The *Saccharomyces cerevisiae* W303-K6001 cross-platform genome sequence: insights into ancestry and physiology of a laboratory mutt

Markus Ralser¹,²,†, Heiner Kuhl²,†, Meryem Ralser², Martin Werber², Hans Lehrach², Michael Breitenbach³ and Bernd Timmermann²

¹Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK
²Department of Vertebrate Genomics and Next Generation Sequencing Core Facility, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany
³Department of Cell Biology, University of Salzburg, 5020 Salzburg, Austria

1. Summary

*Saccharomyces cerevisiae* strain W303 is a widely used model organism. However, little is known about its genetic origins, as it was created in the 1970s from crossing yeast strains of uncertain genealogy. To obtain insights into its ancestry and physiology, we sequenced the genome of its variant W303-K6001, a yeast model of ageing research. The combination of two next-generation sequencing (NGS) technologies (Illumina and Roche/454 sequencing) yielded an 11.8 Mb genome assembly at an N50 contig length of 262 kb. Although sequencing was substantially more precise and sensitive than whole-genome tiling arrays, both NGS platforms produced a number of false positives. At a 378× average coverage, only 74 per cent of called differences to the S288c reference genome were confirmed by both techniques. The consensus W303-K6001 genome differs in 8133 positions from S288c, predicting altered amino acid sequence in 799 proteins, including factors of ageing and stress resistance. The W303-K6001 (85.4%) genome is virtually identical (less than equal to 0.5 variations per kb) to S288c, and thus originates in the same ancestor. Non-S288c regions distribute unequally over the genome, with chromosome XVI the most (99.6%) and chromosome XI the least (54.5%) S288c-like. Several of these clusters are shared with *S. cerevisiae* S1278B, another widely used S288c-related model, indicating that these strains share a second ancestor. Thus, the W303-K6001 genome pictures details of complex genetic relationships between the model strains that date back to the early days of experimental yeast genetics. Moreover, this study underlines the necessity of combining multiple NGS and genome-assembling techniques for achieving accurate variant calling in genomic studies.

2. Introduction

Ageing is common to all living organisms, and knowledge on biochemical and genetic components that accelerate or delay this process are of immense
medical interest. Because the lifespan of mammalian organisms is considerably long, short-living species such as the yeast *Saccharomyces cerevisiae* are popular models in experimental ageing research [1,2]. Widely used measures of yeast ageing are (i) chronological lifespan, defined as survival of a stationary culture at 30°C [3], or in its special case ‘hibernating lifespan’ at 4°C [4]; and (ii) replicative lifespan (RLS), defined as the number of cell cycles an individual yeast cell can complete [5]. Determination of RLS is time-consuming and technically challenging, as it requires continuous micromanipulation of the target strains [6,7], or single cell trapping and microscopy [8]. To simplify RLS analysis, a genetic assay based on the yeast strain W303-K6001 was introduced around a decade ago, and has become popular [9–14]. The W303-K6001 RLS assay bases on differential expression of the essential CDC6 gene. Placed under control of two promoters, CDC6 is always expressed in mother cells (HO promoter), but expressed in daughters only when they grow on galactose (GAL1 promoter) [11]. Thus, on glucose, daughters arrest, whereas mothers continue to divide until senescence. The cell number in a W303-K6001 glucose microcolony is therefore a direct—and the stationary biomass of a W303-K6001 glucose culture an indirect—measure of RLS [9].

W303-K6001 is a direct descendant of the yeast strain W303-1A, which is commonly used in biomedical research laboratories around the world [15]. This strain is a laboratory mutt that was generated through a series of strain crosses, mainly conducted by Rodney Rothstein during his PhD thesis. W303 derivatives therefore have a complex and not thoroughly documented ancestry. The founding W303 strain W303-1A was derived from W301-18A [16], which was transformed by a plasmid containing the HO gene [17]. W301-18A itself originated from crosses of W87 derivatives [18,19], which are themselves partially descended from yeast strain S288c, the source of the S. cerevisiae reference genome [20]. W303-1A further contains genetic material from yeast strains D311-3A [21,22] and a historical yeast strain, D190-9C. Nothing seems to be known about D190-9C, except that it has originated in the laboratory of Jack Szostak (personal communication of R. Rothstein to the *Saccharomyces* genome database, SGD [23]). This complex genealogy mixes with the ancestry of other laboratory strains commonly used today, such as Z1278B and SK1 [24].

Proteomic profiling and tiling microarrays indicated that W303-1A derivatives maintain high similarity to S288c [24,25]. Population genomics confirmed these large genetic similarities, but also revealed the presence of substantial non-S288c material in the W303 genome. For example, on chromosome 2 on the left arm, there is a region similar to west African yeast strains, while a region on the right arm clustered with European strains. Surprisingly, regions that resembled Japanese sake strains were also found in the W303 genome [26]. This genetic divergence probably contributes to physiological differences that have been reported for W303 and S288c derivatives BY4741 and BY4742, the most widespread S288c descendants [27]. These strains differ not only in important physiological parameters such as cell size and volume, but also in their relative plasma-membrane potential and tolerance to alkali-metal cations [28]. Moreover, although S288c strains and W303 have a relatively similar RLS, RLS of W303-K6001 is shortened on glucose media [9,29], and aged W303 cells have considerably larger volumes. As the average protein concentration was reversely changed from 64 pg μm⁻³ (BY4741) to 24 pg μm⁻³ (W303), it was concluded that senescent W303 cells possess larger vacuoles [30].

To provide a comprehensive basis for the interpretation of ageing experiments performed with W303-K6001, and to understand genetic origins and physiological properties of W303-1A derivatives, we constructed a reference genome sequence for W303-K6001. Both high-coverage pyrosequencing (Roche/454) and sequencing by synthesis (Illumina) technologies were then used for highly accurate variant calling against the S288c reference genome. The high sequencing depth (in total 378× average coverage) achieved with both platforms on this 11.8 Mbp haploid genome facilitated a profound comparison on their performances for de novo assembly and variant calling. Roche/454 sequencing performed better in de novo and reference-guided assembly, indicating that the much higher coverage of the short-read technology could not compensate for read length. Both sequencing strategies, however, called a significant number of false positives in variant detection. Merging the outputs of both platforms reduced this number markedly, indicating that, at present, parallel sequencing with more than one NGS technology is essential for generating precise reference genomes and for avoiding false positives in variant detection.

The proportion of the W303-K6001 genome that is highly similar to its main ancestor, S288c, was 85.4 per cent, whereas the remaining genetic material is of different genetic origin. In part, these non-S288c clusters are shared with Σ1278B, another commonly used mutt yeast strain. These regions encode for 799 proteins that have altered amino acid sequence compared with S288c.

### 3. Results and discussion

#### 3.1. Sequencing and assembly of the W303-K6001 genome

We have previously presented a draft genome sequence for W303-K6001 and one of its variants, K6001-B7. Isolated after chemical mutagenesis, K6001-B7 is a W303-K6001 progeny that exhibits a dominantly inherited increase in resistance to oxidative stress, while having a premature ageing phenotype [13]. This genome sequence was generated by massive parallel pyrosequencing (Roche) and led to the identification of 13 single-nucleotide exchanges that distinguish the parent and progeny. One of these, a C→T transition in the peroxiredoxin locus *TSAT* (*tsa1-B7*), was shown to be responsible for the stress-resistance phenotype. Here, this W303-K6001 genome draft was improved by including the newest version of the S288c genome (EF4, Ensembl annotated version of S288c genome R64-1-1, GenBank GCA_000146045.2) for reference-based genome assembly. Moreover, we resequenced W303-K6001 using sequencing by synthesis technology (Illumina). We combined the sequencing information obtained with both technologies to correct for platform-dependent sequencing errors (false-positive variant calls) as reported to occur in NGS experiments [31]. To keep experimental variation at a minimum, we used the very same DNA purifications for Roche and Illumina library preparation. At a median read length of 527 bp, we had collected 662.12 Mbp of sequence information with the Roche FLX pyrosequencing system. Analysing the W303-K6001 DNA on a Genome Analyzer Ix (GA Ix) with a 120 bp paired-end protocol yielded 3.9 Gb of sequencing data.
after quality clipping; on average, 101.22 bp per 120 bp GAIIx read were used. The insert size of the paired ends was 225 ± 60 bp (table 1).

Obtained sequencing information was then assembled using the Newbler mapper/assembler v. 2.6, CLC bio reference mapper (included in CLC Genomics Workbench v. 5.1) and SOAPdenovo (63mer v. 1.05). Using Newbler, we tested different strategies for de novo assembly of the W303-K6001 genome (table 2). First, we compared the performance of de novo assemblies using 454 data alone and in combination of 454 and Illumina data. Interestingly, de novo assembly of pure 454 data yielded the largest N50 contig size of 262 kb; adding higher coverage of the Illumina platform did not result in larger contigs. Indicated by a high number of shorter contigs, this might signify that the assembler software was limited in estimating the accurate genome size at this unusual high coverage. Moreover, the insert size distribution of the Illumina paired-end library was below the average read length of the 454 data, and thus did not provide significant additional information for resolving repetitive structures in the yeast genome. Nevertheless, scaffolds produced by the paired-end information of Illumina data resulted in better long-range continuity than the assembly of 454 single reads alone. To exclude that these findings were specific for the Newbler de novo assembler, we assembled the Illumina reads also by SOAPdenovo. After optimizing the k-mer value to 57, the results were still lagging behind the Newbler results, however (table 3).

Next, we assembled the different W303-K6001 datasets by mapping to the S. cerevisiae S288c reference sequence (table 3). We obtained higher contig sizes and better long-range contiguity than by the different de novo assembly approaches, reflecting the high similarity of the two strains. Also, short Illumina reads resulted in lower length coverage of the reference genome than using 454 reads. Additionally, we did apply the CLC bio reference mapper, which gave similar results for 454 and 454 + Illumina reference-guided assemblies, but was better than Newbler when assembling Illumina-only data (table 3).

For the generation of a W303-K6001 reference genome sequence, we combined de novo and reference-guided assembly. We did simulate paired-end reads 400 bp in length with insert sizes ranging from 2000 to 4000 bp, based on the reference-guided assembly (Newbler/454 data only) by applying the ‘simulate_reads’ tool (CLC bio). The simulated reads comprised a 10× genome coverage and were de novo assembled together with the 454 data. In this way, we did obtain the best de novo results, but they were still behind the reference-guided strategy. For this reason, we finally did scaffold the

### Table 1. Sequencing of W303-K6001 on two different NGS platforms.

<table>
<thead>
<tr>
<th></th>
<th>454 sequencing</th>
<th></th>
<th>Illumina sequencing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K6001</td>
<td>K6001-B7</td>
<td>total</td>
<td>K6001</td>
</tr>
<tr>
<td>high-quality reads</td>
<td>641 083</td>
<td>684 545</td>
<td>1 325 628</td>
<td>20 407 268</td>
</tr>
<tr>
<td>average read length</td>
<td>507.22</td>
<td>492.26</td>
<td>499.48</td>
<td>101.16</td>
</tr>
<tr>
<td>median read length</td>
<td>529.0</td>
<td>524.0</td>
<td>527.0</td>
<td></td>
</tr>
<tr>
<td>average insert size</td>
<td>single read</td>
<td>225.0</td>
<td>221.3</td>
<td>223.15</td>
</tr>
<tr>
<td>(paired end)</td>
<td>bases (Mb)</td>
<td>325 157</td>
<td>336 965</td>
<td>662 122</td>
</tr>
</tbody>
</table>

### Table 2. De novo assembly of the W303-K6001 genome.

<table>
<thead>
<tr>
<th></th>
<th>de novo assembly by SOAPdenovo 63mer</th>
<th>de novo assembly by Newbler v. 2.6</th>
<th>combination of reference-guided and de novo assembly by Newbler v. 2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>assembled data</td>
<td>Illumina</td>
<td>454</td>
<td>454 + 454 mapped contigs</td>
</tr>
<tr>
<td>contig statistics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of contigs</td>
<td>3095</td>
<td>477</td>
<td>2846</td>
</tr>
<tr>
<td>number of bases</td>
<td>11 819 873</td>
<td>11 637 892</td>
<td>13 865 032</td>
</tr>
<tr>
<td>average contig size</td>
<td>3819</td>
<td>24 398</td>
<td>4871</td>
</tr>
<tr>
<td>N50 contig size</td>
<td>39 717</td>
<td>66 531</td>
<td>42 713</td>
</tr>
<tr>
<td>largest contig size</td>
<td>165 547</td>
<td>260 577</td>
<td>164 236</td>
</tr>
<tr>
<td>scaffold statistics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of scaffolds</td>
<td>374</td>
<td>—</td>
<td>357</td>
</tr>
<tr>
<td>number of bases</td>
<td>11 351 824</td>
<td>—</td>
<td>11 856 226</td>
</tr>
<tr>
<td>average scaffold size</td>
<td>30 352</td>
<td>—</td>
<td>33 210</td>
</tr>
<tr>
<td>N50 scaffold size</td>
<td>68 612</td>
<td>—</td>
<td>102 849</td>
</tr>
<tr>
<td>largest scaffold size</td>
<td>226 549</td>
<td>—</td>
<td>386 246</td>
</tr>
</tbody>
</table>
contigs of the reference-guided assemblies according to their positions in the S288c genome. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ALAV00000000. The functional analysis bases on its first version, ALAV01000000. Moreover, the genome sequence in total, and gene-by-gene-wise, is accessible through the web interface of SGD (http://www.yeastgenome.org [23]).

3.2. Combining two next-generation sequencing technologies for SNV calling eliminates a surprisingly high number of platform-specific false positives

To detect variations between S288c and W303 genomes, we performed two independent variant callings using GSMapper 2.6 (Roche). Mapping the pyrosequencing data revealed 11 324 variations, whereas sequencing by synthesis predicted 10 130 differences between the genomes. To eliminate platform-specific errors, both result files were combined. This strategy confirmed solely a number of 9073 differences between the W303-K6001 and the S288c reference genome. These split into 8471 single nucleotide polymorphisms (SNPs), 280 single base insertions/deletions and 322 more complex variants (table 4).

Thus, combining both NGS techniques, but also mapping algorithms, markedly reduces the number of false-positive variant calls.

3.3. Comparison of the W303-K6001 genome assembly with whole-genome tiling arrays

The W303 genome has previously been analysed using whole-genome tiling arrays [32]. To be able to compare our SNP calling with this array, we used Newbler to map the W303-K6001 genome also to the SC genome version EF2 (Ensembl annotated version of R63-1-1, GCA_000146045.1), on which this study was based. Of the 9334 mutant positions, 583 (9.2%) were identified in the tiling array (figure 1). The overlap of the W303-K6001 SNVs with all other analysed yeast strains was around 2 per cent, except for the S288c reference genome.

Thus, the tiling array correctly identified the yeast strain and a number of its specific SNV positions. However, a lot of SNVs found by the tiling array study could not be confirmed by our data. In part, this can be explained by the limited accuracy of tiling arrays in detecting the exact position of a SNV in the genome (5 bp window). Interestingly, all analysed yeast strains (except the reference genome strain) on the tiling array share a similar number of variations to the W303-K6001 genome. This number might indicate errors in the EF2 version of the S288c reference genome, or in the design of the tiling array, and thus could represent noise. Consistent with this observation, mapping to EF4 instead of EF2 eliminated 261 SNV calls.
3.4. Physiological differences between BY4741 and W303-K6001 explained by its genome sequence

W303-K6001 has a short RLS [9]. This property might have an impact on ageing studies in general, but facilitates a quick assay to determine RLS phenotypes by simply counting the cell numbers in glucose microcolonies. The Newbler variant calling identified several mutations within ageing factors, identifying proteins that might contribute to this phenotype. Compared with S288c, 799 K6001-W303 proteins have altered sequence. In 432 proteins, only one or two residues differ. Large, these might represent natural allelic variations that cause only minor physiological effects. However, in 239 proteins, three or more amino acids are exchanged (see the electronic supplementary material, table S2), and the list of mutations continues with 41 small insertions/deletions, and 87 more complex variations. Several of the latter may lead to a loss of the protein function as they affect the reading frame. Identified mutant genes include ade2-1, trp1-1, can1-100, leu2-3,112 and several others (table 6). We found two neutralizing frameshifts within the telomere (table 6). Telo-

meregulation has been closely associated with ageing in vertebrates, and in yeast their length is kept constant during replicative ageing [35]. MTM1 is required for the mitochondrial activation of superoxide dismutase (SOD2) and oxidative stress resistance [36]. Thus, this complex mutation appears to have occurred more than once in the history of W303. It is likely that the altered amino acids sequence within the frameshifts has influence on the functionality of these genes.

Table 5. Auxotrophic marker mutations found in W303-K6001.

<table>
<thead>
<tr>
<th>allele name</th>
<th>locus</th>
<th>detected mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ade2-1</td>
<td>YOR128C</td>
<td>nonsense, glu64STOP</td>
</tr>
<tr>
<td>trp1-1</td>
<td>YDR007W</td>
<td>nonsense, glu83STOP</td>
</tr>
<tr>
<td>can1-100</td>
<td>YEL063C</td>
<td>frameshift, lys47</td>
</tr>
<tr>
<td>leu2-3,112</td>
<td>YCL020W</td>
<td>frameshift, gly83</td>
</tr>
<tr>
<td>his3-11,15</td>
<td>YOR202W</td>
<td>2x frameshift, ala70 and glu106</td>
</tr>
</tbody>
</table>

pathway is used for auxotrophic selection, and closely connected to the oxidative stress defence, glutathione as well as homocysteine metabolism [34]. The nonsense mutation terminates Met1p at its penultimate amino acid, and the two frameshifts are in close proximity so that the second one restores the open reading frame. Indeed, met1-W303 appears to be at least partially functional, as the strain is methionine prototroph (data not shown).

In this context, we noticed that the phenomena of co-occurring frameshift mutations that neutralize each other was not unique among gene ontology terms related to ageing (table 6). We found two neutralizing frameshifts within the telomere organizers TEL2 (12 SNPs and two neutralizing frameshifts), EST1 (two neutralizing frameshifts) and the manganese carrier MTM1 (two neutralizing frameshifts). Telomere organization has been closely associated with ageing in vertebrates, and in yeast their length is kept constant during replicative ageing [35]. MTM1 is required for the mitochondrial activation of superoxide dismutase (SOD2) and oxidative stress resistance [36]. Thus, this complex mutation appears to have occurred more than once in the history of W303. It is likely that the altered amino acids sequence within the frameshifts has influence on the functionality of these genes.

Other genes related to ageing that contain frameshift mutations include eight proteins involved in translation and the biogenesis of the small ribosomal subunit, and...
YHL008C, a protein that is involved in chloride ion uptake [37,38]. Thus, although the artificial expression of CDC6 might explain large parts of the W303-K6001 lifespan, the strain carries several other mutations within genes involved in this biological property.

### 3.5. Phylogeny of W303-K6001

As mentioned earlier, W303 is a mutt of different laboratory strains, including S288c/W87, D311-3A and D190-9C [18,19,21,22]. Proteomic profiling, tiling arrays and what is known about its history indicated that most of the W303 background is S288c-like [24,25]. The W303-K6001 genome sequence allowed us to define the regions that derived from S288c, as they are virtually identical to the reference genome (less than 0.5 SNVs per kb; figure 2a). In contrast, the sequence reveals distinct clusters of much higher genetic variability, identifying the genetic material derived from other parents. In total, the clusters with sequence divergence larger than 1 SNV per kb span 1744 kb, corresponding to 14.6 per cent of the W303-K6001 genome. The differences between the chromosomes are, however, relatively large. W303-K6001 chromosome XVI, for instance, is virtually identical to S288c, whereas just half of chromosome XI is S288c-like
The 454-genome draft sequence of W303-K6001 [11] (MATα, ade2-1, trpl-1, trp1-100, can1-100, leu2-3,112, his3-11,15, GAL, psi+, hcr:HO::CDC6 (at HO), cdc6:hisG, ura3::URA3 GAL-ubiR-CDC6 (at URA3) and K6001-B7 [13] (MATα, ade2-1, trpl-1, can1-100, leu2-3,112, his3-11,15, GAL, psi+, hcr:HO::CDC6 (at HO), cdc6:hisG, ura3::URA3 GAL-ubiR-CDC6 (at URA3) was published earlier [13].

5. Material and methods

The K6001-W303 genome sequence facilitates comprehensive insights into its physiology and genealogy. In comparing W303-K6001 and S288c genome sequences, cross-platform sequencing and mapping eliminated 3308 false-positive nucleotide calls and 2389 mapping artefacts. The resulting consensus genome sequence differs in 8133 positions (including 8049 SNPs) from the S288c genome. These data demonstrate that it is vital to reproduce genetic variations identified by different sequencing strategies. Generated by crossing mutt yeast strains W87, D311-3A and D190-9C, W303-K6001 remains closely related to S288c, sharing 85.4 per cent of its genome. Remaining genetic differences cause altered amino acid composition or reading frame in 799 proteins, some of high relevance for physiology and ageing. Individual studies now have to clarify to what extent these mutations contribute to the physiological differences between these common yeast backgrounds. Unequal genomic SNV distribution allowed conclusions on the W303 genealogy, and identified a close genetic relationship of W303 with Σ1278B. This strain also shares 47 per cent of its background with S288c, but overall differs from it in 3.2 SNPs per kb, which cause 44 genes to be uniquely essential to Σ1278B and 13 to S288c [40]. Because W303-K6001 and Σ1278B share genomic regions not found in their common ancestor, it is likely that they have a second mutual ancestor. Thus, W303 represents a partial hybrid of S288c and Σ1278B, shedding new light into the complex relationships of today’s widely used laboratory strains.

4. Concluding remarks

The W303 genome switches from a non-S288c and non-one exemplary breakpoint of the Chr XI cluster, where the genome sequence of another yeast strain commonly used in research, W303, was published earlier [13]. Confirming the results from Liti et al. [24], we illustrate CLUSTALW-generated alignments (figure 3c) and a corresponding distance diagram (figure 3b) are illustrated. Extending our search, we could then identify several other regions in the W303 genome that were similar to the Σ1278B genome (figure 3c). We illustrate one exemplary breakpoint of the Chr XI cluster, where the W303 genome switches from a non-S288c and non-Σ1278B region to a Σ1278B-like genome, indicating differing phylogenetic origins of this cluster (figure 3d). As Σ1278B shares part of its genealogy with S288c (47% of its genome does not differ from S288c [24]), these results indicate that W303 and Σ1278B share a second ancestor, or (less likely) that a third strain contributed to the genome of S288c after W303/Σ1278B split from the lineage.

Figure 3. W303-K6001 contains clusters that are identical to Σ1278B, but differ in S288c. (a) S288c is the main ancestor parent of W303 and Σ1278B; however, part of the non-S288c-derived W303-K6001 genome is also found in Σ1278B. Shown are two exemplary multiple alignments each from Chr. XIV, XIII and XI, and the breakpoint of the cluster on Chr XI. (b) Distance diagram of S288c, Σ1278B and W303-K6001 for the non-S288c cluster on Chr XIV 730 000 – 760 000. (c) W303-K6001: non-S288c sequence clusters with high sequence similarity to Σ1278B.
5.1. Illumina sequencing

Genomic DNA from both strains was sheared by sonication to fragment sizes of around 225 bp, cleaned (Zymo Research) and universal sequencing adaptors were ligated. After library quantification at a Qubit (Invitrogen), a 10 nmol stock solution of the amplified library was created. We loaded 8 pM of the stock solution onto the channels of a 1.4 mm flow cell, and cluster amplification was performed. Sequencing-by-synthesis was performed on an Illumina Genome Analyzer (GAIIX). After quality control of the first base incorporation (signal intensities, cluster density), the run was started. All samples were subjected to 120 bp paired-end sequencing.

5.2. Data analysis

5.2.1. Raw data processing of 454 reads

After default raw data processing, we used a reséquencing trimming filter to increase the data output. (parameters: doValleyFilterTrimBack = false, vfBadFlowThreshold = 6, vfLastFlowToTest = 168, errorQscoreWindowTrim = 0.01). With these parameters, we got an average quality score of greater than Q30 per base.

5.2.2. Raw data processing of Illumina GAIIX sequence reads

Illumina data were provided as qseq files generated by the Bustard 1.8.0 pipeline. High-quality data were extracted using homemade scripts (perl/awk). As a first step, reads were trimmed in such a way that only the longest sequence range of the reads, which did not contain bases of quality lower than Phred 12, was used. Additionally, adaptor sequences were clipped, if at least 15 bp of the adaptor’s 3’ end was found in each read. After trimming and adaptor removal, only reads equal to or longer than 64 bp were used in the mapping/de novo assemblies. In a final step, most of the duplicate reads resulting from amplification bias during library construction were removed, if the first 64 bases of the reads were identical. Finally, sequence data were stored in fasta files with Newbler-compatible headers.

5.2.3. Assembly and mapping

Assemblies were computed by the Roche/454 Newbler v. 2.6 assembler or mapper software applying default parameters. Additional de novo assemblies were performed by the 127mer or 63mer version of SOAPdenovo v. 1.05, downloaded from http://soap.genomics.org.cn). After running several assemblies, we found that a kmer size of 57 was giving the best results in terms of N50 contig sizes after scaffolding and gap filling and in terms of total consensus length. Additional reference-guided assemblies were carried out by the CLC Genomics Workbench v. 5.1 (CLC BIO, Aarhus, Denmark), and its reference mapper and probabilistic variant detection modules (default parameters for the mapping algorithm, for variant detection the ploidy parameter was set to 1).

6. Acknowledgements

We thank Steve Oliver (University of Cambridge) and our laboratory members for support and critical discussions. We acknowledge funding from the Max Planck Society, the Wellcome Trust (no. RG 093735/Z/10/Z) and the ERC (Starting grant 260809). Markus Ralser is a Wellcome Trust Research Career Development and Wellcome-Beit prize fellow.

References

11. Piatti S, Lengauer C, Nasmyth K. 1995 Cdc6 is an unstable protein whose de novo synthesis in G1 is important for the onset of S phase and for preventing a ‘reductive’ anaphase in the budding yeast Saccharomyces cerevisiae. EMBO J. 14, 3788 – 3799.
17. Stern M, Jensen R, Hershkowitz I. 1984 Five SWI genes are required for expression of the HO gene in...


