Evaluation of genome assemblers for *de novo* genome assembly of two dikaryotic litter-decomposing fungi

Jiawei Zhao[1,*]

[1] Department of Biology, Lund University, Lund, Sweden

Supervisor: Dag Ahren, Dimitrios Floudas

* Corresponding author

E-mail: ji8842zh-s@student.lu.se (JZ)

Abstract

Litter decomposing fungi play key role in soil carbon cycling across terrestrial ecosystems ranging from tundra to tropical forests. However, their decomposition mechanisms are not well-known, which might be partly due to the diverse plant cell wall decomposition systems that these fungi harbor. Therefore, nine species of litter decomposers with diverse habitat preferences from the species-rich fungal order *Agaricales* were recently sequenced using the PacBio RS II sequencer in a previous research. Though seven of these nine species were successfully assembled, the unpolished genome assemblies of *Mycetinis scorodonius* and *Leucopaxillus gentianeus* in this research project did not pass strict quality standards. Since then, the development of new long read assemblers as well as improvements of existing software has made an attempt to create possible solutions to this issue. Therefore, multiple *de novo* long-read genome assemblers and polishers were used to generate and polish new genome assemblies from subread data of these two species with different parameters. As a result, 16 genome assemblies of *Leucopaxillus gentianeus* and 19 genome assemblies of *Mycetinis scorodonius* were generated. Subsequently, I assessed the quality of these new assemblies with regards to contig and scaffold length and gene content completeness by using QUAST and BUSCO. According to assessment by QUAST on the best new assembly of *Leucopaxillus gentianeus* which is generated by Falcon, Falcon-unzip and then polished by genomicconsensus, no improvement of the N50 was observed compared to the original assembly, but 31.1% improvement in the number of complete genes in BUSCO was accomplished. By contrast, the best new assembly of *Mycetinis scorodonius* which was produced using the same software as the assembly above, has a considerably lower N50 value was observed compared to the original assembly according to QUAST assessment, but 29.7% improvement in the number of complete genes in BUSCO was accomplished.

In conclusion, Falcon-unzip serves an essential role on improving quality of dikaryotic fungal assembly by phasing. In addition, genomicconsensus further improves assembly quality by polishing. However, due to the limitations of RSII sequencing, the quality of Falcon assemblies were unable to be further improved. Therefore, in order to generate raw reads with higher quality, sequencing by more advanced third generation long-read sequencers, such as Sequel II HiFi-read sequencer or Oxford Nanopore Technology (ONT) ultra-long read sequencer are recommended to generate more successful genome assemblies of dikaryotic genomes.

Introduction

Microbial litter decomposition plays a major role in linking the two enormous carbon pools in terrestrial ecosystems (soils and plant biomass) with atmospheric CO_2 [1]. Large parts of litter decomposition takes place in the upper layers of soil through the activity of litter decomposers (LD) found in mushroom-forming fungi [1]. A considerable number of those LDs are found in the fungal order *Agaricales* (*Basidiomycota*) and phylogenetically related to white-rot (WR) wood decayers [2]. The decomposition mechanisms that underlie the adaptations in diverse habitats of LDs remain unclear, but genome sequencing of a few LDs has revealed variations in the plant-cell-wall degradation (PCWD) machinery between species [3, 4]. Cellulose is a major carbohydrate of the PCW and is found in amorphous and crystalline form [5]. The crystalline form of cellulose renders cellulose recalcitrant to microbial decomposition [1]. Despite recent research of fungal genomes from comparative genomics, LDs remain an evolutionary and functional black box, which limits researchers' understanding of LDs' role in the soil carbon cycle and their adaptation to diverse habitats [1].

As an attempt to unravel the black box mentioned above, the genomes of nine LDs across three major clades in *Agaricales* were sequenced by PacBio sequencing in a recent research [1]. The PacBio libraries were produced using the SMRTbell™ Template Prep Kit 1.0 according to manufacturer's instructions. One SMRTbell™ with library was sequenced on the PacBio RSII instrument using the P6-C4 sequencing chemistry and 600 minute movie time (the time specified for collecting data from a SMRT cell). After sequencing, genomes of monokaryotic species were assembled by SciLifeLab using HGAP3 [6] whereas genomes of dikaryotic species were assembled by PacBio FALCON assembler [7].

The quality and the completeness of the polished genome assemblies were examined by QUAST (4.4) [8] and BUSCO (2.0) [9]. For the latter, the fungi *basidiomycota_odb9* dataset was used.

However, two of these genomes didn't pass strict quality standards and were discarded. Thus, only seven of the nine newly sequenced LD genomes were compared with 35 published genomes across mushroom-forming fungi about their core plant-cell-wall degradation (PCWD) gene networks [1].

Leucopaxillus gentianeus (*L. gentianeus*) and *Mycetinis scorodonius* (*M. scorodonius*) are the two dikaryotic litter decomposers (LDs) whose genomes proved particularly challenging to assemble and hence did not pass strict quality standards. Thus, this project was set up in order to figure out possible approaches to generate more qualified genome assemblies of these two LDs.

L. gentianeus belongs to *Tricholomataceae* family of *Agaricales* order. It is an inedible mushroom with unpleasant, pungent and farinaceous odor, whereas being non-toxic [10, 11]. It is commonly known as the bitter false funnelcap, or the bitter brown *leucopaxillus*. The diameter of its cap is between 40mm to 120mm. *L. gentianeus* grows singly or in groups on ground, usually under oak, pine and other conifers, often forming fairy rings. The fungus is widely distributed over Europe and Northern and Central America.

M. scorodonius belongs to *Omphalotaceae* family of *Agaricales* order, formerly in the genus *Marasmius*. It is an edible mushroom with strong garlic or onion-like odor. It has a beige, broadly convex cap and a tough slender stipe. The diameter of its cap is between 2-30mm [12, 13, 14]. *M. scorodonius* grows saprotrophically. It is mainly

found on the fallen needles of conifers, growing scattered or gregariously in both North America (east of the Great Plains) and in Europe [15].

High-quality genomic resources that enable haplotype comparisons are essential to reliably answer the long-standing biological question on how evolution has shaped the genomic architecture of dikaryotic fungi, whereas short-read genome assemblies for dikaryotic fungi are often highly fragmented and lack haplotype-specific information due to the high heterozygosity. Furthermore, short-read assemblies are prone to consist of repetitive contents. Therefore, genome assembly for dikaryotic fungi should be based on long reads [16].

Genome assembly of dikaryotic fungal genomes has been a challenging problem for researchers since they belong to the category of non-inbred or rearranged heterozygous genomes [7]. The main problem is that most available genome assemblies are unable to capture the heterozygosity of dikaryotic genomes, thus rendering to mosaic genomic sequence that arbitrarily alternates between parental alleles [17]. Consequently, the variation between the homologous chromosomes will be lost. By 2016, no universal and scalable solution for assembling diploid and polymorphic genomes had been developed [7]. The computational methods for genome assembly of dikaryotic fungi tend to produce highly fragmented results that have average contig lengths shorter than 10 Kbps [18, 19, 20]. Meanwhile, other non-computational approaches are labor intense, costly and are often limited in assembly contiguity [7].

To address this critical need for efficient methods to assemble dikaryotic genomes, the open-source FALCON and FALCON-Unzip algorithms were developed to assemble Single Molecule Real-Time (SMRT®) Sequencing data into highly accurate, contiguous, and correctly phased diploid genomes [7].

The unsuccessful reference genome assemblies of *L. gentianeus* and *M. scorodonius* were unphased [1]. In order to improve the qualities of genome assembly, Falcon-unzip was used a genome assembly phasing tool in this project. Falcon-unzip has shown its promising phasing capabilities in successful genome-assembly of the dikaryotic wheat stripe rust fungus *Puccinia striiformis* f. sp. *tritici* by generating a near-complete phased genome [16].

In this study, transcriptome sequencing data sets were used to infer high-quality gene models and identify virulence genes involved in plant infection which were referred to as effectors. Thus, the genome assembly by FALCON-Unzip assembler in this study was represented as the most complete *Puccinia striiformis* f. sp. *Tritici* genome assembly to date (83Mb, 156 contig, N_{50} of 1.5 Mb). This assembly provides phased haplotype information for over 92% of the genome. High interhaplotype diversity of over 6% was revealed by comparisons of the phase blocks [16]. What's more, candidate effectors which lack an allelic counterpart are more distant from conserved genes than allelic candidate effectors and less likely to be evolutionarily conserved within the *P. striiformis* species complex and *Pucciniales*. In summary, the haplotype-phasing of by FALCON-Unzip assembler renders novel features in its assembly which were previously hidden in collapsed and fragmented unphased genome assemblies of this dikaryotic fungus [16].

To identify other haploid *de novo* long read assemblers, we took advantage of a benchmark study on prokaryotic whole genome sequencing [21]. The researchers assessed the performance of seven long-read assemblers (Canu, Flye, Miniasm/Minipolish, NECAT, Raven, Redbean and Shasta) across a wide variety of genomes and read parameters. Assessment was based on structural accuracy/completeness, sequence identity, contig circularisation and computational

resources used. The results of the study showed that Canu v1.9 produced moderately reliable assemblies but had the longest runtimes. Flye v2.7 was more reliable and did particularly well with plasmid assembly. Miniasm/Minipolish v0.3 and NECAT v20200119 were the most likely to produce clean contig circularisation. Raven v0.0.8 was the most reliable for chromosome assembly, though it did not perform well on small plasmids and had circularisation issues. Redbean v2.5 and Shasta v0.4.0 were computationally efficient but more likely to produce incomplete assemblies [21]. Thus, Flye, Miniasm/Minipolish and Raven were deemed to be the three best-performing non dikaryon-specific *de novo* long-read genome assemblers. As a result, these three assemblers were considered as possible alternatives to the combination of FALCON assembler and FALCON-Unzip, and were subsequently evaluated in our study.

Materials and methods

Sequenced raw data and its pre-processing

L. gentianeus and *M. scorodonius* were both sequenced on PacBio RSII instrument [1].

PacBio RSII sequencing consists of three major steps: Library Preparation (no amplification required), Instrument Run (sequencing time is 30-120 minutes per SMRT cell) and Data Analysis (open source, open standards). In 2013, RSII system was able to provide raw reads that were subsequently generated into highly accurate bacterial genomes with the highest N50, the fewest contigs compared to genome assemblies generated from raw reads that were sequenced from other contemporary sequencing systems [22]. However, RSII has become disadvantaged on its limited number of sequenced reads since several more advanced PacBio sequencing systems have become available in the recent years.

Raw data of each of those two dikaryotic fungi consist of a compressed .fastq.gz file of filtered subreads (which were converted into corresponding .fasta files later as well) , 24 .bax.h5 files and 8 .bas.h5 files. .bas.h5 files contain the information necessary to dereference by hole number the ZMW-level data. One .bas.h5. file corresponds to three PulseData-containing .bax.h5 files. PulseData consists of data about BaseCalls, ConsensusBaseCalls and Regions [23].

Thus, since FALCON-Unzip requires unaligned .bam files as input, it was necessary to convert every three .bas.h5 files of a same movie (there are 8 movies in total for each of the two fungus) into an unaligned .bam file. Thus, bax2bam (0.0.9 was used in order to convert the legacy PacBio basecall format (bax.h5) into the BAM basecall format. After conversion, both *L. gentianeus* and *M. scorodonius* had 8 unaligned .bam files each. What's more, samtools (1.9) was used for generating index files from the .fasta subread files converted from the .fastq.gz filtered subread files of the fungi.

Genome assembly and evaluation

As described in the introduction, four different combinations of genome assembler/phasing tool/polisher/auxiliary tools were run to generate genome assemblies from raw data. These four combinations are listed in the table below:

Table 1. All software used for genome assembly in this project

Combination	Software	Version
Falcon & Falcon unzip	pb-falcon	0.2.7
Falcon & Falcon unzip	falcon-unzip	0.1.0
Falcon & Falcon unzip	genomicconsensus	2.3.3

Combination	Software	Version
Falcon & Falcon unzip	pbmm2	1.3.0
Falcon & Falcon unzip	pbindex	0.23.0
Flye	flye	2.7.1
Miniasm & Minipolish	minimap2	2.17
Miniasm & Minipolish	miniasm	0.3_r179
Miniasm & Minipolish	minipolish	0.1.2
Raven	raven-assembler	1.1.5

Since falcon-unzip failed during the alignment step due to an unsolved issue (see https://github.com/Orthologues/LUfungiProject#intro1), pbmm2 was used for aligning raw .bam basecall files with the reference of a corresponding phased but unpolished assembly from FALCON-Unzip. Thus, pbindex was used for generating an index file of the aligned .bam file. At the final step, genomicconsensus was run as an assembly polisher in arrow algorithm which takes in an aligned.bam file as input and uses its corresponding phased but unpolished assembly as the reference. As for *L. gentianeus,* two different FALCON configuration files were used for generating two different polished FALCON assemblies whereas three different polished FALCON assemblies of *M. scorodonius* were assembled from three different FALCON configuration files. The .cfg files were configured according to a template FALCON .cfg file for fungal genome assembly (For more detailed parameters in .cfg, please see https://pb-falcon.readthedocs.io/en/latest/_downloads/fc_run_fungal.cfg) .

Both Flye and Miniasm/Minipolish were run only with their default parameters whereas Raven was run with 6 different versions of parameters by adjusting its rounds of polishing, match score, mismatch penalty and gap penalty. The reason that Flye and Miniasm/Minipolish were only run with default parameters was that both assemblers did not provide computational parameters that are available for change other than sequencing instrument of raw reads and expected genome size.

After completion of forementioned genome assembly, QUAST (5.0.2) and BUSCO (4.0.6) were used for assembly evaluation. In order to visualize and summarize the results of evaluation, MultiQC (1.9) was run consequently. Linear regression models were used to test if changes in configurations significantly affect vital statistical results in QUAST and BUSCO analysis.

Since it would be biased if we choose to set assembly type as independent variable and statistical results in QUAST and BUSCO as dependent variable due to non-optimized and incomparable configurations, we choose to separate different assembly types in the analyses to focus on different configurations within Raven assemblies and Falcon assemblies respectively. In these linear regression models, different versions of configuration/parameters were set as categorical independent variables, whereas N50, largest contig size and the number of complete/missing BUSCOs of assemblies were set as numerical dependent variables. Subsequently, ANOVA analysis were performed in order to evaluate the effect of parameter/configuration change on assembly results.

In analysis of Raven assemblies, the number of polishing rounds was set as one independent factor and 6 different combinations of alignment parameters were set as another independent factor, whereas the percentage of missing BUSCOs, the percentage of complete BUSCOs, the percentage of complete & single-copy BUSCOs and the percentage of complete & duplicated BUSCOs were regarded as dependent variables respectively in their corresponding linear regression models. Thus, 8 different linear regression models were established.

In analysis of FALCON/FALCON-Unzip assemblies, the stage of assemblies and the version of FALCON configuration files were set as two independent factors respectively whereas the percentage of missing BUSCOs, the percentage of complete

BUSCOs, the percentage of complete & single-copy BUSCOs and the percentage of complete & duplicated BUSCOs were regarded as dependent variables respectively in their corresponding linear regression models. Assemblies have three different stages: unpolished & unphased, unpolished & phased, and polished & phased. Regarding *L. gentianeus*, with two different versions of FALCON configuration files, P-values of its corresponding linear regression models are listed below:

Results and discussion

The results of assembly evaluation were based on both QUAST and BUSCO. In order to analyze the impact of altering parameters and configurations on vital statistical results of QUAST and BUSCO, ANOVA was performed later. The .csv files and scripts used for ANOVA is available at https://github.com/Orthologues/LUfungiProject/tree/master/anova_analysis.

QUAST analysis

Although the output of QUAST analysis consists of many different statistical numbers, only N50, N75, lengths of the largest contig and lengths of the assemblies were included in the following tables in order to simplify the assessment and focus the discussion to the most relevant values for this study.

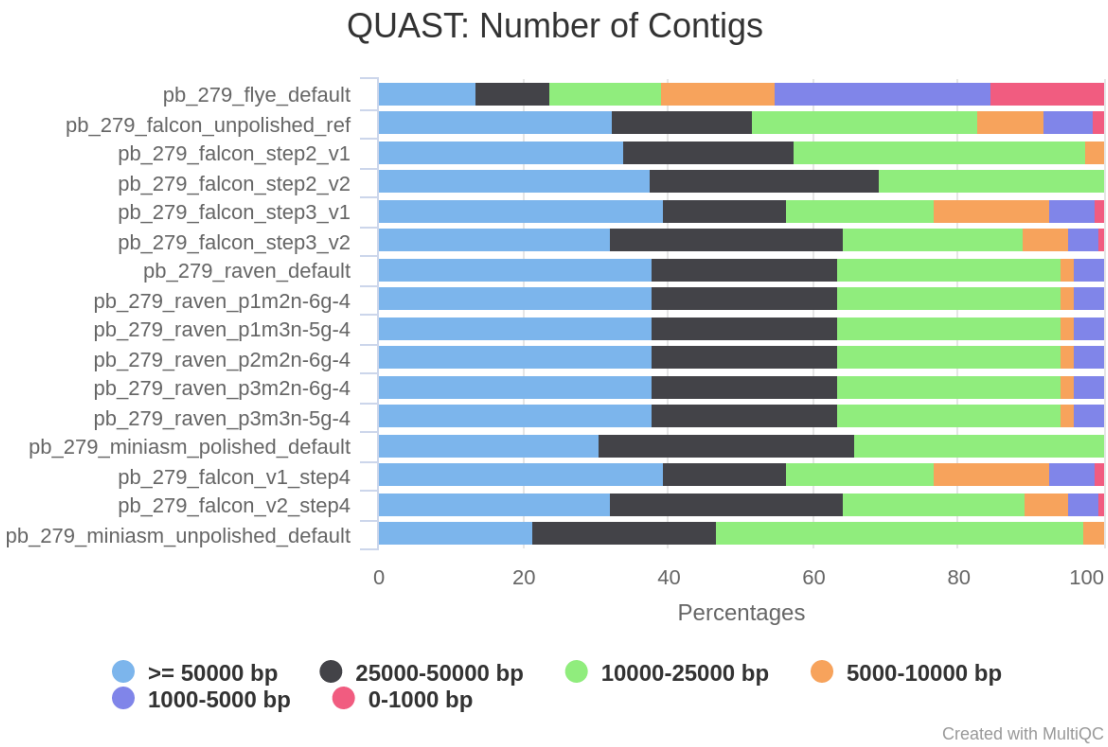

Fig 1. Overall length distribution of contigs in *de novo* genome assemblies of *L. gentianeus* (pb_279). In the names of raven assemblies, the suffix “p1m2n-6g-4” means one round of polishing, match score as 2, mismatch penalty as -6, gap penalty as -4 etc whereas the parameters of the default assembly was p2m3n-5g-4. In the names of falcon assemblies, “v1” means its FALCON configuration file is of the first version etc, whilst “step2” means the unphased & unpolished assemblies, “step3” means the phased but unpolished assemblies, and “step4” means the phased & polished assemblies. “unpolished_ref” means the unphased & unpolished FALCON assembly of *L. gentianeus* provided by Johan Bentzer. For exact values of N50 etc, see Table 2.

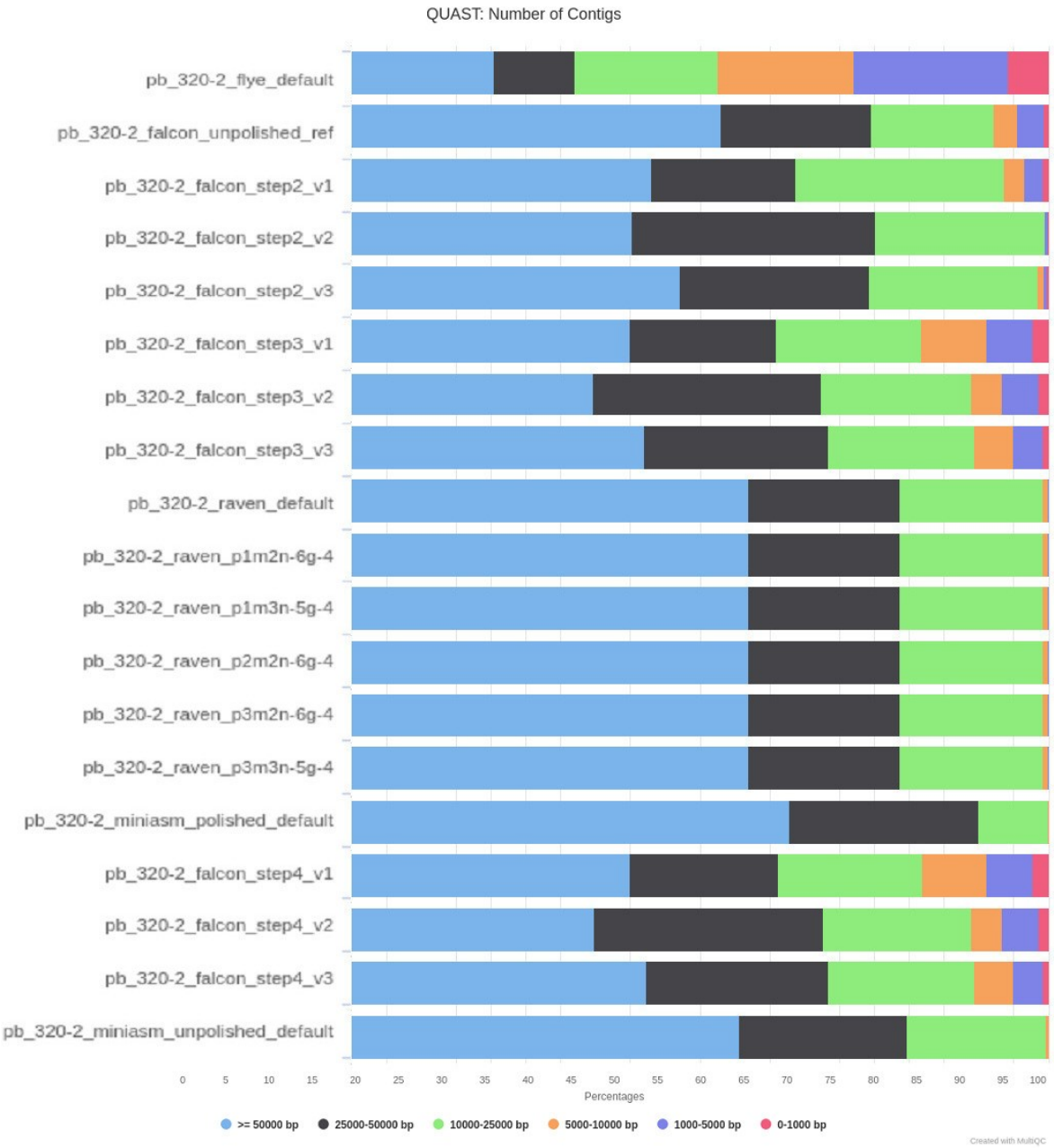

Fig 2. Overall length distribution of contigs in *de novo* genome assemblies of *M. scorodonius* (pb_320-2). The naming patterns here are the same as Fig 1. “unpolished_ref” means the unphased & unpolished FALCON assembly of *M. scorodonius* provided by Johan Bentzer. For exact values of N50 etc, see Table 3.

Table 2. N50 values, N75 values, the length of each assembly’s largest contig, the total length of

each assembly of *L. gentianeus* (pb_279)

Sample Name	N50 (Kbp)	N75 (Kbp)	Largest contig (Kbp)	Length (Mbp)
pb_279_flye_default	122.1	43.8	1711.2	54.6
pb_279_falcon_unpolished_ref	226.0	73.8	2447.5	77.0
pb_279_falcon_step2_v1	305.4	103.8	2239.2	67.6
pb_279_falcon_step2_v2	113.6	49.0	1021.5	39.2
pb_279_falcon_step3_v1	224.7	100.0	2224.5	90.1
pb_279_falcon_step3_v2	86.3	42.1	1015.1	58.4
pb_279_raven_default	130.6	58.2	1846.3	75.4
pb_279_raven_p1m2n-6g-4	131.1	58.3	1848.0	75.6
pb_279_raven_p1m3n-5g-4	131.2	58.3	1848.2	75.6
pb_279_raven_p2m2n-6g-4	130.4	58.1	1846.2	75.4
pb_279_raven_p3m2n-6g-4	130.2	58.0	1845.7	75.4
pb_279_raven_p3m3n-5g-4	130.3	58.0	1846.0	75.4
pb_279_miniasm_polished_default	89.8	39.6	1686.4	76.6
pb_279_falcon_v1_step4	225.1	100.5	2229.7	90.3
pb_279_falcon_v2_step4	86.6	42.1	1017.0	58.6
pb_279_miniasm_unpolished_default	77.3	32.5	1749.3	87.4

* In the names of raven assemblies, the suffix “p1m2n-6g-4” means one round of polishing, match score as 2, mismatch penalty as -6, gap penalty as -4 etc whereas the parameters of the default assembly was p2m3n-5g-4. In the names of falcon assemblies, “v1” means its FALCON configuration file is of the first version etc, whilst “step2” means the unphased & unpolished assemblies, “step3” means the phased but unpolished assemblies, and “step4” means the phased & polished assemblies. “unpolished_ref” means the unphased & unpolished FALCON assembly of *L. gentianeus* provided by Johan Bentzer.

Table 3. N50 values, N75 values, the length of each assembly’s largest contig, the total length of

each assembly of *M. scorodonius* (pb_320-2)

Sample Name	N50 (Kbp)	N75 (Kbp)	Largest contig (Kbp)	Length (Mbp)
pb_320-2_flye_default	144.5	59.5	2249.5	122.7
pb_320-2_falcon_unpolished_ref	432.9	165.8	2323.7	149.9
pb_320-2_falcon_step2_v1	231.7	108.6	2067.9	156.9
pb_320-2_falcon_step2_v2	123.1	55.5	1441.0	119.3
pb_320-2_falcon_step2_v3	171.5	82.8	2176.5	144.4
pb_320-2_falcon_step3_v1	216.3	96.0	2075.5	168.0
pb_320-2_falcon_step3_v2	114.5	49.6	1441.2	130.2
pb_320-2_falcon_step3_v3	160.9	71.4	2185.3	154.4
pb_320-2_raven_default	215.4	104.7	2670.7	140.2
pb_320-2_raven_p1m2n-6g-4	215.2	104.8	2671.4	140.3
pb_320-2_raven_p1m3n-5g-4	215.3	104.8	2671.5	140.3
pb_320-2_raven_p2m2n-6g-4	215.4	104.7	2670.9	140.2
pb_320-2_raven_p3m2n-6g-4	215.2	104.7	2670.9	140.2
pb_320-2_raven_p3m3n-5g-4	215.2	104.7	2671.0	140.2
pb_320-2_miniasm_polished_default	152.2	82.3	1503.7	143.0
pb_320-2_falcon_v1_step4_busco	217.6	96.3	2083.3	168.8
pb_320-2_falcon_v2_step4_busco	115.4	49.9	1446.1	130.8
pb_320-2_falcon_v3_step4_busco	161.4	71.7	2193.2	155.1
pb_320-2_miniasm_unpolished_default	150.7	80.2	1521.1	147.9

* The naming patterns here are the same as Table 1. “unpolished_ref” means the unphased & unpolished FALCON assembly of *M. scorodonius* provided by Johan Bentzer.

Anova results: The genome assemblies of Raven and FALCON/FALCON-Unzip were analyzed separately.

In analysis of Raven assemblies, the number of polishing rounds was set as one independent factor and 6 different combinations of alignment parameters were set as another independent factor, whereas N50 value and the largest contig size were regarded as two dependent variables respectively. Regarding *L. gentianeus*, P-values of its corresponding linear regression models are listed below (values below 0.05 are

considered to be significant):

Table 4. P-values of ANOVA analysis on Linear regression models assessing the results of raven assemblies of *L. gentianeus* (pb_279)

Independent variable	Dependent variable	P value	Level of significance
Number of polishing rounds	N50	0.00385	**
Alignment parameter	N50	0.0572	.
Number of polishing rounds	Largest contig size	0.00173	**
Alignment parameter	Largest contig size	0.0742	.

For Raven assemblies of *L. gentianeus,* we found that the number of polishing rounds significantly improved N50 ($F_{2,2}$=259; P=0.00385) and Largest contig size ($F_{2,2}$=576.33; P=0.00173).

However, the corresponding linear regression models of Raven assemblies of *M. scorodonius* didn't satisfy the assumption of normally distributed residuals, which is one of the assumptions of ANOVA. Thus, P-values are not meaningful to be calculated.

In analysis of FALCON/FALCON-Unzip assemblies, the stage of assemblies and the version of FALCON configuration files were set as two independent factors respectively whereas N50 value and the largest contig size were considered as dependent variables. Assemblies have three different stages: unpolished & unphased, unpolished & phased, and polished & phased. Regarding *L. gentianeus*, with two different versions of FALCON configuration files, P-values of its corresponding linear regression models are listed below:

Table 5. P-values of ANOVA analysis on Linear regression models assessing the results of Falcon/Falcon-Unzip assemblies of *L. gentianeus* (pb_279)

Independent variable	Dependent variable	P value	Level of significance
Stage of assemblies	N50	0.197	None
Version of Falcon .cfg files	N50	0.0127	*
Stage of assemblies	Largest contig size	0.132	None
Version of Falcon .cfg files	Largest contig size	3.95×10^{-6}	***

For Falcon assemblies of *L. gentianeus,* we found that the first version of Falcon .cfg file significantly improved N50 ($F_{1,2}$=77.18; P=0.0127) and Largest contig size ($F_{1,2}$=2.53×10^{5}; P=3.95×10^{-6}).

In analysis on *M. scorodonius*, with three different versions of Falcon .cfg files, P-values of its corresponding linear regression models are listed below:

Table 6. P-values of ANOVA analysis on Linear regression models assessing the results of Falcon/Falcon-Unzip assemblies of *M. scorodonius* (pb_320-2)

Independent variable	Dependent variable	P value	Level of significance
Stage of assemblies	N50	0.00337	**
Version of Falcon .cfg files	N50	8.66×10^{-7}	***
Stage of assemblies	Largest contig size	0.3498	None
Version of Falcon .cfg files	Largest contig size	9.35×10^{-5}	***

For Falcon assemblies of *M. scorodonius,* we found that increased stage of assembly significantly improved N50 ($F_{2,4}$=32.45; P=0.00337). Furthermore, the first version of Falcon .cfg file significantly improved N50 ($F_{2,4}$=2147.47; P=8.66×10^{-7}). In

addition, the first and the third version of Falcon.cfg files significantly improved Largest contig size ($F_{2,4}$=204.80; P=9.35 × 10^{-5}).

BUSCO analysis

L. gentianeus and *M. scorodonius* were sequenced on PacBio RSII instrument and therefore, genome assemblies of *L. gentianeus* were with the prefix pb_279 while genome assemblies of *M. scorodonius* were with the prefix pb_320-2. BUSCO analysis was carried out with the reference dataset fungi_odb10. This dataset consists of Agaricales and other 11 fungal orders. The output of BUSCO analysis consists of the percentages of four different types of BUSCOs: complete and single-copy BUSCOs, complete and duplicated BUSCOs, fragmented BUSCOs and missing BUSCOs. Thus, the percentages of type-I BUSCOs and type-II BUSCOs were summed together to form the percentage of complete BUSCOs in order to gain more statistical insight. According to the dikaryotic nature of the two fungi in this project, high percentages of complete and duplicated BUSCOs were expected due to the homozygotes.

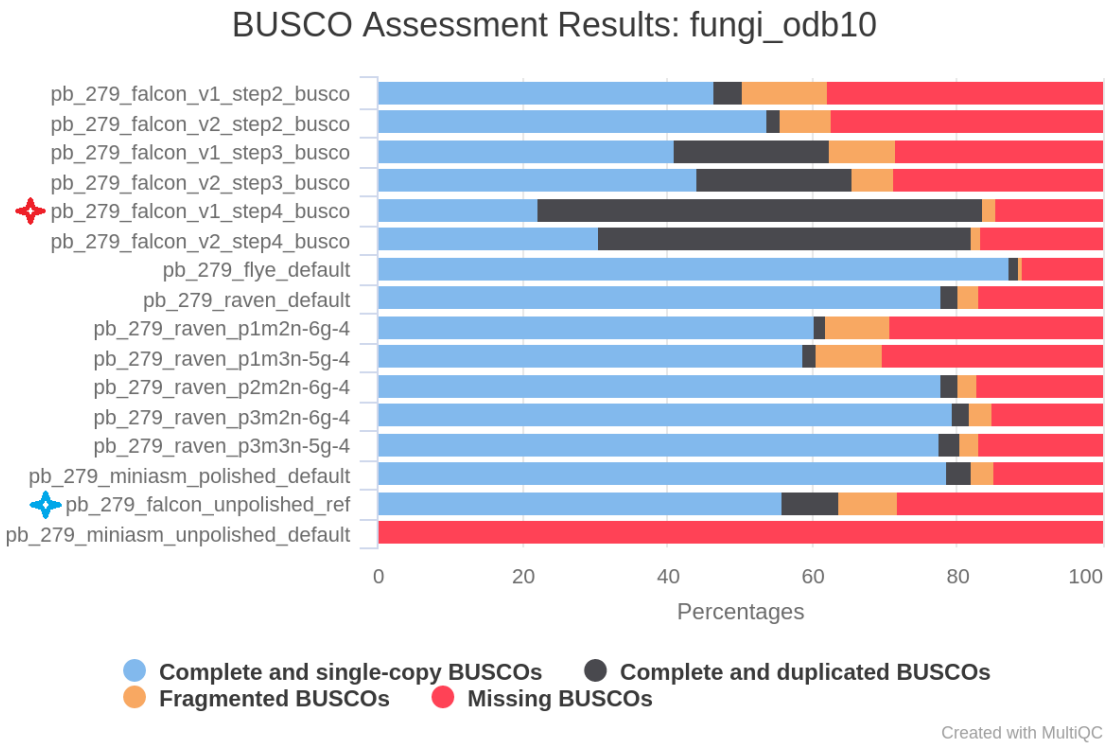

Fig 3. Percentages of four types of BUSCOs in *de novo* genome assemblies of *L. gentianeus* (pb_279). In the names of raven assemblies, the suffix “p1m2n-6g-4” means one round of polishing, match score as 2, mismatch penalty as -6, gap penalty as -4 etc whereas the parameters of the default assembly was p2m3n-5g-4. In the names of falcon assemblies, “v1” means the version of its FALCON configuration file is one etc, whilst “step2” means the unphased & unpolished assemblies, “step3” means the phased but unpolished assemblies, and “step4” means the phased & polished assemblies. “unpolished_ref” means the unphased & unpolished FALCON assembly of *L. gentianeus* provided by Johan Bentzer. The best assembly is labeled in red color, whereas the reference assembly is labeled in blue color.

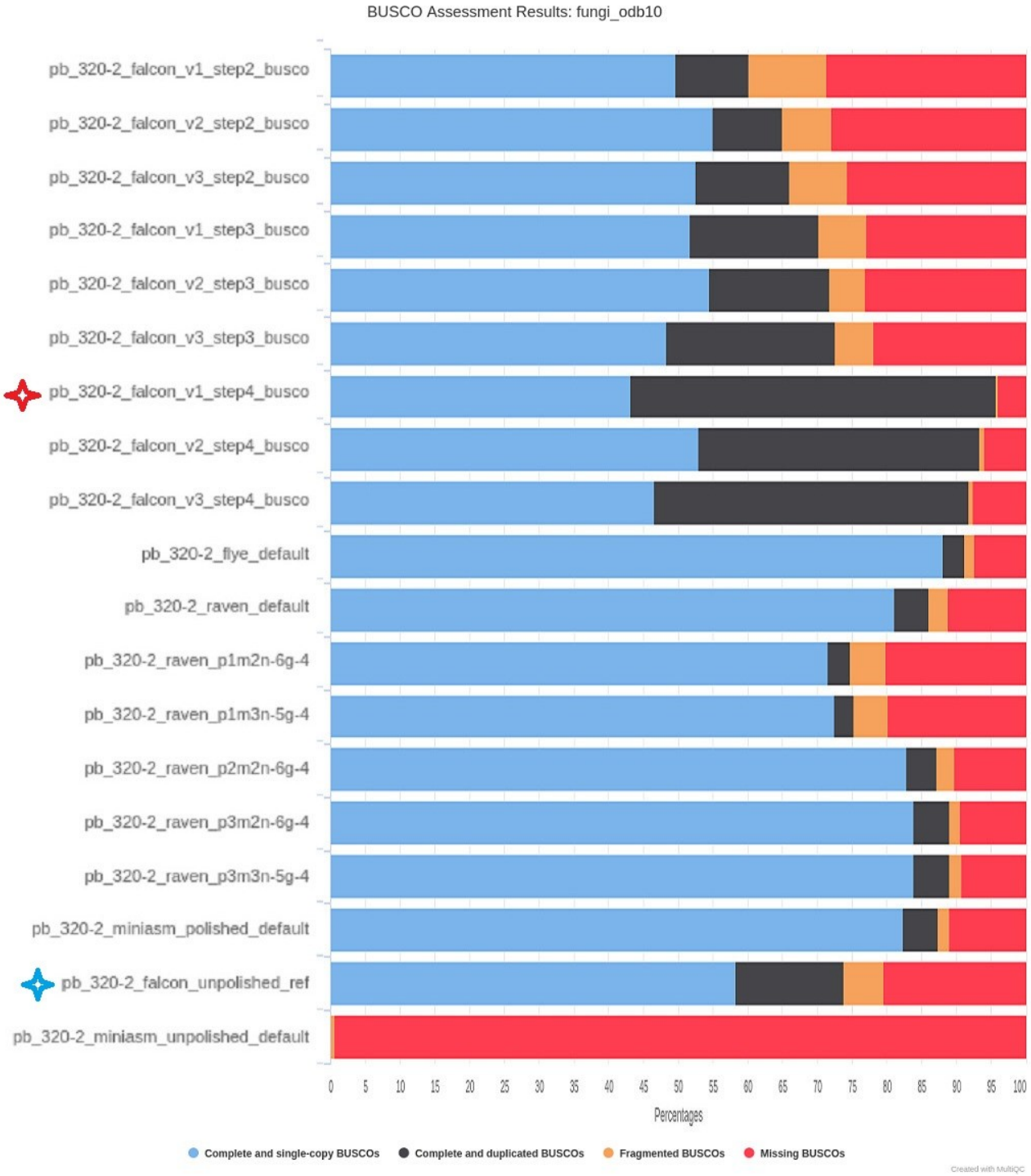

Fig 4. Percentages of four types of BUSCOs in *de novo* genome assemblies of *M. scorodonius* (pb_320-2). The naming patterns here are the same as Fig 3. “unpolished_ref” means the unphased & unpolished FALCON assembly of *M. scorodonius* provided by Johan Bentzer. The best assembly is labeled in red color, whereas the reference assembly is labeled in blue color.

Anova results: The genome assemblies of Raven and FALCON/FALCON-Unzip were analyzed separately.

Table 7. P-values of ANOVA analysis on Linear regression models assessing the results of raven assemblies of *L. gentianeus* (pb_279)

Independent variable	Dependent variable	P value	Level of significance
Number of polishing rounds	Complete BUSCOs	0.00128	**
Alignment parameters	Complete BUSCOs	0.184	None
Number of polishing rounds	Complete & Single-copy BUSCOs	0.00186	**
Alignment parameters	Complete & Single-copy BUSCOs	0.186	None
Number of polishing rounds	Complete & Duplicated BUSCOs	0.0455	*
Alignment parameters	Complete & Duplicated BUSCOs	0.423	None
Number of polishing rounds	Missing BUSCOs	0.00387	**
Alignment parameters	Missing BUSCOs	0.235	None

For Raven assemblies of *L. gentianeus,* we found that the number of polishing rounds significantly improved Complete BUSCOs ($F_{2,2}$=780.03; P=0.00128), Complete & Single-copy BUSCOs ($F_{2,2}$=535.86; P=0.00186) and Complete & Duplicated BUSCOs ($F_{2,2}$=21; P=0.0455), whereas significantly decreased Missing BUSCOs ($F_{2,2}$=257.64; P=0.00387).

Table 8. P-values of ANOVA analysis on Linear regression models assessing the results of raven assemblies of *M. scorodonius* (pb_320-2)

Independent variable	Dependent variable	P value	Level of significance
Number of polishing rounds	Complete BUSCOs	0.00338	**
Alignment parameter	Complete BUSCOs	0.707	None
Number of polishing rounds	Complete & Single-copy BUSCOs	0.0108	*
Alignment parameter	Complete & Single-copy BUSCOs	0.765	None
Number of polishing rounds	Complete & Duplicated BUSCOs	0.0399	*
Alignment parameter	Complete & Duplicated BUSCOs	0.885	None
Number of polishing rounds	Missing BUSCOs	0.00394	**
Alignment parameter	Missing BUSCOs	0.853	None

For Raven assemblies of *L. gentianeus,* we found that the number of polishing rounds significantly improved Complete BUSCOs ($F_{2,2}$=295.09; P=0.00338), Complete & Single-copy BUSCOs ($F_{2,2}$=91.76; P=0.0108) and Complete & Duplicated BUSCOs ($F_{2,2}$=24.03; P=0.0399), whereas significantly decreased Missing BUSCOs ($F_{2,2}$=252.54; P=0.00394).

Table 9. P-values of ANOVA analysis on Linear regression models assessing the results of Falcon/Falcon-Unzip assemblies of *L. gentianeus* (pb_279)

Independent variable	Dependent variable	P value	Level of significance
Stage of assemblies	Complete BUSCOs	0.0128	*

Independent variable	Dependent variable	P value	Level of significance
Version of Falcon .cfg files	Complete BUSCOs	0.362	None
Stage of assemblies	Complete & Single-copy BUSCOs	0.0119	*
Version of Falcon .cfg files	Complete & Single-copy BUSCOs	0.0544	.
Stage of assemblies	Complete & Duplicated BUSCOs	0.00863	**
Version of Falcon .cfg files	Complete & Duplicated BUSCOs	0.306	None
Stage of assemblies	Missing BUSCOs	0.00357	**
Version of Falcon .cfg files	Missing BUSCOs	0.430	None

For Falcon assemblies of *L. gentianeus,* we found that increased stage of FALCON assembly significantly improved Complete BUSCOs ($F_{2,2}$=77.08; P=0.0128), Complete & Single-copy BUSCOs ($F_{2,2}$=83.16; P=0.0119) and Complete & Duplicated BUSCOs ($F_{2,2}$=114.90; P=0.00863), whereas significantly decreased Missing BUSCOs ($F_{2,2}$=279; P=0.00357).

Table 10. P-values of ANOVA analysis on Linear regression models assessing the results of Falcon/Falcon-Unzip assemblies of *M. scorodonius* (pb_320-2)

Independent variable	Dependent variable	P value	Level of significance
Stage of assemblies	Complete	0.000305	***

Independent variable	Dependent variable	P value	Level of significance
	BUSCOs		
Version of Falcon .cfg files	Complete BUSCOs	0.755	None
Stage of assemblies	Complete & Single-copy BUSCOs	0.120	None
Version of Falcon .cfg files	Complete & Single-copy BUSCOs	0.0651	.
Stage of assemblies	Complete & Duplicated BUSCOs	0.000921	***
Version of Falcon .cfg files	Complete & Duplicated BUSCOs	0.312	None
Stage of assemblies	Missing BUSCOs	0.000225	***
Version of Falcon .cfg files	Missing BUSCOs	0.910	None

For Falcon assemblies of *M. scorodonius,* we found that increased stage of FALCON assembly significantly improved Complete BUSCOs ($F_{2,4}$=112.53; P=0.000305) and Complete & Duplicated BUSCOs ($F_{2,4}$=63.92; P=0.000921), whereas increased stage of FALCON assembly significantly decreased Missing BUSCOs ($F_{2,4}$=131.39; P=0.000225).

Comparison between the best assemblies of this project and the reference assemblies

Judged by the comprehensive quality standard of genome assembly which consists of N50-value, percentage of complete BUSCOs and percentage of missing BUSCOs, *pb_279_falcon_v1_step4* and *pb_320-2_falcon_v1_step4* would be evaluated as the best assembly of *L. gentianeus* and the best assembly of *M. scorodonius* respectively.

Regarding *L. gentianeus*, compared to the unphased & unpolished reference assembly, *pb_279_falcon_v1_step4* has a nearly identical N50 whereas it had a

31.1% improvement in the number of Complete BUSCOs in BUSCO analysis.

Regarding *M. scorodonius*, compared to the unphased & unpolished reference assembly, *pb_320-2_falcon_v1_step4* has an approximately 49% worse N50 whereas its had a 29.7% improvement in the number of Complete BUSCOs in BUSCO analysis.

Conclusions

Statistical significance and assumptive explanations

Assessed from the results of BUSCO & QUAST analysis on *L. gentianeus* & *M. scorodonius* assemblies as well as its corresponding ANOVA statistics, four

conclusions were drawn:

Firstly, among Raven assemblies, neither N50 nor Largest contig size was significantly improved by altering alignment parameters, whereas the number of polishing rounds improved N50 and Largest contig size significantly. Secondly, among Raven assemblies, none of BUSCOs in different types was significantly increased or decreased by altering alignment parameters, whereas the number of polishing rounds increased the percentage of complete BUSCOs, the percentage of complete and duplicated BUSCOs significantly, whilst it decreased the percentage of missing BUSCOs significantly. Thirdly, among FALCON/FALCON-Unzip assemblies, both the step of phasing by FALCON-Unzip and the subsequent step of polishing by Genomicconsensus increased the percentage of complete BUSCOs, the percentage of complete and duplicated BUSCOs significantly, whilst it decreased the percentage of missing BUSCOs significantly. The final conclusion made was that among FALCON/FALCON-Unzip assemblies, N50 and largest contig size were significantly changed by altering Falcon .cfg files.

As for the first and the second conclusion, the working mechanisms of FALCON-Unzip and assembly polishing have to be described in order to provide a possible explanation.

Falcon is a diploid-aware assembler which follows the hierarchical genome assembly process (HGAP). It produces a set of primary contigs (a-contigs), which represent diverfent allelic variants. Each a-contig is associated with a homologous genomic region on a p-contig. The first round of HGAP involves the selection of seed reads according to user-defined *length_cutoff* [7]. Every assembly on step two in this project was a concatenation of its corresponding .fasta file of p-contigs and .fasta file of a-contigs.

In contrast, Falcon-Unzip is a novel, true diploid assembler. It takes the contigs

from FALCON and phases the reads based on heterozygous SNPs identified in the initial assembly. It then produces a set of partially phased primary contigs and fully phased haplotigs which represent divergent haplotypes [7]. Every assembly on step three in this project was a concatenation of its corresponding .fasta file of p-contigs and .fasta file of haplotigs.

In long (especially noisy) read assembly, *polishing* refers to the correction step which takes all of the underlying data and the raw quality values inherent to SMRT sequencing into account to improve the base accuracy of contig sequences [6].

One possible explanation to the third conclusion could be that inter-haplotype diversity of dikaryotic fungi would be considered as a high value at 6% [16], thus rendering considerable increase in Complete & duplicated BUSCOs. Another possible explanation is linked to concatenation of p-contigs and haplotigs after running FALCON-Unzip. After the step of concatenation, a large number of highly similar or even completely identical contigs were generated, thus rendered significant increase in the percentage of complete & duplicated BUSCOs. Accompanied by increase in complete BUSCOs, the percentage of missing BUSCOs decreased significantly.

After the phasing step, Genomicconsensus took in phased but unpolished assemblies, then improved the base accuracy of input considerably. Thus, the dikaryotic nature of the fungal species in this project was further identified by polishing, therefore rendering the further significant improvements in BUSCO statistics. Furthermore, base correction in the polishing step also contributed to significant decrease in the percentage of missing BUSCOs.

A possible explanation to the fourth conclusion could be that assumed from the value of *length_cutoff*, which is most predominant difference between between the FALCON configuration files of *L. gentianeus* and of *M. scorodonius*, (for details of the configuration files, please see

https://github.com/Orthologues/LUfungiProject/tree/master/pb-assembly/mycfgs).

Since the configuration files in version one have the lowest *length_cutoff*, a considerable number of PacBio RSII subreads with medium lengths were counted as seed reads instead of being discarded. During the subsequent process, those subreads might become intermediate fragments and render conjunctions between longer fragments to form longer contigs, thus rendering a higher N50 value of its assembly.

Discussion on possible improvement in sequencing

Limitations of PacBio RSII sequencer may have considerably contributed to the failure to improve N50 of the best assemblies of this project compared to the unphased & unpolished reference assemblies. Therefore, re-sequencing by more recent and advanced long-read sequencing techniques in the future may be a potential solution.

Sequel is a second generation PacBio sequencer and generates up to 7x more reads per SMRT-cell compared to the first generation PacBio sequencer RSII. The size and the surface of Sequel SMRT-cells has grown allowing to increase the number of ZMWs from 150,000 to one million per cell. Sequel would be the instrument of choice when speedy and high-throughput long-read sequencing is required. However, due to the significant advantages conveyed by the subset of the longest sequencing reads, the RSII sequencer was likely the preferable instrument for most gnome-wide studies of large eukaryotic genomes, especially for *de novo* genome assemblies [24]. Thus, Sequel I sequencer doesn't seem to be an appropriate alternative.

Sequel II is a third generation PacBio sequencer released in year 2019. It

generates up to 8x more data than the original Sequel System, provides access to even more highly accurate long reads (HIFI Reads) and reduces project time for faster results while making sequencing more affordable [24]. Sequel II enables high throughput HIFI reads with base-level resolution with >99% [25]. Sequel II can be applied to create high-quality whole genome *de novo* assemblies of eukaryotic organisms [24]. Therefore, Sequel II sequencer would possibly be a great alternative for sequencing improvement.

Oxford Nanopore Technologies (ONT) is another third generation long-readsequencer which generated reads as long as 2Mb [25]. What's more, the portable MinION device (ONT) has received much attention because of its small size and possibility of rapid analysis at reasonable cost [26].

Two research papers comparing Sequel II sequencer with third-generation ONT sequencers have been recently published [25] [26]. In the paper which compared HiFi reads of PacBio Sequel II system with ultra-long reads of ONT by applying the two platforms to one single rice individual and subsequently comparing the two assemblies to investigate the advantages and limitations of each, the results showed that ONT ultra-long reads were generally more superior. ONT ultra-long reads deliverd higher contiguity producing a total of 18 contigs of which 10 were assembled into a single chromosome compared to that of 394 contigs and three chromosome-level contigs for the PB Sequel II assembly. The ONT ultra-long reads also prevented assemblies errors caused by long repetitive regions for which the researchers observed a total 44 genes of false redundancies and 10 genes of false losses in the Sequel II assembly whereas the PB assembly from Sequel II HiFi reads had less errors at the level of single nucleotide and small InDels than the ONT assembly [26].

In another paper which compared MinION (ONT) sequencer versus Sequel II (PacBio) third-generation sequencer in identification and diagnostics of fungal and

oomycete pathogens from conifer and potato leaves and tubers, the researchers demonstrated that Sequel II is efficient for metabarcoding of complex samples, whereas MinION is not suited for this purpose due to a high error rate and multiple biases. However, the use of MinION (ONT) for rapid diagnostics of pathogens and potentially other organisms while taking care to control or account for multiple potential technical biases was advocated [27].

Acknowledgement

I sincerely thank Dag Ahrén for providing me with the informative suggestions in software selection and in genome assembly, Johan Bentzer for providing me with all the sequencing data and reference assemblies needed in this project, and Dimitrios Floudas for providing me with his academic papers and supplementary materials in biological background.

References

1. Floudas, D., Bentzer, J., Ahrén, D. *et al.* Uncovering the hidden diversity of litter-decomposition mechanisms in mushroom-forming fungi. *ISME J* (2020). https://doi.org/10.1038/s41396-020-0667-6

2. Matheny PB, Curtis JM, Hofstetter V, Aime MC, Moncalvo J-M, Ge Z-W, et al. Major clades of Agaricales: a multilocus phylogenetic overview. Mycologia. 2006;98:982–95.

3. Bao D, Gong M, Zheng H, Chen M, Zhang L, Wang H, et al. Sequencing and comparative analysis of the straw mushroom (*Volvariella volvacea*) genome. PLoS ONE. 2013;8:e58294.

4. Floudas D, Held BW, Riley R, Nagy LG, Koehler G, Ransdell AS, et al. Evolution of novel wood decay mechanisms in Agaricales revealed by the genome sequences of *Fistulina hepatica* and *Cylindrobasidium torrendii*. Fungal Genet Biol. 2015;76:78–92.

5. Albersheim P, Darvill A, Roberts K, Sederoff R, Staehelin A. Plant cell walls: from chemistry to biology. New York: Garland Science, Taylor & Francis group; 2011.

6. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–9.

7. Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13(12):1050-1054. doi:10.1038/nmeth.4035

8. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.

9. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

10. Phillips, Roger (2010). Mushrooms and Other Fungi of North America. Buffalo, NY: Firefly Books. p. 67. ISBN 978-1-55407-651-2.

11. Wood WF, Brandes ML, Watson RL, Jones RL, Largent DL. Trans-2-nonenal, the cucumber odor of mushrooms. Mycologia 1994 86(4):561-3.

12. "*Mycetinis scorodonius* page". Species Fungorum. Royal Botanic Gardens Kew. Retrieved 2018-10-27.

13. Courtecuisse, R.; Duhem, B. (2013). Champignons de France et d'Europe (in French). Delachaux et Niestlé. p. 254. ISBN 978-2-603-02038-8. Also available in English.

14. Antonín, V.; Noordeloos, M. E. (2010). A monograph of marasmioid and collybioid fungi in Europe. Postfach 1119, 83471 Berchtesgaden, Germany: IHW Verlag. pp. 400–404. ISBN 978-3-930167-72-2.

15. Kuo, M. (2013, January). *Mycetinis scorodonius*. Retrieved from the *Mushroom-Expert.Com* Website: http://www.mushroomexpert.com/mycetinis_scorodonius.html

16. Schwessinger B, Sperschneider J, Cuddy WS, et al. A Near-Complete Haplotype-Phased Genome of the Dikaryotic Wheat Stripe Rust Fungus *Puccinia striiformis* f. sp. *tritici* Reveals High Interhaplotype Diversity. *mBio*. 2018;9(1):e02275-17. Published 2018 Feb 20. doi:10.1128/mBio.02275-17

17. Church DM, et al. Extending reference assembly models. Genome biology. 2015;16:13.

18. Vinson JP, et al. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. Genome Res. 2005;15:1127–1135.

19. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5:e254.

20. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature genetics. 2012;44:226–232.

21. Wick RR and Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 2; peer review: four approved]. F1000Research 2020, 8:2138 (https://doi.org/10.12688/f1000research.21782.2)

22. Pacific Biosciences. (2013, April 16). *PacBio RSII Sequencing System*. Retrieved August 17, 2020, from https://www.mscience.com.au/upload/pages/pacbio/pacbio_rs_ii_brochure.pdf

23. Pacific Biosciences. (2019, August). *SMRT Tools Reference Guide Version 06*. Retrieved August 17, 2020, from https://www.pacb.com/wp-content/uploads/SMRT_Tools_Reference_Guide_v700.pdf

24. Froenicke, L. (2016, November 10). *New Service: Long-Read Sequencing on the PacBio Sequel.* UC Davis Genome Center. https://dnatech.genomecenter.ucdavis.edu/2016/11/10/new-service-long-read-sequencing-on-the-pacbio-sequel/

25. University of Georgia. (n.d.). *PacBio Sequel II Sequencing*. Retrieved July 19, 2020, from https://dna.uga.edu/pacbio-sequel-sequencing-2/

26. Lang, et al.Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacbio Sequel II system and ultralong reads of Oxford NanoporebioRxiv (2020)p. 2020.02.13.948489

27. Loit K, Adamson K, Bahram M, et al. Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens. *Appl Environ Microbiol*. 2019;85(21):e01368-19. Published 2019 Oct 16. doi:10.1128/AEM.01368-19