **heterozygous regions in alternative contigs**.

The boundaries between homozygous and heterozygous regions result in multiple paths that are hard to resolve, which leads to **highly fragmented** assemblies with a total size **larger than expected**.

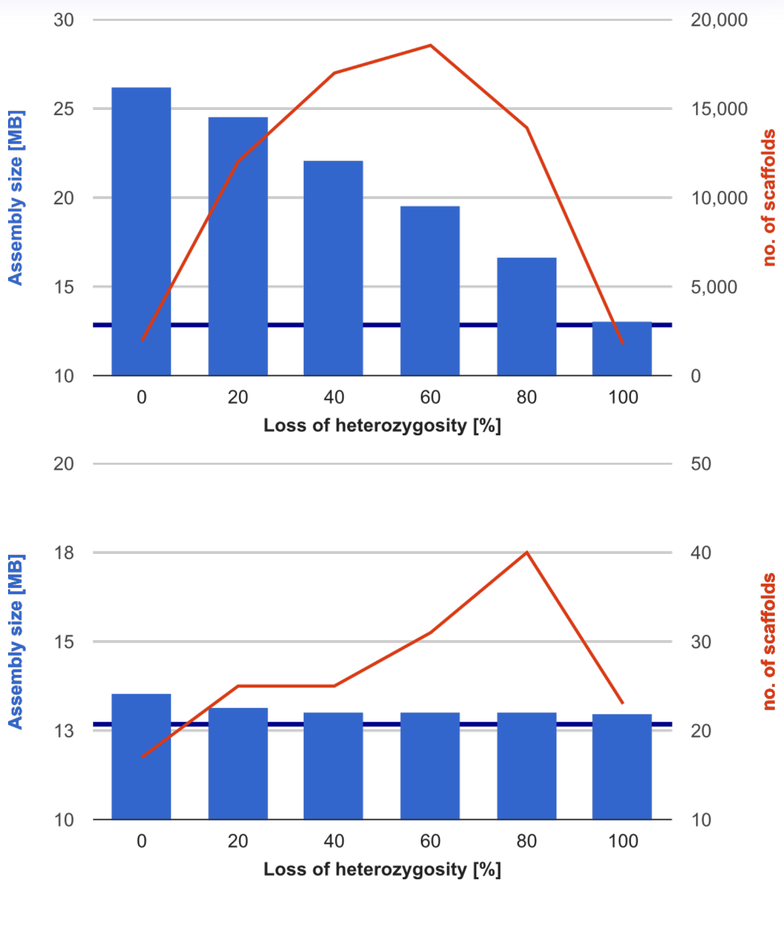
We have developed a **pipeline** that specifically deals with the assembly of heterozygous genomes by introducing a step to **recognise** and **selectively remove** alternative heterozygous contigs.



Simulations

We simulated **six diploid genomes** in which the two haploid sequences had **5% sequence divergence** and with **varying levels of loss of heterozygosity (LOH)**.
Next, we simulated **short reads** from these genomes, **paired-end** and **mate pairs**, which included typical Illumina-related errors.
Finally, we **assembled** these genomes from the simulated reads.

Standard *de novo* genome assemblers, SPAdes & SOAPdenovo, returned **very fragmented** assemblies with **total size larger than expected**.

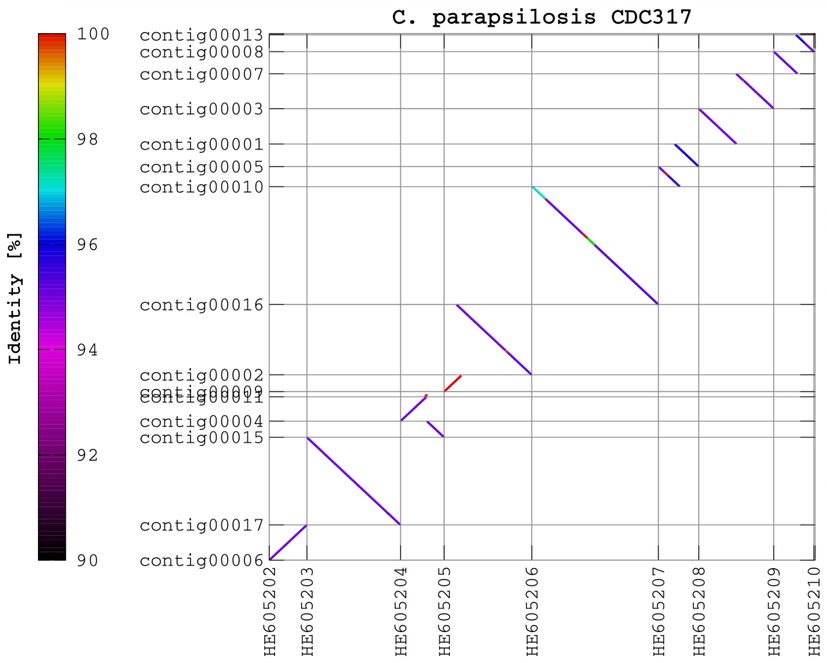


Redundans recovered assemblies with **less than 100 contigs** and with **size close to expected** for all simulated genomes.

Performance & accuracy

Reduction step removes heterozygous regions **reducing the assembly size** and fragmentation. Scaffolding further **reduce fragmentation** of assembly.

Assembled contigs were aligned back onto *Candida parapsilosis* chromosomes to **evaluate the correctness** of each assembly.



Heterozygous genome assembly pipeline, that started from thousands of contigs/scaffolds, returned **full size chromosomes**.

No large inversions and **deletions** were observed.
We have identified a few **translocations**, most of which were tracked back to initial assembly.

https://github.com/lpryszcz/redundans

Conclusions

Redundans **reduces the heterozygous regions** with substantial divergence.
It deals well with **various levels of loss of heterozygosity**.
Redundans allows further scaffolding, resulting in **full size chromosomes**.
Redundans is **superior** to existing tools, while uses fewer resources.

Future work

Imporve **sensitivity** of heterozygous contigs detection.
Recognition of **structural variants**.
Testing on larger, **polyploid** genomes.