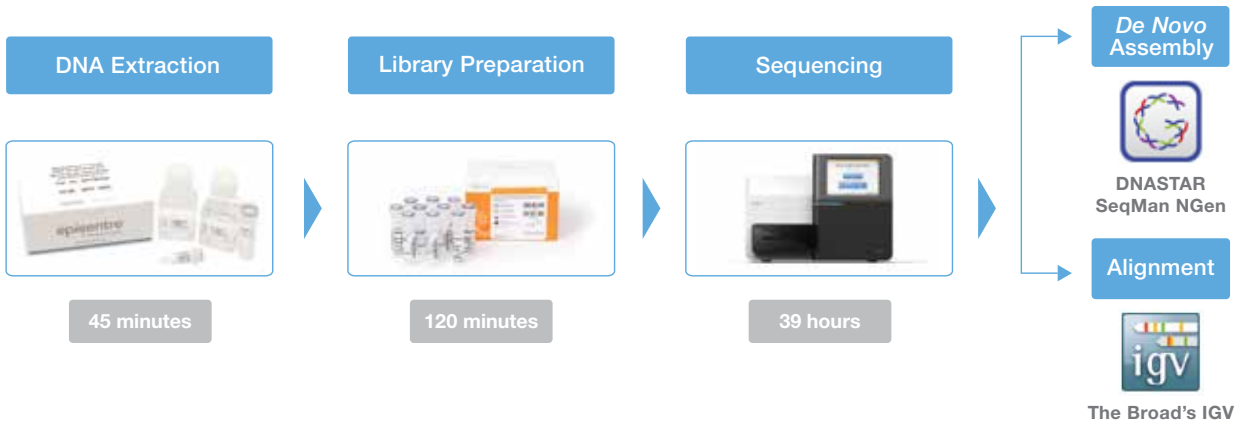
illumina®

De Novo Assembly and Resequencing of the *Saccharomyces cerevisiae* Genome

An end-to-end workflow for sequencing small genomes, from library prep to data analysis in the BaseSpace® environment.

Figure 1: Illumina's End-to-End Solution for Sequencing Small Genomes



Introduction

Saccharomyces cerevisiae is one of the most studied eukaryotic model organisms, representing all the basic gene functions for eukaryotes. In collaboration with White Labs, Inc. (San Diego, CA), Illumina sequenced eight closely related *S. cerevisiae* strains on the MiSeq® system to capture their biological diversity and gain insight into the differences among the strains. These differences include copy number and ploidy variations, genome rearrangements, and polymorphism changes. The data generated in this study represents a significant contribution to ongoing yeast research and is available for public access on the BaseSpace cloud computing platform. This application note describes a complete workflow (Figure 1), from DNA extraction to simple *de novo* genome assembly using SeqMan NGen software from DNASTAR[1] (Madison, WI) within the BaseSpace environment.

Methods and Results

Genomic DNA Isolation

Yeast cultures were obtained through collaboration with White Labs. Genomic DNA from the eight yeast cultures was isolated using the MasterPure™ Yeast DNA Purification Kit from Epicentre, an Illumina company[2] (Madison, WI). This kit employs a non-enzymatic approach to cell lysis, followed by protein precipitation and subsequent nucleic acid isolation. This process results in high yields of purified, high-molecular-weight DNA. After isolation, the purified DNA was quantified using fluorimetric methods and diluted to the optimal concentration for library construction.

Library Construction

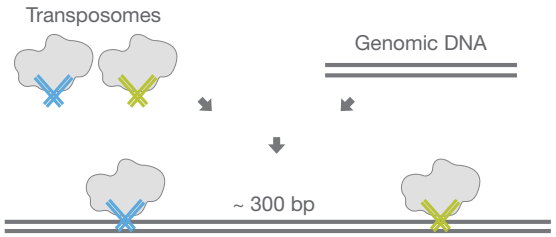
Genomic DNA libraries were prepared using the Nextera® XT Library Preparation Kit, which provides a rapid, easy, and efficient workflow optimized for small genomes (Figure 2). By simultaneously fragmenting and tagging the DNA with adapters (a process called "tagmentation"), this protocol enables the preparation of sequencing-ready libraries in less than 90 minutes.

The Nextera transposome tagmented 1 ng of yeast DNA, resulting in adapter-ligated DNA fragments. The tagmented DNA was purified and then amplified by PCR before cluster generation and sequencing. Because the Nextera XT Library Preparation Kit uses bead-based sample normalization, library quantification was not necessary before pooling. Thus, the kit provides a streamlined workflow from library preparation to sequencing.

Sequencing

Sequencing took advantage of the flexibility and throughput afforded by the MiSeq system's on-board cluster generation. For rapid genome sequencing, 5 µl aliquots from each of the eight strains were pooled together, loaded onto a MiSeq reagent cartridge, and clustered on the MiSeq instrument. Samples were sequenced using 2 × 250 bp paired-end reads and the FASTQ MiSeq Reporter workflow. Sequencing generated a total of 18 million reads, resulting in an average coverage depth of 39× per sample. Base call (*.bcl) files generated by the MiSeq system were converted to FASTQ files and uploaded automatically to the BaseSpace environment for secondary analysis.

Figure 2: Nextera XT Library Preparation



The kit uses a "tagmentation" reaction—simultaneous fragmentation and adapter-tagging of DNA—to enable fast and efficient library preparation. Indices and motifs required for clustering and sequencing are added during subsequent low-cycle amplification.

Data Analysis

Data analysis poses a challenge for any microbiologist analyzing next-generation sequencing data. The BaseSpace platform provides a solution for storage and streamlined analysis, with ready-to-use software applications.

Read Assembly Using *De Novo* Pipeline and SeqMan NGen Software

For this application, Illumina used SeqMan NGen *de novo* assembly software, which allows quick and accurate assembly of either single-read or paired-end data. Sequencing data from the yeast samples were imported into SeqMan NGen and reads were assembled defining the parameters as 350 bp for the pair distances and 12 Mb for the estimated genome size. Across the eight samples, an average N50 length of 63,246 bp was obtained. These draft genome assemblies are ready to use in a range of analyses, including gene correlations, genome rearrangements, or copy number variations. The assembly metrics for each sample are shown in Table 1. All strains were compared with the same *de novo* parameters, with no changes in function of size or coverage. For further investigation, the estimated genome size should be considered for better assemblies, because some strains have variable ploidy and genome size.

Alignment Using Resequencing Pipeline and Integrative Genomics Viewer

The data were also analyzed using the resequencing pipeline of the MiSeq system, which uses the Burrows-Wheeler Algorithm (BWA) to create alignment (*.bam) files. These files can then be opened and visualized using the Integrative Genomics Viewer[3] (IGV), a secondary analysis application developed by the Broad Institute and available on the BaseSpace platform. IGV allows researchers to align BAM files to a reference genome. Three of the *S. cerevisiae* samples were selected for analysis of variations in general, with an emphasis on chromosome 9 (Table 2). In the samples labeled 2591 and 2542, phenotypically, one flocculates normally while the other does not. Flocculation refers to the ability of yeast to aggregate and form fibrous interconnections, or clumps. Comparison between the flocculation genes *FLO1* and *FLO8* did not reveal significant genetic changes, but the *FLO11* gene was shown to have an insertion of 1,000 bp in sample 2542 (Figure 3). This gene encodes a glycolipid (GPI anchor) that can be attached to the C-terminus of a protein during post-translational modification. To understand whether the observed insertion can affect the strain's ability to flocculate, further analysis is required.

Conclusions

The workflow outlined in this application note—including DNA extraction from culture samples, sequencing, and simple data analysis—is broadly applicable to the analysis of any small genome. The data obtained from *S. cerevisiae* strains is available for public use on the BaseSpace platform[4]. As new applications are developed within the BaseSpace environment, the analysis can be extended to other areas of microbiology research. The high accuracy and resolution of the sequencing workflow allows researchers to sequence the genomes of bacteria, archaea, fungi, or lower eukaryotes with a convenient and easy-to-use assembly pipeline.

Table 1: Assembly Output for *S. cerevisiae* Strains

Sample ID	Base Count	N50	Number of Contigs	Clusters Passing Filter	Coverage Based on 12 Mb	Location
2521	10,146,523	27K	1571	1,018,471	21	England
2535	22,353,560	35K	1035	7,395,956	154	San Diego, USA
2542	21,913,088	30K	739	1,189,456	25	Belgium
2543	30,685,603	27K	1551	1,258,847	26	Belgium
2545	39,965,870	34K	1493	1,286,525	27	Belgium
2590	21,367,407	62K	523	1,602,957	33	—*
2591	21,870,148	55K	615	1,803,178	38	California, USA
2592	22,222,588	—*	—*	2,957,468	62	England

*Information not available.

Figure 3: Alignment to Reference Genome

Genome assemblies from two samples were aligned to the *S. cerevisiae* reference genome. A region of chromosome 9 is shown here.

Learn More

Go to www.illumina.com/miseq to learn more about the next revolution in personal sequencing. For more information about the use of Illumina sequencing technology in microbial genomics, visit www.illumina.com/microbiology.

References

1. www.dnastar.com
2. www.epibio.com
3. www.broadinstitute.org/igv
4. www.basespace.illumina.com/projects/White_Labs

Table 2: BWA Whole-Genome Sequencing Output

Sample ID	Number of Reads	% Coverage	% Aligned
2542	2.378,912	93%	98%
2590	1,602,957	54%	88%
2591	3,606,356	94%	98%

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 1270-2013-005 Current as of 11 November 2014

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA