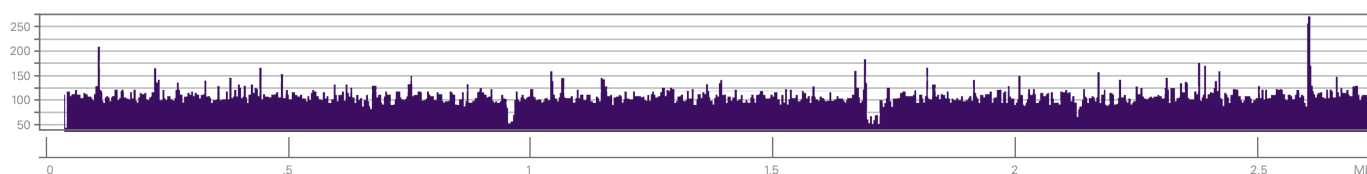


# Whole Genome Resequencing



**Figure 1.** High coverage across chromosome 2 of the *P. stipidis* genome.

## Introduction

Whole genome sequence analysis was traditionally done by shotgun approaches using automated Capillary Electrophoresis (CE) Systems. Once the initial sequence for a particular genome is available, it is then possible to perform comparative sequencing or resequencing to identify polymorphisms, mutations, and structural variations between organisms. Whole genome resequencing requires a highly parallel system to provide the depth of coverage required for variant detection. Library preparation is also critical as the complexity and time involved are multiplied when analyzing multiple genomes.

Automated CE sequencing remains the gold standard for *de novo* analysis where long read lengths are desired, and for smaller targeted resequencing experiments where throughput requirements are lower. The SOLiD™ System is Applied Biosystems' next-generation platform and is well suited for whole genome resequencing projects. The SOLiD System possesses sequencing throughput capacity of more than three gigabases (dual-slide), which allows for the large multiples of coverage necessary to correctly identify SNPs and structural variations such as insertions, deletions, copy number variations and rearrangements, within a single run. Furthermore, clonal amplification by

emulsion PCR dramatically simplifies the library preparation and eliminates the need for bacterial cloning.

## Case Study: Whole Genome Resequencing of *Pichia stipidis*

Whole genome microbial resequencing is conducted in a variety of fields, including biodefense, food testing, bioreactors, biofuels, and phylogenetics. *P. stipidis* (genome size = 15.4 megabases) is a haploid, native xylose-fermenting yeast whose fermentation is important for the bioconversion of plant biomass and exhibits potential for the development of alternative fuels. In collaboration with the Joint Genome Institute (JGI), USDA, Agencourt Biosciences and Applied Biosystems, an engineered mutant of *P. stipidis* was resequenced in order to identify polymorphisms associated with increased ethanol production.

## Methods

A well characterized strain of *P. stipidis* underwent mutagenesis at JGI. One specific strain exhibiting increased ethanol production was selected for and sequenced using the SOLiD System. Both a fragment library and mate-paired library, with 3 kb inserts, were constructed and sequenced at Applied Biosystems. The mate-paired library was utilized to characterize structural variations such as insertions or deletions. Both libraries were

sequenced with the SOLiD System, and analyzed using the Applied Biosystems SOLiD™ ColorSpace Alignment Tool Suite.

## Summary of Results

The SOLiD System generated 600 megabases of mappable data in a single run and accurately identified SNPs in the *P. stipidis* genome. One particular SNP of interest was a functional SNP in a gene involved in ethanol production. The utilization of a mate-paired library also enabled the detection of insertions and deletions in the genome.

## Coverage

Figure 1 illustrates the high level of coverage achieved across chromosome 2 in a single run. Similar coverage was achieved for the rest of the eight chromosomes in the *P. stipidis* genome (data not shown). A restricted analysis requiring both tags to map uniquely to the genome was also performed and identified the position of a large deletion where the tags were missing or deleted. (Ref: Study Poster "Sequencing *Pichia stipidis* Mutants with AB SOLiD™ Technology.")

## SNP Discovery

The SOLiD System detected many SNPs and a representative number of identified SNPs are summarized in Table 1. Green fields indicate SNPs that fit the desired profile of being present in the mutant but not in the wild type. The SNP identified in the alcohol

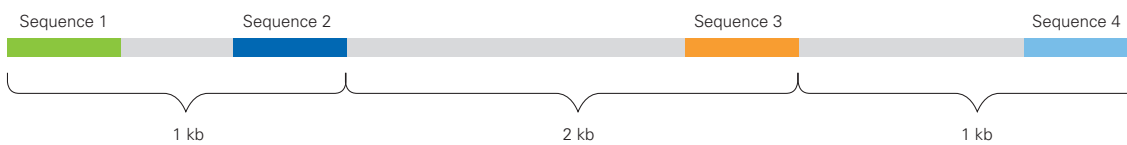
dehydrogenase gene (ALD7) was of particular interest due to its role in ethanol production. Other SNPs were identified that did not fit the predicted profile as they were present in both mutant and wild type strains (red boxes). Many of these SNPs were later identified as regions of genome duplication.

### Detection and Characterization of Structural Variation

Fragment library analysis is useful in identifying sequence variations but mate-paired analysis is required for accurate characterization and mapping of structural changes. Figure 2 describes how a mate-paired analysis effectively

identifies insertions and deletions based on deviations in the distance between tags from the predicted fragment size when mapped back to the reference genome. Multiple reads are mapped for each region, and if the mate-paired insert is larger than the insertion or deletion, it is possible to accurately map

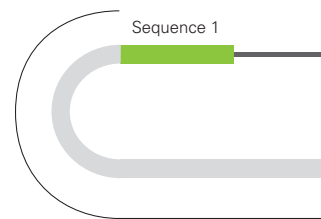
#### Reference Sequence



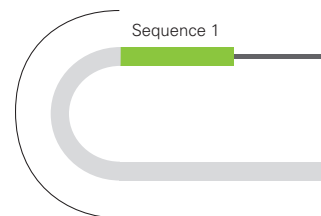
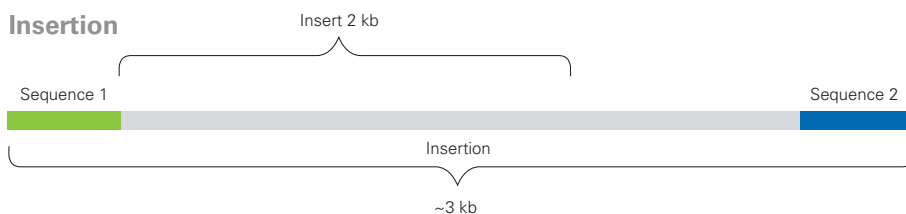
1. Fragment DNA and size select for 2–3 kb fragments.

2. Ligate to internal adaptor and circularize.

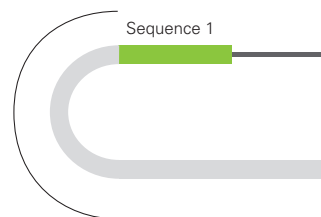
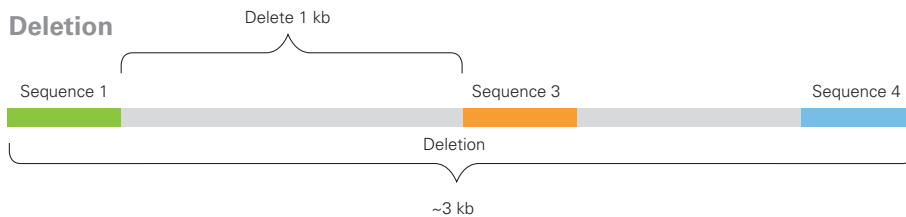
#### Wild Type



#### Insertion



#### Deletion



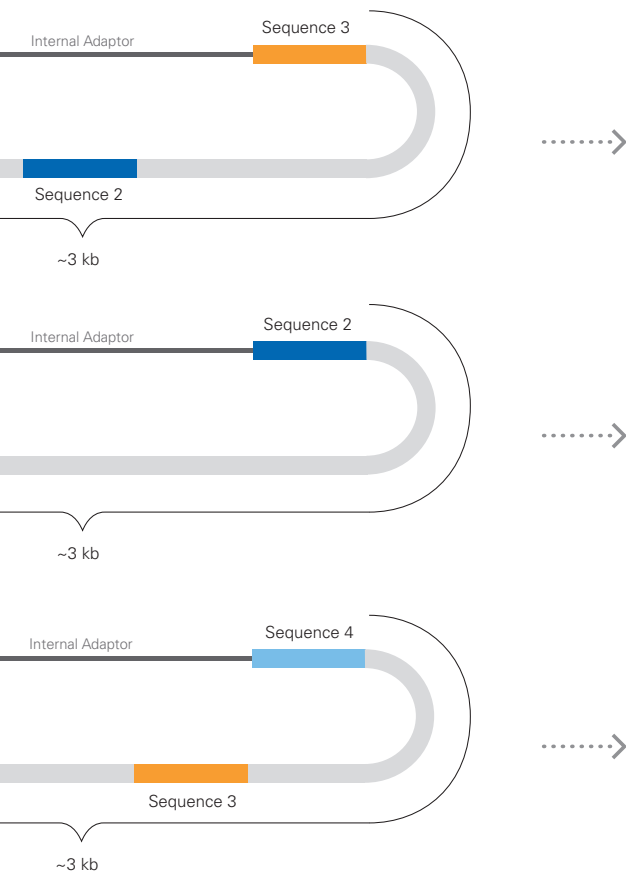
**Figure 2.** Characterization of insertions and deletions using mate-paired analysis.

**TABLE 1.** List of SNPs identified in the *P. stipidis* genome.

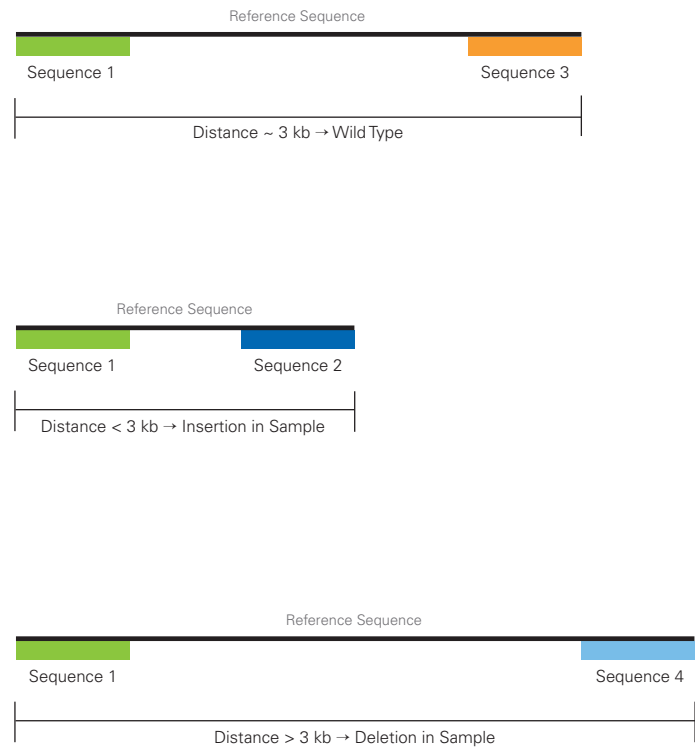
Coverage	SNP Starts	Wild Type Starts	Variant	Description
9	6	0	G/A	ALD7 (aldehyde dehydrogenase)
25	14	0	T/C	NOC2 (Nucleolar Complex 2 involved in nuclear export)
30	14	1	C/T	SEC31 (component of the COPII coat of ER-golgi Vesicle)
14	7	1	G/A	Hypothetical protein ID 54919
20	12	0	C/T	Calcium ion binding protein ID 31101
17	10	0	C/G	MDM34 (mitochondrial outer membrane protein)
11	9	0	C/T	YHN8
14	8	0	A/G	POT11 (3-ketoacyl-CoA thiolase B)
140	7	2	A/C	UQC2 (Ubiquinol-cytochrome-c reductase complex)
31	21	0	A/T	FBX1 (Leucine rich repeat protein. Contains F-box)
96	20	31	C/A	Intergenic region
86	15	19	C/A	Intergenic region
28	10	8	T/A	URA3

**Table 1.** SNP identification using the SOLiD™ System. SNPs that fit the predicted profile of being mutated in mutant, but not in wild type, are shaded in green. Red boxes indicate SNPs that did not fit the predicted profile, but were later identified as being in a region of duplication.

**3.** Digest restriction enzyme and sequentially sequence from P1 and internal primers.



**4.** Map sequences back to REFERENCE GENOME — expected distance is 3 kb based on library.

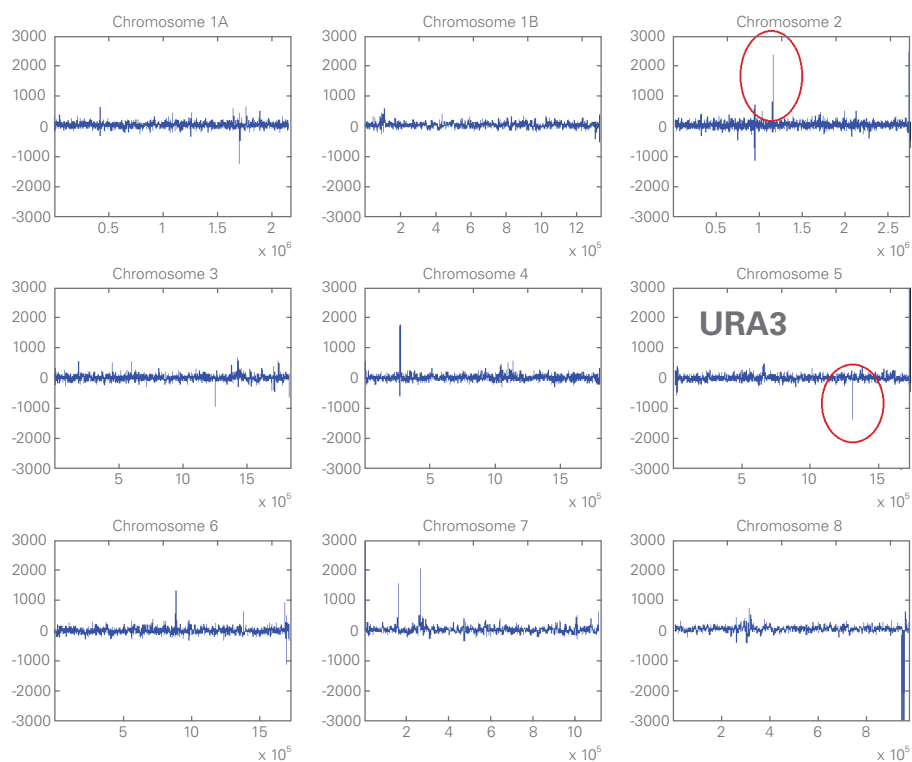


that change back to its location and identify break points. The SOLiD System is uniquely suited to mapping structural variation as it is the only next-generation platform to support mate-paired libraries with inserts up to 10 kb.

Deviations in the distance between mate-paired sequences from the predicted distance were graphed for each chromosome (Figure 3) providing a whole genome view of structural variation for this system. Deletions are depicted as large positive spikes as highlighted on chromosome 2. Insertions are depicted as negative spikes as seen in chromosome 5. Interestingly, the URA3 SNP identified in the fragment analysis mapped to the insertion on chromosome 5. Mate-paired analysis later revealed that this region is actually diploid and heterozygous, due to genetic manipulation, whereby the original gene was mutated and an additional selectable copy was added.

### Conclusion

Large-scale whole genome resequencing is possible with the SOLiD System. The system's three gigabase throughput provided the necessary coverage to accurately identify SNPs across the entire genome. Construction and analysis of a mate-paired library with 3 kb inserts provided the resolution required to accurately characterize insertions and deletions. The SOLiD System enables whole genome resequencing with substantially less time and resources than previously possible.



**Figure 3.** Deviation from expected value of the distance between mate-paired sequence tags. Deletions appear as positive spikes in deviation and insertions are depicted as negative spikes in the deviation.

For Research Use Only. Not for use in diagnostic procedures.

© 2007 Applied Biosystems. All rights reserved. All other trademarks are the property of their respective owners. Applera, Applied Biosystems, and AB (Design) are registered trademarks and SOLiD is a trademark of Applera Corporation or its subsidiaries in the U.S. and/or certain other countries.

Printed in the USA. 10/2007 Publication 139AP02-01