Redundans: an assembly pipeline for highly heterozygous genomes

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

[1] Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG).
Dr. Aiguader, 88. 08003 Barcelona, Spain
[2] Universitat Pompeu Fabra (UPF). 08003 Barcelona, Spain
[3] Instituci Catalana de Recerca i Estudis Avanats (ICREA),
Pg. Llus Companys 23, 08010 Barcelona, Spain
tgabaldon@crg.es
`http://www.crg.eu/en/toni_gabaldon`

Abstract. Many genomes display high levels of heterozygosity (i.e. presence of different alleles at the same loci in homologous chromosomes), being those of hybrid organisms an extreme such case. The assembly of highly heterozygous genomes from short sequencing reads is a challenging task because it is difficult to accurately recover the different haplotypes. When confronted with highly heterozygous genomes, the 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. This, in turn, causes numerous problems in downstream analyses i.e. fragmented gene models, wrong gene copy number, broken synteny. To circumvent these caveats 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. We tested our pipeline on simulated and naturally-occurring heterozygous genomes and compared its accuracy to other existing tools.

Keywords: heterozygous, genome, assembly, hybrid, polymorphism, scaffolding

1 Introduction

The assembly of genomes from short sequencing reads is a complex computational problem. Numerous genome assemblers have been developed to address this task (Bankevich et al., 2012; Luo et al., 2012; Simpson et al., 2009; Zerbino et al., 2009). Typically, when there is some heterogeneity in the sequence (e.g. non haploid organisms, population of cells, etc), a single reference sequence is recovered. In the particular case of non-haploid organisms that are highly polymorphic, the standard genome assemblers produce fragmented assemblies with a total size larger than expected (Pryszcz et al., 2014; Small et al., 2007). This is because short reads are generally not sufficient to accurately recover the different haplotypes in heterozygous regions, which are reported as alternative contigs. In contrast homozygous (or low heterozygosity) regions from the two homeologous chromosomes are collapsed into a single contig. The boundaries between these two types of contigs cannot be resolved by a unique path and, therefore left unlinked. The final result is typically an assembly that is highly fragmented and contains redundant contigs (i.e. same region in homeologous chromosomes). Such assemblies mislead downstream analyses, from gene prediction (i.e. fragmented gene models, apparent paralogs) to comparative genome analysis (i.e. apparent duplicated blocks, synteny breaks).

Because heterozygous contigs represent the sequence of each haploid genome and homozygous contigs represent a consensus between two or more haploid genomes, these two categories of contigs can be recognised by differences in their depth-of-coverage (i.e. the number of sequencing reads that align to a given position). That is, when the reads are aligned back to the assembly, the consensus, homozygous contigs will have a higher number of reads aligned per a given length interval than haploid, heterozygous contigs (roughly double, for diploid organisms). We took advantage of this fact to design a novel assembly strategy that is able to cope with highly heterozygous genomes. In brief our approach consists of three main steps: i) detection and selectively removal of redundant contigs from an initial standard assembly, ii) scaffolding of such non-redundant assembly using paired-end, mate-pair and/or fosmid-based reads, and iii) gap closing. Our strategy (and pipeline) is flexible and can be implemented on top of several software tools for the assembly, mapping, scaffolding, and gap closing steps. We have applied our methodology to both, real and simulated data sets, in order to evaluate its efficacy and accuracy.

2 Results and discussion

2.1 Rationale and design

In the course of our past and ongoing research in genomics we have often encountered difficulties in producing high quality assemblies for highly heterozygous genomes. This problem is shared by many other colleagues, given the abundance in nature of highly heterozygous species, including hybrid species. For instance, in fungi, the number of reported hybrids has increased in the last years, of which many have been discovered in the process of genome sequencing (Morales and Dujon, 2012). The assemblies of genomes from these highly heterozygous species are highly fragmented, which complicates downstream analyses. For instance the genome assemblies of recognized hybrids such as *Dekkera bruxellensis* LAMAP2480 (AZMW01) or *Wickerhamomyces anomalus* NRRL Y-366 (AEGI01), are highly fragmented (9,167 contigs, and 3,133 scaffolds, respectively) and larger (26.9 Mb and 26.2 Mb, respectively) than those of closely related homozygous species i.e. *D. bruxellensis* AWRI1499 is 12.6 Mb in 324 contigs (AHIQ01) and *Wickerhamomyces ciferrii* is 15.9 Mb in 364 contigs (Supplementary table S1). In the framework of the sequencing project of a hybrid strain from the emerging pathogen *Candida orthopsilosis* MCO456 (Pryszcz et al., 2014), we obtained similar low quality initial assemblies. To solve this we devised an ad-hoc strategy to recognize and selectively remove one of the two haploid contigs from heterozygous regions. Subsequent scaffolding and gap closing steps yielded a high quality genome reducing from 5,577 to 116 contigs and 14.4 Mb size to 13.2 Mb.

Here we describe the procedure in more detail and present a programmatic environment to facilitate its application. Although we have chosen a specific set of available tools to perform each step of the pipeline, it must be noted that the strategy itself is flexible and modular and can be implemented on top of different tools. In brief (see Materials and Methods for more details), our pipeline is similar to the standard *de novo* assembly methodology (Figure 1 A): overlapping reads are assembled into contigs (a), contigs are subsequently joined into supercontigs using information from paired-end reads (d) and finally the remaining gaps are closed again utilising paired-end reads (e). We recognised that the standard *de novo* assembly tools fail at the scaffolding step, as in the case of heterozygous genomes there are multiple redundant contigs (marked in colours) that could be connected to any of the homozygous neighbours (see b in Figure 1 B).

Our pipeline proceeds in three distinct steps (Figure 1 B). Firstly, the draft assembly is simplified by removing heterozygous contigs (c). These redundant contigs represent distinct haplotypes from polymorphic chromosomal regions. To circumvent this, the clusters of redundant contigs are recognised and only the longest contig from each cluster is kept. Such reduction of complexity allows for further scaffolding. This is conducted by our in-house solution, *fasta2homozygous.py* v1.0. In the second step, non redundant contigs are joined using SSPACE2 (Boetzer et al., 2011) (d). Finally, the gaps in the scaffolds are closed using GapCloser (Luo et al., 2012). Noteworthy, scaffolding and gap closing are iteratively repeated in order to improve scaffolding with another sequencing library or reduce the number of gaps. We next assessed the accuracy of our pipeline by assembling simulated and naturally-occurring heterozygous genomes.

2.2 Performance on simulated heterozygous genomes

The underlying difficulty of heterozygous genome reconstruction is the lack of a “golden” reference that would allow the identification of possible pitfalls of the genome reconstruction process. To circumvent this, we simulated six diploid genomes in which the two haploid sequences had 5% sequence divergence and with varying levels of loss of heterozygosity (LOH). LOH is a recombination event that renders homozygous regions with the sequence of only one of the two haplotypes (Bennett et al., 2014). Subsequently, we simulated short reads from these genomes, which included typical Illumina-related errors (see Materials and Methods). Then we assembled these genomes from the simulated short reads with either an standard pipeline and the Redundans pipeline. As expected, standard approaches (SPAdes and SOAPdenovo) obtained very fragmented genome assemblies (2,237-3,743 scaffolds) with increased size (119-198% of the original

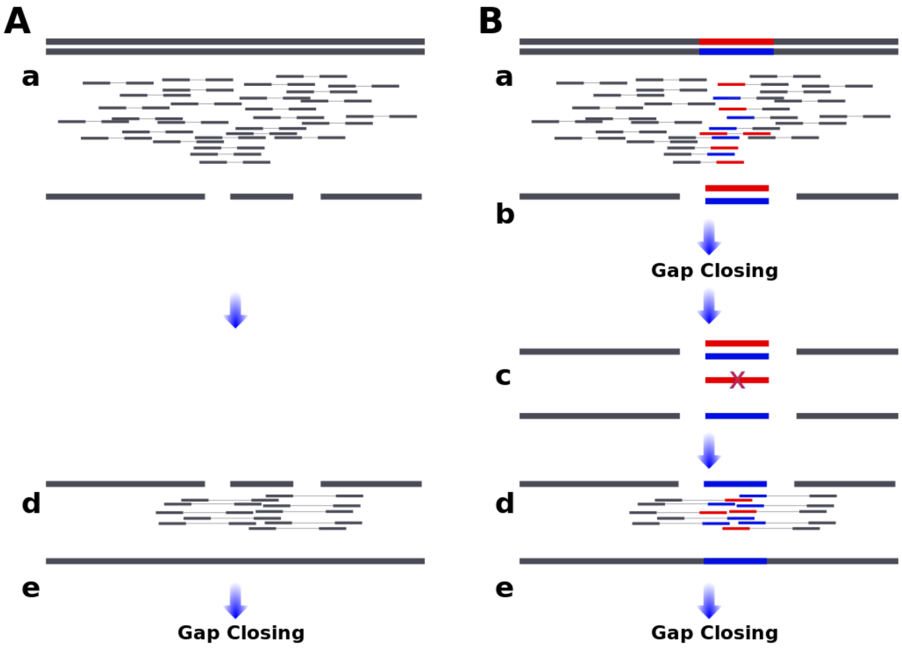

Fig. 1. Genome assembly from short reads. Standard (A) and heterozygous (B) genome assembly pipelines are compared. Heterozygous regions in diploid chromosome are marked in red and blue. Heterozygous genome assembly pipeline consists of five steps. a) Standard de novo assembly is performed and b) optionally gaps are closed. Obtained assembly is larger than expected and fragmented because two alternative contigs are recovered from heterozygous region (blue and red), while single contig is recovered from homozygous regions (grey). Further scaffolding of such assembly is impossible, as homozygous contigs can be joined to any of heterozygous contigs (blue and red). c) To overcome this, redundant contigs from heterozygous regions are removed (here the red contig) and d) reduced assembly is further scaffolded. e) Finally, gaps are closed.

genome) for five heterozygous genomes (Figure 2). In contrast, a fully homozygous genome (LOH of 100%) was recovered in 250 scaffolds with roughly the expected assembly size (99% of the original assembly). Interestingly, the size of the genome assembly was negatively correlated to the LOH level (Pearson r=-0.9939) (Supplementary table S2).

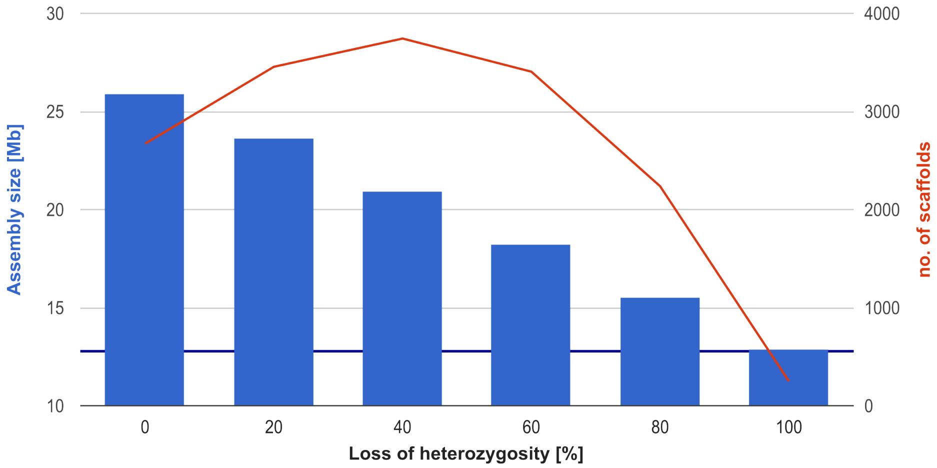

Fig. 2. Heterozygous genome assemblies characteristics. The total SOAPdenovo2 assembly size (blue bars), as well as number of scaffolds longer than 1 kb (red plot) are given for one homozygous (LOH of 100%) and five heterozygous genome assemblies with 5% divergence between haplomes and varying loss of heterozygosity level: 0%, 20%, 40%, 60%, 80%. Expected genome size is marked with purple baseline.

We next applied our heterozygous genome assembly pipeline to the contigs (Supplementary table S3) and scaffolds (Supplementary table S4) reconstructed by SPAdes from the simulated reads. Firstly, we removed heterozygous contigs from all assemblies. The resulting non-redundant assemblies based on SPAdes contigs and scaffolds were very close to the expected size (99%-104%). The non-redundant assemblies based on contigs and scaffolds from SOAPdenovo2 were slightly larger than expected (99%-125%), suggesting SOAPdenovo2 may report wrongly resolved or not overlapping heterozygous contigs/scaffolds that cannot be recognised as heterozygous by our program. Because of this observation, we decided to use SPAdes assemblies for further analysis. Subsequently, the non-redundant contigs/scaffolds were further scaffolded using paired-end (two iterations) and mate-pairs (three iterations) reads. While single pass of heterozygous reduction is enough, we noticed that multiple iterations of scaffolding worked the best. Indeed, the fragmentation of the assembly decreases with each iteration of scaffolding, especially in the case of contig-based reconstruction (Supplementary table S3). Finally, the gaps were closed. The resulting assemblies decreased their fragmentation from several thousands of contigs to less than 110. Interestingly, the assemblies started from contigs that were less fragmented (17-40 contigs) and more similar in size to the target simulated genome (99%-101%, except 0% LOH) than those started from scaffolds (25-225 contigs and 101%-104% of the expected size). Such an observation suggests, that *de novo* assemblers may produce wrongly resolved scaffolds for the heterozygous genomes, as every homozygous region (usually recovered as single consensus contig) can be joined ambiguously with two or more polymorphic neighbour regions on both ends (see b in Figure 1 B).

Unexpectedly, the heterozygous genome assembly pipeline, that started from thousands of contigs/scaffolds, returned full size chromosomes in nearly all reconstructions (Supplementary figure S1). In the case of a simulated genome with 0% LOH, the full size chromosomes were reconstructed for three reference chromosomes (including the longest 3 Mb chromosome), while the remaining five were reconstructed in two scaffolds. This was true for the reconstruction started from both, contigs and scaffolds. The simulated genome with 20% LOH was reconstructed with six full or nearly full size chromosomes and the remaining two were represented in 2-4 scaffolds. On the other hand, the assemblies reconstructed for simulated genomes with a higher level of LOH had fewer full size chromosomes (1-2), if any. In general, the assemblies reconstructed starting from contigs represented higher number of full size chromosomes. In order to evaluate the correctness of each assembly, we aligned the obtained contigs/scaffolds onto *C. parapsilosis* CDC317 chromosomes, as this genome was used to simulate the heterozygous genomes and short reads (Figure 3). Notably, four assemblies for the most heterozygous simulated genomes, these with LOH of 0% and 20%, were resolved correctly (Supplementary figure S1). The remaining eight assemblies with a larger LOH level were carrying between 1 to 4 translocations as compared to the *C. parapsilosis* CDC317 reference (Supplementary figure S1). No large inversions and deletions were observed.

We were interested to know whether the above mentioned translocations were present in the original SPAdes contigs/scaffolds or were introduced during the scaffolding process. Therefore we aligned the SPAdes contigs and scaffolds onto the *C. parapsilosis* CDC317 chromosomes (Supplementary figure S2). As standard *de novo* assemblies were highly fragmented, it was difficult to trace large rearrangements. For this reason, we assumed that most of the observed translocations were introduced during the scaffolding step in our pipeline. Nevertheless, two deletions or translocations are present in the scaffolds from the assembly with 100% LOH, suggesting that at least some of the observed incongruencies may be attributed to errors in the *de novo* contigs or scaffolds from SPAdes. Although, our pipeline returns superior assemblies compared to standard *de novo* assemblers, we need to stress that the resulting scaffolds are chimeric, representing a mixture of both haplotypes (Figure 3).

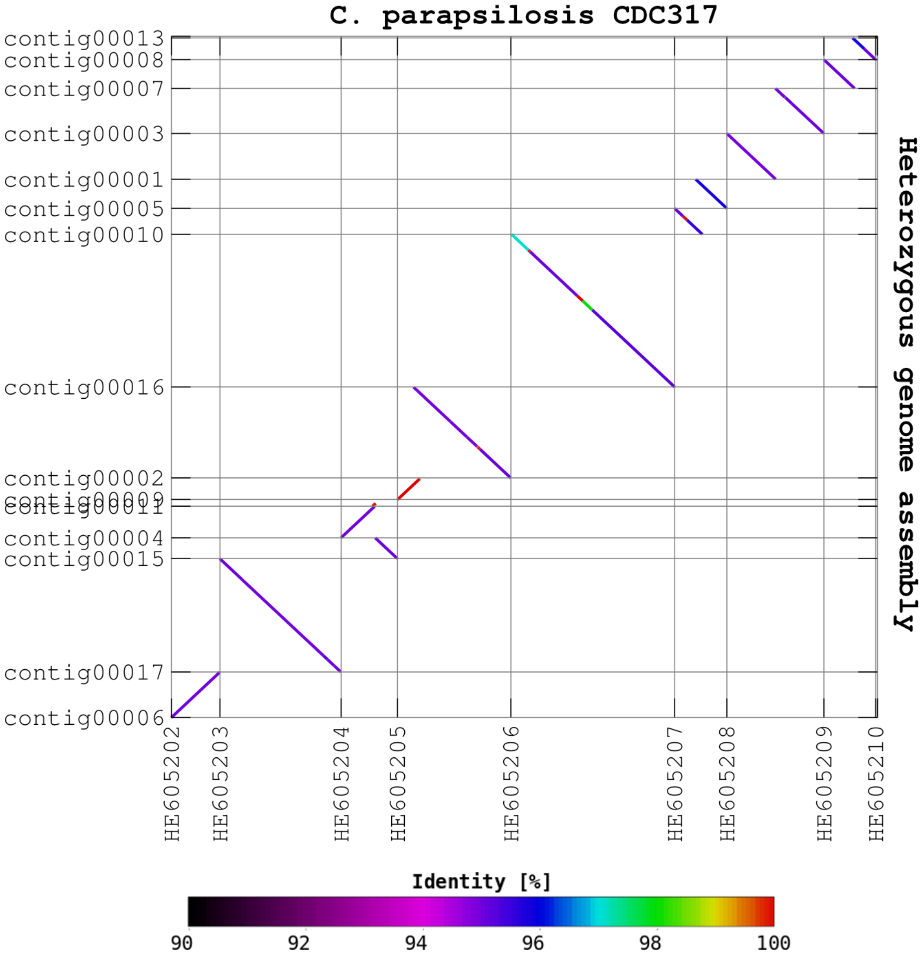

Fig. 3. The assembly of simulated heterozygous genome. Pairwise genome alignment of the final assembly for simulated heterozygous genome with 0% LOH and its reference, *C. parapsilosis* CDC317. Synteny blocks have been coloured accordingly to the identity level between pair of query and target sequences. The assembled genome represent a mixture of two haplomes: 5% diverged (blue or violet) and identical (red) to reference genome. In addition, two short regions with divergence of 2-3% (green and cyan) are present in HE605206. The regions with intermediate divergence were likely assembled from very short contigs from both haplomes.

2.3 Performance on real datasets

As mentioned above, we had earlier applied our pipeline to the *C. orthopsilosis* MCO456 genome, an intraspecies hybrid with 4.5% divergence between parental genomes and over 80% LOH (Pryszcz et al., 2014). To test the generality of our approach we here applied our pipeline to improve the assemblies of the highly heterozygous genomes of another *C. orthopsilosis* heterozygous strain AY2 (NCBI Assembly ID: AMDC01), *D. bruxellensis* LAMAP2480 (AZMW01) and that of *W. anomalus* (AEGI01). In the case of *Candida orthopsilosis* AY2 with a 14.5 Mb genome assembled in 4,152 contigs represents a heterozygous genome with a high level of LOH. 1,293 contigs representing 1.7 Mb were found to be heterozygous in this genome. Again, the haploid assembly of AY2 (12.6 Mb in 255 contigs) is very similar in size to the genome of the highly homozygous strain of the same species, *C. orthopsilosis* 90-125 (12.6 Mb in 8 chromosomes).

In the case of *W. anomalus*, the first, standard assembly contained 26.2 Mb of sequence in 3,133 scaffolds. This represents an heterozygous genome with low level of LOH. Nearly half of this assembly (11 Mb in 1,247 scaffolds) was redundant with an average divergence of 8.6%, suggesting that it is an heterozygous genome with approximately 15% LOH. Similar numbers were obtained for the second version of the *W. anomalus* assembly (AEGI02). Our pipeline identified and removed 11 Mb of sequence in redundant contigs (Supplementary table S5). The remaining 2,801 non-redundant contigs were further scaffolded using publicly available paired-end (SRR072086, SRR072088) and mate-pair (SRR073582, SRR073583, SRR073584) libraries. Noteworthy, the mate-pair libraries were generated by Roche 454, but we were able to use these data as our methodology is highly flexible toward any type of sequencing technology, insert size and read length. We obtained well resolved assembly (85 scaffolds), with 41 times larger N50 and 80 times larger N90 than in the original contigs (Supplementary table S5). Interestingly, the haploid assembly of *W. anomalus* (15.1 Mb) obtained after reduction of heterozygous scaffolds is similar in size to the genome of the closely-related and homozygous *Wickerhamomyces ciferrii* (15.9 Mb) (Schneider et al., 2012). Importantly, our improved *W. anomalus* assembly is less fragmented, than that of *W. ciferii* (Supplementary table S1 and S5, Supplementary figure S4). Similarly, *D. bruxellensis* LAMAP2480 (AZMW01) assembly was improved from 26,9 Mb in 9,167 contigs to 13.6 Mb in 146 contigs using just two mate-pair libraries (SRR1222155, SRR1222162).

2.4 Comparison with heterozygosity-aware tools

Recently, the developers of SPAdes implemented a mode (dipSPAdes) specialised in the assembly of polymorphic genomes. We ran dipSPAdes on our simulated datasets. The big advantage of this program is its simplicity. It requires only sequencing reads as input and no other information like insert sizes, ploidy, expected size, etc, is required. The resulting assemblies were neither fragmented nor larger than expected (Supplementary table S2). Surprisingly, the most heterozygous genome (0% LOH) was the least fragmented with 161 contigs, while the least heterozygous genomes (80% and 100% of LOH) were the most fragmented with 282 and 281 contigs, respectively. Importantly, dipSPAdes produced genome assemblies from 8% to 12% smaller than expected for the genomes that are not 100% heterozygous (Supplementary table S2). In line with this, dipSPAdes assembled only the fully heterozygous (0% LOH) simulated genome correctly, while the remaining genomes were fragmented and contained incongruencies as compared to the reference (Supplementary figure S3). This is an important obstacle as, due to the existence of recombination, the heterozygosity rarely reach 100% in fungal and other genomes (Supplementary table S1). Importantly, although our pipeline is as fast as dipSPAdes (Supplementary table S3), it returned more complete and less fragmented assemblies than dipSPAdes. Typical computation and memory requirements of our pipeline including complexity reduction, scaffolding and gap closing is just a fraction of those necessary for *de novo* assembly. Thus, contig assembly is the most time and memory consuming step.

Recently, support for polymorphic genomes have been also implemented in ALLPATHS-LG, as so-called 'haploidify' mode. We have tested this mode on simulated dataset with 40% LOH of heterozygosity. ALLPATHS-LG requires an additional overlapping paired-end library in order to run, so we needed to simulated such library (2x100 bp with 150 bp insert size). ALLPATHS-LG produced assembly with the size close to expected (13.0 Mb), but fragmented (401 scaffolds) (Supplementary table S2) given more genomic libraries (3) than used with the previous programs (2). Importatntly, ALLPATHS-LG assembly process took over 6 hours using 32 cores (over 5 CPU days!) and over 111 GB of RAM. Moreover, nearly 102 GB of output files were created. In contrast, our heterozygous genome assembly pipeline dealt with the same data in less than two hours using eight cores and less than 16 GB of RAM (including contigs assembly). Importantly, the assembly obtained with our pipeline was less fragmented (57 scaffolds) than the one from ALLPATHS-LG. Such high computational demands prevented us from evaluating ALLPATHS-LG on entire simulated dataset.

2.5 Concluding remarks

We have introduced Redundans, a pipeline that improves the genome assembly of heterozygous genomes. We show that our approach reduces the heterozygous regions with substantial divergence from the genomes under various levels of loss of heterozygosity in both, simulated and real data sets. Moreover, we showed that such reduced assembly can be further scaffolded with success, resulting in full size chromosomes if mate-pair libraries are available. Noteworthy, the assemblies reconstructed from the *de novo* contigs were less fragmented and more accurate than those started from scaffolds. This can be attributed to possible mis-assemblies during the scaffolding of heterozygous contigs by the standard *de novo* assemblers.

We proved our method to be at least as good as (and sometimes superior to) existing tools, resolving complete and correct assemblies by using fewer resources. We need to emphasize, however, that the resulting assembly does not represent individual haplomes, but it is a mosaic of segments from each haplome (i.e. each of the haploid genomes present in a polyploid organism). Thus, one of many haplomes is randomly chosen to fill a given heterozygous region. This is common feature of all heterozygosity-aware methods. In order to identify individual haplomes, sequencing reads need to be realigned onto the genome assembly and re-analysed. Although, such an assembly is somewhat chimeric, in the same way as genomes derived from several individuals as the reference human genome (Lander et al., 2001), it simplifies downstream analysis. In contrast, the typical heterozygous genome assembly is a mixture of consensus and haploid-contigs, which misleads subsequent analyses.

Admittedly, our method still has some limitations. First of all, Redundans relies on BLAT (Kent, 2002) to detect heterozygous regions, and this program has an upper limit of sequence divergence were homology can be detected efficiently (20%). Secondly, large rearrangements may also impede the correct identification of homologous contigs. This limits the usage of our approach to hybrids of somewhat closely-related species. In addition the presence of large segmental duplications that are recovered in different contigs may result in their removal. Finally, our tool has been designed with heterozygous diploid genomes in mind. In principle, it could be applied to polyploid genomes like plants, but we have not tested this so far. To circumvent these limitations, we plan to redesign the heterozygosity reduction step and incorporate depth of coverage information to detect apparent segmental duplications, as well as heterozygous regions with larger divergence and rearrangements.

3 Materials and methods

3.1 Genomes and short reads simulations

We used real data from Illumina-based whole genome shotgun sequencing *C. orthopsilosis* AY2 (AMDC01) and MCO456 (Pryszcz et al., 2014), *D. bruxellensis* (AZMW01), and *W. anomalus* (AEGI01). In addition, we simulated heterozygous genomes based on the 13 Mb *C. parapsilosis* CDC317 genome, which is organised in eight nuclear chromosomes and one mitochondrial chromosome. Six genomes with 5% divergence between haploid genomes and increasing loss of heterozygosity (LOH) levels (0%, 20%, 40%, 60%, 80% and 100%) were generated using *fasta2diverged.py* v1.0. Inserted LOH sizes were modelled based on the real size distributions observed in *C. orthopsilosis* MCO456 and *C. metapsilosis* PL429 (in preparation).

Afterwards, we simulated two Illumina libraries for each simulated genome: i) 100 bp paired-end reads with 600 bp insert size (±50 bp) and 200X coverage, and ii) 50 bp mate-pair reads with 5,000 bp insert size (±1,200 bp) and 20X coverage using GemSIM v1.6 (McElroy et al., 2012). The accuracy of the simulations was confirmed by comparing the estimates of heterozygous fraction, as well as the estimates of divergence between redundant contigs in the resulting assemblies.

3.2 Heterozygous genome assembly pipeline

Reads were pre-processed before assembly to trim at the first undetermined base or at the first base having a PHRED quality below 10. We filtered out pairs with one (or both) reads shorter than 31 bases after trimming using *filterReads.py* v1.0. Two assemblers were used to assemble paired-end reads into contigs and scaffolds: SOAPdenovo v2.04 (Luo et al., 2012) with K-mer ranging from 71 to 91 and SPAdes v3.1.0 (Bankevich et al., 2012) with default parameters. In addition, dipSPAdes, an extension designed to handle polymorphic genomes was used for comparison with our pipeline (**?**). We have also tested ALLPATHS-LG v.R44837 (Gnerre et al., 2011) in haploidify mode.

Heterozygous contigs were identified and removed by *fasta2homozygous.py* v1.0. Afterwards, non-redundant contigs/scaffolds were further scaffolded by SSPACE2 (Boetzer et al., 2011) using pair-end and mate-pair reads. Several iterations of scaffolding were applied. We automatised the scaffolding process in the program named *fastq2sspace.bwamem.py.* It aligns the reads using the fastest short read mapper, BWA MEM (Li, 2013), instead of the standalone SSPACE2 mapper (bowtie) and only a subset of each library (5-10%) was aligned in order to speed-up the scaffolding process and limit the number of intermediate files. Finally, remaining gaps were filled using GapCloser from the SOAPdenovo package (Luo et al., 2012). Redundans pipeline and all programs mentioned in the text can be access publicly (`https://github.com/lpryszcz/redundans`).

3.3 Assembly quality estimation

We assessed the quality of the assemblies by using several parameters of general use (Bradnam et al., 2013). These include, number of contigs, N50, genome completeness expressed as the ratio of the observed versus expected assembly size (Bradnam et al., 2013). For assemblies of simulated data the expected assembly size was that of *C. parapsilosis* CDC317. We inferred the presence of redundant contigs/scaffolds if the assembly had a size larger than the reference. On the other hand, a smaller than expected assembly informed about the extent of missing reference genome regions.

In addition, we analysed the accuracy of each assembly by visual inspection of the alignments of its contigs/scaffolds and the reference chromosomes. The pairwise genome alignments were created and visualised using NUCmer v3.1 (Kurtz et al., 2004). The resulting alignments were filtered, keeping only the best alignment for each region from the query sequence, so called many-to-one mode. Subsequently, we counted large rearrangements, namely deletions, inversions or translocations, between every assembly and the reference genome. Additionally, we marked the reference sequences missing from each assembly. Finally, we checked, whether observed rearrangements originated from the original contigs/scaffolds (SPAdes or SOAPdenovo2) or were introduced during scaffolding with SSPACE2, by pairwise alignment of the *de novo* assemblies against the reference chromosomes. If a particular rearrangement was absent from the respective de novo assembly, we concluded it was introduced during the SSPACE2 scaffolding process of the heterozygous genome assembly pipeline.

Supplementary materials

Supplementary materials can be found at `http://bit.ly/redundans`.

Supplementary figure S1. Scaffolds returned by the heterozygous genome assembly pipeline for various level of LOH were aligned onto *C. parapsilosis* CDC317 chromosomes. The reference chromosomes are denoted on X axis, while query contigs/scaffolds are denoted on Y axis. Best query-to-reference matches are indicated with red (forward) and blue (reverse) dots. The regions of similarity spanning larger regions are denoted by lines. Subsequently, the alignments have been scanned for potential rearrangements (marked by arrows on reference axis).

Supplementary figure S2. Contigs or scaffolds returned by the SPAdes assembler. Indications as in Supplementary figure S1.

Supplementary figure S3. Consensus contigs returned by the dipSPAdes polymorphic genome assembly pipeline for various level of LOH have been aligned onto *C. parapsilosis* CDC317 chromosomes. Indications as in Supplementary figure S1.

Supplementary figure S4. Scaffolds from homozygous *W. ciferrii* genome (Y axis) were aligned against *W. anomalus* scaffolds (X axis) reconstructed by heterozygous genome assembly pipeline. Synteny blocks have been coloured accordingly to the identity level between pair of query and target sequences.

Supplementary table S1. Examples of heterozygous and homozygous genome assemblies retrieved from GenBank. For each analysed assembly, the table provides: species name, accession with the link to GenBank, type (contigs or scaffolds), size, number of contigs/scaffolds, cummulative size and number of identified heterozygous contigs/scaffolds, and cummulative size and number of non-redundant contigs/scaffolds.

Supplementary table S2. Simulated heterozygous genomes with various level of loss of heterozygosity were assembled using SPAdes, SOAPdevono and dipSPAdes. For each assembly, the table provides: the tool and parameters used, assembly type (contigs or scaffolds), loss of heterozygosity level, cumulative size, number of contigs/scaffolds, cummulative size and number of identified heterozygous contigs/scaffolds, cummulative size and number of non-redundant contigs/scaffolds. Finally, the ratio of observed versus expected size is given as percentage for each assembly.

Supplementary table S3. Basic assembly statistics for simulated heterozygous genomes recovered by heterozygous genome assembly pipeline. The reconstructions started from contigs produced by SPAdes. Number of contigs, cumulative assembly size, percentage of GC content, number of contigs longer than 1 kb and the cumulative size of these contigs, N50, N90, the cumulative size of gaps and the length of the longest contigs are given for each step and iteration of heterozygous genome assembly pipeline. Finally, the ratio of observed versus expected size (percentage), runtime, number of CPU cores and peak memory usage are given for each step of heterozygous genome assembly pipeline.

Supplementary table S4. Basic assembly statistics for simulated heterozygous genomes recovered by heterozygous genome assembly pipeline. The reconstructions started from scaffolds produced by SPAdes. Number of contigs, cumulative assembly size, percentage of GC content, number of contigs longer than 1 kb and the cumulative size of these contigs, N50, N90, the cumulative size of gaps and the length of the longest contigs are given for each step and iteration of heterozygous genome assembly pipeline. Finally, the ratio of observed versus expected size is given as percentage for each assembly.

Supplementary table S5. Heterozygous genome assembly pipeline was applied to *Wickerhamomyces anomalus* contigs and scaffolds (AEGI01). Number of contigs, cumulative assembly size, percentage of GC content, number of contigs longer than 1 kb and the cumulative size of these contigs, N50, N90, the cumulative size of gaps and the length of the longest contigs are given for each step and iteration of heterozygous genome assembly pipeline. Finally, the ratio of observed versus expected size of each assembly is given as percentage. The assembly size of closely related homozygous genome of *Wickerhamomyces ciferrii* is taken as expected size.

Bibliography

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.

Bennett, R. J., Forche, A., and Berman, J. (2014). Rapid Mechanisms for Generating Genome Diversity: Whole Ploidy Shifts, Aneuploidy, and Loss of Heterozygosity. *Cold Spring Harbor perspectives in medicine*, 4(10).

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27:578–579.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. a., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. a., Ganapathy, G., Gibbs, R. a., Gnerre, S., Godzaridis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., Macmanes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., and Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2:10.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108:1513–1518.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome research*, 12:656–664.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5:R12.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 00:3.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y. Y. Y. Y. Y. Y. Y. Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J. J., and Lam, T.-w. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18.

McElroy, K. E., Luciani, F., and Thomas, T. (2012). GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC genomics*, 13:74.

Morales, L. and Dujon, B. (2012). Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiology and molecular biology reviews : MMBR*, 76:721–39.

Pryszcz, L. P., Németh, T., Gácser, A., and Gabaldón, T. (2014). Genome comparison of Candida orthopsilosis clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome biology and evolution*, 6(5):1069–78.

Schneider, J., Andrea, H., Blom, J., Jaenicke, S., Rückert, C., Schorsch, C., Szczepanowski, R., Farwick, M., Goesmann, A., Pühler, A., Schaffer, S., Tauch, A., Köhler, T., and Brinkrolf, K. (2012). Draft genome sequence of Wickerhamomyces ciferrii NRRL Y-1031 F-60-10. *Eukaryotic cell*, 11(12):1582–3.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–23.

Small, K. S., Brudno, M., Hill, M. M., and Sidow, A. (2007). A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome biology*, 8:R41.

Zerbino, D. R., McEwen, G. K., Margulies, E. H., and Birney, E. (2009). Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PloS one*, 4(12):e8407.