Genetic Diversity in Yeast Assessed With Whole-Genome Oligonucleotide Arrays

Elizabeth A. Winzeler,*[,1] Cristian I. Castillo-Davis,[†] Guy Oshiro,* David Liang,* Daniel R. Richards,[‡] Yingyao Zhou* and Daniel L. Hartl[†]

**Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, [†]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138 and [‡]Department of Genetics, Stanford University School of Medicine, Stanford, California 95307*

Manuscript received August 4, 2002
Accepted for publication October 21, 2002

ABSTRACT

The availability of a complete genome sequence allows the detailed study of intraspecies variability. Here we use high-density oligonucleotide arrays to discover 11,115 single-feature polymorphisms (SFPs) existing in one or more of 14 different yeast strains. We use these SFPs to define regions of genetic identity between common laboratory strains of yeast. We assess the genome-wide distribution of genetic variation on the basis of this yeast population. We find that genome variability is biased toward the ends of chromosomes and is more likely to be found in genes with roles in fermentation or in transport. This subtelomeric bias may arise through recombination between nonhomologous sequences because full-gene deletions are more common in these regions than in more central regions of the chromosome.

WITH few exceptions, only one strain or an individual of a particular species is sequenced while hundreds of other variants, which may be important to public health, scientific research, or commercial applications, remain undeciphered. In the baker's yeast, *Saccharomyces cerevisiae*, a derivative of strain S288c was sequenced. Despite the availability of the sequence information for this strain, many full-genome studies, including gene expression studies (Chu *et al.* 1998), genome-wide chromatin-binding studies (Wyrick *et al.* 2001), and studies of the replication dynamics of the yeast genome (Raghuraman *et al.* 2001), have been conducted using alternative but commonly used yeast strains. In some cases, the strain may be directly related to S288c (A364, W303, and Σ1278b derivatives) and in some cases the strain may be completely unrelated (SK1). As many of these studies rely on oligonucleotide or cDNA arrays that were derived from S288c sequence information, the quality of the data may differ depending on the region of the genome under investigation and on whether or not the region is identical by descent to that of S288c. These strain differences could contribute to some of the inconsistencies in genome-wide data sets (Grunenfelder and Winzeler 2002). Mortimer and Johnston (1986) have traced the pedigrees of some laboratory yeast strains as far as is known, but direct ascertainment of the relationships between laboratory strains, in the absence of full-genome sequencing, has been impossible.

[1]*Corresponding author:* Department of Cell Biology, The Scripps Research Institute, 10550 Torrey Pines Rd., La Jolla, CA 92037.

Single-base changes between two sequences 25 bp in length, especially those found in the central region of a 25mer, disrupt their hybridization (Chee *et al.* 1996; Gingeras *et al.* 1998; Troesch *et al.* 1999; Lockhart and Winzeler 2000). Thus, with oligonucleotide arrays carrying large numbers of probes of this length (termed features), the approximate location of allelic variation between two strains can be discovered (Winzeler *et al.* 1998). Since the locations of all the features in the genome are generally known, the approximate position of a predicted polymorphism between the two strains can also be ascertained when the genomic DNA hybridization patterns are compared (Figure 1). Such hybridization differences have been termed "single-feature polymorphisms" (SFPs; Borevitz *et al.* 2003).

The Affymetrix S98 oligonucleotide array contains 285,156 different 25mers from the yeast genomic sequence. Although this array was designed primarily to be a tool for gene expression analysis, it was also created to maximize the amount of yeast genome sequence covered. In addition to probes to all annotated genes in the yeast genome, probes to small nonannotated genes (Oshiro *et al.* 2002), probes to nontranslated RNA, and probes to noncoding regions were included in the design. Although a percentage of these probes overlap one another, altogether ~16% of the yeast genome is probed by this array design. We reasoned that because of the high degree of coverage, these arrays could be used to identify a significant proportion of the genetic variation existing between strains and that this information could then be used to determine strain relationships. In addition, through the inclusion of several wild isolates, we have characterized the distribution of allelic

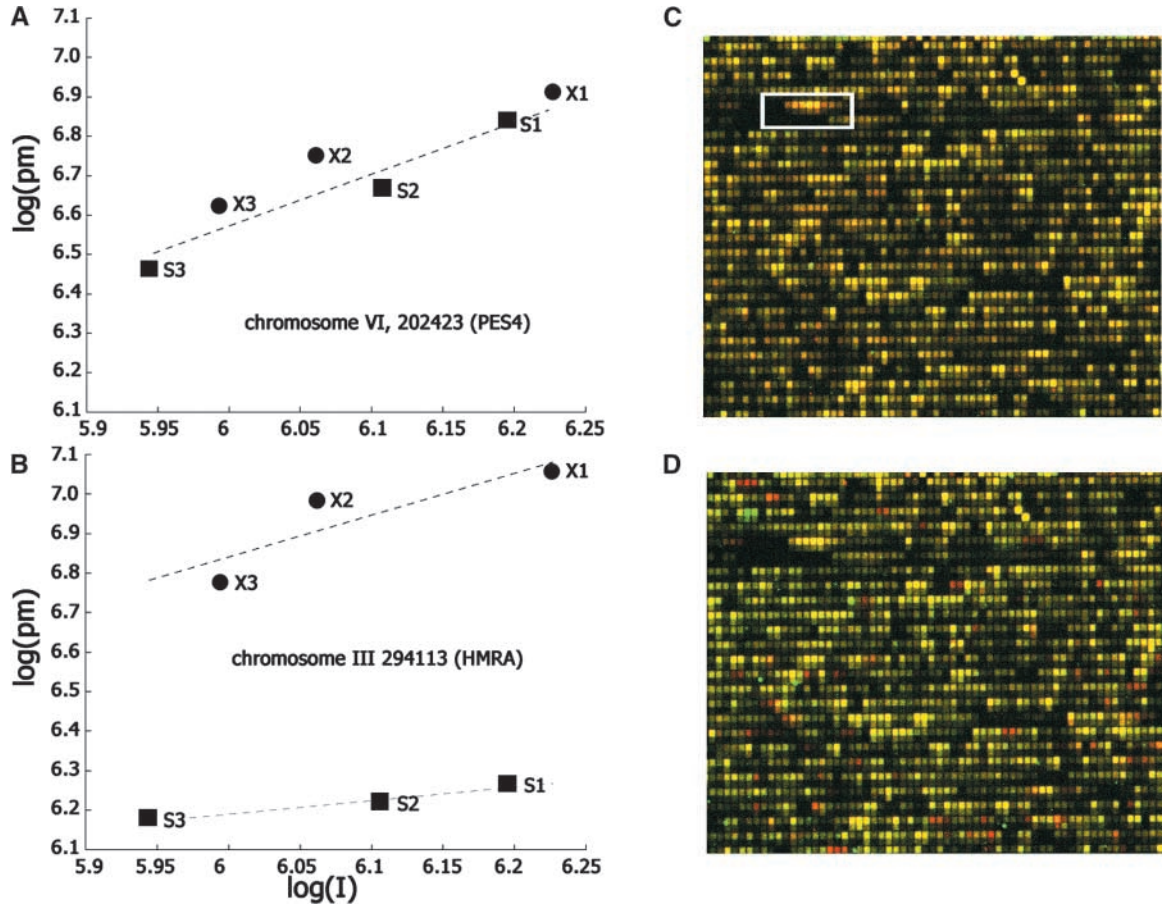

FIGURE 1.—Example of hybridization pattern for probes that detect nonvariant (A) or variant (B) alleles. The logarithm of the probe intensity (pm) for six different hybridizations relative to array intensity (I), three from S288c (solid square) and three from X2180A (solid circle), is shown. Polymorphic probes were identified as those that were best fit by a two-signal model with high confidence. Not every polymorphism is expected to be detected, because of poor probe performance, hybridization saturation, or differences that were below the thresholds used in the analysis. (C) Comparison of X2180-1A *vs.* S288c hybridization pattern. (D) SK1 *vs.* X2180-1A. In both cases the X2180-1A signal is depicted in red while the second signal is green. Probes to the mating-type locus (HMRA) are highlighted. Only a portion of the array is shown.

variability in the genome itself to determine whether or not particular regions of the genome, or classes of genes, might show higher rates of variability.

MATERIALS AND METHODS

Yeast strains: Most yeast strains were obtained from the American Type Culture Collection (ATCC). Strains M1-2, M2-8s2, and M2-8f2 were obtained from Duccio Cavaliero and Y102, YJM789, Σ1278b, and 3962c were kindly provided by John McCusker.

Sample preparation: Yeast strains were routinely grown in yeast extract, peptone, and dextrose (YEPD) medium. Genomic DNA was purified from 20 ml YEPD cultures using a QIAGEN (Chatsworth, CA) genomic extraction kit according to the manufacturer's protocol with slight modifications. Zymolyase and Protease K digestion times were extended from 30 to 45 min. The genomic DNA was ethanol precipitated and resuspended in 100 μl 10 mm Tris-Cl (pH 8.5). Ten micrograms of yeast genomic DNA was fragmented to an average size of 25 bp with 1 unit of DNase I (Promega, Madison, WI) in 1× One-Phor-All Buffer (Pharmacia, Piscataway, NJ) and 1.5 mm cobalt chloride (Boehringer Mannheim, Indianapolis) for 5 min at 37°. DNase I was inactivated by incubation at 99° for 15 min. After heat inactivation of DNase I, the DNA fragments were end-labeled in the same buffer by the addition of 20 units of terminal deoxynucleotidyl transferase (Promega) and 1 nmol Biotin-N6-ddATP (New England Nuclear, Boston) for 1 hr at 37°. Each sample was hybridized to the array in 260 μl containing 1× MES buffer [100 mm MES, 1 m (Na^+), 20 mm EDTA, 0.01% Triton X-100], 30 μg herring sperm DNA (Promega), 150 μg BSA (GIBCO-BRL, Gaithersburg, MD), and 15 nmol of 213B 3-biotin control oligonucleotide that hybridizes to control features on the gene array. Samples were heated to ≥95° for 10 min, placed on ice for 5 min, and then applied to the gene array. Hybridizations were carried out at 45° for 20 hr with mixing on a rotisserie at 60 rpm. Following hybridization, the solutions were removed, the arrays were washed with nonstringent wash A buffer [6× SSPE-T (0.9 m NaCl, 60 mm NaH_2PO_4, 6 mm EDTA, 0.01% Triton X-100, pH 7.6, 25°)], followed by stringent wash in B buffer (1× MES buffer, 0.1 m NaCl, 0.01% Triton-X 100, 50°). Arrays were then stained with R-phycoerythrin-streptavidin (10 μg/ml; Molecular Probes, Eugene, OR) in 1× staining buffer [100 mm MES, 1 m (Na^+), 0.05% Triton X-100] with BSA (2 mg/ml) for 10 min at 25°, followed by rinsing with wash A buffer. The signal was amplified with a biotinylated antistrep-

tavidin antibody (2.25 μg biotinylated antistreptavidin antibody; Vector Laboratories, Burlingame, CA) in 1× staining buffer, with 1.5 mg BSA and 75 μg normal goat IgG (Sigma Chemical, St. Louis) in 750 μl, followed by a second streptavidin-phycoerythrin staining, according to standard Affymetrix protocols. All washes were automated on a fluidics station (Affymetrix). Gene arrays were then scanned at an emission wavelength of 560 nm at 3 μm resolution using a specially designed confocal scanner (Affymetrix). The hybridization intensity for each 25-bp probe from each scan was computed using Affymetrix GeneChip software and then the scanned images were normalized (Li and Wong 2001).

Data analysis: Only perfect match (PM) values were used in the analysis. The computational strategies for detecting polymorphisms were essentially as described (Winzeler *et al.* 1998): An adjusted hybridization intensity value I was determined for each hybridization as the mean of the log(PM) signals of all unique probes (126,645) that showed minimal variation across all hybridizations for the sequenced strains (6) and the test strain (3) in that they showed a standard deviation of <0.15. Then, for each unique probe on the array the linear regression of log (PM) on I for all hybridizations was determined by the least-squares method, first under the null hypothesis that the sequenced strain and the test strain probe show the same response and then under the alternative hypothesis that the responses were different. The models were compared using an *F*-test and the same signal model was rejected in favor of a polymorphic model with 98% confidence ($F = 35$). Then the regression equations for the two strains were compared. Probes were discarded if the sequenced strain regression equation evaluated for both the lowest PM and the highest PM for all hybridizations was not higher than when evaluated with the test strain regression equation. Finally, P was computed as $P(\text{sequenced strain})/[P(\text{sequenced strain}) + P(\text{test strain})]$, where $P(X)$ is the probability from the t distribution that a probe has genotype X based on the observed PM hybridization signal for the probe and the expected signal and on the estimated variance from the regression equation. Probes with $P > 0.05$ were discarded. The number of probes detecting allelic variation is an estimation of the number of polymorphisms, because one oligonucleotide probe may cover two distinct polymorphisms and some polymorphisms are probed by multiple oligonucleotide probes. If probes that overlap by >15 bases are merged, the number of SFPs is reduced to 9670. The parameters used to find variation were conservative and it is expected that some of the existing variability might have been undetected. By increasing the number of hybridizations and loosening the stringency, it is expected that a larger number of cases of allelic variation could be identified.

The locations of 435 Ty or Ty long terminal repeats were obtained at ftp://genome-tp.stanford.edu/pub/yeast/tables/Other_Features_Locations/other_features_table.txt.

Identification of clusters: All polymorphic probes within a distance of 30 kb were grouped into one cluster. Clusters with fewer than three probes were dismissed to reduce statistical fluctuation. Tightness of probe distribution could then be measured by comparing total regions covered by all clusters to the full chromosome length. The P values for clustering patterns were estimated using the following computer simulation procedures: (1) The same number of probes was redistributed to locations randomly sampled from the whole chromosome; (2) the same clustering routine described above was applied to calculate the total length of all clusters; and (3) the simulation was repeated 900 times. If n random simulations resulted in a total group length shorter than the observed value, the P value was estimated as n/N.

Sequencing: To verify the polymorphisms, oligonucleotide primers from ~300 bases upstream and downstream of the probe location were selected. The regions were amplified and both strands were cycle sequenced. The polymorphism in many cases was located within the central region of the 25-bp probe, but in some cases it was within 5 bases of either end.

RESULTS

Identification of SFPs: For this study, 14 different strains of yeast were chosen, either because of their relevance to yeast researchers or because they were wild isolates. For example, strain A364A is widely used in studies of the cell cycle (Hartwell 1967), while Σ1278b is often employed when examining pseudohyphal growth (Liu *et al.* 1993). SK1, W303, and Y102 are used in the analysis of meiosis because they undergo sporulation much more readily than the sequenced strain (McCusker and Haber 1988; Chu *et al.* 1998; Primig *et al.* 2000). In addition, strains that were known to be virtually identical to one another (S288c and X2180-1A) were included as controls. To identify SFPs that could be used to distinguish strains, ~10 μg of genomic DNA from each of the strains listed in Table 1 was digested with DNAse I, end labeled using terminal transferase and biotin-ddATP, and then hybridized to the high-density oligonucleotide array (Wodicka *et al.* 1997; Winzeler *et al.* 1999; see materials and methods). This procedure was repeated three times with each strain for a total of 42 hybridizations. Individual oligonucleotide probes that exhibited statistically significant differential hybridization between the reference strain (X2180-1A, S288c, or S288c and X2180-1A, considered together for added statistical confidence) and any other strain were designated as SFPs (Figure 1; see materials and methods). Altogether 11,115 SFPs were identified between the reference strain and at least one of the other strains included in this study (Table 2).

Validation of SFPs: To validate that these SFPs were genetic polymorphisms several tests were performed. First, to assess the false-positive rate, sequencing was performed on 24 SFPs. Pairs of oligonucleotide primers from ~300 bases upstream and downstream of the SFP were selected. The regions were amplified using DNA from W303, SK1, and Y102 or from X2180 and then both strands were cycle sequenced. As expected, in all cases (72) where good sequence was obtained, a sequence polymorphism was or was not found, as predicted within the 25-bp probe region. The polymorphism was in all cases located within the central 20-base region of the 25-bp probe.

In addition, we calculated the false-negative rate: High-quality genome sequence is available for ~30 kb for four of the strains in the study (GenBank accession nos. AF458975, AF458977, AF458977, and AF458969). There are 254 probes to this 30-kb region on the array. Of the 254 probes, 32 are to regions that contain sequenced polymorphisms that distinguish one of the three strains (W303, YJM789, or SK1) from the se-

TABLE 1

Strains of yeast used in this study

Strain	SFPs	Frequency (per base pair)[a]	Genotype[b]	Reference
S288c	3	~0	*MAT*α *SUC2 mal mel gal2 CUP1 flo1 flo8-1*	Mortimer and Johnston (1986)
A364A	1051	633	*MAT***a** *ade1 ade2 ura1 his7 lys2 tyr1 gal1 SUC mal cup (S)*	Hartwell (1967)
SK1	3754	177	*MAT***a**/*MAT*α *HO can1 (R) gal2 cup(s)*	Kane and Roth (1974)
W303	885	751	*MAT***a** *ade2-1 ura3-1 his3-11 trp1-1 leu2-3 leu2-112 can1-100*	Thomas and Rothstein (1989)
YJM789	1901	350	*MAT*α *ho::hisG lys2 cyh*	McCusker *et al.* 1994; Winzeler *et al.* (1998)
X2180-1A	0	0	*MAT***a** *SUC2 mal mel gal2 CUP1*	Mortimer and Johnston (1986)
Y102	3029	219	*MAT*α *lys2 MAL1 SUC1 gal3*	McCusker and Haber (1988)
EM93	414	1607	*MAT***a**/α *SUC2/SUC2 GAL/gal2 CUP1/cup FLO1/flo1 MAL/MAL mel/mel can (r)*	Mortimer and Johnston (1986)
3962c	1558	427	*MAT***a**	Bechet *et al.* 1970
Σ1278b	1512	440	*MAT*α	Grenson *et al.* 1966; Liu *et al.* (1993)
m28s2	3087	215		Cavalieri *et al.* (2000)
m28f2	2708	245		Cavalieri *et al.* (2000)
M1-2	5401	123	*MAT***a**/α	Cavalieri *et al.* (2000)
99r	214	3109	*MAT*α *SUC2 GAL CUP1 MAL mel ade1 can (r) FLO*	Mortimer and Johnston (1986)

[a] Numbers of polymorphisms per genome are estimations. Sequencing of selected regions of SK1 and Y102 revealed polymorphisms at ~1/200 bases and 1/350 bases in YJM789. Representative regions may not have been chosen.

[b] In most cases haploid strains were used. SK1 is an isogenic diploid, while EM93 and M1-2 are heterothallic wild strains.

quenced strain (S288c). Seven of these polymorphisms were found in the hybridization assay. There were no false positives in this region and the distribution of polymorphisms between the strains was as predicted. In 18 of the 25 misses the polymorphism was outside of the central 10 bases of the probe, and in 16 of the misses it was outside of the central 15 bases. The remaining false negatives were most likely due to poor probe performance. For example, in one S288c hybridization, the average intensity of the 11,115 probes classified as SFPs was 353 units over background, while the average intensity for non-SFP probes was 229 units over background. The average intensity for the remaining seven probes that failed to detect polymorphisms mapping to their central region was 161 units. While the false-negative rate may be high, the resolution of the assay could be improved by performing more hybridizations. In addition, even with this false-negative rate, the overall results described here are unlikely to be affected because most probes behave consistently. For example, if a probe detected a polymorphism for one strain, and the same polymorphism was found in another strain, then the probe would also be classified as an SFP in 9/10 cases for the second strain (64 of 70 examined).

As further confirmation, we expected that very few instances of allelic variation would be found between two closely related strains. Strain X2180-1A was created by the self-diploidization of S288C (Mortimer and Johnston 1986), and only 3 of the 126,645 unique probes on the array were classified as SFPs when the scans of S288C and X2180-1A were compared. One of these probes was located within the mating-type locus, HMRA1 (293,828–294,314), the source of the only known phenotypic difference between the strains. This result is not surprising because X2180-1A (*MAT***a**) carries two copies of this gene, one in the active mating-type locus and one in the silent mating-type locus, while S288C (*MAT*α) harbors only one copy at the inactive site. A total of 214 polymorphic probes were found in strain 99R, the next most closely related strain to the reference strain. In contrast, in strain M1-2, a wild isolate from Tuscany (Table 1), 5401 SFPs were detected. Altogether these data suggest a false-positive rate that is low enough (<2%, based on an *F*-test) to have a negligible effect on the results described here.

Distribution of variation in related strains: To further validate the method as well as to determine the exact relationships between different strains of yeast we asked if SFPs from descendants, ancestors, or collateral relatives of S288C-like strains would show a nonrandom distribution on the chromosome because of genetic linkage. We examined DNA from several such strains including A364A, W303, EM93, 99R, and Σ1278b. Both A364A and W303 were created through crosses between S288c and other strains and are thought to be closely related to S288c (Thomas and Rothstein 1989; Esposito 1993). The progenitor strains of X2180-1A, which

TABLE 2

Distribution of SFPs among different genomic classes in lab strains and natural isolates

Frequency class	SFP			
	No. loci	ORFs	Intergenic	Transposable elements
			Lab strains	
(0,5)	120,697	96,337	7,558	3,277
(1,4)	3,769	2,980	236	139
(2,3)	1,717	1,301	61	55
			Natural isolates	
(0,4)	119,415	95,279	7,521	3,196
(1,3)	3,976	3,131	222	203
(2,2)	1,172	930	61	27

Polymorphism frequency class denotes number of strains that are polymorphic within each group. The number of loci in each polymorphism frequency class is indicated under "No. loci." The subset of these loci falling in open reading frames, intergenic areas, and in transposable elements is indicated under "ORFs," "Intergenic," and "Transposable element" headings, respectively. Loci counts are out of a possible 126,645 loci examined.

account for 88% of the X2180-1A genome, are EM93 and 99R (Mortimer and Johnston 1986). The parents of Σ1278b and 3962c were the baking strains, "yeast foam" and 1422-11D (Grenson *et al.* 1966; Bechet *et al.* 1970). "Yeast foam" also contributed part of its genome to S288c, and 1422-11D is thought to be derived from the same progenitor strains as S288c (Mortimer and Johnston 1986). In all cases, the exact lineages are not publicly available or are not known. Examination of the genomic DNA hybridization patterns revealed that the distribution of polymorphic probes is indeed nonrandom and tends to form clusters. For A364A, 99% of the variation was found in 25% of the genome (Figure 2); for W303, 99% in 14% of the genome; for EM93, 92% in 14% of the genome; and for Σ1278b, 99.9% in 53% of the genome. The 25% for A364A includes most of chromosome III and portions of chromosomes V and XII. The probability of these distributions occurring by chance is essentially zero (Figure 2; see materials and methods). In contrast, when wild isolates such as M2-8s2, M2-8f2, or M1-2 were examined (as opposed to clonal laboratory strains), variation was found distributed throughout the genome.

Relatedness of different strains: The large amount of data allows us to investigate the relatedness of different strains. We used the combined set of probes (11,115), to construct a Fitch-Margoliash distance tree from the data set (Figure 3). As expected, S288c was most closely related to X2180-1A, followed by 99R, EM93, W303, and A364A. In addition, Σ1278b and 3962c (an isogenic pair) are placed as sisters to one another. The laboratory strains SK1 and Y102 are as distantly related to S288c as are most natural isolates and are more distantly related than other natural isolates (such as YKM789, which was derived from a pathogenic strain in San Francisco). Finally, M2-8s2 and M2-8f2, which are the diploid progeny of a wild Tuscan homothallic isolate, were located together in the same branch of the tree. Another strain (M1-2) that was located from the same geographical region was also found on this branch. The branch leading to M1-2 is very long. M1-2 is a homothallic diploid strain isolated from a vineyard in Montalcino, and the long branch may reflect the heterozygosity that has accumulated in this strain since the last "genome renewal" (sporulation; Mortimer *et al.* 1994). Likewise, the branches to M28s2 and M28f2 are shorter, probably because they are homozygous diploids that originated as sister spores from the same ascus (Cavalieri *et al.* 2000).

Variation and recombination: The large number of SFPs scattered throughout the genome allowed us to ask questions about genome-wide distributions of genetic variation. Data from Drosophila, and to a lesser extent from humans, have indicated a significant relationship between level of polymorphism and level of recombination, with higher local recombination rates associated with higher levels of polymorphism (Begun and Aquadro 1992; Nachman *et al.* 1998). However, these yeast data show no consistent pattern for the individual strains, for the whole set, or for individual chromosomes. In particular, for each individual chromosome, the correlation between local rate of recombination for regions across the chromosome (logarithm of centimorgan per kilobase) and frequency of polymorphic probes within each region is significantly negative for chromosome V and significantly positive for chromosome VIII; the remaining chromosomes have nonsignificant correlations, among which half are negative (chromosomes VII, IX, X, XII, XIII, XIV, and XV) and half are positive (I, II, III, IV, VI, XI, and XVI). One reason for this lack of correlation could be that the recombination per kilobase in yeast overwhelms the effects of selective sweeps or background selection on levels of regional polymorphism. Alternatively, the sexual cycle in yeast may be so rare relative to vegetative propagation (especially in homothallic strains) that most of the SFPs result from mutation in particular lineages, so that no correlation with recombination rate would be expected.

Genome-wide distribution of variation: Although we found little relationship between recombination and the distribution of SFPs, we did find that the variation was unevenly distributed within chromosomes. Subtelomeric regions are known to exhibit variability at the sequence level in many organisms (Valgeirsdottir *et al.* 1990; Broun *et al.* 1992; *C. elegans* Sequencing Consortium 1998). Indeed, we found a much higher proportion of polymorphic probes in the subtelomeric regions. Regions within 25 kb of the chromosome ends

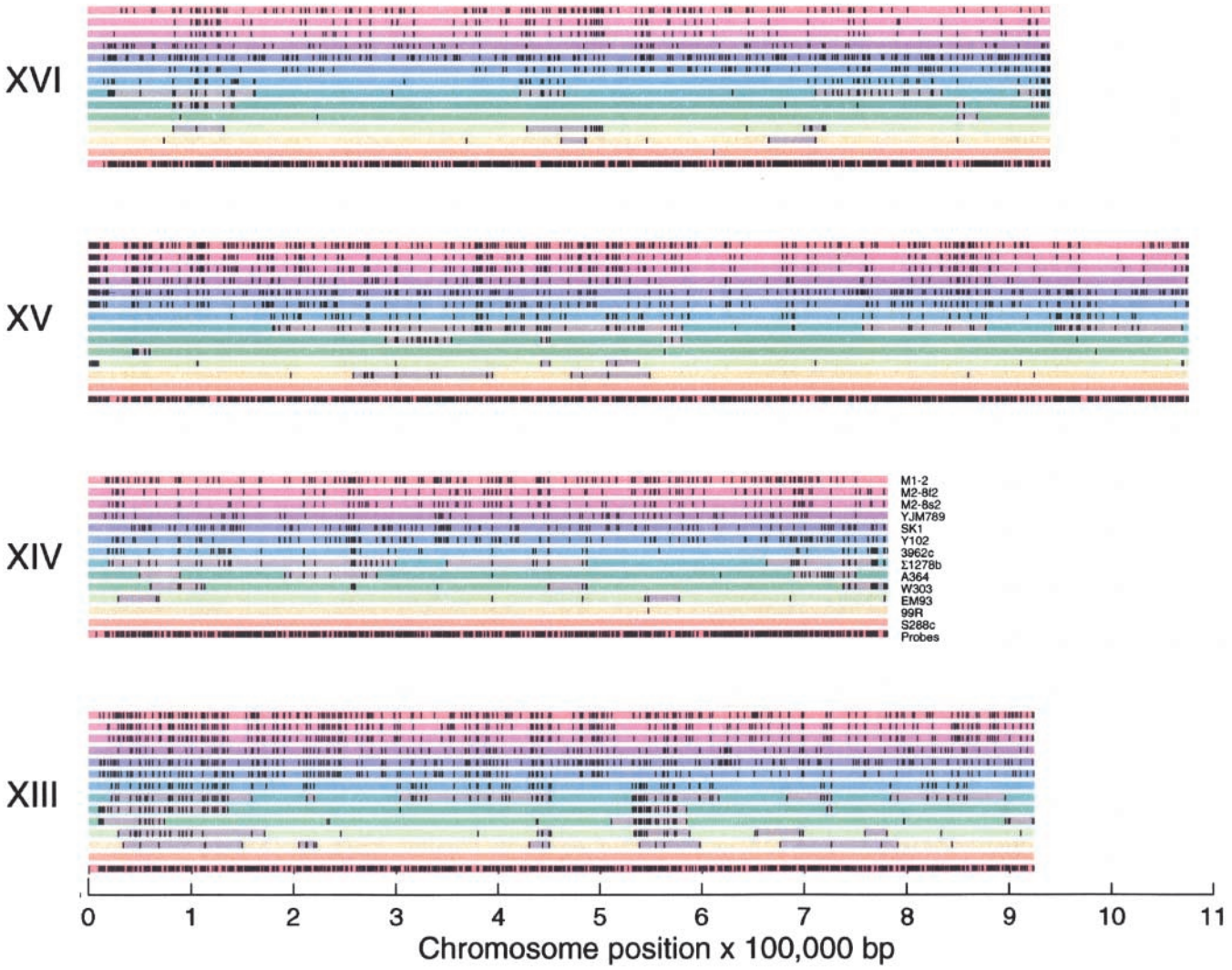

FIGURE 2.—Distribution of variation for 13 strains for four chromosomes. The position of every unique probe is shown at the bottom of each chromosome group and the locations of SFPs for 13 strains is shown above. In addition, genomic regions that are likely to be inherited from an S288C-like strain (determined as described in MATERIALS AND METHODS) are highlighted in gray only for strains related to S288c (EM93, W303, A364, 99R, and Σ1278b). The line labeled S288c shows the location of polymorphisms determined between S288c and X2180. Data for the rest of the genome can be obtained from supplementary Figure S2, A–C, at http://www.scripps.edu/cb/winzeler/genetics_supplement/supplement.htm. Although sequencing showed a very low frequency of false positives, false negatives were occasionally found. This is a result of the stringency associated with the prediction routine.

had fewer probes than average, because of low gene density and the presence of repetitive elements (Figure 4). However, on average, 1 in 3 probes was classified as an SFP in at least one strain in the subtelomeric regions *vs.* <1 in 10 probes in the central regions of the chromosome.

A model that could explain this type of distribution of SFPs is that the copy number of genes conferring an adaptive advantage could be increased or decreased through recombination between nonsister chromatids. Although recombination between nonhomologous chromosomes could occur anywhere, it is less likely to result in the loss of an essential gene if the event takes place in the subtelomeric region, and the recombinants are more likely to be transmitted to future generations. Furthermore, subtelomeric regions are rich in redundant sequences such as the X element or the Y′, which could serve as initiation points for nonhomologous recombination (LOUIS *et al.* 1994).

To provide more support for this model, we asked if variability in hybridization of subtelomeric probes was likely to be the result of an underlying deletion. Whole-gene deletions can be detected if all probes to a gene exhibit a loss of hybridization. Our sequencing revealed that SFPs in which a single probe per gene shows strain-to-strain variability, whereas most other probes in the gene are unaffected, are often the result of a single-nucleotide polymorphism (data not shown). We ex-

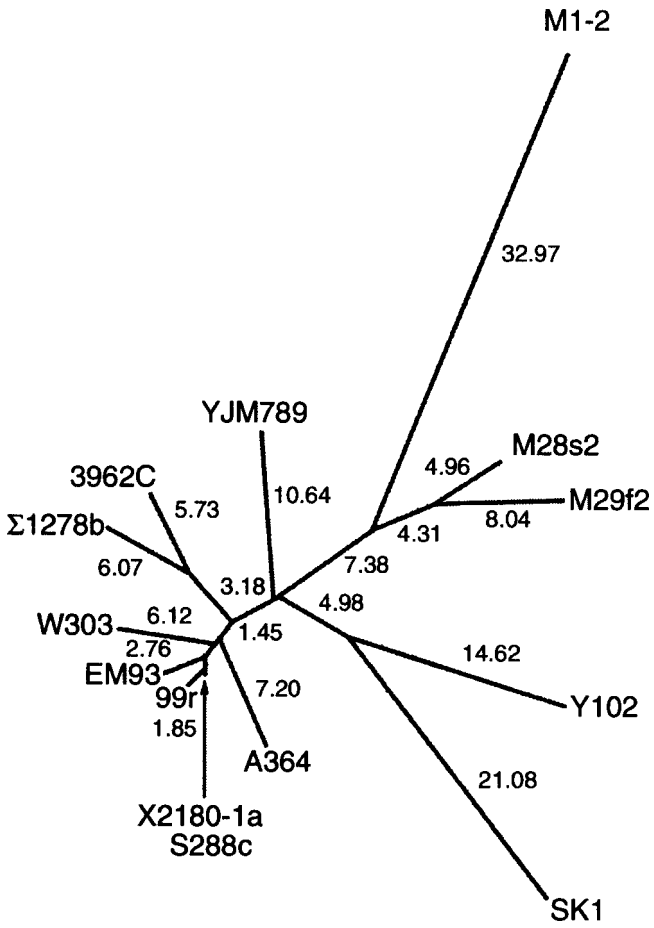

FIGURE 3.—Geneological relationship of different strains. A total of 11,115 informative probes that detected variation in at least 1 of the 14 strains were selected. The distances between each strain were determined using the Fitch Margoliash method (FELSENSTEIN 1993) with the following assumptions: If two strains showed hybridization at the same oligonucleotide, they were assumed to be identical. If one strain showed hybridization and the other did not, they were assumed to be different. Finally, if both strains failed to hybridize, the oligo was considered to be ambiguous and treated as missing data. The tree topology was stable upon recoding of ambiguous oligos as matches or mismatches.

pected that variation in the subtelomeric regions would take the form of deletions whereas variation in the central regions would consist of single-nucleotide changes. To identify deletions the Affymetrix GeneChip program was used on the DNA hybridization data. This program calls genes "absent" if the signals for the perfect match probes to a gene are no different from those for the mismatch probes (predicted to be at background). Ninety genes were considered "absent" in at least one strain by the GeneChip program. Partial-gene deletions were not identified and were excluded from this analysis but are expected to contribute to the subtelomeric bias. Considering all 14 strains, 195 out of 3710 genes within 25 kb of either chromosome end (265 × 14) were deleted, giving a deletion rate of 5.2%. In contrast, only 101 out of 82,488 genes in the central chromosome regions (5892 × 14) were deleted, giving a deletion rate of 0.12% (101/82,488). If gene deletion is location independent and deletion events can be described by a binomial distribution [with a mean of 0.12% and a standard deviation of 0.01% (std(p) = Sqrt(0.0012 × (1 − 0.0012)/82,488) = 0.01%] (GLANTZ 1997), the chance of a disparity in deletion rates this large or larger due to chance alone is essentially zero. By subtracting variation due to the full-gene deletions (Figure 4) we found that the bias toward subtelomeric variation was significantly reduced, providing further support that the nonhomologous exchange model might be correct.

Two other observations support a model in which the telomeres act as reservoirs for genetic material subject to rapid change. First, the yeast sequencing project confirmed that many subtelomeric genes are duplicated within the genome (YEAST GENOME DIRECTORY 1997). Second, subtelomeric regions are enriched for genes with known functions in transport, facilitation, fermentation, and C-compound metabolism [$P <$ 1.8 E-7, 1.16 E-5, and 1.5 E-5, respectively (MEWES *et al.* 1999; TAVAZOIE *et al.* 1999)], including genes encoding maltases, alcohol dehydrogenases, and sodium-phosphate antiporters. Of the 17 hexose transporter genes in yeast, 9 are also located near within 25 kb of the chromosome end and 12 of the 17 are within 50 kb. As a free-living organism, it is likely that yeast must often adjust to life on new food sources—a grape, a fig, or in fermenting grain. Thus, the ability to shuffle a complement of genes involved in carbon uptake and metabolism during mitosis or meiosis may be important as mechanisms for generating potentially adaptive variation.

Interestingly, many nontelomeric regions that demonstrated substantial variability were associated with transposable elements. One of the regions with the greatest amount of variability was the region around Ty4 on chromosome VIII. However, not all Tys could be probed because of the repetitive nature of their DNA, and probes mapping to more than one region in the genome were discarded. Around Ty4 there were 46 variable probes in 30 kb of sequence. As with telomeric variation, variation in transposable elements tended to be associated with deleted genes. Of the 34 nontelomeric deletions detected in this study, 20 were located within 5 kb of either a known transposable element or a transposable element long terminal repeat. This observation is statistically significant (P = 5 E-11), as only 2 Mb of the 12-Mb genome is located within 5 kb of one of these elements.

There were also genomic regions with lower-than-average variability (Figure 2). For example, in the region between base pairs 52,000 and 70,000 on chromosome VI, only two variant probes among all 14 strains were identified. This chromosome VI region contains several essential genes, including *YPT1*, *TUB1*, and *ACT1*, which encode the RAB small monomeric GTPase, tubulin, and

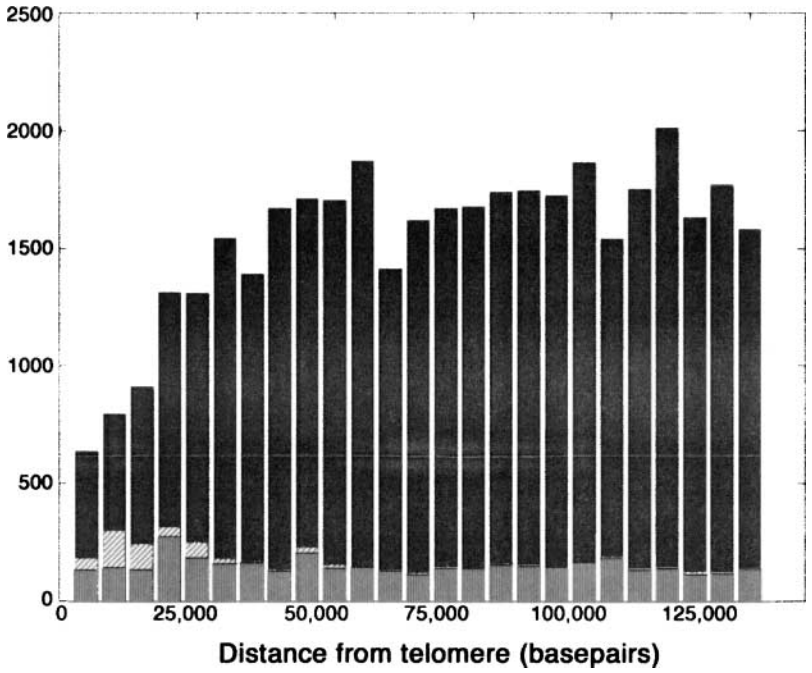

FIGURE 4.—Distribution of variation in the genome. Each 5-kb bin represents the average of the total number of probes (shaded), the total number of SFPs (cross-hatched) for all strains, or the SFPs after subtraction of SFPs within deleted genes (solid) within the 5-kb region starting at the telomeres and moving toward the centromere for both ends of all chromosomes. Only data from the 125 kb at each chromosome end are considered.

actin, respectively. Both *ACT1* and *TUB1* are essential in yeast and encode two of the most conserved proteins in eukaryotes. We also examined variation with respect to functional classification of open reading frames (ORFs) using categories published by the Munich Information Center for Protein Sequences (Table 3). Genes involved in transport were highly variable although the sample size was small. Those genes having roles in functional classes such as fermentation, extracellular secretion, and the cell wall were on average more variable than those having roles in translation, ribosome biogenesis, gluconeogenesis, and the pentose phosphate pathway categories.

There was also more variation in nonessential genes: of the 126,645 unique probes, 17,733 probes were mapped to genes identified as essential by tetrad dissection (WINZELER *et al.* 1999; GIAEVER *et al.* 2002). A total of 76,472 probes mapped to nonessential genes or to genes whose status could not be determined, and 28,118 probes mapped to the intergenic regions or to nonannotated open reading frames, untranslated RNA, transposons, or other features. Of the 11,115 variable probes, 1119 were found in genes known to be essential, 7094 in undetermined or nonessential genes, and 2799 in intergenic regions. Thus, we found that 14% of the probes on the array map to known essential genes, but only 10.8% of the probes detecting allelic variation were found in the same class of genes. Regions with lower-than-average variability did not appear to be associated with centromeres.

DISCUSSION

The characterization of genome-wide genetic variability has several important consequences. First, it allows relationships between different strains to be determined with a level of detail previously impossible. This has great practical importance. Many scientists are wedded to a genetic background with which they may have 30 years of experience. It will be very valuable for these researchers to know whether particular genomic regions are likely to be nearly identical to those in a reference strain. Knowing where differences exist will be particularly important for studies of gene expression in divergent strains, because observed expression changes could result from underlying coding sequence variability rather than from differential gene regulation. As a result of probe richness in coding regions, the majority of the polymorphisms we detected are likely to have some impact on the apparent transcript levels measured by oligonucleotide or cDNA arrays. In addition, knowing which functional classes of genes or genome regions are likely to be variable may also allow better interpretation of gene expression data collected for divergent strains.

Second, the genome-wide characterization of genetic diversity permits identification of many genetic markers (11,115 different ones in this case) that can be used in the analysis of quantitative traits, genetic mapping, linkage analysis, and population studies (STEINMETZ *et al.* 2002). In some cases the polymorphisms may result in changes in protein structure or function. Of course this method cannot be used to identify all variability, in part because of incomplete array coverage and probe performance. Newer generations of arrays will contain on the order of 500,000 unique oligonucleotides that, in theory, could probe almost 100% of the nonredundant regions in the yeast genome. Using oligonucleotide arrays, thousands of new markers can be identified in

TABLE 3

Functional classification of SFPs

Functional category	SFPs	Probes	Ratio
Cell wall (2 ORFs)	3	3	1
Other transport facilitators (1 ORF)	13	16	0.81
Ion transporters (2 ORFs)	13	32	0.41
Electron transport and membrane-associated energy conservation (2 ORFs)	5	32	0.16
Endoplasmic reticulum (4 ORFs)	10	63	0.16
Cytoplasm (7 ORFs)	20	132	0.15
Fermentation (33 ORFs)	64	464	0.14
Degradation of foreign (exogenous) compounds (9 ORFs)	18	144	0.13
Extracellular/secretion proteins (19 ORFs)	36	295	0.12
Nucleus (11 ORFs)	22	191	0.12
Secondary metabolism (5 ORFs)	9	80	0.11
Other cell division and DNA synthesis activities (16 ORFs)	27	256	0.11
Protein folding and stabilization (59 ORFs)	105	947	0.11
Other protein fate-related activities (8 ORFs)	14	128	0.11
Detoxification (102 ORFs)	169	1547	0.11
Cell wall (39 ORFs)	58	505	0.11
Mitochondrion (6 ORFs)	11	96	0.11
Metabolism of energy reserves (glycogen, trehalose; 37 ORFs)	53	539	0.10
Cell differentiation (6 ORFs)	10	96	0.10
Metabolism of vitamins, cofactors, and prosthetic groups (84 ORFs)	118	1325	0.09
Oxidation of fatty acids (7 ORFs)	10	112	0.09
Other protein-synthesis activities (16 ORFs)	20	219	0.09
Nonvesicular cellular import (100 ORFs)	136	1472	0.09
Intracellular signalling (134 ORFs)	196	2196	0.09
Stress response (175 ORFs)	223	2587	0.09
Ionic homeostasis (117 ORFs)	164	1807	0.09
Cytoskeleton (2 ORFs)	3	32	0.09
Nitrogen and sulfur metabolism (74 ORFs)	89	1135	0.08
Nucleotide metabolism (144 ORFs)	173	2233	0.08
Phosphate metabolism (33 ORFs)	38	465	0.08
C-compound and carbohydrate metabolism (415 ORFs)	525	6239	0.08
Lipid, fatty-acid and isoprenoid metabolism (214 ORFs)	281	3474	0.08
Tricarboxylic-acid pathway (citrate cycle, Krebs cycle, TCA cycle; 25 ORFs)	32	400	0.08
Respiration (88 ORFs)	111	1328	0.08
DNA processing (76 ORFs)	99	1230	0.08
Cell cycle (11 ORFs)	14	176	0.08
mRNA transcription (579 ORFs)	702	9017	0.08
RNA transport (27 ORFs)	34	448	0.08
Other transcription activities (57 ORFs)	74	908	0.08
Protein targeting, sorting, and translocation (144 ORFs)	177	2288	0.08
Assembly of protein complexes (95 ORFs)	116	1540	0.08
Nuclear transport (59 ORFs)	76	964	0.08
Mitochondrial transport (80 ORFs)	99	1244	0.08
Other cell rescue activities (9 ORFs)	12	157	0.08
Plasma membrane (144 ORFs)	174	2163	0.08
Cytoskeleton (106 ORFs)	136	1694	0.08
Intracellular transport vesicles (41 ORFs)	50	652	0.08
Nucleus (755 ORFs)	930	12080	0.08
Endosome (12 ORFs)	15	192	0.08
Amino acid metabolism (204 ORFs)	226	3210	0.07
Glyoxylate cycle (6 ORFs)	7	96	0.07
Translational control (31 ORFs)	34	496	0.07
Aminoacyl-tRNA-synthetases (37 ORFs)	41	591	0.07
Protein modification (187 ORFs)	220	3016	0.07
Proteolytic degradation (160 ORFs)	180	2572	0.07
Vacuolar transport (56 ORFs)	57	864	0.07

(*continued*)

TABLE 3

(Continued)

Functional category	SFPs	Probes	Ratio
Extracellular transport, exocytosis, and secretion (39 ORFs)	40	605	0.07
Cytoskeleton-dependent transport (27 ORFs)	30	432	0.07
Other intracellular transport activities (32 ORFs)	36	512	0.07
Endoplasmic reticulum (153 ORFs)	168	2457	0.07
Golgi (80 ORFs)	94	1272	0.07
Mitochondrion (360 ORFs)	385	5580	0.07
Vacuole or lysosome (59 ORFs)	69	928	0.07
rRNA transcription (109 ORFs)	113	1773	0.06
tRNA transcription (83 ORFs)	84	1358	0.06
Vesicular transport (Golgi network, etc.; 129 ORFs)	126	2106	0.06
Peroxisomal transport (16 ORFs)	14	254	0.06
Cytoplasm (547 ORFs)	453	8172	0.06
Peroxisome (37 ORFs)	37	590	0.06
Prokaryotic cell membrane (inner membrane of gram−; 1 ORF)	1	16	0.06
Other control of cellular organization (8 ORFs)	8	128	0.06
Plasma membrane (1 ORF)	1	16	0.06
Centrosome (1 ORF)	1	16	0.06
Target of regulation (1 ORF)	1	16	0.06
Pentose-phosphate pathway (9 ORFs)	7	144	0.05
Translation (64 ORFs)	44	973	0.05
Centrosome (30 ORFs)	26	480	0.05
Glycolysis and gluconeogenesis (35 ORFs)	21	478	0.04
Other energy generation activities (16 ORFs)	10	256	0.04
Ribosome biogenesis (215 ORFs)	134	3047	0.04
Cell growth/morphogenesis (3 ORFs)	2	48	0.04
Golgi (2 ORFs)	1	32	0.03
Intracellular transport vesicles (1 ORF)		32	0
Peroxisome (1 ORF)		16	0
Transport mechanism (1 ORF)		16	0

Functional classifications were downloaded from http://mips.gsf.de/proj/yeast/CYGD/db/index.html (Mewes *et al.* 1999) on 11/31/01. Some genes may be found in multiple functional categories.

a single day in a cost-effective manner. Because of this ease, wild natural isolates, which usually contain few useful conventional auxotrophic markers, can now be used in genetic experiments.

A further application of this approach is accurate strain identification in cases of microorganisms associated with foodborne illness or bioterrorism. Here we showed that when two strains that were virtually identical to one another were compared across >100,000 probes, few genotypic differences were discovered. Such genome-wide analysis of allelic variation provides much more confidence in assigning relatedness than does the sequencing of relatively small regions of the genome.

Additionally, analysis of genetic variability is relevant to the process of drug discovery and vaccine development. The malaria and African sleeping sickness parasites both use antigenic variation to evade the host's immune responses (Roberts *et al.* 1992), and nonhomologous recombination between chromosome ends has been proposed as a method to increase genetic variability in Plasmodium (Freitas-Junior *et al.* 2000). The method described here could be used to identify genomic regions that are under evolutionary pressure in different parasite isolates and thus the most likely to interact with the host immune system; these regions could then serve as candidate antigenic targets for vaccine development (Conway *et al.* 1999).

In short, the characterization of genome-wide diversity permits novel observations that would be impossible if only a small region of the genome were examined. Through this study we have shown that a higher proportion of genetic variability is located at chromosome ends. The notion that the ends of chromosomes serve as the places where new genes evolve has been proposed on the basis of the distribution of nonconserved genes, transposable elements, and tandem repeats in the genome of the nematode *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium 1998). In the absence of extensive sequencing, such hypotheses are difficult to test. However, preliminary data, based on work with *Arabidopsis thaliana*, an organism with a genome similar in size to that of *C. elegans*, suggest that the method described here could be used to examine variability in any number of organisms. To this end, we expect that

the use of high-density oligonucleotide arrays to study genome evolution and population genetics in any fully sequenced organism will become commonplace.

We thank Lars Steinmetz, John McCusker, Jeff Townsend, and Duccio Cavalieri for providing strains and helpful discussions; Peter Dmitrov and Ruben Abagyan for computer support; David Lockhart and Steve Kay for support of this project; and Joseph Heitman, Rodney Rothstein, and Rochelle Esposito for helpful advice about yeast strains.

LITERATURE CITED

Bechet, J., M. Greenson and J. M. Wiame, 1970 Mutations affecting the repressibility of arginine biosynthetic enzymes in Saccharomyces cerevisiae. Eur. J. Biochem. **12:** 31–39.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature **356:** 519–520.

Borevitz, J. O., D. Liang, D. Plouffe, H. Chang, T. Zhu *et al.*, 2003 Large scale identification of single feature polymorphisms in complex genomes. Genome Res. (in press).

Broun, P., M. W. Ganal and S. D. Tanksley, 1992 Telomeric arrays display high levels of heritable polymorphism among closely related plant varieties. Proc. Natl. Acad. Sci. USA **89:** 1354–1357.

C. elegans Sequencing Consortium, 1998 Genome sequence of the nematode C. elegans: a platform for investigating biology. Science **282:** 2012–2018.

Cavalieri, D., J. P. Townsend and D. L. Hartl, 2000 Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc. Natl. Acad. Sci. USA **97:** 12369–12374.

Chee, M., R. Yang, E. Hubbell, A. Berno, X. C. Huang *et al.*, 1996 Accessing genetic information with high-density DNA arrays. Science **274:** 610–614.

Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein *et al.*, 1998 The transcriptional program of sporulation in budding yeast. Science **282:** 699–705.

Conway, D. J., C. Roper, A. M. Oduola, D. E. Arnot, P. G. Kremsner *et al.*, 1999 High recombination rate in natural populations of Plasmodium falciparum. Proc. Natl. Acad. Sci. USA **96:** 4506–4511.

Esposito, R. E., 1993 Humble beginnings, pp. 417–433 in *The Early Days of Yeast Genetics*, edited by P. L. M. Hall. Cold Spring Harbor Laboratory Press, Plainview, NY.

Felsenstein, J. P., 1993 *Phylogeny Inference Package*. Department of Genetics, University of Washington, Seattle.

Freitas-Junior, L. H., E. Bottius, L. A. Pirrit, K. W. Deitsch, C. Scheidig *et al.*, 2000 Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. Nature **407:** 1018–1022.

Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles *et al.*, 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature **418:** 387–391.

Gingeras, T. R., G. Ghandour, E. Wang, A. Berno, P. M. Small *et al.*, 1998 Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic Mycobacterium DNA arrays. Genome Res. **8:** 435–448.

Glantz, S. A., 1997 *Primer of Biostatistics*. McGraw-Hill, New York.

Grenson, M., M. Mousset, J. M. Wiame and J. Bechet, 1966 Multiplicity of the amino acid permeases in Saccharomyces cerevisiae. I. Evidence for a specific arginine-transporting system. Biochim. Biophys. Acta **127:** 325–338.

Grunenfelder, B., and E. Winzeler, 2002 Treasures and limitations contained in genome-wide data sets. Nat. Rev. Genet. **3** (9): 653–661.

Hartwell, L. H., 1967 Macromolecule synthesis in temperature-sensitive mutants of yeast. J. Bacteriol. **93:** 1662–1670.

Kane, S. M., and R. Roth, 1974 Carbohydrate metabolism during ascospore development in yeast. J. Bacteriol. **118:** 8–14.

Li, C., and W. H. Wong, 2001 Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA **98:** 31–36.

Liu, H., C. A. Styles and G. R. Fink, 1993 Elements of the yeast pheromone response pathway required for filamentous growth of diploids. Science **262:** 1741–1744.

Lockhart, D. J., and E. A. Winzeler, 2000 Genomics, gene expression and DNA arrays. Nature **405:** 827–836.

Louis, E. J., E. S. Naumova, A. Lee, G. Naumov and J. E. Haber, 1994 The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. Genetics **136:** 789–802.

McCusker, J. H., and J. E. Haber, 1988 Cycloheximide-resistant temperature-sensitive lethal mutations of *Saccharomyces cerevisiae*. Genetics **119:** 303–315.

McCusker, J. H., K. V. Clemons, D. A. Stevens and R. W. Davis, 1994 Saccharomyces cerevisiae virulence phenotype as determined with CD-1 mice is associated with the ability to grow at 42 degrees C and form pseudohyphae. Infect. Immun. **62:** 5447–5455.

Mewes, H. W., K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer *et al.*, 1999 MIPS: a database for genomes and protein sequences. Nucleic Acids Res. **27:** 44–48.

Mortimer, R. K., and J. R. Johnston, 1986 Genealogy of principal strains of the yeast genetic stock center. Genetics **113:** 35–43.

Mortimer, R. K., P. Romano, G. Suzzi and M. Polsinelli, 1994 Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of Saccharomyces cerevisiae derived from natural fermentation of grape musts. Yeast **10:** 1543–1552.

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics **150:** 1133–1141.

Oshiro, G., L. M. Wodicka, M. P. Washburn, J. R. Yates, III, D. J. Lockhart *et al.*, 2002 Highly parallel identification of new genes in Saccharomyces cerevisiae. Genome Res. **12:** 1210–1220.

Primig, M., R. M. Williams, E. A. Winzeler, G. G. Tevzadze, A. R. Conway *et al.*, 2000 The core meiotic transcriptome in budding yeasts. Nat. Genet. **26:** 415–423.

Raghuraman, M. K., E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka *et al.*, 2001 Replication dynamics of the yeast genome. Science **294:** 115–121.

Roberts, D. J., A. G. Craig, A. R. Berendt, R. Pinches, G. Nash *et al.*, 1992 Rapid switching to multiple antigenic and adhesive phenotypes in malaria. Nature **357:** 689–692.

Steinmetz, L. M., H. Sinha, D. R. Richards, J. I. Spiegelman, P. J. Oefner *et al.*, 2002 Dissecting the architecture of a quantitative trait locus in yeast. Nature **416:** 326–330.

Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, 1999 Systematic determination of genetic network architecture. Nat. Genet. **22:** 281–285.

Thomas, B. J., and R. Rothstein, 1989 Elevated recombination rates in transcriptionally active DNA. Cell **56:** 619–630.

Troesch, A., H. Nguyen, C. G. Miyada, S. Desvarenne, T. R. Gingeras *et al.*, 1999 Mycobacterium species identification and rifampin resistance testing with high-density DNA probe arrays. J. Clin. Microbiol. **37:** 49–55.

Valgeirsdottir, K., K. L. Traverse and M. L. Pardue, 1990 HeT DNA: a family of mosaic repeated sequences specific for heterochromatin in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **87:** 7998–8002.

Winzeler, E. A., D. R. Richards, A. R. Conway, A. L. Goldstein, S. Kalman *et al.*, 1998 Direct allelic variation scanning of the yeast genome. Science **281:** 1194–1197.

Winzeler, E., D. Shoemaker, A. Astromoff, H. Liang, K. Anderson *et al.*, 1999 Functional characterization of the *Saccharomyces cerevisiae* genome by precise deletion and parallel analysis. Science **285:** 901–906.

Wodicka, L., H. Dong, M. Mittmann, M.-H. Ho and D. J. Lockhart, 1997 Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat. Biotechnol. **15:** 1359–1367.

Wyrick, J. J., J. G. Aparicio, T. Chen, J. D. Barnett, E. G. Jennings *et al.*, 2001 Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. Science **294:** 2357–2360.

Yeast Genome Directory, 1997 The yeast genome directory. Nature **387:** 5.

Communicating editor: M. Johnston