nature genetics

Intragenic tandem repeats generate functional variability

Kevin J Verstrepen[1,2], An Jansen[1], Fran Lewitter[1] & Gerald R Fink[1]

Tandemly repeated DNA sequences are highly dynamic components of genomes[1]. Most repeats are in intergenic regions, but some are in coding sequences or pseudogenes[2]. In humans, expansion of intragenic triplet repeats is associated with various diseases, including Huntington chorea and fragile X syndrome[3,4]. The persistence of intragenic repeats in genomes suggests that there is a compensating benefit. Here we show that in the genome of *Saccharomyces cerevisiae*, most genes containing intragenic repeats encode cell-wall proteins. The repeats trigger frequent recombination events in the gene or between the gene and a pseudogene, causing expansion and contraction in the gene size. This size variation creates quantitative alterations in phenotypes (*e.g.*, adhesion, flocculation or biofilm formation). We propose that variation in intragenic repeat number provides the functional diversity of cell surface antigens that, in fungi and other pathogens, allows rapid adaptation to the environment and elusion of the host immune system.

The sequenced and annotated genome of *S. cerevisiae* provides a unique opportunity to determine the function of intragenic repeat sequences. To identify the *S. cerevisiae* open reading frames (ORFs) that contain intragenic tandem repeats, we scanned all 6,591 ORFs for long (>40 nucleotide (nt)) or short (3–39 nt) repeats and identified 44 ORFs: 29 with repeats longer than 40 nt (**Fig. 1a**) and 15 with short repeats (**Fig. 1b**). These 44 genes showed unexpected functional similarities. Eighteen of the 29 ORFs (62%) with conserved long repeats encode cell-wall proteins. By comparison, only 1.3% of all *S. cerevisiae* ORFs are cell-surface proteins (88 of 6,591). An additional four genes (*CTR1*, *MNN4*, *MSB2* and *HKR1*) encode plasma membrane proteins with extracellular domains. Hence, more than 75% (22 of 29) of all genes with long intragenic tandem repeats encode cell-surface proteins. The group of 15 genes with short repeats contains only one cell-wall gene (*SCW11*), but several genes in this group encode regulators of cell-wall synthesis and maintenance, such as *MSS11* (regulator of adhesion), *WSC3* (regulator of cell-wall integrity) and *CHS5* (regulator of chitin biosynthesis).

All the repeats were in-frame, so that deletion or addition of repeat units would not alter the reading frame. To verify that the intragenic repeat regions vary in size between yeast strains owing to expansion or contraction of the repeats, we amplified each of the identified repetitive regions by PCR and compared their sizes in six different *S. cerevisiae* strains. The length of the repeat region in 35 of the 44 genes with intragenic repeats varied from strain to strain (**Fig. 2** and **Supplementary Figs. 1** and **2** online). Virtually all cell-surface genes with conserved repeats showed size variation. Moreover, strains that had a ploidy greater than haploid often contained several different alleles of the same gene. The size difference between the genes in different *S. cerevisiae* strains is noteworthy, as the size of most genes has been conserved over millions of years in different yeast species[5]. To confirm that genes in these six strains do not generally vary in size, we analyzed 16 genes without repeats: 8 cell-surface genes, 4 long genes (>3 kb) and 4 genes encoding various enzymes. None of these 16 genes lacking repeats showed any length differences among the six *S. cerevisiae* strains (**Supplementary Fig. 3** online).

To characterize the events leading to expansion and contraction of intragenic repeats, we designed a system that permitted us to detect events occurring in the repeat region in one of the genes with repeats (*FLO1*). *FLO1* is a homolog of the human mucin genes and encodes a cell-surface adhesin, a mannoprotein responsible for adherence to other yeast cells (flocculation) as well as certain surfaces[2,6]. We inserted a single copy of the gene *URA3* among the repeats of the genomic copy of *FLO1* in strain S288C (**Fig. 3a**). In this strain, *FLO1* is 4.6 kb long and contains 18 repeats of ~100 nt, separated by a less-conserved 45-nt sequence. We grew the *FLO1*::*URA3* strains on medium without uracil and then spread them on plates containing 5-fluoro-orotic acid (5-FOA), which selects for mitotic segregants that have lost the *URA3* marker (**Fig. 3b**).

The *FLO1*::*URA3* strains gave rise to Ura− segregants at a high frequency ($\sim 1 \times 10^{-5}$; **Fig. 3b**,**c**). Moreover, the frequency of segregants gradually increased with increasing numbers of repeated DNA motifs surrounding the *URA3* marker (**Fig. 3c**). The Ura− segregants had alterations in the number of repeats relative to the starting strain. Most of the new *FLO1* alleles obtained after loss of the *URA3* marker had fewer repeat units than did the wild-type *FLO1* allele. But ~15% (7 of 50) of the alleles gained extra repeats, causing gene size to increase to as much as 1 kb. This indicates that the *URA3* marker is not just 'looped out' by unequal crossover between repeat units surrounding the marker. Moreover, the wide range of different alleles generated in this procedure indicates that different repeat units

[1]Whitehead Institute for Biomedical Research/Massachusetts Institute of Technology, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Centre for Malting and Brewing Science, Department of Microbial and Molecular Systems, Faculty of Applied Bioscience and Engineering, K.U. Leuven, Kasteelpark Arenberg 22, B-3001 Leuven (Heverlee), Belgium. Correspondence should be addressed to G.R.F. (GFink@wi.mit.edu).

Published online 7 August 2005; doi:10.1038/ng1618

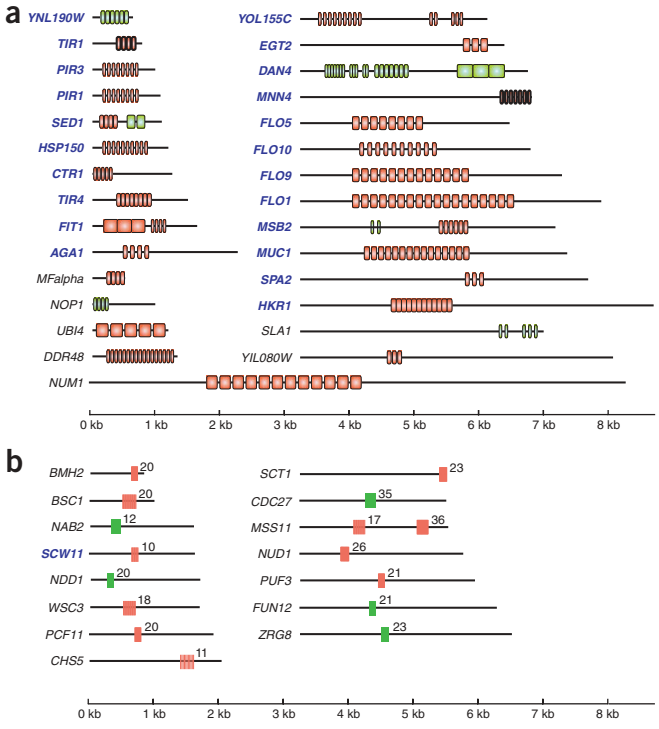

Figure 1 *S. cerevisiae* genes containing conserved intragenic repeats. A screen of all ORFs in the *S. cerevisiae* genome for those containing conserved intragenic tandem repeats identified 29 genes with large (≥40 nt) repeats (**a**) and 15 genes with short (<40 nt) repeats (**b**). Repeats (vertical boxes) that vary in size among 6 different *S. cerevisiae* strains are shown in red (**Supplementary Figs. 1** and **2** online); repeats that do not show size variation among these strains are shown in green. The names of cell-surface genes are shown in boldface blue type. The numbers in **b** indicate the number of repeats. More information about the repeats is given in **Supplementary Table 1** online. The repeat units in most genes are distinct from those in others except in *FLO1*, *FLO5* and *FLO9*, which share the same repeat unit, and in *PIR1*, *PIR3* and *HSP150*, which share the same repeat unit. *SED1*, *FIT1*, *SLA1* and *MSB2* contain two intragenic repeat regions with different repeat sequences. *YOL155C* and *DAN4* contain three distinct repeat regions.

can freely recombine with each other. Sequence analysis of the wild-type *FLO1* allele and three of the new short *FLO1* alleles confirmed that each of these short alleles had lost several repeat units. Moreover, because all different repeats in the *FLO1* genes have slight sequence differences[2], it is possible to determine which repeat units were lost by aligning the sequences of the new alleles with that of the wild-type *FLO1* allele (**Supplementary Fig. 4** online). This analysis shows that, in all cases, an upstream repeat unit had fused with one of the downstream units, removing several repeat units in between but preserving the ORF.

Both the *PIR* and the *FLO1* gene families have pseudogenes containing repeats that are similar to those in the functional copies[2,7]. These pseudogenes may provide additional genetic information that could be incorporated by a recombination event. In fact, two independent strains contain a new *FLO1* allele formed by the fusion of the first repeat unit of *FLO1* with a repeat unit similar to those in *FLO1* found in the *FLO1* pseudogene *YAR062W*. This pseudogene is located ~12 kb downstream from the *FLO1* termination codon[2] (**Fig. 4a**). Sequence analysis of the *FLO1-YAR062W* fusion showed that the first *FLO1* repeat had recombined with the repeat in the pseudogene, looping out the complete 12-kb sequence between the repeats in *FLO1* and the pseudogene. Southern blotting and clamped homogeneous electrical field (CHEF) chromosome analysis confirmed the loss of ~12 kb between *FLO1* and the pseudogene on chromosome I (**Fig. 4b,c**).

To determine the functional consequence of continued variation in cell-wall genes carrying intragenic repeats, we compared the effects of eight newly generated *FLO1* alleles (2.9–5.4 kb; **Fig. 5a,b**) on various adhesion phenotypes associated with *FLO1* (ref. 6). We fused each *FLO1* size variant to the inducible *GAL1* promoter in the S288C background. In strain S288C, all five *FLO* genes are transcriptionally silent[2,8], and so the ectopic expression of these *GAL1*p-*FLO1* constructs permits evaluation of the contribution of the particular *FLO1* allele. As expected, none of the strains had any adhesion phenotypes on glucose medium. When we grew these strains carrying the *GAL1*p-*FLO1* fusion on galactose medium (YPGal), however, there was a linear correlation between gene size and the extent of adhesion: as the Flo1 proteins became longer (carrying more repeats), the adhesion properties gradually became stronger (**Fig. 5c,d**). Flocculation (*i.e.*, adhesion to other yeast cells) showed the same quantitative relationship to the repeat number: the more repeats, the greater the fraction of flocculating cells (**Fig. 5e**). The observed correlation between the number of repeats and gain of function of Flo1 relied on the specific amino acids encoded by the repeats, because insertion of *URA3* in the *FLO1* repeat region totally abolished adhesion (data not shown).

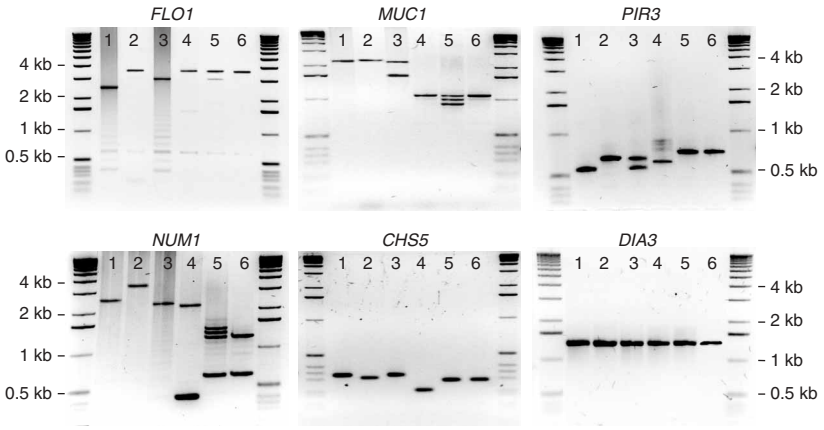

Figure 2 Intragenic repetitive domains vary in size. The repetitive domains of all 44 genes carrying intragenic repeats (**Fig. 1**) and the ORFs of 16 control genes without repeats were amplified by PCR for six different *S. cerevisiae* strains. Results for five genes with repeats (*FLO1*, *MUC1* (also called *FLO11*), *PIR3*, *NUM1* and *CHS5*) and one gene without repeats (*DIA3*) are shown. Lane 1, S288C (haploid); lane 2, Sigma1278b (haploid); lane 3, EM93 (diploid); lane 4, CMBS355 (polyploid); lane 5, CMBS DL16 (polyploid); lane 6, CMBS33 (polyploid). Results for other genes are shown in **Supplementary Figures 1–3** online. The variability in some cell-surface genes has been used for the genotyping of wine yeasts[30].

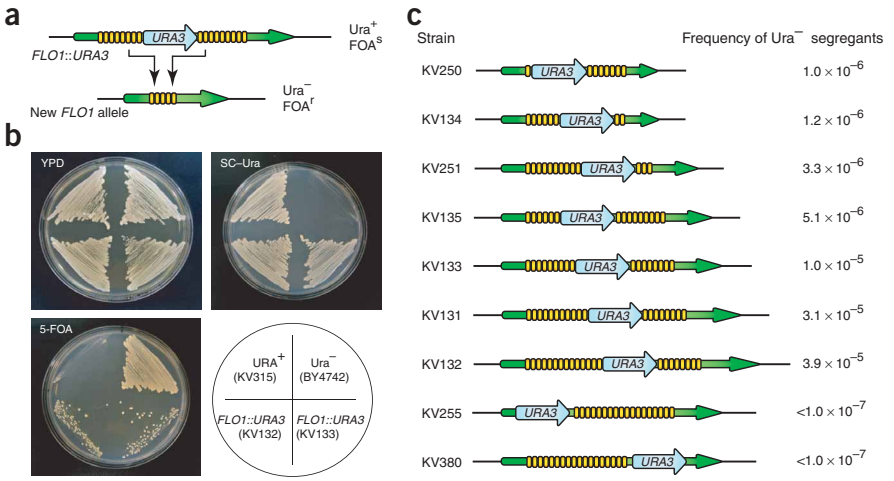

Figure 3 Intragenic repeats are hot spots for recombination. (**a**) To monitor recombination between intragenic repeats in *FLO1*, a *URA3* expression cassette was integrated at various positions in the *FLO1* repeats. As a consequence of the numerous recombination events in the repeats, the *URA3* marker is lost at exceptionally high frequencies, resulting in a 5-FOA-resistant (Ura⁻) strain containing a new *FLO1* allele. (**b**) Assay for loss of the *URA3* marker. Ura⁺ strains (KV315, *URA3* integrated at its native locus in the genome) grow on minimal medium (SC –Ura) but not on 5-FOA medium. Ura⁻ strains (BY4742) grow on 5-FOA but not on minimal medium. Strains KV132 and KV133, with a *URA3* cassette in the *FLO1* repeats (*FLO1::URA3*), grow on minimal medium. Owing to recombination events in the repeats, the *URA3* cassette is looped out at high frequencies, yielding many 5-FOA-resistant segregants. (**c**) *S. cerevisiae* strains with the *URA3* cassette integrated into various positions in the genomic *FLO1* repeats *(FLO1::URA3)* were grown on medium lacking uracil and then plated onto 5-FOA medium. The numbers indicate the frequencies of Ura⁻ segregants.

To analyze the mechanism involved in the recombination of intragenic *FLO1* repeats, we measured the stability of the repeats in various key DNA repair and recombination mutants (**Table 1**). In most cases, replication slippage or the repair of doubled-stranded breaks during DNA replication are the main mechanisms for repeat expansion and contraction[9–13]. Various *RAD* genes influence mutation frequencies in repeats[9–13]. We found that loss of the *RAD27*-encoded flap endonuclease, which causes the formation of double-stranded breaks during replication[12–15], increased the instability of *FLO1* repeats by a factor of almost 40. The increased recombination frequency in *rad27*Δ mutants suggests that *FLO1* repeat instability is associated with the occurrence of double-stranded breaks due to defective DNA replication[12,14,15]. Deletion of *RAD52* and *RAD50* severely reduced the frequency of rearrangements, whereas deletion of the RecA homolog *RAD51* did not affect the frequency. Rad51 is required for ATP-dependent strand invasion during conservative DNA repair and recombination processes[16]. The absence of an effect in *rad51*Δ mutants suggests that *FLO1* recombination does not require strand invasion; therefore, gene conversion, break-induced replication and crossing-over are unlikely recombination mechanisms. Instead, the decrease in recombination observed in *rad50*Δ, *rad52*Δ and *rad1*Δ *rad52*Δ mutants suggests that the process depends on break repair by single-strand annealing[11], a conclusion further supported by the decrease in *FLO1* recombination in the *rad59*Δ mutant, which is known to be deficient in this type of DNA repair[17]. Moreover, in contrast to many other possible models, the proposed model also accounts for the expansion in the number of repeats found in some of the Ura⁻ segregants. Taken together, the recombination frequencies observed in the various mutants indicate that recombination between the *FLO1* repeats is caused by a replication slippage process similar to that observed in intergenic repeats (**Supplementary Fig. 5** online).

Our data show that expansion and contraction of repeats result in gradual, quantitative and fully reversible functional changes that permit existing features of the organism to be rapidly attuned to a particular environment. The presence of repeats in the *FLO* adhesins, for example, enables *S. cerevisiae* to adapt its adhesion behavior,

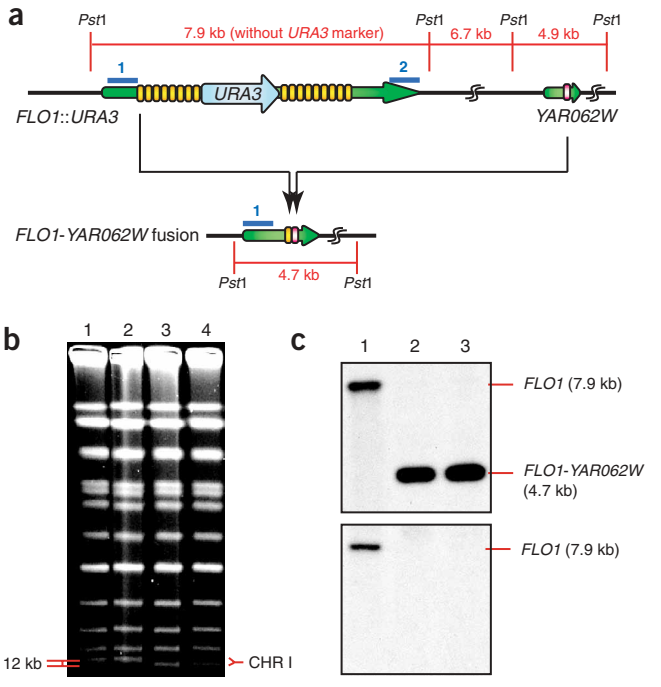

Figure 4 Repeats in pseudogenes provide an additional source of variability. (**a**) Most of the recombination events between *FLO1* repeats are strictly intragenic (*i.e.*, only repeats in *FLO1* recombine with each other), but in some cases, the repeats in *FLO1* recombine with a similar repeat unit found in the *FLO1* pseudogene *YAR062W*, which is located ~12 kb downstream of *FLO1*. Fusion of a repeat in *FLO1* with that in *YAR062W* results in deletion of the 3′ end of *FLO1* and the entire 12 kb of DNA separating the two ORFs. (**b**) The *FLO1-YAR062W* deletion and fusion results in altered mobility of chromosome I (231 kbp) using CHEF electrophoresis. Lane 1, wild-type *S. cerevisiae* S288C; lane 2, control Ura⁻ segregant (KV291) that has lost only intragenic *FLO1* repeats; lanes 3 and 4, *FLO1-YAR062W* fusion strains (KV360 and KV361). (**c**) Southern-blot analysis confirms the deletion of the 12-kb region between *FLO1* and *YAR062W*. Genomic DNA of wild-type cells (lane 1) and the *FLO1-YAR062W* fusion strains (lanes 2 and 3) was cut with *Pst*1 and used for Southern blotting with probes that bind to the 5′ portion of *FLO1* (probe 1, top) and the 3′ portion of *FLO1* (probe 2, bottom). Other probes were used to confirm the fusion (data not shown).

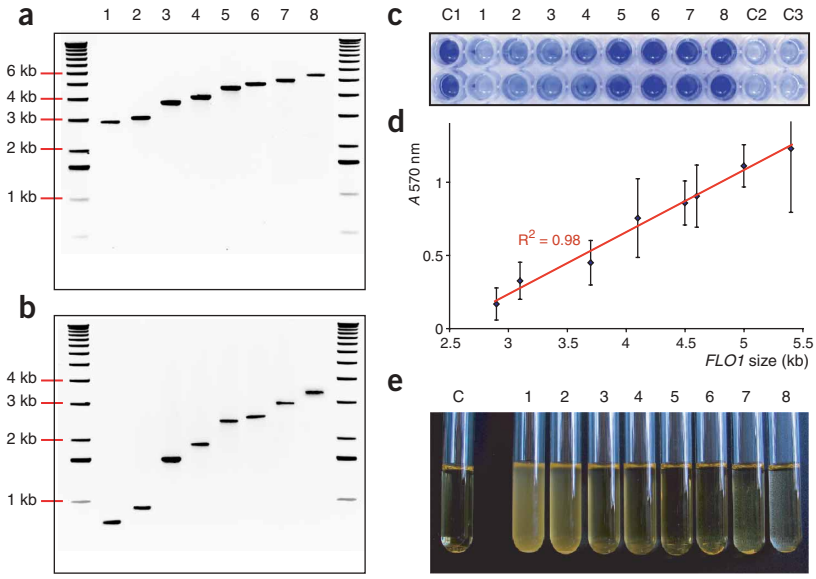

Figure 5 Instability of the *FLO1* repeats generates functional variability. *S. cerevisiae* strain KV133 (*FLO1*::*URA3*) was plated onto 5-FOA medium to select Ura$^-$ segregants. (**a**) The Ura$^-$ segregants contain *FLO1* alleles of different lengths, ranging from 2.9 kb to 5.4 kb. (**b**) PCR amplification of the *FLO1* repetitive domains shows that the differences in length of the alleles are due to corresponding differences in the length of the *FLO1* repeat region. Lane 1, strain KV298 (*FLO1* ORF = 2.9 kb); lane 2, KV308 (3.1 kb); lane 3, KV220 (3.7 kb); lane 4, KV219 (4.1 kb); lane 5, KV224 (4.5 kb); lane 6, KV211 (4.6 kb); lane 7, KV312 (5.0 kb); lane 8, KV311 (5.4 kb). (**c**) Expansion of *FLO1* repeats leads to increased adherence to polystyrene. The *FLO1* genes of the S288C strain BY4742 (KV210, lane C1) and of eight strains containing different-sized *FLO1* alleles (**a**) were fused to the *GAL1* promoter. Cells were grown on galactose medium and then tested for adhesion to polystyrene by staining with crystal violet[29]. Cells expressing a long allele of *FLO1* showed strong adhesion to polystyrene, whereas cells expressing shorter alleles did not adhere. When grown on glucose, strain KV311 failed to adhere (lane C2). Strain KV306, which contains the same *FLO1* allele as strain KV311 but lacks the *GAL1* promoter, failed to adhere when grown on galactose medium (lane C3). (**d**) Linear relationship between adherence to polystyrene and number of repeats. The error bars represent the standard deviation of three independent experiments. (**e**) Expansion of *FLO1* repeats results in stronger cell-cell adhesion. The *GAL1*p-*FLO1* fusion strains (tubes 1–8; **b**) were tested for flocculation. Cells expressing a long allele of *FLO1* showed extremely strong cell-cell adhesion (all cells sedimented on the bottom of the test tube). Tube C contains a strain (KV210) carrying the wild-type *FLO1* allele of *S. cerevisiae* strain BY4742 fused to the *GAL1* promoter.

finding an optimal balance between adherent cells and free cells that can escape from the mass and explore new surfaces. For pathogenic fungi like *Candida albicans* and *Candida glabrata*, such recombination events in their adhesin genes (*ALS* and *EPA* genes, respective homologs of *FLO1*) could enable the cells to adhere to new host tissues. Variability at the cell surface of these pathogens may also permit evasion of the host immune system[2]. Notably, intragenic repeats are also present in cell-surface genes of nonfungal pathogens, including *Haemophilus influenzae*[18], *Bacillus anthracis*[19], *Leishmania infantum*[20] and various *Plasmodium* species[21]. Hence, recombination of intragenic repeats may be a widespread mechanism among microorganisms to generate cell-surface diversity from a single gene. This mechanism differs from that in trypanosomes, where diversity arises from the expression of different, unlinked members of a large library of genes[22].

In multicellular eukaryotes, repeat expansion and contraction may have relevance for the generation of variability in genes other than those that function in the cell surface. For example, the rapid yet topologically conservative evolution of canine skeletal morphology has been attributed to the expansion and contraction of intragenic repeats in developmental genes[23]. In humans, the mucin (*MUC*) genes, which are homologs of the *S. cerevisiae FLO* genes, contain variable numbers of a 60-bp intragenic tandem repeat. Elevated expression of *MUC* genes induces tumorigenesis[24] and is currently used as a marker for malignant tumors. Extensive size differences in *MUC* genes have been reported[25], but the relationship of this variation to malignancy is not yet known.

Table 1 Frequency of recombination between intragenic repeats in selected DNA repair and recombination mutants

Relevant genotype	Ratio relative to wild-type
FLO1::*URA3* (wild-type)	1
FLO1::*URA3* Δ*rad1*	0.7
FLO1::*URA3* Δ*rad27*	37*
FLO1::*URA3* Δ*rad50*	0.09*
FLO1::*URA3* Δ*rad51*	0.94
FLO1::*URA3* Δ*rad52*	0.05*
FLO1::*URA3* Δ*rad59*	0.24*
FLO1::*URA3* Δ*rad1* Δ*rad52*	0.01*
FLO1::*URA3* Δ*dnl4*	0.9
URA3 inserted at *URA3* locus	<0.03*

*Significant differences from the Rad$^+$ control as measured by Student's *t*-test at a confidence level of 99%.
The frequency of recombination in the intragenic repeats in *FLO1* was determined by measuring the frequency of Ura$^-$ segregants. The frequency measured in various DNA recombination and repair mutants was compared with that of an otherwise isogenic Rad$^+$ strain (KV133). The recombination frequency in strain KV133 is 1×10^{-5} (**Fig. 3c**). All measurements were repeated at least three times.

METHODS

Bioinformatics. To find intragenic repeats, we used the EMBOSS ETANDEM software[26] to screen the sequences of all *S. cerevisiae* ORFs. Two separate screens identified the short (3–39 nt) and long (>40 nt) repeats. We set the ETANDEM threshold score to 20. The first screens identified 323 ORFs with long repeats and 859 ORFs with short repeats (**Supplementary Table 1** online). A second screen refined the results of the initial screens by excluding dubious ORFs and ORFs with poorly conserved (degenerated) repeats. A sequence was considered to be an intragenic repeat if it met three conditions: (i) the ORF was not a dubious or hypothetical ORF according to the Saccharomyces genome database; (ii) repeat conservation was at least 85%; and (iii) the number of repeats was at least 20 for trinucleotide repeats, 16 for repeats between 4 and 10 nt, 10 for repeats between 11 and 39 nt and 3 for repeats of at least 40 nt.

Strains and molecular biology. All yeast strains used are listed in **Supplementary Table 2** online. We grew yeast cultures as described[27]. YPGal medium contained 2% raffinose, 2% galactose (Sigma Chemical Co.), 2% peptone (Difco) and 1% yeast extract (Difco). We used standard procedures and reagents for molecular biology. We inserted the *URA3* marker into the

intragenic repeats in *FLO1* by transformation. We amplified a *URA3* cassette by PCR using primers containing 5′ tails with sequences homologous to the consensus repeated motif found in *FLO1* and the plasmid pRS306 (ref. 28) as a template. We used these constructs to transform a Ura$^-$ recipient. Owing to the similarity between the repeats, the construct integrated at various positions in the *FLO1* repeats, thereby replacing a variable number of repeats. In some cases, insertion of *URA3* led to an increase in the number of repeat units. We carried out real-time PCR using the ABI 7500 system (Applied Biosystems) with the appropriate enzymes and chemicals from Applied Biosystems as recommended by the supplier. All PCR primers are listed in **Supplementary Table 3** online. We carried out CHEF chromosome separation using a BioRad CHEF-DRII using the protocol supplied by BioRad. We tested flocculation and adhesion to polystyrene as described previously[6,29].

Recombination analysis. To measure the recombination frequency in the various *FLO1*::*URA3* strains, we grew single colonies on SD −Ura plates, inoculated them in SD −Ura medium and used them to inoculate a 50-ml culture with an initial cell concentration of 1×10^6 cells ml^{-1}. We shook this culture for 14 h at 28 °C, collected cells, washed them with sterile distilled water and resuspended them in water to a concentration of 5×10^8 cells ml^{-1}. We used this cell suspension to make a dilution series, of which we plated 150 µl onto SD plates containing 1 g l^{-1} 5-FOA to select for Ura$^-$ segregants. Because there is no growth on nonselective medium, frequencies measured by this method provide a good estimate of actual recombination rates (number of events per cell division). Loss of the *URA3* marker was confirmed by PCR. We repeated all experiments at least three times and used the average number of colonies to calculate the recombination frequency. We estimated statistical significance using the Student's *t*-test.

Accession codes. GenBank: short *FLO1* alleles, AY949845–AY949848; *FLO1-YAR062W* fusion genes, DQ029324 and DQ029325.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank A. Sinskey, A. Amon, T. Petes and all members of the laboratory of G.R.F. for discussions; K. Walker for assistance with the bioinformatics; and T. DiCesare for help with the graphics. K.J.V. is a post-doctoral fellow of the Fund for Scientific Research Flanders (F.W.O. Vlaanderen) and a D. Collen Fellow of the Belgian American Educational Foundation (B.A.E.F.). G.R.F. is an American Cancer Society Professor of Genetics. This research was supported by a grant from the US National Institutes of Health to G.R.F.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 21 April; accepted 12 July 2005
Published online at http://www.nature.com/naturegenetics/

1. Hartl, D.L. Molecular melodies in high and low C. *Nat. Rev. Genet.* **1**, 145–149 (2000).
2. Verstrepen, K.J., Reynolds, T.B. & Fink, G.R. Origins of variation in the fungal cell surface. *Nat. Rev. Microbiol.* **2**, 533–540 (2004).
3. Li, Y.C., Korol, A.B., Fahima, T. & Nevo, E. Microsatellites within genes: structure, function and evolution. *Mol. Biol. Evol.* **21**, 991–1007 (2004).
4. Jin, P., Alisch, R.S. & Warren, S.T. RNA and microRNAs in fragile X mental retardation. *Nat. Cell Biol.* **6**, 1048–1053 (2004).
5. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
6. Guo, B., Styles, C.A., Feng, Q. & Fink, G. A *Saccharomyces* gene family involved in invasive growth, cell-cell adhesion, and mating. *Proc. Natl. Acad. Sci. USA* **97**, 12158–12163 (2000).
7. Harrison, P. *et al.* A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.* **316**, 409–419 (2002).
8. Liu, H.P., Styles, C.A. & Fink, G.R. *Saccharomyces cerevisiae* S288C has a mutation in *FLO8*, a gene required for filamentous growth. *Genetics* **144**, 967–978 (1996).
9. Lovett, S.T. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* **52**, 1243–1253 (2004).
10. Viguera, E., Canceill, D. & Ehrlich, S.D. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.* **20**, 2587–2595 (2001).
11. Paques, F. & Haber, J.E. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **63**, 349–404 (1999).
12. Richard, G.F., Dujon, B. & Haber, J.E. Double-strand break repair can lead to high frequencies of deletions within short CAG CTG trinucleotide repeats. *Mol. Gen. Genet.* **261**, 871–882 (1999).
13. Kokoska, R.J. *et al.* Destabilization of yeast micro- and minisatellite DNA sequences by mutations affecting a nuclease involved in Okazaki fragment processing (rad27) and DNA polymerase delta (pol3-t). *Mol. Cell. Biol.* **18**, 2779–2788 (1998).
14. Callahan, J.L., Andrews, K.J., Zakian, V.A. & Freudenreich, C.H. Mutations in yeast replication proteins that increase CAG/CTG expansions also increase repeat fragility. *Mol. Cell. Biol.* **23**, 7849–7860 (2003).
15. Debrauwère, H., Loeillet, S., Lin, W. & Nicolas, A. Links between replication and recombination in *Saccharomyces cerevisiae*: a hypersensitive requirement for homologous recombination in the absence of Rad27 activity. *Proc. Natl. Acad. Sci. USA* **98**, 8263–8269 (2001).
16. Krogh, B.O. & Symington, L.S. Recombination proteins in yeast. *Annu. Rev. Genet.* **38**, 233–271 (2004).
17. Davis, A.P. & Symington, L.S. The yeast recombinational repair protein Rad59 interacts with Rad52 and stimulates single-strand annealing. *Genetics* **159**, 515–525 (2001).
18. Hood, D.W. *et al.* DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. USA* **93**, 11121–11125 (1996).
19. Keim, P. *et al.* Molecular diversity in *Bacillus anthracis*. *J. Appl. Microbiol.* **87**, 215–217 (1999).
20. Boceta, C., Alonso, C. & Jimenez-Ruiz, A. Leucine rich repeats are the main epitopes in *Leishmania infantum* PSA during canine and human visceral leishmaniasis. *Parasite Immunol.* **22**, 55–62 (2000).
21. Sakihama, N. *et al.* Relative frequencies of polymorphisms of variation in Block 2 repeats and 5′ recombinant types of *Plasmodium falciparum msp1* alleles. *Parasitol. Int.* **53**, 59–67 (2004).
22. Pays, E., Vanhamme, L. & Berberof, M. Genetic controls for the expression of surface antigens in African trypanosomes. *Annu. Rev. Microbiol.* **48**, 25–52 (1994).
23. Fondon, J.W. III & Garner, H.R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. USA* **101**, 18058–18063 (2004).
24. Schroeder, J.A. *et al. MUC1* overexpression results in mammary gland tumorigenesis and prolonged alveolar differentiation. *Oncogene* **23**, 5739–5747 (2004).
25. Patton, S., Gendler, S.J. & Spicer, A.P. The epithelial mucin, *MUC1*, of milk, mammary gland and other tissues. *Biochim. Biophys. Acta* **1241**, 407–423 (1995).
26. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
27. Sherman, F., Fink, G.R. & Hicks, J. *Methods in Yeast Genetics* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1991).
28. Sikorski, R.S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* **122**, 19–27 (1989).
29. Reynolds, T.B. & Fink, G.R. Bakers' yeast, a model for fungal biofilm formation. *Science* **291**, 878–881 (2001).
30. Marinangeli, P., Angelozii, D., Ciani, M., Clementi, F. & Mannazzu, M. Minisatellites in *Saccharomyces cerevisiae* genes encoding cel wall proteins: a new way towards wine strain characterisation. *FEMS Yeast Res.* **4**, 427–435 (2004).