Genome-Wide Detection of Polymorphisms at Nucleotide Resolution with a Single DNA Microarray

David Gresham, *et al.*

Science **311**, 1932 (2006);
DOI: 10.1126/science.1123726

Foundation, Sumitomo Foundation, Tokyo Biochemical Research Foundation, Uehara Memorial Foundation (N.S.), Boninchi Foundation (T.C.), Terry Fox Foundation from National Cancer Institute of Canada (NCIC) (T.C.), the Canadian Institutes of Health Research (W.-C.Y. and P.S.O.), and a Canadian Network for Vaccines and Immunotherapeutics (CANVAC) Network Centres of Excellence grant (P.S.O.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/311/5769/1927/DC1
Materials and Methods
SOM Text
Figs. S1 to S13
References

23 December 2005; accepted 23 February 2006
10.1126/science.1124256

Genome-Wide Detection of Polymorphisms at Nucleotide Resolution with a Single DNA Microarray

David Gresham,[1,2]* Douglas M. Ruderfer,[1,3] Stephen C. Pratt,[1,3] Joseph Schacherer,[1,3] Maitreya J. Dunham,[1] David Botstein,[1,2] Leonid Kruglyak[1,3]*

A central challenge of genomics is to detect, simply and inexpensively, all differences in sequence among the genomes of individual members of a species. We devised a system to detect all single-nucleotide differences between genomes with the use of data from a single hybridization to a whole-genome DNA microarray. This allowed us to detect a variety of spontaneous single–base pair substitutions, insertions, and deletions, and most (>90%) of the ~30,000 known single-nucleotide polymorphisms between two *Saccharomyces cerevisiae* strains. We applied this approach to elucidate the genetic basis of phenotypic variants and to identify the small number of single–base pair changes accumulated during experimental evolution of yeast.

Despite the ongoing development of DNA sequencing technology (*1*, *2*), it remains technically and financially infeasible for individual laboratories to sequence whole genomes. Moreover, for global comparisons of genomes within species, where one expects a relatively small number of sequence differences throughout the genome, determining the entire sequence is unnecessary. In such cases, it is sufficient to assess the extent and location of sequence variation in a manner analogous to comparative genomic hybridization, which compares copy number changes between closely related genomes at genic resolution (*3*).

DNA microarrays of short oligonucleotides designed to interrogate each base individually (i.e., resequencing arrays) have been applied to the analysis of individual human genes (*4*) and small genomes such as the human mitochondrial (*5*) and the SARS coronavirus (*6*) genomes. However, extension of this approach to whole genomes of most organisms is currently impractical because of the large number of probes required for complete coverage.

An alternative approach uses microarrays that detect mismatches, exploiting the fact that hybridization to a short oligonucleotide is quantitatively sensitive to the number and position of mismatches (*7*). Sequence-level differences are detected, without allele-specific probes, by comparing hybridization intensities of individual features on the microarray [referred to as single-feature polymorphisms (SFPs) (*8*)]. This method has been successfully applied to studies of genetic diversity (*9*–*11*) and gene mapping (*12*–*17*). Until recently, comprehensive detection of single–base pair differences has been limited by probe density across the genome, which is typically a few oligonucleotides per gene. Even complete single-copy coverage of the genome is unlikely to be sufficient for finding all mutations, because statistically detectable decreases in hybridization intensity usually require that a variant nucleotide fall within the central 15 bases of a 25-base probe (*18*).

We used high-density Affymetrix yeast tiling microarrays (YTMs) with overlapping 25-nucleotide oligomers spaced an average of 5 base pairs (bp) apart to provide complete and ~5-fold redundant coverage of the entire *S. cerevisiae* genome. This array design was previously used to discover novel expressed sequences and to precisely map sites of transcription in humans (*19*). This design provides five to seven measurements of a given nucleotide's effect on hybridization efficiency, which we exploited to predict the presence and location of SNPs and deletion breakpoints throughout the entire yeast genome.

Each YTM has ~2.6 million perfect match (PM) probes and ~2.6 million corresponding mismatch (MM) probes. We modeled the decrease in PM probe intensity caused by a single SNP as a function of the SNP's position within the probe, the probe's GC content, the nucleotide sequence surrounding the SNP, and the hybridization intensity obtained using a nonpolymorphic reference (S288C) genome [strain FY3 (*20*)]. To fit the model, we used hybridization data for a training set of nearly 25,000 high-quality SNPs in strain RM11-1a, all identified by direct comparison of the genomic sequences (*20*). The model predicts the intensity of a probe in the presence of a specified SNP (*20*) (figs. S1 and S2) and is used in our algorithm, SNPscanner, which calculates the log of the likelihood ratio (the "prediction signal") for the presence of a SNP at each nucleotide position in the genome using measurements from all probes that cover that site. By scanning the entire genome, we identify SNPs as regions of elevated signal in which the position of the peak value is considered the predicted polymorphic site.

We tested the performance of SNPscanner on a set of 981 high-quality SNPs from RM11-1a that were not included in the training set. We assessed the false-positive rate by using SNPscanner to predict SNPs from an independent hybridization of the reference strain, where no true polymorphisms are expected. At a prediction signal of 1, we detected 915 (93.3%) known SNPs in RM11-1a and called 177 false positives in the reference strain (fig. S3). By increasing the prediction signal to 5 and applying a heuristic filter (*20*), we elim-

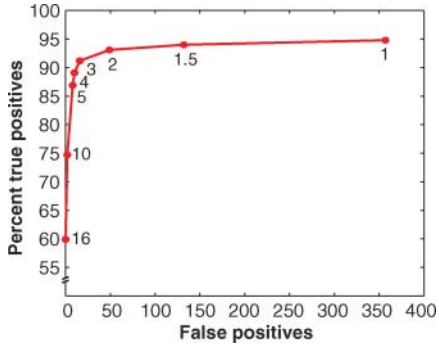

Fig. 1. Nucleotide-level comparison with a genome divergent from the sequenced reference genome. We applied our approach to test how many of 30,303 known SNPs in the yeast strain YJM789 we were able to detect. Numbers on the graph indicate prediction signal thresholds. On the basis of data from a single hybridization experiment, we were able to correctly identify as many as 28,737 SNPs at a prediction signal of 1. At prediction signals of >5, the number of false-positive predictions is reduced to 8 in a test of the reference genome and 86.9% of true positives are still predicted.

[1]Lewis-Sigler Institute for Integrative Genomics, [2]Department of Molecular Biology, [3]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA.

*To whom correspondence should be addressed. E-mail: leonid@genomics.princeton.edu (L.K.); dgresham@genomics.princeton.edu (D.G.)

inated all false positives and retained 77.5% (760) of real SNPs. Analysis of this set of correctly predicted SNPs showed the sequence-confirmed SNP to be within 2 bp of the predicted site 87.1% of the time (*20*).

To test our ability to predict a large number of SNPs, we analyzed the highly diverged sequenced strain YJM789, originally recovered from an AIDS patient (*21*). We selected a set of 30,303 sequence-confirmed SNPs in YJM789 that were isolated from each other by at least 25 bp and were covered by probes on the YTM. Analysis of a single hybridization with SNPscanner yielded 28,737 (94.8%) correctly predicted SNPs at a prediction sig-

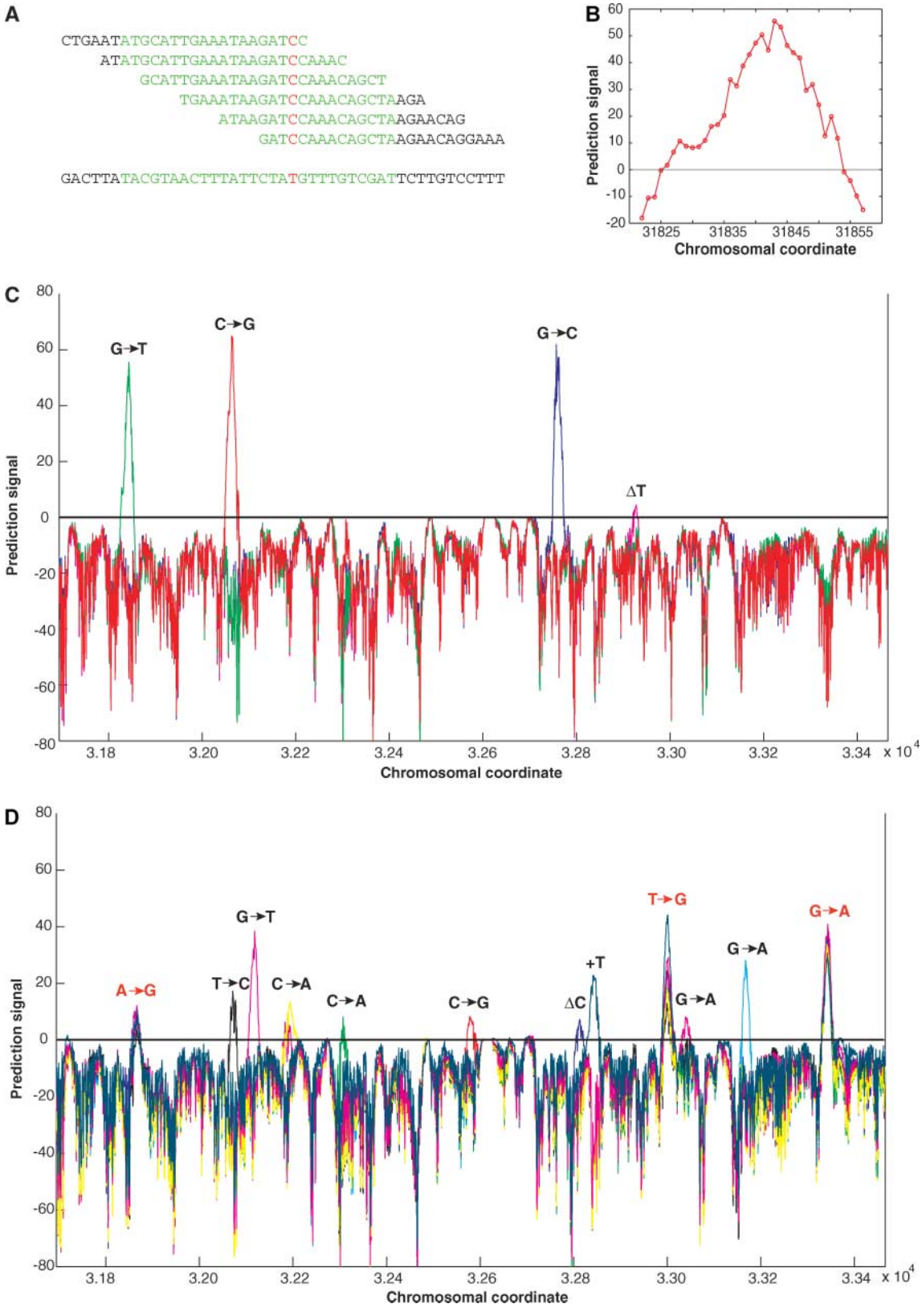

Fig. 2. SNPscanner accurately predicts SNPs in *CAN1* for independent CAN^R mutants. (**A**) Multiple overlapping probes cover each nucleotide. A mutation at the site indicated in red perturbs hybridization of the sample to all probes. (**B**) The decrease in observed hybridization is used to estimate the log of the likelihood ratio of the presence of a polymorphism versus the absence of a polymorphism (the prediction signal). The presence of a SNP typically results in a region of positive prediction signal with a peak defined as the predicted SNP; for the confirmed mutation indicated in red text in (A), the entire sequence in green has a positive prediction signal shown in (B). (**C**) Using this approach, we detected single–base pair substitutions and a 1-bp deletion in four independent spontaneous CAN^R mutants isolated in a reference genome background (each color represents a different experiment). (**D**) SNPscanner accurately predicts mutations and SNPs in a nonreference genome. The results of nine independent CAN^R mutants in the CEN.PK strain background are shown for the entire *CAN1* gene. We confirmed unique nucleotide substitutions for seven of the mutants, as well as a single-base insertion in one mutant and a single-base deletion in another. At common polymorphisms, indicated in red text, the SNPscanner signal is highly reproducible across multiple samples, allowing intrastrain comparisons of nonreference genomes.

nal threshold of 1 (Fig. 1). At a prediction signal threshold of 5, we detected 86.9% of known SNPs and called only eight false positives in a similar analysis of the reference genome. These false positives were readily excluded by our heuristic filter.

To test our ability to detect accurately a very small number of sequence differences that distinguish two genomes, we analyzed spontaneous mutants in the strain FY3. Independent clones from the same archival isolate were grown, and mutants in the *CAN1*, *GAP1*, and *FCY1* genes were selected on plates containing canavanine sulfate, D-serine and D-histidine, or 5-fluorocytosine, respectively (*20*). For each mutant we hybridized total genomic DNA to a single YTM and analyzed the data with SNPscanner (Fig. 2, A and B). In each of four *can1* mutants, we detected a single peak at the *CAN1* locus that fulfilled our prediction criteria for a SNP (Fig. 2C). Amplification and sequencing of the *CAN1* locus identified a single-base substitution in each of three mutants (31844G → T; 32064C → G; 32757G → C) and deletion of a single thymine in a run of four thymines in the fourth (32924ΔT). Although the prediction signal for this deletion was comparatively low, its detection is noteworthy because no insertions or deletions (indels) were included in the set of SNPs used to train the model.

Analysis of DNA from a mutant resistant to D-histidine and D-serine predicted a mutation in *GAP1* (chromosome XI), which we confirmed as a 514919C → G substitution by sequence analysis (fig. S4). Similarly, we accurately predicted a mutation in *FCY1* (chromosome XVI) for a mutant resistant to 5-fluorocytosine (677256C → T; fig. S5). Thus, we were able to detect a variety of single-base changes, including a single-base deletion, at several different loci in the genome and map them to within 2 bp of the verified substitution (table S1).

In addition to the anticipated mutations, our analysis yielded 12 to 414 additional predictions per genome (table S2). We identified two main causes of experimental noise: (i) false positives that fell within repetitive genomic features, such as retrotransposons and telomeres, which we subsequently excluded (table S1); and (ii) manufacturing defects in microarrays, which we computationally removed (*20*) (fig. S6). We ranked the remaining predictions on the basis of signal strength for each mutant and found the expected mutation in the top five predictions for all mutants except the one resulting from an indel (table S1). One SNP prediction (chromosome IV, position 548,350, sequence confirmed as 548348G → C) was common to all samples, suggesting an early mutation event that preceded later experiments (perhaps during single-colony purification from the archived stock culture). Sequence confirmation of high-quality predictions passing our filtering criteria identified additional unique mutations in three of the six spontaneous mutants (table S1). Thus, our algorithm is sufficiently sensitive to detect a small number of base changes that distinguish two genomes with no a priori knowledge of the variants' location. These results indicate that only a small number of mutations (<5) are associated with the generation of spontaneous drug resistance mutants.

We extended our approach to characterize the genome of the unsequenced laboratory yeast strain CEN.PK, commonly used in continuous culture experiments. CEN.PK shares ancestry with the reference strain, S288C, but some genes are absent in CEN.PK (*22*). We obtained a nucleotide-resolution comparison with the reference sequence by analyzing data with SNPscanner from a single hybridization of CEN.PK DNA. CEN.PK has a strikingly mosaic structure, with large portions of the genome sharing essentially complete sequence identity with FY3 interspersed with regions of sequence divergence and large deletions (fig. S7).

We investigated whether we could detect single mutations on a genome-wide scale in a nonreference genome; this was expected to be a more difficult statistical problem (*20*). We selected 10 spontaneous Can^R mutants in the CEN.PK strain background and hybridized genomic DNA to the YTM. SNPscanner predictions correctly identified a mutation in 9 of 10 mutants, as well as three polymorphic sites present in the wild-type CEN.PK background (Fig. 2D). We confirmed the sequences of all *CAN1* mutations and polymorphisms in the 10 Can^R mutants. Whereas seven of the nine detected mutants had base substitutions in *CAN1*, one mutant contained a 1-bp insertion and another had a 1-bp deletion. All mutations were confirmed as lying within 7 bp of the predicted site (table S2).

SNPscanner prediction signals were highly reproducible across multiple experiments. We compared genome-wide SNP predictions for each CEN.PK *can1* mutant to SNP predictions for CEN.PK wild-type DNA and applied our heuristic filter (*20*). This resulted in the prediction of fewer than 100 SNPs genome-wide that were not predicted to exist in wild-type CEN.PK for 9 of 10 mutants (table S2). In most cases, excluding those predictions that fell in repetitive regions further reduced the total number. By using this approach, we retained the identified *can1* mutation for seven of nine mutants. We ranked the remaining predictions and observed that the sequence-confirmed mutation was in the top 10 predictions for all seven mutants. So even in this somewhat more challenging case, our system succeeded in detecting most of the single-nucleotide sequence differences and mapping them within a few nucleotides. Mutations predicted in our

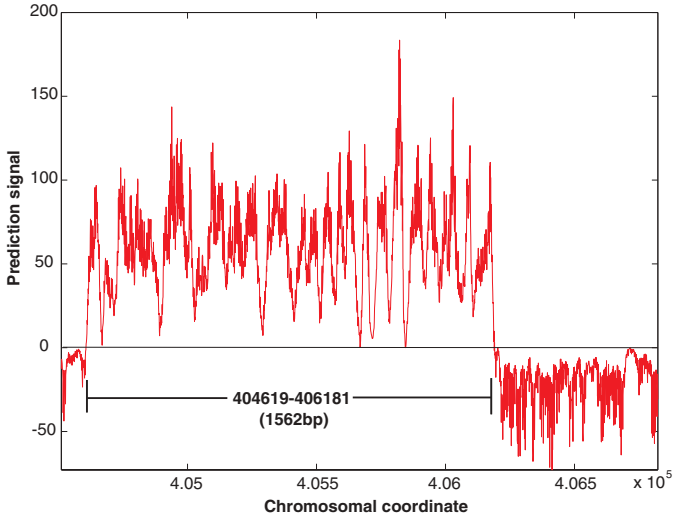

Fig. 3. Genome-wide mutation detection facilitates a genomic approach to genetics. Whole-genome analysis of a strain in which *AMN1* was deleted but that failed to demonstrate the expected nonclumpy phenotype predicted the presence of a 1562-bp deletion (defined by the outermost peak values in prediction signal) in *ACE2* (shown in its entirety). Sequence analysis confirmed the deletion—which spans 1558 bp and is flanked by the nucleotide sequence CTG—and mapped the breakpoints to nucleotides 404,621 to 406,179.

Table 1. Predicted SNPs detected in a yeast strain subjected to experimental evolution.

Strain	Generations under sulfur limitation	Number of SNPscanner predictions	Sequence-confirmed mutation
DBY11130	63	19 unique SNPs	Chromosome IV, 498631C → A in *REG1* (D749Y)
DBY11131	123	6 unique SNPs	Chromosome VII, 858403G → C in *TIM13* (A38P)
DBY11130 and DBY11131	—	12 shared SNPs	—

collection of CEN.PK and FY3 spontaneous mutants corresponded to 9 of the 12 possible base substitutions that resulted in six of the eight possible mismatches between probe and sample. Thus, our method can detect single–base pair indels in addition to virtually all base substitutions.

We sought to apply our genome-wide mutation detection approach to biological questions that had remained refractory to traditional genetic techniques. We complemented a positional cloning project to predict and confirm mutations in *AEP3*, a peripheral mitochondrial inner membrane protein (*23*), that are causative of a growth defect on a nonfermentable carbon source (*20*) (table S3). We also used our method to determine the genetic basis of an unusual phenotype. Deletion of *AMN1* results in upregulation of daughter-specific genes and a nonclumpy growth phenotype (*17*). However, when we deleted *AMN1* in an S288C-like strain (BY4716), we recovered a transformant that displayed low expression of daughter-specific genes and a clumpy phenotype (strain YEF1695). Deletion of *AMN1* in YEF1695 was confirmed by sequence analysis, and independent deletions of *AMN1* in both BY4716 and RM11-1a yielded the expected phenotype, which suggested the presence of a suppressor mutation in YEF1695. Preliminary genetic analysis tended to indicate the presence of an unlinked suppressor mutation. We hybridized genomic DNA from YEF1695 to the YTM. Analysis using SNPscanner confirmed the deletion of *AMN1* (*24*) and identified an additional deletion on chromosome XII (Fig. 3A). The predicted deletion spans ~1.5 kb and includes the majority of the coding region of *ACE*2. Subsequent sequence analysis confirmed that the predicted breakpoints were within 2 bp of the actual sites (Fig. 3).

The deletion of *ACE2* provides a plausible explanation for both aspects of the aberrant phenotype of YEF1695. *ACE2* encodes a transcription factor that is thought to drive the transcription of genes with daughter-specific expression (*25*). Its absence in YEF1695 probably causes a low expression of the daughter cell–specific genes, some of which are required for cell separation after budding (e.g., *CTS1*, which encodes chitinase). Moreover, deletion of *ACE2* alone results in a clumpy phenotype (*26*), and clumpiness segregates with the *ACE2* locus in a cross between YEF1695 and RM11Δ*amn1* (*24*).

Previous studies have shown the occurrence of characteristic gene expression patterns (*27*) and large-scale gene duplication and deletion (*28*) in yeast cultures that are experimentally evolved under a nutrient-limiting condition. However, the extent and nature of nucleotide changes that occur during this process have remained completely unknown. We sought to assess the degree of sequence variation that had accumulated in a strain of yeast subjected to experimental evolution under sulfur limitation in continuous culture. We compared the SNPscanner signals obtained from DNA of the ancestral strain, CEN.PK, to those signals obtained from DNA of two clones from the same population that had undergone experimental evolution under sulfur limitation for 63 (DBY11130) and 123 (DBY11131) generations. We compared our set of predictions to those made for CEN.PK CAN^R mutants to exclude common predictions that were the result of systematic error. SNP predictions that fell within repetitive regions were considered to be unreliable and were excluded from further analysis.

At a prediction signal of >5 we called a small number of predicted SNPs in strains DBY11130 and DBY11131, 12 of which were common to the two strains (Table 1). We confirmed the sequences of single strain-specific mutations found in DBY11130 and DBY11131 (Table 1). The relatively small number of mutations strongly suggests that the events associated with adaptive evolution in chemostats do not involve even transient genome-wide mutagenesis; this number is also consistent with the experience that in yeast, evolved strains are rarely if ever found to have mutator phenotypes (*24*). This is in contrast to studies of *Escherichia coli* grown in batch conditions, in which mutator phenotypes have been observed in numerous independent cultures (*29*). The small number of mutations identified in our experiments means that it will be feasible to comprehensively identify and experimentally verify mutations that are important for adaptation during studies of experimental evolution.

On the basis of a single experimental hybridization, we are able to accurately detect the single-nucleotide changes that distinguish two genomes. Recently, a similar microarray design has been used as a preliminary screen to identify possible mutations in the pathogen *Helicobacter pylori* (*30*). In this method, the initial screen is followed by the manufacture of targeted resequencing microarrays. Our method relies on only a single experiment to derive a statistical measure of the likelihood of a polymorphism at a particular site. Our approach is optimal when direct comparisons are made to the reference strain represented on the microarray. However, we are also able to compare two nonreference genomes and identify the SNPs that distinguish them with only minimal added cost in terms of false negatives and false positives. Although our algorithm is trained on a set of known base substitutions, we found that it also detected single-base deletions and insertions, as well as large deletions with near-nucleotide accuracy in the prediction of breakpoints. Any genomic variation that results in novel sequence (such as inversions or retrotransposon insertions) should, in principle, be detectable by SNPscanner.

We expect that the simplicity and affordability of this method will enable individual laboratory groups to devise and use new and truly comprehensive genomic approaches to Mendelian and complex genetics and to the characterization of mutants obtained through genetic and suppressor screens. In addition, complete knowledge of nucleotide diversity will allow us to address questions regarding the mutagenic effect of phenomena such as aging and recombination on a genome-wide scale. By representing entire genomes of other organisms on oligonucleotide microarrays with a similar redundant design, it is likely that our approach may be extended to higher organisms. Although increased genome complexity presents a challenge, reports of successful SFP-based genotyping in *Arabidopsis* (*12*, *31*), which has a genome of 125 Mb, suggest that genome-wide prediction of all sequence variants may be possible in larger genomes, including those of model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster*. We analyzed haploid genomes and a single homozygous diploid genome; as with all sequencing technologies, identifying heterozygosity in diploid genomes represents the ultimate challenge.

References and Notes

1. M. Margulies *et al.*, *Nature* **437**, 376 (2005).
2. J. Shendure *et al.*, *Science* **309**, 1728 (2005); published online 4 August 2005 (10.1126/science.1117389).
3. J. R. Pollack *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12963 (2002).
4. J. G. Hacia, L. C. Brody, M. S. Chee, S. P. Fodor, F. S. Collins, *Nat. Genet.* **14**, 441 (1996).
5. A. Maitra *et al.*, *Genome Res.* **14**, 812 (2004).
6. C. W. Wong *et al.*, *Genome Res.* **14**, 398 (2004).
7. U. Maskos, E. M. Southern, *Nucleic Acids Res.* **20**, 1675 (1992).
8. E. A. Winzeler *et al.*, *Science* **281**, 1194 (1998).
9. T. L. Turner, M. W. Hahn, S. V. Nuzhdin, *PLoS Biol.* **3**, e285 (2005).
10. S. K. Volkman *et al.*, *Science* **298**, 216 (2002).
11. E. A. Winzeler *et al.*, *Genetics* **163**, 79 (2003).
12. J. O. Borevitz *et al.*, *Genome Res.* **13**, 513 (2003).
13. R. B. Brem, G. Yvert, R. Clinton, L. Kruglyak, *Science* **296**, 752 (2002); published online 28 March 2002 (10.1126/science.1069516).
14. A. M. Deutschbauer, R. W. Davis, *Nat. Genet.* **37**, 1333 (2005).
15. J. Ronald, R. B. Brem, J. Whittle, L. Kruglyak, *PLoS Genet.* **1**, e25 (2005).
16. L. M. Steinmetz *et al.*, *Nature* **416**, 326 (2002).
17. G. Yvert *et al.*, *Nat. Genet.* **35**, 57 (2003).
18. J. Ronald *et al.*, *Genome Res.* **15**, 284 (2005).
19. J. Cheng *et al.*, *Science* **308**, 1149 (2005); published online 24 March 2005 (10.1126/science.1108625).
20. See supporting material on *Science* Online.
21. Z. Gu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1092 (2005).
22. P. Daran-Lapujade *et al.*, *FEMS Yeast Res.* **4**, 259 (2003).
23. T. P. Ellis, K. G. Helfenbein, A. Tzagoloff, C. L. Dieckmann, *J. Biol. Chem.* **279**, 15728 (2004).
24. D. Gresham *et al.*, data not shown.
25. A. Colman-Lerner, T. E. Chin, R. Brent, *Cell* **107**, 739 (2001).
26. W. J. Racki, A. M. Becam, F. Nasr, C. J. Herbert, *EMBO J.* **19**, 4524 (2000).
27. T. L. Ferea, D. Botstein, P. O. Brown, R. F. Rosenzweig, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9721 (1999).
28. M. J. Dunham *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 16144 (2002).
29. P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, *Nature* **387**, 703 (1997).
30. T. J. Albert *et al.*, *Nat. Methods* **2**, 951 (2005).

31. S. P. Hazen *et al.*, *Plant Physiol.* **138**, 990 (2005).
32. We thank E. Foss for creation of strain YEF1695, E. Smith for analysis of YEF1695 mapping data, and D. Storton and J. Matese for technical support. Supported by National Institute of General Medical Sciences grant R01 GM046406 and center grant P50 GM071508 (D.B.), NIH grant R37 MH059520 and James S. McDonnell Foundation grant 99-11T (L.K.), and Pew Charitable Trusts award 2000-002558 (S.C.P.). The raw data and software are available at http://genomics-pubs.princeton.edu/SNPscanner/.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1123726/DC1
Materials and Methods
SOM Text
Figs. S1 to S7
Tables S1 to S3
References

12 December 2005; accepted 24 February 2006
Published online 9 March 2006;
10.1126/science.1123726
Include this information when citing this paper.

Rice Domestication by Reducing Shattering

Changbao Li, Ailing Zhou, Tao Sang*

Crop domestication frequently began with the selection of plants that did not naturally shed ripe fruits or seeds. The reduction in grain shattering that led to cereal domestication involved genetic loci of large effect. The molecular basis of this key domestication transition, however, remains unknown. Here we show that human selection of an amino acid substitution in the predicted DNA binding domain encoded by a gene of previously unknown function was primarily responsible for the reduction of grain shattering in rice domestication. The substitution undermined the gene function necessary for the normal development of an abscission layer that controls the separation of a grain from the pedicel.

Cereals, the world's primary food, were domesticated from wild grass species. Because wild grasses naturally shed mature grains, a necessary early step toward cereal domestication was to select plants that could hold on to ripe grains to allow effective field harvest (*1*, *2*) (fig. S1). The selection process might have been mainly unconscious because grains that did not fall as easily had a better chance of being harvested and planted in the following years. Consequently, nonshattering alleles had an increased frequency and eventually replaced the shattering alleles during domestication. The finding that one locus accounted for most phenotypic variance of grain shattering between a cereal crop and its wild progenitor suggested that the domestication process could have been initiated quickly by selection at the locus (*3*–*5*). The molecular genetic basis of the selection, however, has not been characterized.

Rice (*Oryza sativa*) was domesticated from one or both of two closely related species—*O. nivara* and *O. rufipogon*—distributed from southeastern Asia to India (*6*, *7*). Our recent genetic analysis of an F_2 population derived between *O. sativa* ssp. *indica* and the wild annual species *O. nivara* identified three quantitative trait loci (QTL)—*sh3*, *sh4*, and *sh8*—responsible for the reduction of grain shattering in cultivated rice (*5*). Of these QTL, *sh4* explained 69% of phenotypic variance, and the other two explained 6.0% and 3.1% of phenotypic variance. The *sh4* allele of the wild species caused shattering and was dominant.

Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA.

*To whom correspondence should be addressed. E-mail: sang@msu.edu

Two previous QTL studies using crosses between *O. sativa* ssp. *indica* and the wild perennial species *O. rufipogon* detected four and five shattering QTL (*8*, *9*). Both studies identified a QTL at the same location of *sh4* with either the largest or nearly largest phenotypic effect among the detected QTL. Moreover, genetic analyses between *O. sativa* ssp. *japonica* and *O. rufipogon* and two other closely related wild species *O. glumaepetula* and *O. meridionalis* all found that a single dominant allele from each of the three wild species was responsible for grain shattering (*10*, *11*). This locus, named *Sh3*, was mapped to the same chromosomal location as *sh4*.

Our QTL analysis located *sh4* between simple sequence repeat (SSR) markers RC4-123 and RM280 (*5*), which had a physical distance of about 1360 kb in the *O. sativa* genome (*12*) (Fig. 1A). Because of the large and dominant effect of the *O. nivara* allele, we were able to phenotypically distinguish F_2 individuals that were homozygous recessive (ss) from those that had at least one *O. nivara* allele of *sh4* (ns and nn), regardless of the genotypes at the remaining two QTL of small effect. After evaluating a total of 489 F_2 plants genotyped at the three shattering QTL, we consistently found that plants with the ns and nn genotypes at *sh4* shed all mature grains when hand tapped, whereas plants with the ss genotype at *sh4* did not shed grains or only partially shed mature grains under vigorous hand shaking.

With the reliable phenotyping method available, we grew ~12,000 F_2 seedlings and screened for recombinants between RC4-123 and RM280 (*13*). Plants with the genotype of ss at one marker and ns at the other were selected, and a total of 134 individuals were grown for phenotypic evaluation. By progressively examining SSR and SNP (single-nucleotide polymorphism) markers between RC4-123 and RM280, we finally mapped the mutation responsible for the derivation of nonshattering in cultivated rice to a 1.7-kb region of a gene with a previously unknown function (Fig. 1B and table S1). The gene is predicted to be a transcription factor, and its coding region is physically located between 34,014,305 and 34,012,126 base pairs (bp) on assembly LOC_Os04 g57530 of rice chromosome 4 (The TIGR Rice Genome Annotation Database).

The comparison of the 1.7-kb sequences between the mapping parents revealed seven mutations (Fig. 1C). These include one mutation in the intron: (*a*) a 1-bp substitution; three mutations in the first exon: (*b*) a 15-bp or five–amino acid insertion/deletion, (*c*) a 3-bp or one–amino acid insertion/deletion, and (*d*) a 1-bp or an amino acid substitution; and three mutations 5′ upstream of the start codon: (*e*) a 1-bp substitution at site -55, (*f*) a 3-bp insertion/deletion between sites −343 and −344, and (*g*) an 8-bp insertion/deletion between sites −558 and −559.

To assess the polymorphism and evolutionary direction of these mutations, we sequenced this 1.7-kb region from an additional 14 rice cultivars representing the diversity of *O. sativa* (*14*), 21 accessions of *O. nivara* covering the distributional range of the wild species (*15*), 6 accessions of *O. rufipogon*, and 1 accession of each of the four remaining wild A-genome species (Fig. 1C and table S2). The cultivars were polymorphic for mutation *f*, i.e., some of the cultivars had the same sequence as *O. nivara*. At the remaining six mutation sites, all cultivars shared the same sequences, which were different from those of the *O. nivara* parent.

Surprisingly, three accessions of *O. nivara* had the same sequences as *O. sativa* at these six sites. It was then found that plants grown from these accessions had the nonshattering phenotype. Greenhouse observations indicated that these accessions had additional characteristics of cultivated rice that were not found in *O. nivara*, such as upright tillers, short awns, and/or photoperiod sensitivity. This suggests that the three accessions are weedy rice that has received and fixed the *sh4* allele from cultivars.

The remaining accessions of the wild species with confirmed shattering differed invariably from the cultivars by one mutation, *d*, which was a nucleotide substitution of G for T or an amino acid substitution of