

# GENOME RESEARCH

## Rapid whole-genome mutational profiling using next-generation sequencing technologies

Douglas R. Smith, Aaron R. Quinlan, Heather E. Peckham, *et al.*

*Genome Res.* published online Sep 4, 2008;  
Access the most recent version at doi:[10.1101/gr.077776.108](https://doi.org/10.1101/gr.077776.108)

---

<b>Supplementary data</b>	"Supplemental Research Data" <a href="http://genome.cshlp.org/cgi/content/full/gr.077776.108/DC1">http://genome.cshlp.org/cgi/content/full/gr.077776.108/DC1</a>
<b>P&lt;P</b>	Published online September 4, 2008 in advance of the print journal.
<b>Open Access</b>	Freely available online through the Genome Research Open Access option.
<b>Email alerting service</b>	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a>

---

---

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

---

To subscribe to *Genome Research* go to:  
<http://genome.cshlp.org/subscriptions/>

---

## Methods

# Rapid whole-genome mutational profiling using next-generation sequencing technologies

Douglas R. Smith,<sup>1,7,9</sup> Aaron R. Quinlan,<sup>2,7</sup> Heather E. Peckham,<sup>3,7</sup> Kathryn Makowsky,<sup>1</sup> Wei Tao,<sup>1</sup> Betty Woolf,<sup>1</sup> Lei Shen,<sup>1</sup> William F. Donahue,<sup>1</sup> Nadeem Tusneem,<sup>1</sup> Michael P. Stromberg,<sup>2</sup> Donald A. Stewart,<sup>2</sup> Lu Zhang,<sup>2</sup> Swati S. Ranade,<sup>3</sup> Jason B. Warner,<sup>3</sup> Clarence C. Lee,<sup>3</sup> Brittney E. Coleman,<sup>3</sup> Zheng Zhang,<sup>3,4</sup> Stephen F. McLaughlin,<sup>3</sup> Joel A. Malek,<sup>3</sup> Jon M. Sorenson,<sup>3,4</sup> Alan P. Blanchard,<sup>3</sup> Jarrod Chapman,<sup>5</sup> David Hillman,<sup>5</sup> Feng Chen,<sup>5</sup> Daniel S. Rokhsar,<sup>5</sup> Kevin J. McKernan,<sup>3</sup> Thomas W. Jeffries,<sup>6</sup> Gabor T. Marth,<sup>2,9</sup> and Paul M. Richardson<sup>5,8,9</sup>

<sup>1</sup>Agencourt Bioscience Corporation, Beverly, Massachusetts 01915, USA; <sup>2</sup>Boston College Biology Department, Higgins Hall, Chestnut Hill, Massachusetts 02467, USA; <sup>3</sup>Applied Biosystems, Beverly, Massachusetts 01915, USA; <sup>4</sup>Applied Biosystems, Foster City, California 94404, USA; <sup>5</sup>US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA; <sup>6</sup>Institute for Microbial and Biochemical Technology, US Forest Products Laboratory, Madison, Wisconsin 53726, USA

Forward genetic mutational studies, adaptive evolution, and phenotypic screening are powerful tools for creating new variant organisms with desirable traits. However, mutations generated in the process cannot be easily identified with traditional genetic tools. We show that new high-throughput, massively parallel sequencing technologies can completely and accurately characterize a mutant genome relative to a previously sequenced parental (reference) strain. We studied a mutant strain of *Pichia stipitis*, a yeast capable of converting xylose to ethanol. This unusually efficient mutant strain was developed through repeated rounds of chemical mutagenesis, strain selection, transformation, and genetic manipulation over a period of seven years. We resequenced this strain on three different sequencing platforms. Surprisingly, we found fewer than a dozen mutations in open reading frames. All three sequencing technologies were able to identify each single nucleotide mutation given at least 10–15-fold nominal sequence coverage. Our results show that detecting mutations in evolved and engineered organisms is rapid and cost-effective at the whole-genome level using new sequencing technologies. Identification of specific mutations in strains with altered phenotypes will add insight into specific gene functions and guide further metabolic engineering efforts.

[Supplemental material is available online at [www.genome.org](http://www.genome.org). Complete data sets are available at the NCBI Short Read Archive under accession no. SRA 001158 (<ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/>.)]

*Pichia stipitis* (Pinal) is a haploid yeast related to endosymbionts of beetles that degrade rotting wood (Suh et al. 2003). It is an important organism for bioenergy production from lignocellulosic materials because of its high capacity to ferment xylose and cellobiose to ethanol (Parekh et al. 1988). We previously sequenced the reference strain, *Pichia stipitis* CBS-6054, resulting in a completely characterized genome of eight chromosomes totaling 15.4 Mb of sequence (Jeffries et al. 2007). This strain has been subjected to chemical mutagenesis, phenotypic selection, genetic engineering, and adaptive evolution in order to develop strains improved for ethanol production. Chemical mutagenesis and selection resulted in small improvements in ethanol production attributable in part to carbon catabolite derepression (Supple-

mental Fig. 1; Methods). Disruption of *CYC1* (cytochrome *c*, isoform 1) to create strain Shi21 increased the specific ethanol production rate by 50% and the ethanol yield by 10%; however, the nature of additional mutational events leading to this phenotype was uncharacterized.

Traditional methods for identifying mutations are labor- and time-intensive, so we tested the ability of next-generation sequencing technologies to determine the differences in this improved strain's entire genome relative to the reference strain. We generated high-coverage, whole-genome data sets using single fragment end reads from three next-generation sequencing platforms: 454 Life Sciences (Roche) (~225-bp reads), Illumina (formerly Solexa sequencing) (32-bp reads), and Applied Biosystems SOLiD (35-bp reads) (Schuster 2008). We assessed these data to determine the effect of sequence coverage (i.e., data set size) on the accuracy of mutation detection, and to compare the efficiency of the three platforms for this application.

## Results

Genomic DNA from *P. stipitis* (Shi21) was sequenced using the three advanced sequencing platforms according to specifications

<sup>7</sup>These authors contributed equally to this work.

<sup>8</sup>Present address: Progentech Limited, 5885 Hollis St., Suite 155, Emeryville, CA 94608, USA.

<sup>9</sup>Corresponding authors.

E-mail [PaulRichardson@Progentech.com](mailto:PaulRichardson@Progentech.com); fax (510) 655-5840.

E-mail [douglas.smith@agencourt.com](mailto:douglas.smith@agencourt.com); fax (978) 867-2601.

E-mail [marth@bc.edu](mailto:marth@bc.edu); fax (617) 552-2011.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.077776.108>. Freely available online through the *Genome Research* Open Access option.

**Table 1.** Sequencing and mutation discovery statistics

Sequencing technology	Total no. of reads <sup>a</sup>	Total sequence (bp, in millions)	Average sequence coverage from aligned reads <sup>b</sup>	False-positive (spurious) mutations	False-negative (missed) mutations
454 FLX (2 runs)	887,123	199.35	10.78 ×	1	0
454 FLX (1.5 runs)	669,783	150.64	8.15 ×	6	1
454 FLX (1 run)	459,563	103.38	5.62 ×	17	1
Illumina (7 lanes)	25,818,266	826.18	44.24 ×	0	0
Illumina (3 lanes)	11,281,705	361.01	19.40 ×	0	0
Illumina (2 lanes)	7,548,407	241.55	13.00 ×	2	0
Illumina (1 lane)	3,674,253	117.58	6.32 ×	2	2
AB (2 flow cells)	228,191,758	7,986.71	175.09 ×	0	0
AB (30 ×)	39,111,512 <sup>c</sup>	1,368.90	30.01 ×	0	0
AB (20 ×)	26,065,653 <sup>c</sup>	912.30	20.00 ×	0	0
AB (10 ×)	13,045,859 <sup>c</sup>	456.61	10.01 ×	0	0
AB (8 ×)	10,426,261 <sup>c</sup>	364.92	8.00 ×	0	4
AB (6 ×)	7,819,696 <sup>c</sup>	273.69	6.00 ×	0	5

The overall sequence throughput and aligned coverage is shown for each sequencing technology used in the study. We also report the number of spurious and missed mutations observed from each experiment.

<sup>a</sup>For the 454 and Illumina technologies, the total number of reads reflects the number of reads that remained after manufacturer quality controls. The Applied Biosystems (AB) read totals reflect all reads produced by the sequencing run.

<sup>b</sup>The coverage produced by those reads in the second column that passed the mapping filters we used for each technology (Methods).

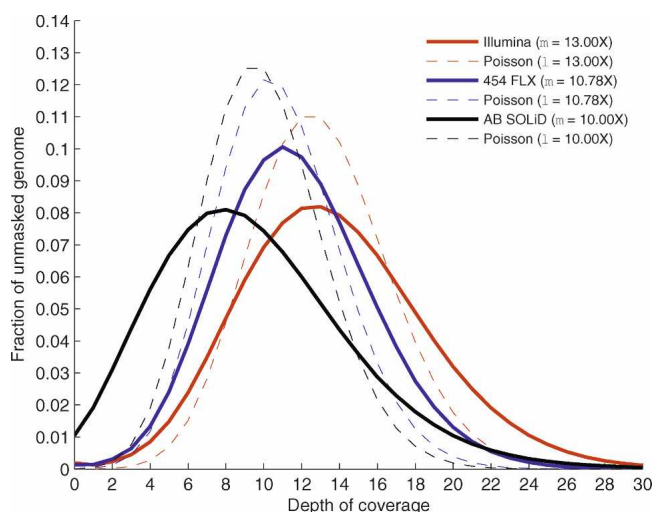
<sup>c</sup>Estimated number of reads based on in silico subsampling of coverage.

of the manufacturers (Methods). Low-quality sequence reads from the 454 Life Sciences and Illumina technologies were excluded by manufacturer quality control filters prior to analysis. Since the Applied Biosystems SOLiD sequencing technology does not exclude low-quality reads prior to data analysis, we instead discarded all SOLiD reads that had too many mismatches when they were mapped (Methods) to the *Pichia* reference genome. We processed the sequence reads from each technology with the manufacturer-supplied base-calling software. We additionally recalled the 454 pyrosequences with the Pyrobayes (Quinlan et al. 2008) program because it produces a lower number of substitution errors and more accurate base quality values than the native base-calling program (Methods). We first identified and masked (i.e., excluded from the genome sequence) all repetitive elements within the *P. stipitis* genome (Jeffries et al. 2007) that would interfere with unique read alignments, including short genomic repeats as well as nuclear mitochondrial DNAs (NUMTs), which are sequences of mitochondrial origin that were inserted into the nuclear genome (Methods; Supplemental Table 1) (Richly and Leister 2004). Due to the nature of the unpaired short reads produced by these methods, this repeat masking prevented shorter SOLiD and Illumina reads from mapping to 6.8% of the genome and prevented the medium-length 454 FLX reads from mapping to 5.3% of the genome (Supplemental Methods). The total number of aligned reads passing alignment quality filters and the corresponding aligned read coverage are shown in Table 1. Alignment of reads from each technology to the repeat-masked reference sequence resulted in 11–175 × coverage of the genome depending on the type of platform and number of runs (Table 1; Supplemental Table 2).

When mapping the Illumina, 454, and Applied Biosystems reads to the masked reference sequence, we allowed one, two, and three mismatches, respectively (Methods). The Illumina and 454 reads were mapped to the reference sequence with the MOSAIK program (Methods). At the time of this analysis, MOSAIK was unable to align reads from the Applied Biosystems SOLiD technology because of the dinucleotide encoding (also termed “color-space” alignments) that this technology uses (Valouev et al. 2008). Therefore, we mapped the Applied Biosystems

SOLiD reads to the *Pichia* genome with the Applied Biosystems SOLiD Alignment Tool. Despite the algorithmic differences owing to color-space alignments, MOSAIK and the SOLiD Alignment Tool use a similar hash-based method to find potential genomic alignment locations for each sequence read.

The distribution of sequence coverage across the *Pichia* genome was similar for each of the sequencing technologies (Fig. 1). The observed coverage distributions are substantially dispersed as compared to the expected Poisson distributions (Fig. 1, dotted lines), indicating that there are regions of the *Pichia* genome that are more facile to sequence than others. The causes and dynamics



**Figure 1.** Distribution of genome sequence coverage. The distribution of sequence coverage across the unmasked portion of the genome is shown for each technology. Here we represent comparable mean coverage levels for Illumina (red line, 13.00 × mean genome coverage), 454 FLX (blue line, 10.78 × mean genome coverage), and Applied Biosystems SOLiD (black line, 10.00 × mean genome coverage) technologies. For each, we compare the observed coverage distribution to the expected Poisson coverage distribution (dotted lines of the same color for each technology).

of these biases are beyond the scope of this study but are an important consideration for genome resequencing studies. Multiple read alignments from the 454 and Illumina platforms were screened for mutations using GIGABAYES, a new version of the POLYBAYES (Marth et al. 1999) SNP discovery program (Methods). Color-space alignments of the SOLiD data were similarly screened using software supplied by Applied Biosystems. The 17 candidate mutations discovered among the three sequencing technologies were resequenced in CBS-6054 and in each of the four derived strains with a capillary sequencing machine and were all confirmed (Table 2). Three of the changes were found to be errors in the reference sequence, as the alternate base is present in the validation traces not only from all sequenced mutants but also from the parent strain. This implies an error rate of 3 nt in the 15-Mb *Pichia* reference genome, far exceeding the established standards for genome finishing (1 error/10 kb). Given that the mutations were discovered in very deep data sets and independently confirmed by four different sequencing methods, it is unlikely that we missed any additional mutations in the unmasked fraction (~95%) of the Shi21 mutant genome. We therefore believe that the remaining 14 mutations comprise the complete set of single nucleotide variants between the mutant and the parent (i.e., reference) *Pichia* strains.

Since the *Pichia* genome is haploid during vegetative growth, all mutations are expected to be homozygous. An apparent heterozygous change at position 358,358 on chromosome 8 is a result of the intentional gene disruption of *CYC1* with a *URA3* selection cassette, which resulted in a *URA3* duplication. This apparent variation represents a paralogous difference between the two copies of a duplicated gene and thus cannot be considered a true point mutation. We screened for small (1–2 bp) INDEL polymorphisms with GIGABAYES, but none were found, which is not surprising considering that the alkylating agents (Methods) used in mutagenesis principally induce base substitutions. However, because we strictly limited the number of mismatches allowed during read mapping (Methods), it is theoretically possible that longer (>2 bp) INDEL mutations were missed. Additionally,

we are currently investigating the use of paired-end sequence data to identify and resolve structural variations as well as larger insertions and deletions.

A primary focus of this study was to evaluate the utility of next-generation sequencing technologies for mutational profiling. We therefore compared the capabilities of the three platforms for the identification of the 14 confirmed point mutations in the *Pichia* mutant. Each of the three sequencing technologies correctly identified all 14 variations with essentially no false positives when all available reads generated on the platform were used (Table 1; Fig. 2). The complete Illumina and Applied Biosystems alignments afforded perfect accuracy: All 14 mutations were found and no false-positive predictions were made. A single false-positive prediction was found in the complete 454 FLX data (which produced lower overall coverage than the other platforms) and was most likely the result of a PCR error during sequence library construction (data not shown). The accuracy we observed is encouraging given that low false discovery (i.e., that is, the fraction of erroneously identified mutations) and false negative (i.e., the fraction of true mutations that were missed) rates are critical considerations for the application of these technologies to rapid forward genetic mutational profiling. These results show that all three technologies are suitable for highly accurate mutation screening (Supplemental Fig. 2).

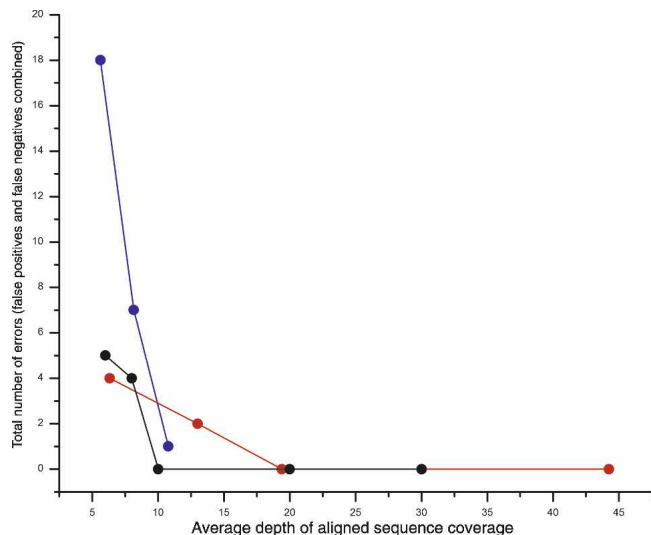
An important consideration for the cost of such experiments is the depth of sequence coverage required to achieve a desired sensitivity and specificity. To determine how the error rate changes as fewer reads are used, we selected subsets of reads of varying size (corresponding to likely use cases for each platform) from each of the three full data sets and subjected the resulting lower-coverage assemblies to our mutation discovery analysis. As shown in Table 1, a combined missed mutation (false negative [FN]) and erroneously called mutation (false positive [FP]) error rate of 50% is achieved with 1.5 454 FLX machine runs (8.15-fold aligned read coverage; six FP and one FN errors), a single lane of Illumina reads (6.32-fold aligned read coverage; two FP and two FN errors), and sixfold coverage of Applied Bio-

systems SOLiD reads (zero FP and five FN errors). The increased number of false positives observed with the lower 454 FLX coverage is the result of local homopolymer misalignments that arise when a nucleotide overcall (that is, calling too many nucleotides) is followed by a nucleotide undercall (that is, calling too few bases), or vice versa. Deeper coverage mitigates such alignment artifacts (Quinlan et al. 2008). The fact that the Applied Biosystems SOLiD technology produced zero false positives is a result of the “di-base encoding” which facilitates the segregation of sequencing error from true mutations (Valouev et al. 2008). It is important to note that we may have missed additional mutations in the Shi21 strain because we masked between 5.3% and 6.8% of the genome. Given the constraints of plate configurations and run conditions on the different platforms, we find that a minimum of 10–15-fold genome coverage is required for the desired error rate.

**Table 2. Summary of discovered point mutations relative to the *Pichia* reference genome**

Chrom.	Location	Nucleotide change	Amino acid change	Functional description of mutation
2	1,339,463	T>C	V>A	Error in reference sequence
2	2,598,869	C>A	-	Error in reference sequence
3	1,769,576	C>T	G>S	Error in reference sequence
1	1,143,120	C>T	G>S	<i>YHN8</i> (predicted GPCR)
2	746,465	C>T	D>N	<i>IFI3</i> (hypothetical protein; ID 29635)
2	1,102,664	G>T	-	Upstream of <i>RAD15</i>
3	104,338	T>G	-	Non-coding interval
4	1,499,156	T>A	K>N	<i>VSP36</i> (vacuolar sorting protein)
7	930,181	A>T	W>R	<i>FBX1</i> (Leucine rich repeat protein, contains F-box)
8	36,439	A>G	D>G	<i>POT11</i> (3-ketoacyl-CoA thiolase B)
1	839,170	C>T	V>I	<i>SEC31</i> (component of the COPII coat of ER-golgi vesicles)
2	617,666	G>A	S>F	<i>SLX8</i> (Zn finger RING domain protein; ID 54919)
1	670,317	G>A	R>K	<i>ALD7</i> (aldehyde dehydrogenase)
8	358,358	T>A	D>V	<i>URA3</i> (orotidine-5'-phosphate decarboxylase)
1	947,086	C>G	L>V	<i>MDM34</i> (mito. outer membrane protein involved in mitochondrial shape)
3	885,477	G>C	-	Intergenic region between <i>LEU3</i> and <i>YXE1</i>
6	1,088,427	G>C	-	Upstream of <i>TSC11</i> (TOR binding protein; ID 84674)

Color coding indicates in which strain each mutation first appeared relative to the parent, CBS-6054. Orange, FPL-061 (rapid growth on L-xylose in the presence of the respiration inhibitors); yellow, FPL-DX26 (2-deoxyglucose resistance); green, FPL-UC7 (FOA resistance); blue, Shi21 (*CYC1:ura3* targeted gene disruption).



**Figure 2.** The effect of sequence coverage on mutation discovery accuracy. The total number of mutation discovery errors is shown for the three sequencing technologies at various levels of aligned sequence coverage. (Blue circles) 454 FLX; (red circles) Illumina; (black circles) Applied Biosystems SOLiD.

## Discussion

All three next-generation sequencing platforms correctly identified nucleotide variations between the reference and mutant strains given sufficient coverage. The fraction of mutations in open reading frames (78%) was slightly higher than the average gene density (56%) (Jeffries et al. 2007). In the absence of selection, about two-thirds of the base changes should have resulted in silent mutations at the amino acid level, due to redundancy in the genetic code. Surprisingly, all mutations retained in open reading frames resulted in amino acid changes, indicating high selective pressure and little or no neutral drift (Table 2). Further characterization of the identified mutational events through physiological and genetic studies will be necessary to determine how they affect cell phenotype.

Overall, our results demonstrate that the new sequencing technologies tested are well suited for mutational analysis of novel yeast strains derived from multistep mutagenesis procedures. For most applications, 10–15-fold redundant genome coverage will allow for accurate and cost-effective mutational profiling. Deeper coverage is likely necessary for similar experiments in diploid organisms (e.g., ENU mutagenesis in mouse), as the discovery of heterozygous loci requires that both alleles be sampled from high-quality reads. The approach is expected to be equally suitable for the analysis of bacterial, fungal, and other organisms derived by directed evolution and natural variation, especially as sequencing costs and throughput continue to improve for all of these technologies. Thus, this approach could help accelerate the development of novel organisms for bioenergy and biotechnology applications as well as facilitate traditional forward and reverse genetic screens.

## Methods

### Derivation of the mutagenized Shi21 strain

The Shi21 derivation of the Shi21 strain of *P. stipitis* is thoroughly described by Shi et al. (1999).

## Sequencing

Chromosomal DNA from *P. stipitis* Shi21 was prepared by standard methods (Burke et al. 2000). For 454 sequencing, a library was prepared and sequenced using manufacturer-supplied protocols and reagents, as follows. Five micrograms of DNA was sheared to an average size of 480 bp. Adaptors were ligated, and the correct products were selected using 454 library immobilization beads. The single-stranded DNA library was quantified using the Invitrogen Ribogreen assay, and 32 emulsion PCR reactions were prepared with a ratio of two molecules per DNA capture bead. After amplification, the emulsions were broken and enriched, resulting in a total of 3.92 million beads containing amplified library fragments. The beads were sequenced in two full 454 FLX sequencing runs, each loaded with 1.8 million beads, yielding a total of ~199 Mb of sequence data.

For Illumina sequencing, 3  $\mu$ g of genomic DNA was fragmented below 800 bp using a nebulizer. Fragments were end-repaired with T4 DNA polymerase. A single dA was added to the ends using Klenow fragment and dATP. Fragments were then ligated with adaptors provided by the manufacturer. Adaptor-ligated fragments were separated from unligated adaptors by running and agarose gel and cutting a band corresponding to ~150–300 bp and purified using a spin column. The fragment library containing adaptors was subjected to 18 rounds of PCR using primers supplied by Illumina. This amplified library was then loaded onto the cluster generation station for single molecule bridge amplification on slides containing attached primers. The slide with amplified clusters was then subjected to step-wise sequencing using four-color labeled nucleotides on the Illumina 1G sequencing system for 32 cycles. A total of 25,818,266 reads were obtained after quality filtering, yielding ~826 Mb of sequence data.

For SOLiD sequencing, five micrograms of DNA was sheared and size-selected to an average size of 100 bp. P1 and P2 adaptors were ligated and amplified for 15 cycles; 0.2 pg/ $\mu$ L of double-stranded library was added to the emulsion with 950 million beads according to manufacturers' instructions. Twenty-nine percent of the beads were P2 positive (contained amplified library fragments) before enrichment and 91% of the beads were P2 positive after enrichment, yielding 277 million beads deposited on two slides; 228 million of these beads fell within the imaged area and were detected in sequencing, yielding 2.7 Gb of aligned 35-mer sequence.

For confirmation sequencing, PCR products were generated from genomic DNA of each strain using M13-tailed primer pairs, the products were sequenced on ABI3730xl instruments, variants were identified using PolyPhred, and confirmed using *consed* (Stephens et al. 2006). Complete data sets are available at the NCBI Short Read Archive under accession no. SRA 001158 (<ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead>).

### Illumina and 454 sequence alignment

We used our general reference sequence-guided alignment and assembly tool, MOSAIK, to process the Illumina and 454 data sets. MOSAIK (Michael Stromberg, Boston University) uses a hashing scheme to seed full Smith-Waterman gapped alignments against the concatenated *P. stipitis* genome. The resulting pairwise alignments are then consolidated into a multiple sequence alignment (assembly) and saved as an ACE assembly file. These assemblies can be viewed by programs such as *consed* (Gordon et al. 1998). To correct for 454 indel alignment errors, the Smith-Waterman scoring algorithm has been augmented to use an alternate gap open penalty when a homopolymer region is detected. For both the Illumina and the 454 reads, we required that

at least 95% of each read align to the reference sequence. In order to ensure that we only aligned high-quality reads from each technology, we also required that the reads from each technology had few sequence differences (i.e., mismatches, insertions, or deletions) relative to the reference genome sequence. We allowed at most one sequence difference in the Illumina reads and two sequence differences in the longer 454 reads.

### SOLiD sequence alignment

The Applied Biosystems SOLiD alignment tool translates the reference sequence to di-base encoding ("color-space") and aligns the reads in color space. The program guarantees finding all alignments between a read and the reference sequence with up to M mismatches (a user-specified parameter). Applied Biosystems SOLiD reads were mapped to the *Pichia* genome allowing up to three mismatches for each read. The alignment tool uses multiple spaced seeds (discontinuous word patterns) to achieve a rapid running time.

### Acknowledgments

Author contributions to this work are as follows: D.R.S., project initiation, design and coordination, mutation analysis. K.M., 454 FLX sequencing. W.T. and L.S., initial 454 data analysis and comparison with SOLiD data. B.W., Sanger sequencing confirmation and analysis. W.F.D., SOLiD library construction. N.T., R&D manager. H.E.P., development of SOLiD consensus algorithm, mutation detection, manuscript preparation. S.S.R., library development. J.B.W., C.C.L. and B.E.C., SOLiD emulsion and sequencing. Z.Z., SOLiD alignment algorithm. S.F.M., J.A.M., and J.M.S., development of SOLiD consensus algorithm. A.P.B. development of 2-base encoding. K.J.M., analysis and manuscript preparation. D.H. and F.C., 454 and Illumina sample prep and data generation. J.C. and D.S.R., Initial Illumina data analysis. P.M.R., Experimental design, coordination, manuscript preparation. A.R.Q., mutation detection, sequence mapping, data analysis, and manuscript preparation. M.P.S., read

mapping. D.A.S., structural variation discovery. L.Z., read mapping. G.T.M., mutation detection and manuscript preparation.

### References

- Burke, D., Dawson, D., and Stearns, T., eds. 2000. *Methods in yeast genetics. Cold Spring Harbor Laboratory course manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Gordon, D., Abajian, C., and Green, P. 1998. *Consed*: A graphical tool for sequence finishing. *Genome Res.* **8**: 195–202.
- Jeffries, T.W., Grigoriev, I.V., Grimwood, J., Laplaza, J.M., Aerts, A., Salamov, A., Schmutz, J., Lindquist, E., Dehal, P., Shapiro, H., et al. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat. Biotechnol.* **25**: 319–326.
- Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitzel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**: 452–456.
- Parekh, S.R., Parekh, R.S., and Wayman, M. 1988. Fermentation of xylose and cellobiose by *Pichia stipitis* and *Brettanomyces clausenii*. *Appl. Biochem. Biotechnol.* **18**: 325–338.
- Quinlan, A.R., Stewart, D.A., Stromberg, M.P., and Marth, G.T. 2008. Pyrobayes: An improved base caller for SNP discovery in pyrosequences. *Nat. Methods* **5**: 179–181.
- Richly, E. and Leister, D. 2004. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* **21**: 1081–1084.
- Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**: 16–18.
- Shi, N.Q., Davis, B., Sherman, F., Cruz, J., and Jeffries, T.W. 1999. Disruption of the cytochrome *c* gene in xylose-utilizing yeast *Pichia stipitis* leads to higher ethanol production. *Yeast* **15**: 1021–1030.
- Stephens, M., Sloan, J.S., Robertson, P.D., Scheet, P., and Nickerson, D.A. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* **38**: 375–381.
- Suh, S.O., Marshall, C.J., McHugh, J.V., and Blackwell, M. 2003. Wood ingestion by passalid beetles in the presence of xylose-fermenting gut yeasts. *Mol. Ecol.* **12**: 3137–3145.
- Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 1051–1063.

Received February 22, 2008; accepted in revised form July 10, 2008.