

RESEARCH ARTICLE

Integrative Meta-Assembly Pipeline (IMAP): Chromosome-level genome assembler combining multiple *de novo* assemblies

Giltae Song¹, Jongin Lee², Juyeon Kim², Seokwoo Kang¹, Hoyong Lee¹, Daehong Kwon², Daehwan Lee², Gregory I. Lang³, J. Michael Cherry⁴, Jaebum Kim^{2*}

1 School of Computer Science and Engineering, Pusan National University, Busan, South Korea,

2 Department of Biomedical Science and Engineering, Konkuk University, Seoul, South Korea,

3 Department of Biological Sciences, Lehigh University, Bethlehem, PA, United States of America,

4 Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America

* jbkim@konkuk.ac.kr



Abstract

OPEN ACCESS

Citation: Song G, Lee J, Kim J, Kang S, Lee H, Kwon D, et al. (2019) Integrative Meta-Assembly Pipeline (IMAP): Chromosome-level genome assembler combining multiple *de novo* assemblies. PLoS ONE 14(8): e0221858. <https://doi.org/10.1371/journal.pone.0221858>

Editor: Maureen J. Donlin, Saint Louis University, UNITED STATES

Received: November 10, 2018

Accepted: August 18, 2019

Published: August 27, 2019

Copyright: © 2019 Song et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. And our pipeline is available online at <http://github.com/jkimlab/IMAP>.

Funding: This work was supported by Pusan National University Research Grant, 2016, and National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017074529 and NRF-2018R1A5A2023879) to GS, and Ministry of Science and ICT of Korea Grant 2014M3C9A3063544 and Ministry of Education of

Background

Genomic data have become major resources to understand complex mechanisms at fine-scale temporal and spatial resolution in functional and evolutionary genetic studies, including human diseases, such as cancers. Recently, a large number of whole genomes of evolving populations of yeast (*Saccharomyces cerevisiae* W303 strain) were sequenced in a time-dependent manner to identify temporal evolutionary patterns. For this type of study, a chromosome-level sequence assembly of the strain or population at time zero is required to compare with the genomes derived later. However, there is no fully automated computational approach in experimental evolution studies to establish the chromosome-level genome assembly using unique features of sequencing data.

Methods and results

In this study, we developed a new software pipeline, the integrative meta-assembly pipeline (IMAP), to build chromosome-level genome sequence assemblies by generating and combining multiple initial assemblies using three *de novo* assemblers from short-read sequencing data. We significantly improved the continuity and accuracy of the genome assembly using a large collection of sequencing data and hybrid assembly approaches. We validated our pipeline by generating chromosome-level assemblies of yeast strains W303 and SK1, and compared our results with assemblies built using long-read sequencing and various assembly evaluation metrics. We also constructed chromosome-level sequence assemblies of *S. cerevisiae* strain Sigma1278b, and three commonly used fungal strains: *Aspergillus nidulans* A713, *Neurospora crassa* 73, and *Thielavia terrestris* CBS 492.74, for which long-read sequencing data are not yet available. Finally, we examined the effect of IMAP parameters, such as reference and resolution, on the quality of the final assembly of the yeast strains W303 and SK1.

Korea Grant 2016R1D1A1B03930209 and 2019R1F1A1042018 to JK.

Competing interests: The authors have declared that no competing interests exist.

Conclusions

We developed a cost-effective pipeline to generate chromosome-level sequence assemblies using only short-read sequencing data. Our pipeline combines the strengths of reference-guided and meta-assembly approaches. Our pipeline is available online at <http://github.com/jkimlab/IMAP> including a Docker image, as well as a Perl script, to help users install the IMAP package, including several prerequisite programs. Users can use IMAP to easily build the chromosome-level assembly for the genome of their interest.

Introduction

As sequencing technology is advancing in terms of cost efficiency and sequencing accuracy, whole genome sequencing is being used more frequently in modern genetics. This sequencing approach provides innovative temporal and spatial resolution to examine fundamental problems in population and evolutionary genomics [1], and empowers the study of human genetic disease, such as studies of cancer [2]. Recently, several laboratories have performed short-term evolution studies in evolving populations of *Saccharomyces cerevisiae* using whole genome sequence data to understand evolutionary patterns associated with cell population growth [3, 4].

Despite the substantial contributions that whole genome sequencing approaches have made to experimental evolutionary studies, some sequence variations might escape detection if they are examined based on the mapping of short-reads to a diverged reference genome. For example, the yeast strain W303, used in many evolution experiments, differs from the reference S288C strain at approximately 9,500 nucleotide positions [5]. If structural or copy number variations occur in strain-specific genomic regions that are not present in the S288C reference strain, the lack of a true chromosome-level genome assembly of a specific strain is especially problematic.

Building a *de novo* sequence assembly for the target genome permits genome analyses of novel sequences beyond the reference. However, *de novo* assembly is a computationally challenging problem because of the short length of the read sequences that must be merged into a complete chromosome-level sequence of the target genome. Instead of a complete continuous chromosome sequence, short-reads are built into fragmented chunks of continuous nucleotides, termed contigs, and into supercontigs, termed scaffolds, especially if frequent repeat elements are present in the genome [6]. Long-read sequencing from third generation sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore, can greatly improve the quality of sequence assembly; however, they are currently expensive, and are more error-prone than short-read sequencing [7, 8]. Instead of using these expensive technologies, we designed a new computational pipeline to resolve this issue using only cost-effective short-read sequencing data, which we term the integrative meta-assembly pipeline (IMAP).

One approach to improving the continuity of the assembly is to combine multiple versions of sequence assemblies into larger scaffolds for a given target genome (for examples, see references [9, 10]). However, this approach cannot produce a chromosome-level genome assembly. Another strategy is to use a reference-guided approach that uses a set of existing genome assemblies that are closely related to the target sequence to fill assembly gaps, and sort the order of contigs or scaffolds [11–13]. Unfortunately, this approach cannot address the issue that different assembly approaches may yield different assembly outputs.

To overcome these limitations, we developed a new hybrid assembly approach that combines two recently published tools, RACA [13] and DESCHRAMBLER [14], to order scaffolds and build a chromosome-level assembly. RACA and DESCHRAMBLER are used to increase the contiguity of initial assemblies, and to generate a meta-assembly that integrates multiple initial assemblies, respectively. We also iteratively refined the assembly to improve its accuracy at the nucleotide level. We integrated this method into a pipeline to serve other groups in the fields of population and evolutionary genomics, especially for studies of evolving populations using unicellular eukaryotic organisms. We built sequence assemblies of *S. cerevisiae* strains, W303, and SK1 automatically, and compared them with assemblies generated using long-read sequencing for validation [5, 15]. We demonstrated the utility of our pipeline by constructing a chromosome-level assembly for *S. cerevisiae* Sigma 1278b, a strain widely used as a model for pseudohyphal growth [16]. In addition to *S. cerevisiae*, we also applied IMAP for other fungal species such as *Aspergillus nidulans*, *Neurospora crassa*, and *Thielavia terrestris* that are used as model organisms in cell biology and genetics studies [17–19].

Our generalizable approach could establish a chromosome-level sequence assembly for the genome of a given strain, which maximizes the unique characteristics of short-read sequencing data from evolving populations based on *de novo* assembly, and covers strain-specific genomic regions. In studies where long-read sequencing is not available, such as those investigating short-term experimental evolution, our pipeline will be a useful tool to elucidate evolutionary dynamics with fine-scale temporal resolution.

Materials and methods

Data preparation and initial assembly

We obtained 40 datasets representing the initial time point of evolution studies using *Saccharomyces cerevisiae* W303 strain (NCBI Bioproject accession: PRJNA205542, [4]). Each set comprised Illumina paired-end short-read sequencing data at approximately 100× fold coverage (read length: 100 bases). After the 40 datasets were merged, the depth of the W303 sequencing data was approximately 4,000× coverage. For strain SK1, we used two sequencing datasets: One with approximately 474× coverage (NCBI Biosample accession: SAMN05792036, [20]), and the other with approximately 56× coverage (available at ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/SGRP2/input/strains/17A_Sc_SK1/, [21]). For yeast strain Sigma 1278b, we used a data set with approximately 191× coverage (Biosample accession: SAMN03020229, [22]). The information of data for three fungal strains is available in [S1 Table](#).

First, we filtered out sequencing reads based on ambiguous bases (Ns) and low-quality reads from the merged dataset. We then applied several *de novo* assembly methods using the collected short-read data. We compared their assembly results using major quality metrics that were used in the assembly evaluation consortium project called Assemblathon [23]. From this validation, we decided to use three *de novo* assembly tools: MaSurCA, SPAdes, and SOAPdenovo2 [24–26], using default parameters. If users wish to include other initial assemblies, they can easily add them at this step.

Meta-assembly approach for chromosome level assembly

We extended the scaffolds of the three initial assemblies to a chromosome-level assembly using RACA [13] and DESCHRAMBLER [14]. In our pipeline, the contiguity of the initial assemblies was increased by RACA, and the meta-assembly from the improved initial assemblies was generated by DESCHRAMBLER (details of command lines and options are available at our GitHub website: <http://github.com/jkimlab/IMAP>). RACA requires a reference sequence that is close to the user's target genome, and an outgroup with a known phylogeny.

The order of scaffolds in each initial assembly was determined under the guidance of the reference sequence and the sequence of the outgroup, and the scaffolds were then enlarged to chromosome-level scaffolds using RACA. For *S. cerevisiae* strain genomes, we used the reference sequence from strain S288C, which was obtained from the *Saccharomyces* Genome Database (<http://www.yeastgenome.org>). For the outgroup in RACA, we used the genome sequence of *Naumovozyma dairenensis* (NCBI RefSeq: GCF_000227115.2). All RACA assembly outcomes in chromosome-level scaffolds were combined together into a consensus sequence using DESCHRAMBLER with the same reference sequence. DESCHRAMBLER was originally developed to reconstruct the ancestral genome, given the genome assemblies of multiple descendants. In this study, we applied DESCHRAMBLER to construct a meta-assembly by treating multiple initial assemblies as descendant assemblies and the meta-assembly as their common ancestor.

We then iteratively refined this chromosome-level assembly using Pilon (with default options in version 1.13; [27]), and GATK (with default options in version 3.7; [28]), using the original raw reads as input, and our assembly as the reference. We expected that a more accurate assembly is likely to have fewer SNPs, when SNPs are called based on the mapping of raw reads data against its own assembly, itself created using the raw reads. After mapping read sequences against an assembly sequence, SNPs were called using GATK. GATK is a pipeline for calling single nucleotide polymorphisms (SNPs). The nucleotides at SNP positions were corrected to the majority base present among the mapped nucleotides. If the majority base at a position was identical to the base in the assembly, the base at that position was unchanged. If the two mapped nucleotides at a position were tied in number, we also left the base at that position unchanged.

Assembly validation

We used several evaluation metrics to assess the quality of the assembly. The metrics included basic assembly statistics, such as N50 (the shortest sequence length at 50% of the genome), MIN (the shortest scaffold), MAX (the longest scaffold), and total length. We also evaluated the assembly results using Bowtie2 [29] and GATK [28]. We calculated the portion of read pairs that were properly mapped out of mapped reads for paired-end sequencing data relative to our assembly sequence using Bowtie2. In addition, we counted the number of false positive SNPs called by GATK in the raw read data relative to its own assembly. If the new assembly is built more correctly, we would expect fewer SNPs than those observed in the original assembly.

In addition, for certain target genomes, such as yeast strains W303 and SK1, another version of their assembly was built by different research groups using long-read sequencing technology, such as PacBio (e.g. [5,15]). We compared our assembly results with those assemblies for validation.

We also used AGAPE (with default options; [22]) for evaluation. AGAPE is a pipeline to annotate chromosomal features in a given sequence. We measured the assembly accuracy based on the number of yeast genes correctly annotated using AGAPE for a given sequence assembly. We assumed that a sequence of higher quality would have more genes annotated accurately by AGAPE. We also investigated paralogous genes using the AGAPE gene annotation results, and compared them between our assembly and the PacBio assembly. Yeast paralogous genes were determined using BLAST [30] against the yeast reference genome. A total of 480 paralogous genes were grouped into 197 sets. The genome sequence segments including these paralogous genes were often missed in *de novo* assembly, because of their repetitiveness. Checking the annotation status for the paralogous genes was another way to assess assembly accuracy.

Synteny-level similarity between our IMAP assembly and the PacBio output was checked using LASTZ alignment (with version 1.02, with options -E = 150 -H = 2000 -K = 4300 -L = 2200 -M = 254 -O = 600 -T = 2 -Y = 15000; [31]). We mapped syntenic blocks that represented conserved genomic segments between two assembly sequences. We visualized the synteny mapping results using SynCircos, implemented in mySyntenyPortal [32]. We used this to verify the agreement of the order and orientation of large-scale genomic segments between the two assemblies.

Results and discussion

Overview of the Integrative Meta-Assembly Pipeline (IMAP)

We developed a software pipeline called IMAP to establish a chromosome-level sequence assembly of a user's target genome given a collection of sequencing raw reads (see Fig 1). First, we generated initial draft assemblies using *de novo* assembly tools. Note that although we built the three initial draft assemblies based on MaSurCA, SPAdes, and SOAPdenovo2, respectively, more tools could be applied. These initial assemblies were combined and extended to larger scaffolds at the chromosome level using RACA and DESCHRAMBLER. The improved assembly was refined using Pilon and GATK error correction. All of these steps were implemented and integrated into one pipeline for automatic construction of the chromosome-level *de novo* assembly. Our pipeline is available on GitHub for users who wish to perform chromosome-level assembly using short-read sequencing data of their genome. This pipeline is ideally suited for laboratories that need to generate a strain specific chromosome-level genome assembly in a cost-effective manner. Once users input their own initial raw sequencing data (in the FASTQ format), the final assembly at the chromosome level can be automatically built by IMAP using a single command.

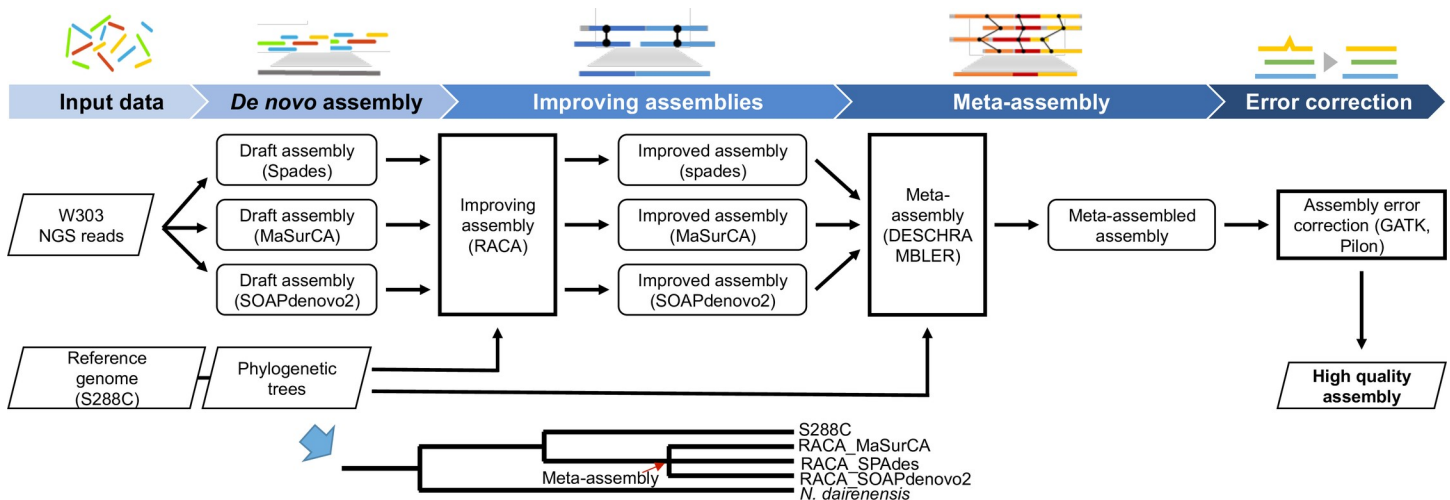


Fig 1. Data flow chart of the integrative meta-assembly pipeline (IMAP). Short raw sequencing data reads for a target genome were extended to initial scaffolds using *de novo* assembly software. We applied several *de novo* assembly tools, and produced several draft assemblies for the target genome. The order of scaffolds in each assembly was determined under the guidance of the reference sequence and the sequence of an outgroup; and the scaffolds were then enlarged to chromosome-level scaffolds using RACA. All RACA assembly outcomes in chromosome-level scaffolds were combined together into a consensus sequence, using DESCHRAMBLER. We then refined this consensus sequence at the nucleotide level using Pilon and GATK, and finished it as a chromosome-level sequence for the target genome. The tree in the bottom of this figure illustrates the phylogeny among the target genome, the reference, and the outgroup sequence. For the target sequence, the consensus of the three RACA assembly results obtained using the three *de novo* assembly tools (MaSurCA, SPAdes, and SOAPdenovo2) was obtained using DESCHRAMBLER.

<https://doi.org/10.1371/journal.pone.0221858.g001>

Evaluation of chromosome-level assembly built by IMAP

We applied IMAP to the *S. cerevisiae* strain W303, one of the most commonly used strains in yeast studies [33]. We relied on W303 short-read sequence data from an experimental evolution study [4] that is available at NCBI SRA with Bioproject ID PRJNA205542. To construct the W303 sequence assembly, we used only samples from time point zero. We downloaded all 40 W303 datasets (each set has approximately 100× coverage) at the initial time point, and merged them to approximately 4,000× coverage data.

We generated three initial draft assemblies for this data using MaSurCA, SPAdes, and SOAPdenovo2. The chromosome-level assembly of the W303 strain genome was constructed using IMAP. We evaluated each step of our pipeline. To validate which step influenced the quality of the final assembly, we measured basic statistics for the outcome of each step in the pipeline, using metrics such as the length of the shortest contig (MIN), the length of the longest contig (MAX), and N50 of the assembly (see Table 1). According to the basic statistics of the assemblies, extending all initial draft assemblies to chromosome-level scaffolds and combining them into a consensus assembly using RACA and DESCHRAMBLER increased the N50 by 2.6–53 times compared with those of the initial *de novo* draft assemblies. In addition, the N50 was about 76% of that of the PacBio assembly. Although our IMAP assembly does not show as large N50 as the PacBio assembly, the chromosome-level assembly outcome that was itself obtained by IMAP, which was generated without additional sequencing to improve the initial draft assembly, can contribute significantly to the community. While an assembly generated using the Illumina sequencing data needs several months of extra labor cost and time to finish an assembly [7], IMAP can shorten this step substantially to only several additional hours that are necessary for running the pipeline. Since IMAP uses only Illumina short read sequencing data to build the chromosome-level assembly, this avoids the considerable cost and material requirements that PacBio needs [7, 8].

We also computed the fraction of mapped reads and the portion of proper read pairs out of the mapped reads. The total length of the assembly and the fraction of mapped reads after the DESCHRAMBLER meta-assembly step slightly decreased, compared to the original SPAdes assembly that initially showed the highest value initially. This occurs because some bases in a RACA assembly may be inconsistent with the majority of other RACA assemblies, and in the meta-assembly step they may be discarded. However, the total length and the fraction of mapped reads were quite close to the initial assembly, and the fraction of the proper read pairs in the final assembly was slightly improved, compared to the initial input.

Table 1. Assembly evaluation metrics and results. The W303 strain assembly in each step of IMAP, including the final one, was validated using evaluation metrics such as N50, the length of the shortest contig (MIN), the length of the longest contig (MAX), and total length. We also calculated the fraction of mapped reads and the portion of read pairs that were properly mapped.

Dataset (W303)		MIN (bp)	MAX (bp)	N50 (bp)	Total length (bp)	Mapped reads (%)	Proper pairs (%)
De novo assembly	SPAdes	80	515,973	187,035	13,901,101	99.40	96.24
	MaSurCa	300	784,921	273,283	11,838,299	82.86	97.88
	SOAPdenovo2	200	61,911	13,286	11,749,637	50.60	88.48
RACA assembly	RACA-SPAdes	80	1,058,428	716,084	13,905,771	99.40	96.24
	RACA-MaSurCa	300	1,436,612	706,991	11,842,202	82.86	97.88
	RACA-SOAPdenovo2	200	1,076,849	69,631	11,772,637	50.60	88.49
Meta assembly	Meta	80	1,448,740	702,641	13,773,679	98.56	96.19
Final assembly	Corrected assembly	80	1,450,556	705,629	13,847,490	98.57	97.10
PacBio	PacBio	3,688	1,575,129	929,095	12,433,409	99.15	98.73

<https://doi.org/10.1371/journal.pone.0221858.t001>

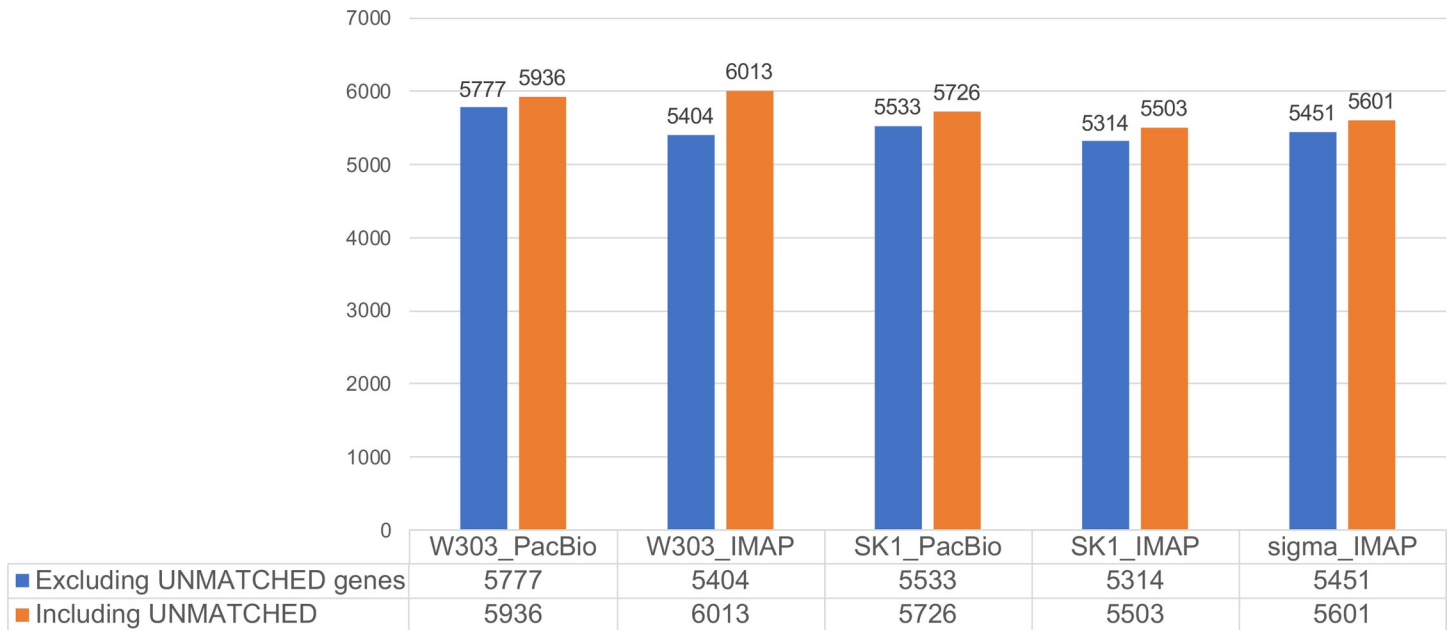


Fig 2. Assembly evaluation using gene annotation. We validated the quality of our assemblies by measuring annotation completeness using AGAPE gene annotation [22]. AGAPE determines genes that are longer than 300 bases, and matches each gene to the NCBI database. If there are no genes found in the NCBI database or genes are partially matched less than a threshold value [22], they are marked as UNMATCHED, and parts of these genes are likely to be missed, because of assembly or annotation errors (note that although UNMATCHED genes could include novel genes, most are partially assembled genes in the *S. cerevisiae* strain genome [22]). We evaluated the annotation completeness for the assemblies of the genomes of three yeast strains: W303, SK1, and Sigma1278b. For W303 and SK1, we compared the gene annotation results of our IMAP assembly with that of the PacBio assembly using AGAPE.

<https://doi.org/10.1371/journal.pone.0221858.g002>

We also applied IMAP for other commonly used yeast strains: SK1 (at approximately 530× coverage; S2 Table) and Sigma1278b (at approximately 191× coverage; S3 Table). For SK1, we also evaluated our IMAP output against a recently available PacBio assembly. For Sigma1278b, our improved assembly could serve as a reference sequence for the strain genome, until any long-read sequencing data are generated. Our assembly of Sigma1278b used the same sequencing data that was built in the original study [22]. Therefore, we could observe a drastic increase in the N50 from 109,268 bp [22] to 719,378 bp in our IMAP results.

We also examined the impact of the choice of a reference genome in our pipeline. We used the SK1 PacBio assembly as another reference for RACA instead of S288C strain genome for W303 sequencing data using IMAP, and the W303 PacBio assembly for SK1. The statistics of these IMAP assembly results for W303 and SK1 (S4 and S5 Tables) were quite similar to the original results (Table 1 and S2 Table) with the S288C strain sequence for the RACA reference. This shows that the assembly output is not much influenced by the choice of the RACA reference as long as it is close enough to the target genome.

We validated the accuracy of the assemblies by measuring annotation completeness using AGAPE gene annotation (Fig 2). We annotated genes with both our IMAP assemblies and PacBio assemblies for the yeast W303, and SK1 strains using AGAPE, and compared their annotation results. AGAPE detects genes longer than 300 bases in a sequence, and searches for their gene names in the NCBI database [22]. If a gene detected by AGAPE is not found in the NCBI database, it is marked as “UNMATCHED”, and part of the gene is likely to be missed because of assembly and annotation errors. We counted the number of genes, both including and excluding the UNMATCHED genes. When the UNMATCHED genes were excluded, our IMAP assemblies for W303 and SK1 had 93.5 and 96%, respectively, of the genes to be determined, relative to the genes annotated using the PacBio assemblies. The initial *de novo* draft

assemblies missed scaffolds that contained paralogous genes (480 yeast genes in 197 paralogous groups); therefore, some of them were missed in our IMAP assembly in the AGAPE annotation: 233 fewer paralogous genes for W303, and 177 fewer for SK1 were detected in the IMAP assemblies than in the PacBio assemblies. In most cases, this explained the difference in the number of annotated genes between the IMAP and PacBio assemblies (note that repetitive genomic segments that include similar gene copies are known to be challenging for *de novo* assembly using short-read sequencing only).

For Sigma1278b, we used the same sequencing data as used by the original study [22]. We compared our annotation results with those of the annotation performed in the original study [22]. Our IMAP assembly had 5,451 genes annotated, while the original study [22] reported 5,224 genes, excluding the UNMATCHED genes. This illustrates the improvement in assembly quality gained using IMAP.

When we compared the number of annotated genes among the three *S. cerevisiae* strains, we expected similar numbers; however, there was some variation. This might have been caused by the variation in the sequencing depth and quality of the original sequencing data. We also applied IMAP for three commonly used fungal strains: *Aspergillus nidulans* A713, *Neurospora crassa* 73, and *Thielavia terrestris* CBS 492.74. S6–S8 Tables summarize the statistics of their assemblies.

Synteny mapping of W303 and S288C

We mapped large-scale conserved segments, called synteny blocks, between the IMAP and PacBio assemblies for the SK1 and W303 strain genomes, using LASTZ and software tools available in mySyntenyPortal [32]. Fig 3 shows that each chromosome in our IMAP result is almost exactly mapped to one in the PacBio assembly, excluding two synteny blocks in W303. The additional panel at the right side of Fig 3 shows the synteny blocks in a different order. The discrepancy in the synteny blocks between the IMAP and PacBio assemblies for W303 requires further investigation. We also calculated the percent identity of the alignment between the IMAP and PacBio assemblies generated using LASTZ, and found 100% identity.

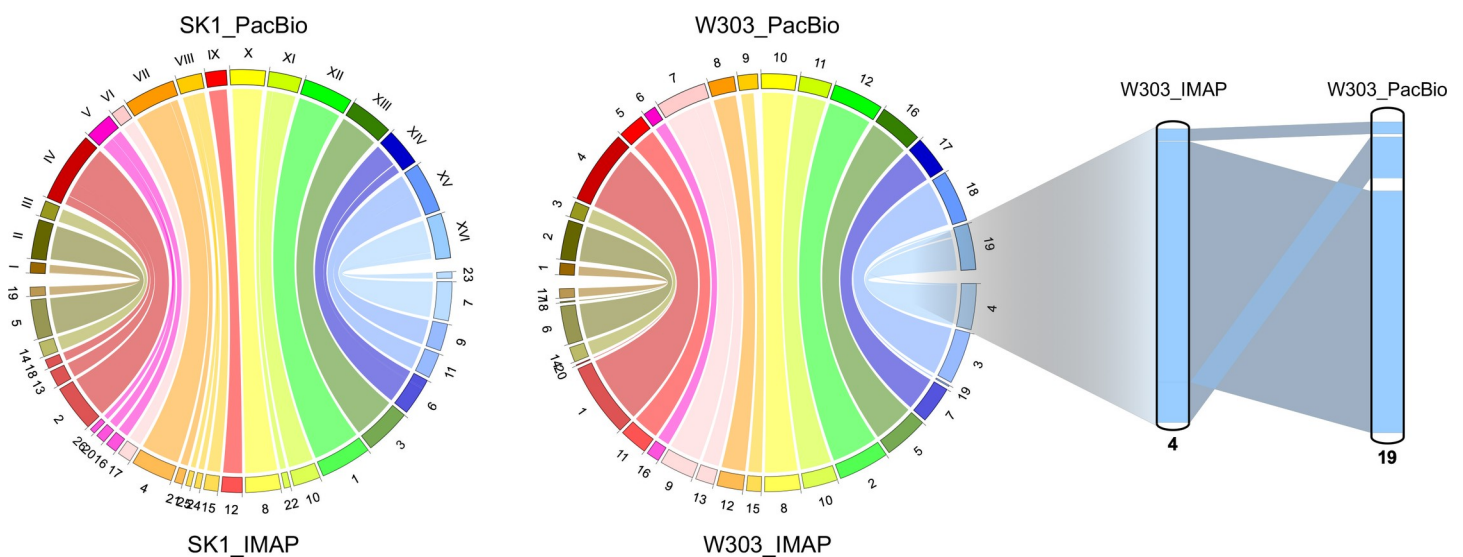


Fig 3. Synteny mapping of the IMAP and PacBio assemblies for strains SK1 and W303. Each color indicates a pair of genomic segments aligned together between the two genomes. Each scaffold at the chromosome level of the IMAP assembly is almost exactly mapped to a chromosome sequence in the PacBio assembly, except for a couple of synteny blocks in W303, as shown in the right panel. In general, there is no significant large-scale structural difference between the IMAP and PacBio assemblies.

<https://doi.org/10.1371/journal.pone.0221858.g003>

In the comparison of the W303 IMAP assembly and the S288C reference sequence using mySyntenyPortal, synteny blocks between the two assemblies were consistent in terms of their order. This agreed with a previous study for the W303 strain genome [34]. In general, there was no significant large-scale difference in terms of the order of synteny blocks between the IMAP and PacBio assemblies.

Correlation of sequencing depth and assembly quality

The W303 sequence datasets had very deep fold coverage (4,000×); therefore, we assessed the effects of sequencing fold coverage on assembly quality. We sampled sequence datasets randomly to test the assembly for 10×, 20×, 50×, 100×, 200×, 500×, 1,000×, and 4,000× coverage, respectively. The initial datasets were composed of 40 population samples, each of which had 100× coverage. Merging datasets for several samples generated data with multiples of 100× coverage. For example, merging two samples produced sequence data with 200× coverage, while merging five samples produced 500× coverage. To generate the 10×, 20×, and 50× coverage data, we randomly sampled reads from the 4,000× coverage data.

For each dataset, we applied two *de novo* assembly pipelines: SPAdes and MaSurCA. The N50 of the SPAdes assembly was highest at 100× coverage, while that of MaSurCA was highest at 20× coverage (Fig 4A; S9 Table). In terms of the number of scaffolds, SPAdes showed the fewest scaffolds at 500× coverage, while MaSurCA showed the fewest at 20× coverage (Fig 4B; S9 Table). In principle, deeper coverage should generate better quality assemblies. However, we found no correlation between assembly quality and sequencing coverage above 10× coverage. This analysis suggests that 10× coverage may be sufficient enough to obtain a high-quality sequence assembly of yeast strains from short-read sequencing data, and that there is high variability in assembly quality, depending on the initial *de novo* assemblers.

Description of the usage of IMAP and influence of parameters

Users can install the IMAP package including all prerequisite programs using a Docker image. Docker can be installed by following the instruction in <https://www.docker.com/>. Once

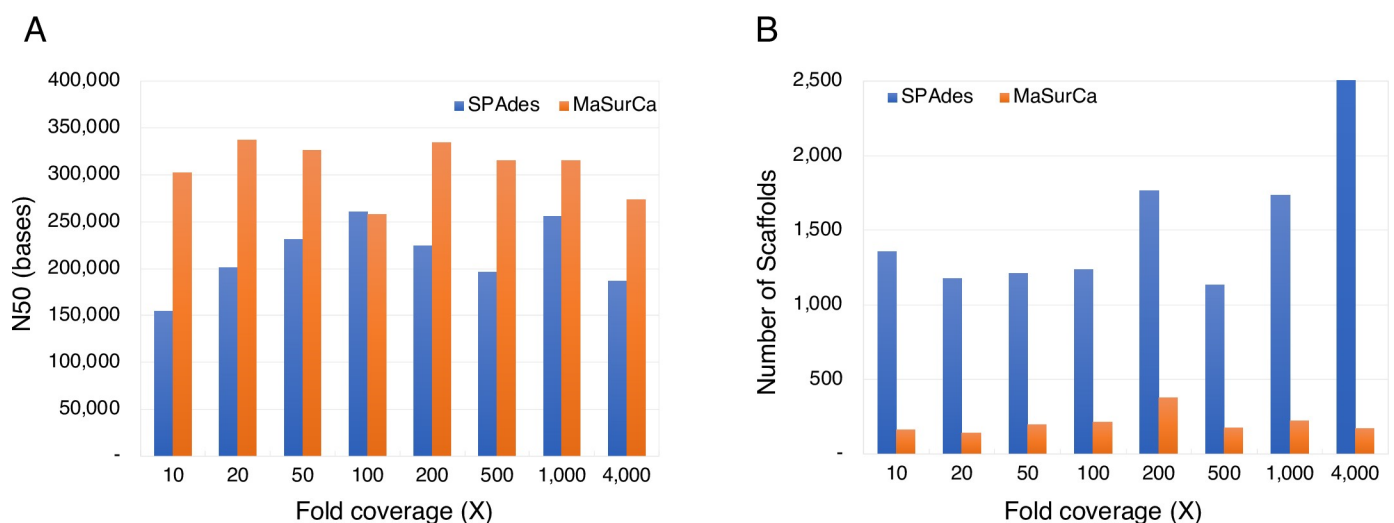


Fig 4. The effects of sequencing fold coverage on assembly quality. We generated sequence datasets with various sequencing fold coverages via random sampling of very deep coverage sequencing data, and applied two assembly pipelines: SPAdes and MaSurCA. (A) The assembly using SPAdes with 100× coverage showed the highest N50 value, while the one using MaSurCA with 20× coverage had the highest N50. (B) Sequencing data at 500× coverage yielded an assembly with the fewest scaffolds using SPAdes, while that with 20× coverage generated the fewest using MaSurCA.

<https://doi.org/10.1371/journal.pone.0221858.g004>

Docker is ready, download the source of IMAP using the command “git clone <https://github.com/jkimlab/IMAP.git>”, build a Docker image at the directory where “Dockerfile” is located via command “docker build -t [image_name]”, and run the image and create a container using command “docker run -it [image_name] /bin/bash”. Then users can obtain an assembly result using the IMAP Perl script. To run the IMAP Perl script, a parameter file needs to be provided, and the file format and details are available in <https://github.com/jkimlab/IMAP>.

There are some major parameters that need to be determined by users such as resolution, which is the minimum length of similar assembly regions (set as 10 Kbp by default), and the genome sequences of a reference species. We examined how these parameters affect the resulting meta-assembly (Fig 5). To see the influence of these parameters, we built the yeast strain W303 and SK1 assemblies using IMAP with different parameter settings: four different resolutions (5 Kbp, 10 Kbp, 25 Kbp, and 50 Kbp), and four different reference genomes (*S. cerevisiae* strain S288C, *S. cerevisiae* strain SK1, *S. paradoxus*, and *S. kudriavzevii* for the W303 assembly, and *S. cerevisiae* strain S288C, *S. cerevisiae* strain W303, *S. paradoxus*, and *S. kudriavzevii* for the SK1 assembly). *S. cerevisiae* strain S288C is closest and *S. kudriavzevii* furthest as the reference for W303 and SK1 [35]. The chromosome-level assemblies of those reference genomes were downloaded from the NCBI database (Accession number GCA_000146045.2 for S288C, GCA_002163515.1 for W303, GCA_002057885.1 for SK1, GCA_002079055.1 for *S. paradoxus*, and GCA_003327635.1 for *S. kudriavzevii*).

We compared our various assembly results with the W303 and SK1 PacBio assemblies using the assembly assessment tools QUILT [36] and GMASS [37]. We measured the number of misassembled contigs (Fig 5A and 5B; the number of contigs that contain misassembly events relative to the PacBio assembly) and NA50 (Fig 5C and 5D; N50 after fixing misassemblies) using QUILT, and a structural similarity score (Fig 5E and 5F; structural similarity in terms of synteny mapping of our meta-assembly against the PacBio assembly) using GMASS. In terms of the number of misassembled contigs, the reference SK1 with 25 Kbp resolution showed the lowest number for the W303 meta-assembly, and the reference S288C with 5 Kbp resolution for SK1. For NA50, references S288C with 5 Kbp resolution and SK1 with 10 Kbp resolution produced results with the highest value for the W303 meta-assembly, and references S288C and W303 with 5 Kbp resolution for SK1. In the case of the GMASS score, reference S288C with 10 Kbp resolution generated the most similar synteny structures to the PacBio assembly for the W303 meta-assembly, and reference W303 with 5 Kbp resolution for the SK1 result.

These results suggest that using closer reference sequences with 5 or 10 Kbp resolution produces assemblies in high quality, but references at further distances can still generate assemblies in reasonable quality. In addition, the resolution affects the quality of meta-assembly results a little more than the reference setting. The default value (10 Kbp) seems reasonable, but we suggest that users build assemblies using a few different resolution settings.

Conclusion

We developed a new pipeline, called IMAP, to integrate various *de novo* assembly tools and post-scaffolding steps to extend and order scaffolds to build a chromosome-level genome assembly. We rigorously validated various combinations of approaches in each step and identified the optimal combination of all steps to build the chromosome-level assembly using only short-read sequencing data. We demonstrated the utility of our pipeline by constructing chromosome level assemblies for *S. cerevisiae* non-reference strains: W303, SK1, and Sigma1278b, which are the commonly used laboratory strains in yeast genetics studies, and for other fungal genomes, which serve as model organisms in cell biology and genetics studies. There are

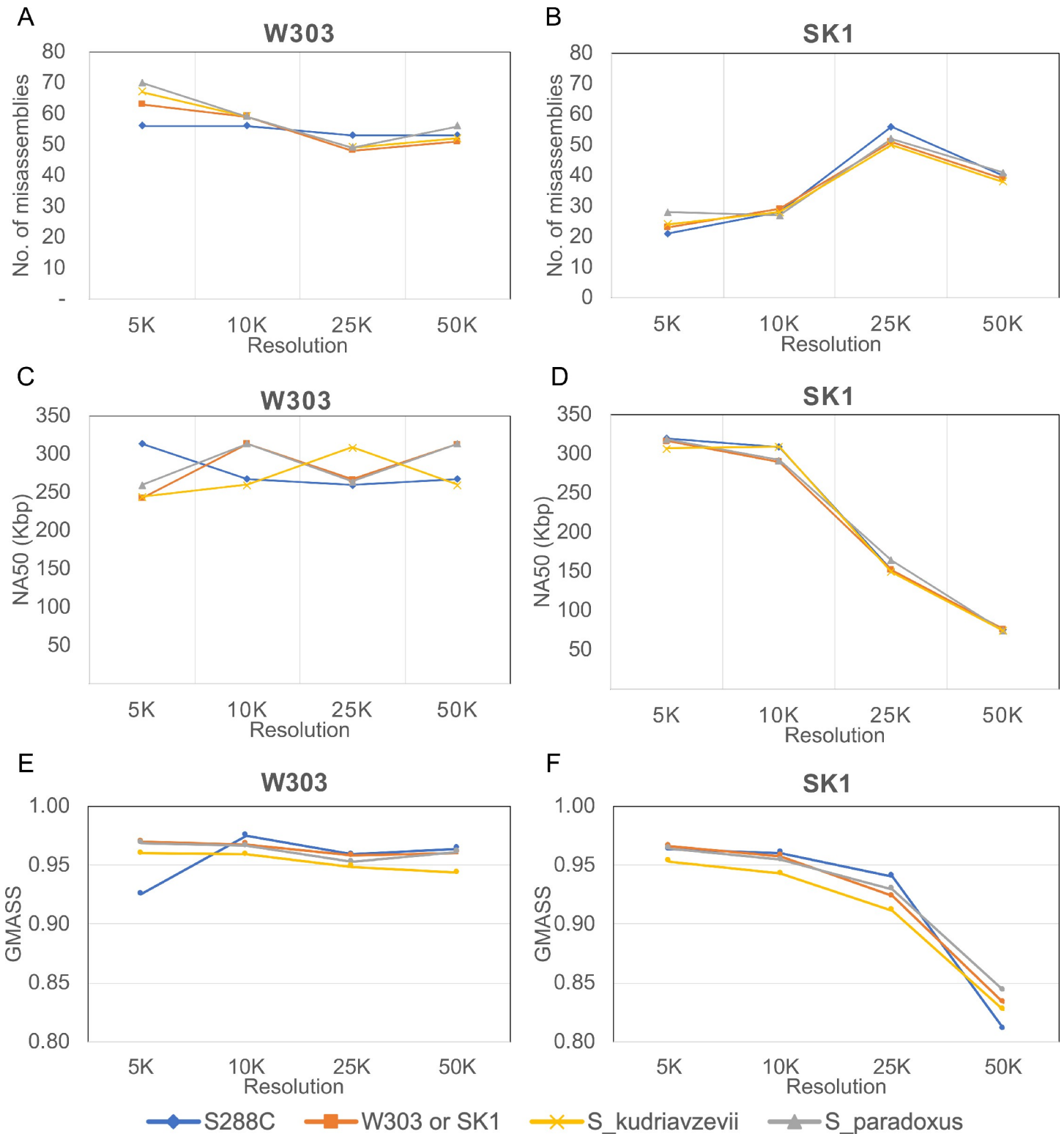


Fig 5. The effects of IMAP parameters on the quality of meta-assemblies. Meta-assemblies of yeast strain W303 and SK1 were generated using different settings of two IMAP parameters (reference species and resolution), and compared with their corresponding PacBio assembly in terms of the number of misassemblies (A and B), NA50 (C and D), and assembly structural similarity (E and F). In this comparison, four different references (S288C, SK1, *S. paradoxus*, *S. kudriavzevii* for the W303 meta-assembly, and S288C, W303, *S. paradoxus*, *S. kudriavzevii* for the SK1 meta-assembly), and four different resolutions (5, 10, 25, and 50 Kbp) were used.

<https://doi.org/10.1371/journal.pone.0221858.g005>

missing bases in our assemblies of repetitive genomic segments, such as those in Ty elements and telomeric regions, which are known to be very challenging for sequencing assembly. However, the scaffolds of the initial *de novo* draft assemblies were combined and ordered accurately.

In general, much time and effort are required for manual curation and additional computation to complete a reference sequence for a target genome, and to confirm that that sequence could serve as the reference for the community via a consortium. Our pipeline can reduce the time spent on the manual tasks required for curation and discussion. We believe that IMAP could serve as a useful pipeline for collaboration between genomics and bio-curation communities to build reference genome sequences for various organisms. Our study could also help to identify novel evolutionary patterns using chromosome-level assemblies at various temporal or spatial points, using short-read sequencing data.

Long-read sequencing technologies, such as PacBio and Nanopore, are becoming more available, and high-quality assemblies can be obtained with the integration of long- and short-read data. However, for some projects, these new methods are still too expensive, such as for short-term evolution studies using sequencing data, and because of budget limitations that dictate genomics analyses using short-read sequencing data only. IMAP will be helpful in these cases. Assemblies from long-read sequencing can also be easily plugged into our pipeline to improve the assembly quality, by combining long-read results with short-read data. In the future, we will continue to improve our pipeline to build chromosome-level assemblies for larger genomes, such as humans and cancer samples. Our pipeline can accelerate the determination of temporal evolutionary dynamics in cancer genomes at fine-scale resolution, using comparative analyses using sequences assembled at the chromosome level.

Supporting information

S1 Table. Information of data for three fungal strains.

(DOCX)

S2 Table. Evaluation results for the SK1 strain assembly.

(DOCX)

S3 Table. Evaluation results for the Sigma 1278b strain assembly.

(DOCX)

S4 Table. Evaluation results for the W303 strain assembly with reference SK1 PacBio instead of S288C.

(DOCX)

S5 Table. Evaluation results for the SK1 strain assembly with reference W303 PacBio instead of S288C.

(DOCX)

S6 Table. Statistics of the assembly of the fungal strain *Aspergillus nidulans* A713.

(DOCX)

S7 Table. Statistics of the assembly of the fungal strain *Neurospora crassa* 73.

(DOCX)

S8 Table. Statistics of the assembly of the fungal strain *Thielavia terrestris* CBS 492.74.

(DOCX)

S9 Table. Statistics of the assemblies of the yeast strain W303 generated using sequencing data with various coverages.

(DOCX)

Author Contributions**Conceptualization:** Giltae Song, Jaebum Kim.**Data curation:** Giltae Song, Jongin Lee, Juyeon Kim, Seokwoo Kang, Hoyong Lee, Daehong Kwon, Daehwan Lee, Jaebum Kim.**Formal analysis:** Jongin Lee, Juyeon Kim, Seokwoo Kang, Hoyong Lee, Daehong Kwon, Daehwan Lee, Jaebum Kim.**Funding acquisition:** Giltae Song, Jaebum Kim.**Investigation:** Giltae Song.**Methodology:** Jongin Lee.**Project administration:** Giltae Song, Jaebum Kim.**Software:** Giltae Song, Jongin Lee, Juyeon Kim, Seokwoo Kang, Hoyong Lee, Daehong Kwon, Daehwan Lee.**Supervision:** Jaebum Kim.**Validation:** Gregory I. Lang, J. Michael Cherry.**Writing – original draft:** Giltae Song, Jaebum Kim.**Writing – review & editing:** Giltae Song, Gregory I. Lang, J. Michael Cherry, Jaebum Kim.**References**

1. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, et al. Global diversity, population stratification, and selection of human copy-number variation. *Science*. 2015; 11:349(6253).
2. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, Marjoram P, Siegmund K, Press MF, Shibata D, et al. A Big Bang model of human colorectal tumor growth. *Nat Genet*. 2015; 47(3):209–16. <https://doi.org/10.1038/ng.3214> PMID: 25665006
3. SF Levy, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*. 2015; 519(7542):181–6. <https://doi.org/10.1038/nature14279> PMID: 25731169
4. Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*. 2013; 500(7464):571–4. <https://doi.org/10.1038/nature12344> PMID: 23873039
5. Matheson K, Parsons L, Gammie A. Whole-Genome Sequence and Variant Analysis of W303, a Widely-Used Strain of *Saccharomyces cerevisiae*. *G3: Genes, Genomes, Genetics* 2017; 7(7):2219–26.
6. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res*. 2010; 20(9):1165–73. <https://doi.org/10.1101/gr.101360.109> PMID: 20508146
7. Teng JLL, Yeung ML, Chan E, Jia L, Lin CH, Huang Y, Tse H, Wong SSS, Sham PC, Lau SKP, et al. PacBio But Not Illumina Technology Can Achieve Fast, Accurate and Complete Closure of the High GC, Complex *Burkholderia pseudomallei* Two-Chromosome Genome. *Frontiers in Microbiology*. 2017; 8:1448. <https://doi.org/10.3389/fmicb.2017.01448> PMID: 28824579
8. Paajanen P, Kettleborough G, Lopez-Girona E, Giolai M, Heavens D, Baker D, Lister A, Cugliandolo F, Wilde G, Hein I et al. A critical comparison of technologies for a plant genome sequencing project. *Giga-science*. 2019; <https://doi.org/10.1093/gigascience/giy163> PMID: 30624602
9. Wences AH, Schatz MC. Metassembler: merging and optimizing *de novo* genome assemblies. *Genome Biol*. 2015; 16:207. <https://doi.org/10.1186/s13059-015-0764-4> PMID: 26403281

10. Mirebrahim H, Close TJ, Lonardi S. *De novo* meta-assembly of ultra-deep sequencing data. *Bioinformatics*. 2015; 31(12):i9–16. <https://doi.org/10.1093/bioinformatics/btv226> PMID: 26072514
11. Kolmogorov M, Raney B, Paten B, Pham S. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*. 2014; 30(12):i302–9. <https://doi.org/10.1093/bioinformatics/btu280> PMID: 24931998
12. Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lió P, Crescenzi P, Fani R, Fondi M. MeDuSa: a multi-draft based scaffolder. *Bioinformatics*. 2015; 31(15):2443–51. <https://doi.org/10.1093/bioinformatics/btv171> PMID: 25810435
13. Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge RL, Auvil L, Capitanu B, Zhang G, Lewin HA, et al. Reference-assisted chromosome assembly. *Proc Natl Acad Sci U S A*. 2013; 110(5):1785–9. <https://doi.org/10.1073/pnas.1220349110> PMID: 23307812
14. Kim J, Farré M, Auvil L, Capitanu B, Larkin DM, Ma J, Lewin HA. Reconstruction and evolutionary history of eutherian chromosomes. *Proc Natl Acad Sci U S A*. 2017; 114(27):e5379–e5388. <https://doi.org/10.1073/pnas.1702012114> PMID: 28630326
15. Yue J, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino M, Fischer G, Durbin R, Liti G. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet*. 2017; 49(6):913–924. <https://doi.org/10.1038/ng.3847> PMID: 28416820
16. Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B, et al. Genotype to phenotype: a complex problem. *Science* 2010; 328(5977):469. <https://doi.org/10.1126/science.1189015> PMID: 20413493
17. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglu S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* 2005; 438: 1105–1115. <https://doi.org/10.1038/nature04341> PMID: 16372000
18. Znameroski EA, Coradetti ST, Roche CM, Tsai JC, Iavarone AT, Cate JH, Glass NL. Induction of lignocellulose-degrading enzymes in *Neurospora crassa* by cellodextrins. *Proc Natl Acad Sci U S A* 2012; 109(16):6012–7. <https://doi.org/10.1073/pnas.1118440109> PMID: 22474347
19. Berka RM, Grigoriev IV, Otillar R, Salamov A, Grimwood J, Reid I, Ishmael N, John T, Darmond C, Moisan MC, et al. Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris*. *Nat. Biotechnol*. 2011; 29:922–927. <https://doi.org/10.1038/nbt.1976> PMID: 21964414
20. Sasaki M, Tischfield SE, van Overbeek M, Keeney S. Meiotic recombination initiation in and around retrotransposable elements in *Saccharomyces cerevisiae*. *PLoS Genet*. 2013; 9(8):e1003732. <https://doi.org/10.1371/journal.pgen.1003732> PMID: 24009525
21. Bergström A, Simpson JT, Salinas F, Barré B, Parts L, Zia A, Nguyen Ba AN, Moses AM, Louis EJ, Mustonen V, et al. A high-definition view of functional genetic variation from natural yeast genomes. *Mol Biol Evol*. 2014; 31(4):872–88. <https://doi.org/10.1093/molbev/msu037> PMID: 24425782
22. Song G, Dickins BJ, Demeter J, Engel S, Gallagher J, Choe K, Dunn B, Snyder M, Cherry JM. AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of *Saccharomyces cerevisiae*. *PLoS ONE* 2015; 10(5):e0129184. <https://doi.org/10.1371/journal.pone.0129184> PMID: 26017550
23. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res*. 2011; (12):2224–41.
24. Zimin A, Marçais G, Puiu D, Roberts M, Salzberg S, Yorke J. The MaSuRCA genome assembler. *Bioinformatics*. 2013; 29(21):2669–77. <https://doi.org/10.1093/bioinformatics/btt476> PMID: 23990416
25. Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A, Lesin V, Nikolenko S, Pham S, Prijbelski A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012; 19(5):455–77. <https://doi.org/10.1089/cmb.2012.0021> PMID: 22506599
26. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*. 2012; 1(1):18. <https://doi.org/10.1186/2047-217X-1-18> PMID: 23587118
27. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014; 9(11):e112963. <https://doi.org/10.1371/journal.pone.0112963> PMID: 25409509
28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. 2010; 20(9):1297–303.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 2009; 25:1754–60. <https://doi.org/10.1093/bioinformatics/btp324> PMID: 19451168

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 1990; 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) PMID: 2231712
31. Harris RS. Improved pairwise alignment of genomic DNA. Ph.D. Thesis 2007; The Pennsylvania State University.
32. Lee Jongin, Lee Daehwan, Sim Mikang, Kwon Daehong, Kim Juyeon, Ko Younhee, and Kim Jaebum. mySyntenyPortal: an application package to construct websites for synteny block analysis. 2018; 19 (1):216.
33. Song G, Balakrishnan R, Binkley G, Costanzo MC, Dalusag K, Demeter J, Engel S, Hellerstedt ST, Karra K, Hitz BC, et al. Integration of new alternative reference strain genome sequences into the *Saccharomyces* Genome Database. *Database (Oxford)*. 2016; pii: baw074.
34. Ralser M, Kuhl H, Ralser M, Werber M, Lehrach H, Breitenbach M, Timmermann B. The *Saccharomyces cerevisiae* W303-K6001 cross-platform genome sequence: insights into ancestry and physiology of a laboratory mutt. *Open Biol* 2012; 2(8):120093. <https://doi.org/10.1098/rsob.120093> PMID: 22977733
35. Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* 2018; 556(7701):339–344. <https://doi.org/10.1038/s41586-018-0030-5> PMID: 29643504
36. Gurevich A, Saveliev V, Vyahhi N and Tesler G. QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 2013; 29(8):1072–1075. <https://doi.org/10.1093/bioinformatics/btt086> PMID: 23422339
37. Kwon D, Lee J and Kim J. GMASS: a novel measure for genome assembly structural similarity. *BMC Bioinformatics* 2019; 20:147. <https://doi.org/10.1186/s12859-019-2710-z> PMID: 30885117