Genome-Wide Single-Nucleotide Polymorphism Map for
* Candida albicans*†

Anja Forche,1,2 P. T. Magee,1 B. B. Magee,1 and Georgiana May2*

*Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, Minnesota 55455,* and *Department of Ecology, Evolution and Behavior, University of Minnesota, St. Paul, Minnesota 55108*†

Received 9 January 2004/Accepted 4 March 2004

Single-nucleotide polymorphisms (SNPs) are essential tools for studying a variety of organismal properties and processes, such as recombination, chromosomal dynamics, and genome rearrangement. This paper describes the development of a genome-wide SNP map for *Candida albicans* to study mitotic recombination and chromosome loss. *C. albicans* is a diploid yeast which propagates primarily by clonal mitotic division. It is the leading fungal pathogen that causes infections in humans, ranging from mild superficial lesions to severe, life-threatening diseases in patients with suppressed immune systems. The SNP map contains 150 marker sequences comprising 561 SNPs and 9 insertions-deletions. Of the 561 SNPs, 437 were transition events while 126 were transversion events, yielding a transition-to-transversion ratio of 3:1, as expected for a neutral accumulation of mutations. The average SNP frequency for our data set was 1 SNP per 83 bp. The map has one marker placed every 111 kb, on average, across the 16-Mb genome. For marker sequences located partially or completely within coding regions, most contained one or more nonsynonymous substitutions. Using the SNP markers, we identified a loss of heterozygosity over large chromosomal fragments in strains of *C. albicans* that are frequently used for gene manipulation experiments. The SNP map will be useful for understanding the role of heterozygosity and genome rearrangement in the response of *C. albicans* to host environments.

Single nucleotide markers are essential tools for studying a variety of organismal properties and processes, such as recombination, chromosomal dynamics, genome rearrangement, and genetic relatedness between individuals. If one compares the same stretch of DNA for two or more individuals, nucleotide polymorphisms are the most frequently observed differences at the nucleotide level within diploid organisms. Single-nucleotide polymorphisms (SNPs) can be located in coding regions of genes or in intergenic regions, where they are most abundant (6, 25). In coding regions, SNPs can alter the function and structure of encoded proteins, e.g., proteins involved in drug metabolism (27, 44). In humans, single-nucleotide substitutions are the cause for most of the known recessively or dominantly inherited monogenic disorders, and missense SNPs also often contribute to common diseases (31, 50). SNPs are estimated to occur once every 1 kb throughout the human genome and are being targeted for association mapping of disease susceptibility genes (42) and used to study traits of diseases, such as cancers, which are often accompanied by a loss of heterozygosity at SNP loci (12). High-density SNP maps were developed for the model organism *Caenorhabditis elegans* to facilitate rapid gene mapping, to characterize natural isolates, and to compare different strains of *C. elegans* (23, 25, 51, 60). To make large-scale studies of SNPs possible, high-throughput methods such as SNP microarrays have been developed to analyze regions of interest in humans (14, 38, 54) and for genetic mapping projects and recombination studies in plants (7, 46).

This paper describes the development of a SNP map for *Candida albicans* to study recombination as a genome-wide process. *C. albicans* is a diploid yeast which propagates primarily by clonal mitotic division (18, 20, 40, 62) and demonstrates high levels of heterozygosity (41, 53, 56, 57). It is the leading fungal pathogen that causes infections in humans, ranging from mild superficial lesions in healthy individuals to severe, life-threatening diseases in patients with suppressed immune systems (5). To date, the causes for the transition from a commensal to a pathogenic organism are not entirely clear. Certainly, factors such as host susceptibility play an important role, as does the developmental transition from a budding yeast form to a filamentous form, which is essential for virulence (11, 15, 22). High levels of heterozygosity and genome rearrangement remain relatively unexplored sources of variation for adaptive responses of *C. albicans* populations to host environments.

Clinical isolates of *C. albicans* exhibiting high levels of variation among strains have been studied in great detail by such techniques as randomly amplified polymorphic DNA analysis (35), restriction fragment length polymorphism (RFLP) analysis (62), karyotyping (32, 33, 43), and DNA fingerprinting (39). Correlations between multilocus genotypes and the evolution of drug resistance have been examined in populations of *C. albicans* with the help of heterozygous genomic loci (10). In addition, SNPs have been used for population genetic studies with *C. albicans* to demonstrate that some form of recombination must also occur, albeit at a low level (18, 20).

SNP loci, as well as larger heterozygous genomic regions,
can be used to study the mechanism and rates of mitotic recombination by an analysis of loss of heterozygosity (LOH). Forche et al. (17) constructed strains that were heterozygous at the GAL1 locus and determined the in vivo frequencies of mitotic recombination at that locus through the loss of the second copy of the GAL1 gene. Others have examined the rate of loop-out in transformation cassettes used for gene disruption in C. albicans and have reported mitotic recombination rates ranging from $6 \times 10^{-4}$ to $6 \times 10^{-8}$/cell/generation (13, 16). In a recent study, Yesland and Fonzi found that the integration of a disruption cassette is allele, and thus sequence-dependent. The failure to disrupt the second copy of a gene was due not to its essential nature but to allelic variation in the flanking regions of that gene (63). Mitotic recombination and LOH occur at observable rates and frequencies in vitro and in vivo, respectively, suggesting that C. albicans is a tractable system for studying the role of recombination in the evolution of virulence.

Our goal was to use SNP loci to study genome-wide mitotic recombination in C. albicans in vitro and in vivo and to understand the role of mitotic recombination and genome rearrangement in the evolution of virulence. With the complete genome sequence of strain SC5314 of C. albicans now available (http://www-sequence.stanford.edu:8080/haploid19.html), abundant heterozygosity is apparent and is represented by approximately 62,000 transitions and transversions across the 32-Mb diploid genome (24). Using these data for strain SC5314, we developed a SNP map representing the entire genome of C. albicans and show here that we can detect LOH by using these markers. We address the concern that LOH, through chromosome loss and reduplication or mitotic recombination, might confound the interpretation of results for targeted manipulations by assessing LOH at a number of SNPs flanking targeted loci (13, 16, 17, 36, 61).

### MATERIALS AND METHODS

#### Strains and DNA extraction

The strains used for this study are listed in Table 1. DNA extraction was performed by a quick prep method that is described elsewhere (1).

#### Development of SNP markers

The goal for the development of the SNP map was to design one SNP marker for every 100 to 200 kb of sequence and to space the SNP markers as evenly as possible across the genome. For this study, we defined a SNP marker as a small PCR product ranging in size between 100 and 400 bp and containing at least one SNP. To design SNP markers, we took advantage of the available genome sequence represented as contigs-6 (http://www-sequence.stanford.edu:8080/haploid19.html) and the contig-19 assembly of sequencing contigs-6 (http://www-sequence.stanford.edu:8080/haploid19.html) at the Stanford C. albicans genome web site (Fig. 1). Our method was as follows. First, two overlapping sequence contigs from a 10× genome coverage, representing the two potentially heterozygous alleles for each genomic region, were selected from the map option on the contig-19 sequence web page for C. albicans (http://www-sequence.stanford.edu:8080/haploid19.html) (Fig. 1A). Combinations of such pairs were further examined only when at least one of the two contigs-6 sequences was mapped to the corresponding chromosome. Complete sequences for both contigs were then obtained from the web site http://www-sequence.stanford.edu:8080/contigs-to-blast6.html (Fig. 1B). The contig sequences were compared by using the BLAST algorithm (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html) to get the exact start-to-end base positions of the overlapping sequence (Fig. 1C). This step was followed by obtaining the actual sequences for the overlapping parts of the contigs at the web site http://candida.cshl.edu/contig_get_orf.html (Fig. 1D). The two sequences were then aligned with ClustalX software, version 1.81 (Fig. 1E), to locate heterozygous positions. Forward and reverse primers flanking apparent SNPs were designed. To test their compatibility and to ensure the amplification of a single product, we performed PCR simulations with the software Amplidyne, version 1.2. Primer pairs that gave a single, strong amplification product in Amplidyne were synthesized by IITDNA (Iowa City, Iowa). To confirm the heterozygous base pair positions, we amplified regions by PCRs, using the same primer pairs and the genomic DNA of strain SC5314, and conducted a sequence analysis (Fig. 1G; also see below). Thirty-one polymorphic markers from previous studies (9, 18, 20, 62) were obtained and evaluated for polymorphisms in strain SC5314 as described below.

#### Confirmation of heterozygosity of SNP markers by sequencing

To confirm the heterozygous polymorphisms obtained from the genome sequence and reported in other studies, we again sequenced each marker for strain SC5314. Internal primers were designed to sequence through larger PCR products (e.g., RFLP markers from reference 62). PCRs were performed as described above. PCR products were purified by using a PCR purification kit (Qiagen Inc., Valencia, Calif.), and 50 to 100 ng of purified DNA was used as a sequencing template. Sequencing reactions were performed with both forward and reverse primers according to the manufacturer’s instructions (ABI Big Dye Chemistry; Perkin-Elmer, Boston, Mass.) and were run on an ABI377 automated sequencer. Sequences were edited and aligned with Sequencer, version 3.1.1, software (Perkin-Elmer).

#### Inter- and intragenic locations of SNP markers and changes in predicted amino acid sequence

For all SNP markers, we determined whether the SNPs were located within open reading frames (ORFs) or in intergenic regions. The sequence of each marker was compared against the entire genome sequence of C. albicans (http://www-sequence.stanford.edu:8080/bmcontig6.html) to obtain the exact nucleotide positions within the sequence contigs. The ORF information available at http://www-sequence.stanford.edu:8080/contigs-to-blast6.html was used to determine whether the SNP marker sequence was located inside, outside, or partially within an ORF. For SNPs located within an ORF, we assessed alterations of the predicted amino acid sequence for the entire ORF.

### TABLE 1. Strains of C. albicans used for this study

<table>
<thead>
<tr>
<th>Strain</th>
<th>Parent</th>
<th>Genotype or description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC5314</td>
<td>SC5314</td>
<td>Clinical isolate</td>
<td>24a</td>
</tr>
<tr>
<td>CAI4</td>
<td>SC5314</td>
<td>ura3Δ::imm434/ura3Δ::imm434</td>
<td>16</td>
</tr>
<tr>
<td>RM1000</td>
<td>CAI4</td>
<td>ura3Δ::imm434/ura3Δ::imm434 his1::hisG/his1::hisG</td>
<td>36</td>
</tr>
<tr>
<td>BWP17</td>
<td>RM1000</td>
<td>ura3Δ::imm434/ura3Δ::imm434 his1::hisG1/hisG4::hisG</td>
<td>61</td>
</tr>
<tr>
<td>AF6</td>
<td>CAI4</td>
<td>ura3Δ::imm434/ura3Δ::imm434 gall::URA3/GAL1</td>
<td>17</td>
</tr>
<tr>
<td>AF14</td>
<td>AF6</td>
<td>ura3Δ::imm434/ura3Δ::imm434 his1::hisG1/hisG4::hisG</td>
<td>17</td>
</tr>
<tr>
<td>AF12</td>
<td>RM1000</td>
<td>ura3Δ::imm434/ura3Δ::imm434 his1::hisG1/hisG4::hisG</td>
<td>17</td>
</tr>
<tr>
<td>AF27</td>
<td>AF12</td>
<td>ura3Δ::imm434/ura3Δ::imm434 his1::hisG1/hisG4::hisG</td>
<td>17</td>
</tr>
</tbody>
</table>
FIG. 1. Flow chart for web-based development of SNP markers in *C. albicans*, with marker 1363/2056 used as an example. Two overlapping sequence contigs representing the two potentially heterozygous alleles for each genomic region were selected (A). Complete sequences for both contigs were then obtained (B) and compared by using the BLAST algorithm (C), followed by obtaining the actual sequence for overlapping parts of the contigs (D). The two sequences were aligned (E), heterozygous positions were identified, and forward and reverse primers flanking apparent SNPs were designed and synthesized. PCR simulations were performed (F), primer pairs were synthesized, and SNPs were confirmed by sequencing (G).
LOH in laboratory strains of *C. albicans*. We were interested in whether processes such as transformation or growth on selective media, such as 5-fluoro- or 2-deoxygalactose (2DG), would lead to LOH in *C. albicans*. Strains that had undergone one or multiple rounds of transformation and passing on selective media were analyzed (Table 1). These strains were derived from previous transformation experiments from the wild-type strain SC5314. SNP markers flanking altered genomic regions (*URA3, HIS1, ARG4, and GAL1*) were used to assess LOH (Table 2; also see the supplemental material) and then were confirmed by resequencing. However, for four markers developed in our study, not all of the SNPs could be confirmed. For markers 26 and 62, the genome sequence of SC5314 indicated these loci to be homozygous, but our sequences for strain SC5314 indeed revealed the presence of SNPs (see the supplemental material). In total, 131 of the SNP marker sequences used for this study were heterozygous and revealed 561 SNPs and 9 insertions-deletions (indels), for a total of 570 polymorphisms (Table 3).

The frequency of SNPs for our data set varied between 1 per 47 bp on chromosome 4 and 1 per 119 bp on chromosome 7. For all of the chromosomes, the average SNP frequency was 1 SNP per 83 bp (Table 3). Disregarding SNP markers 104 to 107 and markers 143 and 144 because of their very close proximities to each other (Fig. 2), on average our map contains 1 SNP marker every 111 kb in the 16-Mb haploid genome. Of the 561 SNPs, 437 were transition events while 126 were transversion events (Table 3), yielding a transition-to-transversion ratio of 3:1, as expected for a neutral accumulation of mutations.

Mapping SNP markers onto the *C. albicans* genome. For a better visualization of the locations of the SNP markers, a SNP map was created and is shown in Fig. 2 as a cartoon of the eight chromosomes of strain SC5314. Since all but four of the SNP markers developed in this study were obtained from contigs with known genome locations (chromosome and/or Sfl fragment), the vast majority of marker sequences, whether homozygous or heterozygous, could be mapped to their corresponding chromosomal locations. Markers that are heterozygous in strain SC5314 are shown as red stars. Homozygous markers are indicated by blue stars.

### RESULTS

**Development and characterization of SNP markers.** Using a web-based strategy to develop SNP markers (PCR fragments that contain one or more SNPs), we generated 119 potential marker sequences that are heterozygous in strain SC5314 (Fig. 1 and Fig. 2). By adding 31 markers from previous studies (10, 18, 20, 62), we obtained a total of 150 potential SNP markers for strain SC5314 (see the supplemental material). The SNP marker sequences in the set of 150 ranged in size from 120 to 1,821 bp, with an average of 322 bp. We confirmed most of the SNPs reported in the genome sequence of strain SC5314 ([http://www-sequence.stanford.edu:8080/contigs-to-blast6.html](http://www-sequence.stanford.edu:8080/contigs-to-blast6.html)) by resequencing. However, for four markers developed in our study, not all of the SNPs could be confirmed. For markers 26 and 63, only half of the reported polymorphisms were detected by sequencing, and for markers 12 and 52, one SNP could not be confirmed for each (data not shown). We found that 19 of the markers from previous studies were homozygous once they were examined for heterozygosity in SC5314 by resequencing.

The SNP markers correspond either to the two contig-6 sequence numbers from which they were generated or to the corresponding gene.

**Table 2. Detection of LOH in strains of *C. albicans***

<table>
<thead>
<tr>
<th>Marker no.</th>
<th>Marker name</th>
<th>Flanking gene</th>
<th>Chromosome</th>
<th>Genotype of strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>76</td>
<td>DLH1</td>
<td>NA</td>
<td>1</td>
<td>Het</td>
</tr>
<tr>
<td>118</td>
<td>HIS1</td>
<td>5I</td>
<td>3</td>
<td>Het</td>
</tr>
<tr>
<td>120</td>
<td>HIS1</td>
<td>5I</td>
<td>3</td>
<td>Het</td>
</tr>
<tr>
<td>141</td>
<td>ARG4</td>
<td>7G</td>
<td>3</td>
<td>Het</td>
</tr>
<tr>
<td>143</td>
<td>ARG4</td>
<td>7G</td>
<td>3</td>
<td>Het</td>
</tr>
<tr>
<td>22</td>
<td>NA</td>
<td>1S</td>
<td>3</td>
<td>Het</td>
</tr>
<tr>
<td>23</td>
<td>NA</td>
<td>1S</td>
<td>3</td>
<td>Het</td>
</tr>
<tr>
<td>24</td>
<td>MNS1</td>
<td>NA</td>
<td>4</td>
<td>Het</td>
</tr>
<tr>
<td>25</td>
<td>NA</td>
<td>1S</td>
<td>4</td>
<td>Het</td>
</tr>
<tr>
<td>26</td>
<td>NA</td>
<td>1S</td>
<td>4</td>
<td>Het</td>
</tr>
<tr>
<td>27</td>
<td>NA</td>
<td>1S</td>
<td>4</td>
<td>Het</td>
</tr>
<tr>
<td>28</td>
<td>NA</td>
<td>1S</td>
<td>4</td>
<td>Het</td>
</tr>
<tr>
<td>29</td>
<td>GAL1</td>
<td>1S</td>
<td>4</td>
<td>Het</td>
</tr>
<tr>
<td>30</td>
<td>GAL1</td>
<td>1E</td>
<td>4</td>
<td>Het</td>
</tr>
<tr>
<td>42</td>
<td>NA</td>
<td>1L</td>
<td>5</td>
<td>Het</td>
</tr>
<tr>
<td>43</td>
<td>NA</td>
<td>1L</td>
<td>5</td>
<td>Het</td>
</tr>
<tr>
<td>46</td>
<td>NA</td>
<td>1</td>
<td>5</td>
<td>Het</td>
</tr>
<tr>
<td>48</td>
<td>NA</td>
<td>2U</td>
<td>5</td>
<td>Het</td>
</tr>
<tr>
<td>103</td>
<td>NA</td>
<td>5M</td>
<td>5</td>
<td>Het</td>
</tr>
<tr>
<td>113</td>
<td>NA</td>
<td>5M</td>
<td>5</td>
<td>Het</td>
</tr>
<tr>
<td>137</td>
<td>NA</td>
<td>7F</td>
<td>6</td>
<td>Hom</td>
</tr>
<tr>
<td>147</td>
<td>NA</td>
<td>Hom</td>
<td>6</td>
<td>Hom</td>
</tr>
</tbody>
</table>

a Numbering corresponds to the supplemental material and Fig. 2.
b The names of markers correspond either to the two contig-6 sequence numbers from which they were generated or to the corresponding gene.
c Het, heterozygous; hom, homozygous; NA, not applicable.
d Flanks HIS1.
e Flanks ARG4.
f Flanks GAL1.
the *C. albicans* genome sequence) are mapped. The order of many contig-19 sequences was determined by the association between previously mapped genes and their approximate locations on the physical map (http://alces.med.umn.edu/Candida.html). For contig-19 sequences for which the orientation along the chromosome is known (e.g., see chromosomes 5 and 7 [6a, 6b]) the 5’-to-3’ direction of the sequence is indicated by an arrow (Fig. 2).

The SNP markers are distributed evenly along chromosomes 1, 2, 4, 5, and 6, but there are large chromosomal regions on chromosomes R, 3, and 7 which are apparently highly or entirely homozygous in *C. albicans* strain SC5314. Chromosomal fragments RB2 and RD1 on chromosome R show no or very little heterozygosity. On chromosome 3, a region covering three contig-19 sequences and making up a total of 464 kb of *Sfi*I fragment P is apparently homozygous. On chromosome 7, *Sfi*I fragments 7A and 7C are completely homozygous, and a large portion at the telomere-proximal end of 7G also lacks heterozygosity (Fig. 2). In contrast, there are some genomic locations with abundant polymorphisms. Two examples are the sequences from markers 104 to 108 on chromosome 5 M (near the MTL loci) and from markers 143 to 145 on chromosome 7G.

We were interested in whether the SNPs were located in intergenic or intragenic regions and whether intragenic SNPs changed the amino acid coding sequence (nonsynonymous substitutions). Of the 131 heterozygous marker sequences, 40 were located in intergenic regions and 91 were partially or completely within coding regions. At least one nonsynonymous substitution was found in 81 of the 91 (89%) intragenic SNP marker sequences. Only synonymous substitutions were found in 10 of the intragenic SNP marker sequences (11%).

**LOH in selected strains of *C. albicans*.** To address concerns that various genetic manipulations such as transformation or growth on selective media might increase the rate of LOH, we examined loci that had been targets of knockout experiments for evidence of LOH in flanking regions. We used strains that are often used for gene manipulations as well as some that were grown on selective media during the transformation procedure. Strains CAI4, RM1000, BWP17, AF14, and AF27 were derived from SC5314 by consecutive transformations to knock out both copies of *URA3* (CAI4 [16]), *HIS1* (RM1000 [36]), and *ARG4* (BWP17 [61]), in that order. Furthermore, we examined four strains, AF6, AF14, AF12, and AF27, in which one or both copies of *GAL1* had been knocked out (17). One copy of *GAL1* was knocked out in AF6, which was derived from CAI4, and in AF12, which was derived from RM1000.
TABLE 3. Summary of SNP marker characterization

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>No. of markers a,b</th>
<th>Total no. of polymorphisms</th>
<th>No. of base pairs sequenced</th>
<th>SNP frequency (1 SNP/no. of bp)</th>
<th>No. of SNPs</th>
<th>No. of indels</th>
<th>No. of transitions</th>
<th>No. of transversions</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>18</td>
<td>3</td>
<td>56</td>
<td>1/98</td>
<td>55</td>
<td>1</td>
<td>47</td>
<td>4 2 2</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>2</td>
<td>103</td>
<td>8,117</td>
<td>1/79</td>
<td>101</td>
<td>2</td>
<td>76</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>3</td>
<td>74</td>
<td>5,690</td>
<td>1/77</td>
<td>74</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>2</td>
<td>38</td>
<td>3,739</td>
<td>1/98</td>
<td>36</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>0</td>
<td>101</td>
<td>4,785</td>
<td>1/47</td>
<td>99</td>
<td>2</td>
<td>76</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>2</td>
<td>75</td>
<td>8,103</td>
<td>1/108</td>
<td>74</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>1</td>
<td>60</td>
<td>2,894</td>
<td>1/48</td>
<td>60</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>4</td>
<td>53</td>
<td>6,288</td>
<td>1/119</td>
<td>53</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>Unknown</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>1,959</td>
<td>1/196</td>
<td>9</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>131</td>
<td>19</td>
<td>570</td>
<td>47,190</td>
<td>1/83</td>
<td>561</td>
<td>9</td>
<td>437</td>
</tr>
</tbody>
</table>

a Markers are ordered and summarized by chromosomal location. For each chromosome, the number of heterozygous and homozygous SNP markers is given followed by the total number of SNP markers. The number of bp sequenced is listed followed by the frequency of SNPs. The number of SNPs and indels is given and the total number of SNPs is categorized into transition and transversion events.

b Heterozygous and homozygous markers were based on sequences obtained for C. albicans strain SC5314.

The second copy of GAL1 was deleted from strain AF6 to generate AF14 by growth on a 2DG (toxic to Gal+) medium, so AF14 is in a CAI4 background. The second copy of GAL1 was deleted from AF12 to generate AF27 by two sequential rounds of transformation without growth on selective medium, so AF27 is in an RM1000 background (17). The early steps in the construction of most of these knockout strains involved growth on 5-fluoroorotic acid to select for the loss of URA3 due to loop-out events (Table 1).

We sequenced SNP markers adjacent to gene knockouts (URA3, HIS1, ARG4, and GAL1) (see Fig. 2 for genomic locations) from strains CAI4, RM1000, BWP17, AF14, and AF27. We did not observe LOH at URA3 in CAI4 or any of the descendant strains. For HIS1, marker 120, located at the distal side of HIS1, demonstrated LOH in strains RM1000, BWP17, and AF27, but not in CAI4 and its descendant strain AF14 (Fig. 3A, arrow). Thus, the LOH in marker 120 might have accompanied HIS1 mutations in RM1000 such that LOH is observed in RM1000 descendant strains but not in CAI4 descendant strains. SNP marker 118, flanking the proximal side of HIS1, was still heterozygous in all five strains. We did not observe LOH at markers flanking ARG4 in BWP17, the strain from which ARG4 was deleted, or in any of the other strains. Thus, ARG4 deletions were not accompanied by LOH.

To analyze the LOH for strains in which GAL1 manipulations had been done, we initially sequenced markers flanking the GAL1 gene on chromosome fragment 1S for strains AF14 and AF27. Marker 29 was homozygous in strain AF14 but heterozygous in strain AF27. In contrast, marker 30, which flanks GAL1 on the right and is located on the 1E chromosome fragment, was heterozygous in both strains (Fig. 3B). To determine whether the entire chromosome fragment 1S had lost heterozygosity in AF14, we sequenced the remaining SNP markers on fragment 1S for this strain. All SNP markers did indeed show LOH (Fig. 3B). We then sequenced the markers on 1S for strain AF6, the parent of AF14, and found that all were still heterozygous. Together, these results suggest that the heterozygosity of marker 29 and all 1S markers distal to marker 29 occurred during continued selection on 2DG in the cell lineage leading from AF6 to AF14.

From the above results, we found that LOH is not directly related to gene manipulations or growth on selective media. To confirm this result, we sequenced all of the SNP markers at other genomic locations for the five strains discussed above, and these results also demonstrated LOH (Table 2). For example, all of the markers sequenced from chromosome fragment 1L demonstrated LOH in strains RM1000, BWP17, AF12, and AF27 (Fig. 3C). Together, our results suggest that transformation per se does not seem to increase LOH but that using a medium such as 2DG to select for a mitotic recombination event can result in the recovery of homozygosity across an entire chromosome region.

**DISCUSSION**

This paper describes a web-based approach to developing a genome-wide SNP map for strain SC5314 of the opportunistic human pathogen *C. albicans*. Heterozygosity in this fungus was demonstrated some 20 years ago (56–58), but only recently have molecular and genomic advances allowed for the exploitation of such variation for the study of the biology of infection. The genome sequence reveals approximately 62,000 SNPs (24), which are a great resource with which to study population genetics, the relationships of strains, and transmission genetics. Heterozygosity can also be exploited to unravel the processes contributing to the diversity and plasticity of the *C. albicans* genome.

For this study, we took advantage of the publicly available diploid genome sequence (http://www-sequenc.stanford.edu:8080/haploid19.html) to develop a genome-wide SNP map for *C. albicans*. The SNP map contains 131 markers comprising 561 SNPs and 9 indels (see the supplemental material), incorporates 12 markers from previous studies (10, 18, 20, 62), and includes about 1% of the total number of SNPs in the *C. albicans* genome. The map has one marker every 111 kb, on average.
The frequencies of SNPs per base pair in our data set are quite high and vary over the genome, ranging from 0 in some locations to 1 per 47 bp on chromosome 4 to 1 per 119 bp on chromosome 7, with a genome-wide frequency of 1 per 83 bp. For \textit{C. elegans}, SNPs have been used to study the relationships among strains and to carry out high-throughput gene mapping (25, 51). In one study using 366 SNPs, the SNP frequencies varied from 1 per 1,445 bp to 1 per 8,750 bp (25). Brumfield et al. (4) examined the utility of SNPs to infer population histories and reviewed SNP frequencies obtained from a variety of different organisms. Studies comparable to ours found SNP frequencies ranging from 1 per 1,001 bp in humans to 1 per 2,119 bp in chickens. Since most organisms covered in that review have complete sexual cycles, the most likely explanation for the high SNP frequency we found in \textit{C. albicans} is the accumulation of mutations in this largely clonal organism (55). Some form of balancing selection across variable environments might also maintain heterozygosity against the mitotic conversion (LOH) we observed, but it is difficult to demonstrate without genealogical data for specific coding regions.

Recently, multilocus sequence typing (MLST) was recommended as a standard method for determining relationships among strains (2, 3, 52). Sequences of several coding regions (housekeeping genes) are combined into multilocus genotypes to determine the relatedness among strains. Because the SNP markers presented here are distributed across all of the chromosomes, while the MLST markers represent only five of the eight chromosomes, the two systems could be used together in studies of genome regions covered by both. We recommend the use of SNP markers for genome-wide analyses of recombination. A combination of SNP markers and MLST would not only increase the discriminatory power of strain typing, but would unite data from noncoding and coding regions across the entire genome.

To study mitotic recombination, including gene conversion at heterozygous loci, we sequenced several SNP markers for several strains (Table 2; Fig. 3) that were derived sequentially from SC5314 by transformation (see Results and Table 1). We found little evidence that gene knockouts directly increase LOH at flanking regions. No LOH was found for markers flanking \textit{URA3} and \textit{ARG4} in knockout strains. However, marker 120 (Fig. 2; also see the supplemental material), next to \textit{HIS1} and located at the distal end of chromosome fragment 5I, was homozygous for all SNPs tested, while marker 118, approximately 50 kb to the left end of \textit{HIS1}, was not. Karyotypic analysis has shown that strain RM1000 carries two differently sized chromosome 5 homologues, whereas its progenitor strain, CAI4, carries chromosome 5 homologues of the same size.
size (data not shown). Further analysis showed that the gene ILV3, which is telomeric to HIS1, is missing from the smaller homologue of chromosome fragment 5I in RM1000 (unpublished data). These facts suggest that a partial loss and reconstitution of the telomeric end of one homologue of chromosome fragment 5I occurred in strain RM1000 (Fig. 3A) and may have occurred during the disruption of HIS1. In the second case of large chromosomal fragment changes that we observed, chromosomal fragment 1S (~1,400 kb) lost heterozygosity during growth on 2DG medium (Fig. 3B and C). Our data indicate that a single mitotic crossover in strain AF6 caused the loss of the second GAL1 allele, generating strain AF14. This event seems to have occurred spontaneously, in line with previous observations of a low but finite rate of mitotic recombination in Candida.

We investigated the genomic locations of the SNP markers with respect to predicted ORFs. More than two-thirds of the heterozygous markers (89%) were located in intragenic regions with the potential to affect the protein sequence or expression level (26). Diverged functions of two alleles at the SAP2 gene were demonstrated by Staib and coworkers (48), who showed that differences in the numbers of pentanumeric repeats in the promoter region led to differential regulation of the alleles. It was recently shown that allelic variation in gene expression is frequent in the human genome (30). From a SNP array, 326 of 602 (54%) genes expressed in kidney and liver tissues showed two- to fourfold differences between the expression levels of alleles and 170 genes exhibited more than fourfold differences in allelic expression levels. Given the high level of heterozygosity in the C. albicans genome, the differential expression of genes seems likely.

The main goal of this paper was to put together tools to study recombinational and genomic changes in C. albicans under different environmental conditions, such as changing media, growth conditions, exposure to antifungal drugs, and infection in animal hosts. For example, several mechanisms might lead to resistance to the antifungal drug fluconazole in C. albicans. One mechanism involves LOH at the ERG11 locus, a gene in the ergosterol biosynthesis pathway that is targeted by fluconazole. White (59) studied 17 serial isolates of C. albicans and discovered that isolate 13 and all subsequent isolates became resistant to fluconazole. An analysis showed that allelic differences within the ERG11 ORF, in the ERG11 promoter, and in the downstream THR1 gene were lost in these isolates, and a single amino acid substitution in ERG11 was also identified. Another group discovered that multiple mechanisms, including LOH, contribute to a stepwise development of antifungal drug resistance in clinical isolates (19). The fact that LOH at and around the ERG11 gene was discovered by two independent groups suggests that homozgyosity at this locus increases fitness in a fluconazole environment. In a second example of the importance of recombination, rates and frequencies of mitotic recombination have been determined for the GAL1 locus in C. albicans in vitro by the use of fluctuation tests and in vivo by the use of a mouse model (17). Strains that were heterozygous at the GAL1 locus were injected into mice, Gal− strains were identified, and the frequencies of mitotic recombination at the GAL1 locus were determined. Although in vitro and in vivo data sets cannot be compared directly, these results suggest that there may be a higher level of mitotic recombination in vivo than in vitro (17).

So far, sequencing has been the method of choice for SNP analysis. However, analyzing large numbers of SNP markers for entire populations of C. albicans by sequencing is very time- and labor-intensive. Recently, SNP microarrays have been developed for a variety of organisms, including humans (14, 21, 37, 45), mice (28), plants (7), and protozoa (47, 49), and they are becoming a fast, accurate, and reproducible method for SNP analysis (7, 12, 29, 34). We have developed a SNP microarray (unpublished data) for C. albicans to study the mitotic recombination of populations of cells in vitro and in vivo to learn more about how and by which genomic processes C. albicans adapts to diverse and constantly changing host environments. This will provide a first glimpse of what makes C. albicans so successful as an opportunistic pathogen of humans.

ACKNOWLEDGMENTS

We thank Christine Ramos for her excellent technical assistance. Sequence data for C. albicans were obtained from the DNA Sequencing and Technology Center web site at http://www-sequence.stanford.edu/group/candida. This work was supported by NIH grant AI46351 awarded to P. T. Magee and G. May and by contract AI054006.

REFERENCES


