5970–5982 *Nucleic Acids Research, 2008, Vol. 36, No. 18* *Published online 23 September 2008*
doi:10.1093/nar/gkn594

Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in *Candida glabrata*

Agnès Thierry[1], Christiane Bouchier[2], Bernard Dujon[1] and Guy-Franck Richard[1,*]

[1]Institut Pasteur, Unité de Génétique Moléculaire des Levures; CNRS, URA2171; Université Pierre et Marie Curie, UFR 927; 25 rue du Dr Roux and [2]Institut Pasteur, Plate-Forme 1-Génomique; 28 rue du Dr Roux, F-75015 Paris, France

Received July 1, 2008; Revised August 26, 2008; Accepted September 3, 2008

ABSTRACT

Minisatellites are DNA tandem repeats that are found in all sequenced genomes. In the yeast *Saccharomyces cerevisiae*, they are frequently encountered in genes encoding cell wall proteins. Minisatellites present in the completely sequenced genome of the pathogenic yeast *Candida glabrata* were similarly analyzed, and two new types of minisatellites were discovered: minisatellites that are composed of two different intermingled repeats (called compound minisatellites), and minisatellites containing unusually long repeated motifs (126–429 bp). These long repeat minisatellites may reach unusual length for such elements (up to 10 kb). Due to these peculiar properties, they have been named 'megasatellites'. They are found essentially in genes involved in cell–cell adhesion, and could therefore be involved in the ability of this opportunistic pathogen to colonize the human host. In addition to megasatellites, found in large paralogous gene families, there are 93 minisatellites with simple shorter motifs, comparable to those found in *S. cerevisiae*. Most of the time, these minisatellites are not conserved between *C. glabrata* and *S. cerevisiae*, although their host genes are well conserved, raising the question of an active mechanism creating minisatellites *de novo* in hemiascomycetes.

INTRODUCTION

As more and more eukaryotic genomes are sequenced, a wealth of new information on gene duplications, evolution of paralogous sets of genes, differential loss of genes and neo-functionalization of paralogues becomes available. With the largest number of species sequenced within a single phylum, hemiascomycetous yeasts stand up as a reference for comparative genomics (1). Minisatellites are a subclass of DNA tandem repeats, that exhibit size polymorphism among different individuals or isolates (2). In previous works, it was found that minisatellites are spread in the *Saccharomyces cerevisiae* genome, and are preferentially encountered in genes encoding proteins involved in cell wall formation (3,4). Such proteins, including those belonging to the FLO family of flocculins, exhibit a variable number of repeats among different yeast strains. The role of such repeats is illustrated by the fact that strains having a larger number of repeats in the *FLO1* gene exhibit better adhesion than those with a smaller number of repeats (5). Similarly, *S. cerevisiae* strains isolated from biofilms formed at the surface of sherry wines contain an increased number of repeats in one of the *FLO11* minisatellites, supporting the importance of such sequences in cell adhesion (6).

We previously reported that *S. cerevisiae* minisatellites are frequently not conserved in the corresponding orthologous gene in other hemiascomycetes, suggesting that minisatellites are created, evolve and disappear at a faster pace that the genes containing them (3), a property shared by microsatellites, another class of DNA tandem repeats with shorter motifs (7). In order to investigate more thoroughly the mechanisms of creation and loss of minisatellites in a pathogenic hemiascomycete, we searched all such repeats in the genome of *Candida glabrata*, a human opportunistic pathogen, responsible for mucosal candidiasis, blood stream infections and vaginitis. *Candida glabrata* is the second cause of nosocomial infections due to yeasts, after *C. albicans*. Its genome was completely sequenced, and revealed its closer relationship to *S. cerevisiae* than to *C. albicans* (8), making comparisons easier. Despite similar genome sizes, we found three times as many minisatellites in *C. glabrata* as compared to *S. cerevisiae*. We also discovered two new species of minisatellites absent from the *S. cerevisiae* genome, including some unusually long minisatellites, composed of several kilobases of a tandemly repeated sequence. We propose to

*To whom correspondence should be addressed. Tel: +33 1 40 61 34 54; Fax: +33 1 40 61 34 56; Email: gfrichar@pasteur.fr

name them 'megasatellites', in order to distinguish them from more regular minisatellites. Megasatellites are present in genes whose sequences suggest that they are involved in cellular adhesion. Some of the peculiar DNA motifs encoded by megasatellites are also found in *Kluyveromyces delphensis*, but not in more distantly related yeast species (nor in any other living organism), suggesting that they are specific to this branch of the hemiascomycetes, and may be involved in creating new gene functions in these yeast species.

MATERIALS AND METHODS

Analysis of the *C. glabrata* genome

The complete sequence of *C. glabrata* strain CBS138 was analyzed using the MREPS program (9), and the following parameters: minimal size of repeat unit (-minp) equal to 10, minimal repeat length (-minsize) equal to 30. Since the resolution parameter (allowing some degree of 'fuzziness' within the repeat) was set at the minimal value, variant repeats could not be detected. Therefore, repeats were individually examined and minisatellites manually extended 5′ and 3′ of the initial repeat detected by MREPS, as described previously (3).

In addition, some minisatellites, corresponding in fact to imperfect microsatellites (10), were detected by the program but not taken into account thereafter. Using this approach, MREPS detected 706 repeats fulfilling the required criteria. After careful examination, some of the repeats found by the program were partially overlapping or were part of the same minisatellite, resulting in a final number of 238 minisatellites used for the present analysis, including 145 detected in coding regions. Since several genes contain more than one repeat array, each of the 145 minisatellites was given a unique identifier, from MS#1 to MS#237. Compound minisatellites are also given a single identifier, followed by a letter for each motif of the minisatellite (e.g. MS#109a represents the 20×12 bp motif and MS#109b represents the 3×168 bp motif of compound MS#109).

Minisatellite size polymorphism was determined by standard PCR and Southern blot methods, using the CBS138 type strain, the laboratory BG2 strain (11), and two strains isolated form infected patients, F11017Blo1 and F15035Blo1, a kind gift of C. Hennequin (Muller,H. *et al.*, manuscript in preparation).

Search for orthologues

The functional annotation of the *C. glabrata* genome, developed during the course of the *Génolevures 2* project, was used [http://cbi.labri.fr/Genolevures; (8)]. Whenever several homologs were found in the *S. cerevisiae* genome, synteny data were used to discriminate among the possible genes. When synteny data were unsufficient to discriminate between two or three possible homologs of a *C. glabrata* gene, all of them were indicated (seven instances, Table 2). Many *C. glabrata* genes exhibit sequence similarities to several *S. cerevisiae* genes belonging to the FLO/STA superfamily of flocculins. These similarities were always limited in size and not sufficient to identify the right ortholog. Synteny data did not allow to discriminate among the possible homologues either. Sequence similarities to large DNA motifs in *K. delphensis* were searched with tblastx, using the motif itself (SHITT, SFFIT or TTITL) as a query, in a *K. delphensis* DNA database of 17 000 sequences (genome coverage $0.8\times$), provided by the Pasteur Genopole DNA sequencing facilities.

Amino acid composition and motif analysis

To determine the global composition of the 93 minisatellites with short motifs, all motifs were concatenated and calculation was performed using the DNA Strider 1.4f6 software (12). Long motifs were aligned using the ClustalW software on the BioWeb interface at the Pasteur Institute (http://bioweb.pasteur.fr/seqanal/interfaces/clustalw-simple.html). GC skews were calculated as $(G-C/G+C)$ or $(A-T)/(A+T)$, using DNA Strider. Both GC content and GC skew of minisatellite-containing genes were calculated on the gene DNA sequence without the minisatellite. Search for known domains in cell wall proteins was performed using the InterProScan software (http://www.ebi.ac.uk/InterProScan). The long motifs (SHITT, SFFIT, TITTL and the three unknown motifs) were used as queries for a Blast search into the NCBI non-redundant GeneBank CDS translations, PDB, Swissprot and PIR databases.

RESULTS

We performed a genome-wide search for minisatellites in the *C. glabrata* genome, using the MREPS software (9), set to the same parameters used previously for the *S. cerevisiae* genome (3) (see Materials and methods section). Given that the two yeast genomes have similar sizes and nucleotide composition [12.1 Mb for *S.cerevisiae*, 12.3 Mb for *C. glabrata*; (8)], a similar number of minisatellites was expected. Instead, a total of 145 minisatellites in 112 protein-coding genes and 93 minisatellites in intergenic regions were found in *C. glabrata*, compared to 55 in genes and 11 in intergenic regions in *S. cerevisiae*. Minisatellites in *C. glabrata* show no obvious bias for specific chromosomal locations (Figure 1).

Minisatellites are, on average, more GC-rich than the genes containing them, but no obvious GC skew was noted, in contrary to *S. cerevisiae* where minisatellites show more cytosines than guanines on the coding strand. In intergenic regions, minisatellites are shorter and contain less repeat units than in coding regions, like in *S. cerevisiae*. All these data are summarized in Table 1.

Unusual minisatellites scatter the *C. glabrata* genome

In addition to the presence of 93 'simple' minisatellites similar in size and composition to those discovered in *S. cerevisiae* (Tables 2, 3, 4, 5, minisatellites numbered from #1 to #93), the *C. glabrata* genome contains two peculiar types of minisatellites. First, 15 minisatellites are made of two different motifs (or even three, in one case) intermingled with each other (Tables 3, 5 minisatellites numbered from #101 to #115). In each case, the two

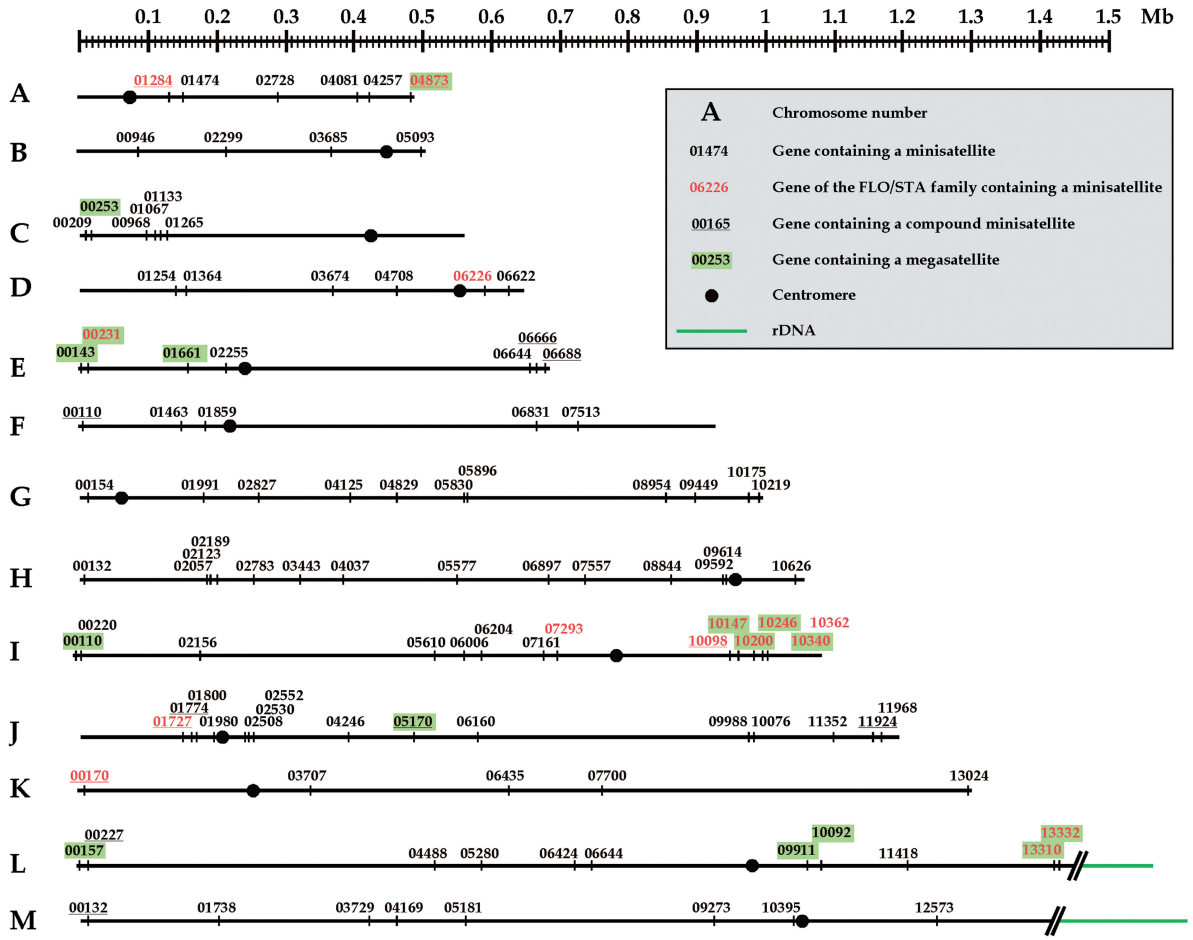

Figure 1. Distribution of minisatellites in the *C. glabrata* genome. Each chromosome is represented by a horizontal line, from the left to the right telomere. Vertical short lines represent the 109 minisatellite-containing genes and pseudogenes. Each gene starts with CAGL0 followed by the chromosome letter (A–M) then by the gene five-digit number and a final 'g' (38). Only the five-digit number is given here (e.g. 01284 on chromosome A stands for CAGL0A01284g). Note that some minisatellites may cumulate several properties, i.e. being a compound minisatellite with a long motif, in which case it is both underlined and colored. Size of the two rDNA arrays is not precisely known.

Table 1. Comparative distributions of minisatellites in the *S. cerevisiae* and *C. glabrata* genomes

Regions	Characteristics	*S. cerevisiae* (1)	*C. glabrata*
	Genome size (Mb)	12.1	12.3
	Total number of minisatellites (2)	66	238
Coding regions	Number of minisatellites	55	145
	Average GC% (3)	44	43
	Minisatellite GC%	48	51
	Minisatellite GC skew (4)	−0.11	0.00
Intergenic regions	Number of minisatellites	11	93
	Average GC%	29–36 (5)	ND
	Minisatellite GC%	29	46

(1) From Richard and Dujon (3).
(2) Excluding the 18 Y′ subtelomeric minisatellites.
(3) Calculated on minisatellite-containing genes only. The average for the complete set of genes is 39% for *S. cerevisiae* and 41% for *C. glabrata*.
(4) Calculated as the difference between the minisatellite GC skew and the gene GC skew (excluding the minisatellite sequence).
(5) GC% in intergenic regions varies between promoter-convergent and promoter-divergent regions.

motifs have different sizes and are repeated a different number of times, with no regular period, as if two decks of playing cards were shuffled with each other (two examples are shown in Figure 2). In 10 cases out of these 15 'compound minisatellites', the two motifs share a common sequence at their 5′-ends (L on Figure 2) but the 3′ ends (R on Figure 2) are different (5- and 3-ends are defined according to the coding DNA strand of the gene that contains the minisatellite) (Tables 3, 5).

The second type of peculiar minisatellites is made by those composed of unusually long motifs (from 126 to 429 bp) repeated from 3 to 32 times. Thirty-seven such minisatellites with long motifs were detected (numbered from #201 to #237), in addition to seven being part of a compound minisatellite, for a total of 44 minisatellites with long motifs. Five of them reach a total length >2 kb, the longest being 9.6 kb long (MS#214 in gene CAGLOI10147g), a length that no *S. cerevisiae* minisatellite reached. Given the unusual size of the repeated motif, these tandem repeats were named 'megasatellites' (Tables 3, 4). Both compound minisatellites and

Table 2. Simple minisatellites in *C. glabrata* genes

MS #	Gene Name	MS	Size (nt)	*S. cerevisiae* homologue	MS in S.c.	Domain (1)
1	A01474g	5 × 15	75	YGL028c (*SCW11*)	8 × 12	
2		18 × 9	162			
3	A02728g	3 × 18	54	YDR363wa (*SEM1*)		
4	A04081g	3 × 12	36	YLR194c		
5	A04257g	6 × 12	72	YBL054w		
6	B00946g	3 × 18	54	YCL028w (*RNQ1*)		
7	B02299g	3 × 12	36	YML114c (*TAF65*)		
8	C00209g (2)	11 × 18	198	YJR151c (*DAN4*)	30 × 18; 7 × 72	TM × 3; Serpaup
9	C00968g (2)	5 × 12	60	YOL155c (*HPF1*)	5 × 39	TM × 6
10		3 × 12	36			
11		42 × 12	504			
12	C01133g (2)	45 × 12	540	YOL155c (*HPF1*)	5 × 39	TM × 8; ABC
13	C01265g	3 × 15	45	YIL115c (*NUP159*)		
14	D01254g	3 × 12	36			
15	D01364g	4 × 12	48	YBR112c (*CYC8*)	3 × 18	
16	D03674g	4 × 9	36	YPL226w (*NEW1*)		
17	D04708g	4 × 9	36	YPR124w (*CTR1*)		
18		3 × 12	36			
19	D06226g	3 × 108	324	YAL063c (*FLO9*)	13 × 135	TM × 9
20	D06622g	4 × 15	60	YLL021w (*SPA2*)		
21	E02255g	3 × 12	36	YOL109w (*ZEO1*)		
22	EO6644g (EPA1)	4 × 120	480	YAR050w (*FLO1*)	10 × 135	TM × 1
23	F01463g	3 × 18	54	YOR010c (*TIR2*) or YER011w (*TIR1*)	5 × 33	
24	F01859g	3 × 12	36	YLR054c (*OSW2*)		
25	F06831g	4 × 18	72	YIR033w (*MGA2*) or YKL020c (*SPT23*)		
26	F07513g	3 × 12	36	YKL093w (*MBR1*)		
27	G00154g	6 × 12	72	YGR285c (*ZUO1*)		
28	G01991g	4 × 12	48	YOR056c (*NOB1*)		
29	G02827g	6 × 12	72	YIL105c (*SLM1*)		
30	G04125g	9 × 12	108	YJR004c (*SAG1*)		
31	G04829g	5 × 12	60	YML017w (*PSP2*)		
32	G05830g	5 × 54	270	YHR146w (*CRP1*) or YNL173c (*MDG1*)		
33	G05896g	4 × 12	48	YHR143w (*DSE2*)		
34	G08954g	6 × 12	72	YOL019w (*TOS7*)		
35	G09449g	3 × 12	36	YGR189c (*CRH1*)	5 × 24	
36	G10175g	5 × 30	150	YJR151c (*DAN4*)	30 × 18; 7 × 72	TM × 4; Serpaup
37	H02057g	5 × 30	150	YHR089c (*GAR1*)		
38	H02123g	9 × 12	108	YHR086w (*NAM8*)		
39	H02189g	13 × 9	117	YMR269w (*TMA23*)		
40	H03443g	3 × 18	54	YGL073w (*HSF1*)		
41	H04037g	3 × 12	36	YOR178c (*GAC1*) or YLR273c (*PIG1*)		
42	H05577g	9 × 18	162	YPL085w (*SEC16*)		
43	H06897g	3 × 12	36	YML098w (*TAF13*)		
44	H07557g	5 × 12	60	YGL254w (*FZF1*)		
45	H08844g	36 × 45	1,620	YMR173w (*DDR48*)	6 × 24; 4 × 24	
46	H09592g	3 × 18	54	YER011w (*TIR1*)	7 × 36	
47	H09614	3 × 18	54	YER011w (*TIR1*)	7 × 36	
48	I02156g	3 × 21	63	YHR161c (*YAP1801*)		
49	I05610g	5 × 12	60	YNR014w		
50	I06006g	4 × 9	36	YJL148w (*RPA34*)		
51	I06204g	4 × 57	228	YKL164c (*PIR1*) or YKL163w (*PIR3*)	8 × 57 or 6 × 54	
52	I07161g	4 × 12	48	YOR141c (*ARP8*)		
53	J01980g	7 × 12	84	YIR002c (*MPH1*)		
54	J02508g	9 × 15	135			TM × 3; Collagen; Antifreeze
55		11 × 18	198			
56	J02530g	9 × 15	135			TM × 1; Collagen
57		26 × 18	468			
58	J02552g	12 × 15	180			TM × 3; Collagen; Antifreeze; PRich × 3
59		11 × 18	198			
60		23 × 18	414			
61	J04246g	3 × 18	54	YMR234w (*RNH1*)		
62	J06160g	8 × 15	120	YNL166c (*BNI5*)		
63	J09988g	4 × 15	60	YNL063w (*MTQ1*)		
64	J10076g	3 × 15	45	YNL058c		
65	J11352g	7 × 12	84	YNL186w (*UBP10*)	7 × 12	

(continued)

Table 2. Continued

MS #	Gene Name	MS	Size (nt)	*S. cerevisiae* homologue	MS in S.c.	Domain (1)
66	J11968g(2) (*EPA15*)	27 × 12	324	YIR019c (*FLO11*)	5 × 30; 5 × 36	TM × 1; PT × 2; S-T Kin.; Plakin
67		17 × 24	408			
68	K03707g	4 × 12	48	YMR124w		
69	K06435g	6 × 12	72	YDR464w (*SPP41*)		
70	K07700g	9 × 27	243	YFL023w (*BUD27*)	6 × 30	
71	L04488g	4 × 15	60	YOR166c (*SWT1*)		
72	L05280g	3 × 15	45	YKL087c (*CYT2*)		
73	L06424g	9 × 27	243	YLR110c (*CCW12*) or YDR134c	3 × 12	
74	L06644g	6 × 12	72	YHR154w (*RTT107*)		
75	L11418g	4 × 21	84	YML071c (*COG8*)		
76	M01738g	3 × 18	54	YBR081c (*SPT7*)		
77	M03729g	3 × 18	54	YNL298w (*CLA4*)		
78	M04169g	4 × 12	48	YNL322c (*KRE1*)		
79	M05181g	4 × 15	60	YMR240c (*CUS1*)		
80	M09273g	3 × 15	45	YJR083c (*ACF4*)		
81	M10395g	3 × 39	117			TM × 3
82	M12573g	3 × 12	36	YIL061c (*SNP1*)		

C. g. gene names are abbreviated, only the chromosome letter and the five-digit number are given. Underlined *C. g.* names bear the signature of cell wall components or involved in cell wall metabolism (see text).
(1) As described by InterProScan. TM, transmembrane span domain. Given the high number of false positive predictions, proteins with only one TM span have a low probability of being transmembrane proteins [ref. (17)]; Serpaup, Seripauperin domain PT, short repeat domain composed on the tetrapeptide XPTX; ABC, ABC transporter type-1 domain; EGF, found in the extracellular domain of membrane-bound proteins or in proteins known to be secreted; Q6CXZ8, lipoprotein GPI-anchor membrane facilitator; α−β Hyd., domain found in the superfamily of α−β hydrolases; PA14, found in bacterial toxins, glucosidases and adhesins, probably involved in carbohydrate binding [ref. (18)]; β-lac., β-lactamase/transpeptidase domain; NADP, NADP-binding domain; Invasin, Invasin/intimin cell-adhesion domain; Ribo, ribosomal protein S14 domain; Collagen, member of the collagen superfamily, involved in connective tissue structure; Antifreeze, insect cysteine-rich antifreeze protein; PRich, highly glycosylated proline-rich cell wall proteins (extensins) in plants, probably involved in interactions with cell-wall carbohydrates [ref. (21)]; S-T Kin., Serine-Threonine kinase domain; Plakin, multiple repeats of beta(2)-alpha(2) motif, found in Ankyrin and Plakin repeats; GLNA, glutamine synthetase domain; CWP, cell wall peptidoglycan-anchor surface signal; Kelch, actin-interacting Kelch domain; ASP, aspartyl protease active site; Dynein, outer arm dynein light chain superfamily domain; GETHR, pentapeptide repeat of unknown function, mainly found in *C. elegans*.
(2) Overall quality of the sequence is not sufficient to determine the precise number of repeat units.

megasatellites are exclusively found in coding regions and their motif is always a multiple of three, raising the intriguing question of their formation in the genes containing them. Note that 12 out of 144 minisatellites are found in pseudogenes (Table 5), a much higher proportion than expected from random distribution, the *C. glabrata* genome containing ∼1% of pseudogenes, compared to active genes (I. Lafontaine and B. D., personal communication).

In order to estimate the degree of polymorphism found in such large arrays, we analyzed megasatellite sizes, by Southern blot hybridization of DNA extracted from three *C. glabrata* strains, isolated from infected patients (Muller,H. *et al.*, manuscript in preparation), and compared them to the same megasatellites in the sequenced strain (CBS138), used as a reference. For two megasatellites out of the three tested, we found polymorphism in at least one of the strains tested (data not shown). One strain shows a large increase in the MS#213 megasatellite size, corresponding to 7–8 additional 135-bp repeat units within gene CAGL0I07293g. In addition, gene CAGL0J05170g shows size increase in this strain, whereas in another strain it exhibits a size decrease. This gene contains three different megasatellites (MS#202, 203 and 109), and we did not determine which one is polymorphic (more than one may exhibit polymorphism). The last minisatellite tested (MS#224) did not show any clear polymorphism among the three strains tested as compared to the reference strain (data not shown).

In addition, we also compared the size of minisatellites found in the EPA gene family, between the CBS138 reference strain and the BG2 strain. The EPA gene family is composed of at least 15 members in the BG2 strain. These genes encode surface glycoproteins involved in cell–cell adhesion and pathogenicity. Eight of them (*EPA1* to *EPA8*) were sequenced in the BG2 strain (13,14). Among them, *EPA7* and *EPA8* do not contain minisatellites, the six other members containing simple minisatellites, as well as compound minisatellites and megasatellites. *EPA4* and *EPA5* were not found in the CBS138 strain, and *EPA6* does not contain any minisatellite in this strain. We, therefore, focused on the three remaining members (*EPA1* to *EPA3*), that contain minisatellites both in the CBS138 and in the BG2 strains. As shown in Figure 3, five out of six minisatellites found within these three genes exhibit polymorphism between the two strains sequenced. An additional megasatellite was detected in the *EPA3* gene, in the BG2 strain, that was absent from the CBS138 strain. This suggests that this megasatellite was inserted or deleted since the separation of the two strains. Note that outside of the regions containing tandem repeats, the *EPA3* genes in both strains show 99.8% identity, at the nucleotidic level. In a more specific analysis, Frieman and colleagues (15) showed that the number of repeat units of the 120 bp

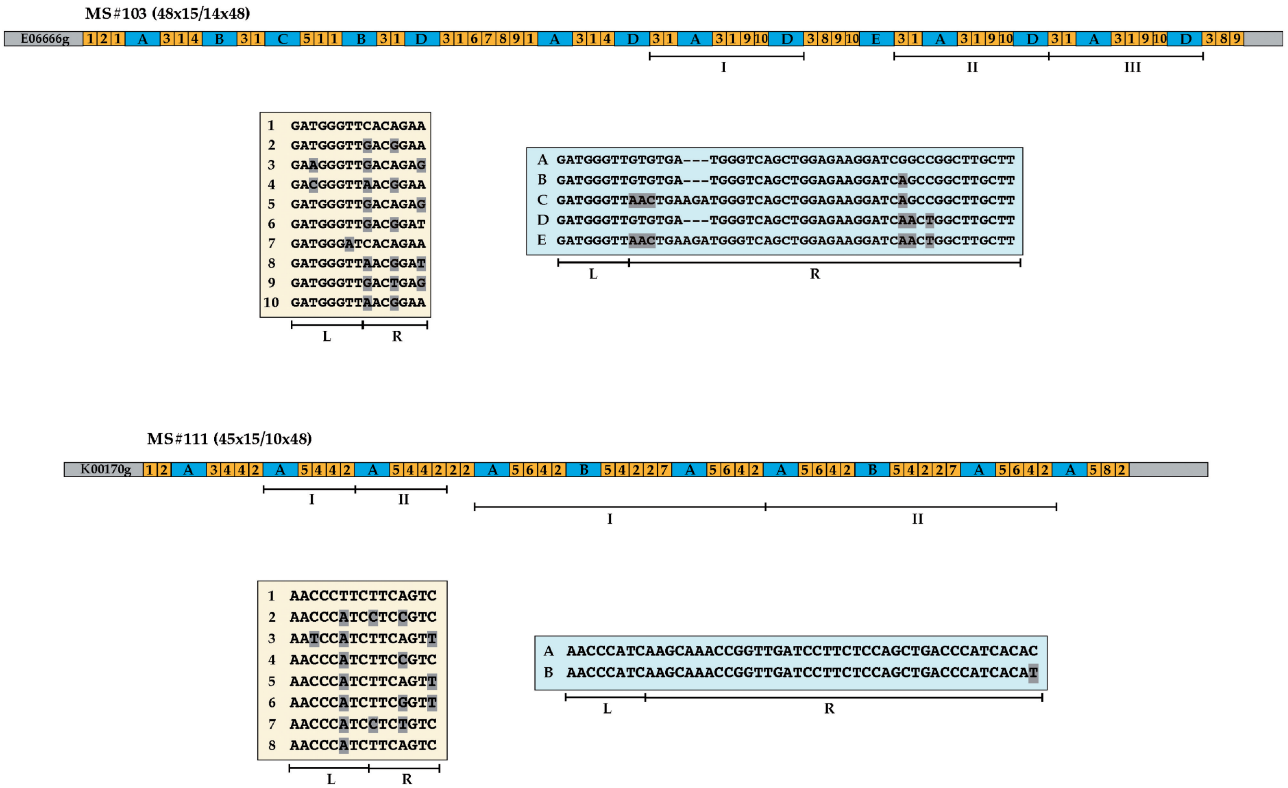

Figure 2. Two examples of compound minisatellites. Minisatellites are shown by color boxes, short motifs in yellow, long motifs in blue. Gray boxes represent partial 5′ and 3′ parts of gene coding sequences, along with gene names for which the first five characters have been ommited (see legend to Figure 1). Short motifs have been numbered from 1 to 10, motif 1 is used as the reference, and point mutations are shaded. Long motifs have been lettered from A to E, motif A is used as the reference, and point mutations are shaded. The 5′ part of each motif (L region) is common to both short and long motifs, whereas the 3′ part (R region) is different between short and long motifs. Duplicated blocks are roman numbered under each minisatellite. Note that for MS#111, the large duplicated block in the middle of the minisatellite contains several shorter internal duplications.

Table 3. Compound minisatellites in *C. glabrata* genes

MS#	Gene Name	MS	Size (nt)	*S. cerevisiae* homologue	MS in S.c.	Domain (1)	Motif (3)
101	A01284g (2)	5 × 27/3 × 30	135/90			TM × 4	
102	B03685g	4 × 15/3 × 18	60/54	YCR004c (*YCP4*) or YDR032c (*PST2*) or YBR052c			
83	E06666g (2)	13 × 42	546	YIR019c (*FLO11*)	5 × 30; 5 × 36	TM × 5; PRich × 7; PA14	
201	(*EPA2*)	4 × 168	672				TTITL
103		48 × 15/14 × 48	720/672				
84	E06688g (2)	10 × 42	420	YKR102w (*FLO10*)	3 × 81	TM × 5; PRich × 5; PA14	
104	(*EPA3*)	26 × 15/9 × 63	390/567				
85	I10098g	8 × 33	264	YIR019c (*FLO11*)	5 × 30; 5 × 36	TM × 10; PA14; PRich × 4; CWP; Kelch	
105		35 × 33/4 × 240	1155/912				-/SHITT-G
106	J01727g	47 × 15/13 × 24	705/312	YIR019c (*FLO11*)	5 × 30; 5 × 36	TM × 8	
86	J01774g (2)	5 × 24	120	YKL112w (*ABF1*)		Collagen	
107		32 × 24/60 × 48	768/2880				
108		6 × 33/2 × 258	198/420				-/SHITT-G
202	J05170g (2)	7 × 135	945				SHITT
203		4 × 270	1080				SFFIT degen.
109		20 × 12/3 × 168	240/504				-/SHITT
110	J11924g (2)	35 × 12/3 × 429	420/1287			TM × 2; ASP	-/SFFIT degen.
87	K00170g (2)	12 × 15	180			TM × 7; Dynein; GETHR × 9	
111		45 × 15/10 × 48	675/480				
204		5 × 168	840				TTITL
88		16 × 42	672				
89	L00227g (2)	39 × 39	1521			TM × 1	
112		31 × 75/44 × 45/4 × 243	2325/1980/924				-/-/SHITT-G
90	M00132g (2)	17 × 24	408	YIR019c (*FLO11*)	5 × 30; 5 × 36	TM × 1; PT × 2; Plakin	
113	(*EPA12*)	31 × 12/4 × 51	372/204				

C. g. gene names are abbreviated, only the chromosome letter and the five-digit number are given. Underlined *C. g.* names bear the signature of cell wall components or involved in cell wall metabolism (see text).
Please refer to Table 2 for cues (1–2).
(3) Only long motifs are indicated (see text for details).

Table 4. Megasatellites in *C. glabrata* genes

MS #	Gene Name	MS	Size (nt)	*S. cerevisiae* homologue	MS in S.c.	Domain (1)	Motif (3)
205	C00253g (2)	6 × 300	1800	YIR019 (*FLO11*)	5 × 30; 5 × 36	TM × 1	SFFIT
206	E00231g (2)	5 × 135	675				SHITT
207		3 × 300	900				SFFIT
208	E01661g	5 × 300	1500	YIR019 (*FLO11*)	5 × 30; 5 × 36	TM × 1	SFFIT
209	G10219g (2)	5 × 138	690	YHR211w (*FLO5*)	7 × 135; 3 × 21	EGF × 1	SHITT
210	H02783g	3 × 135	405	YJL076w (*NET1*)			unknown (4)
211	H10626g (2)	3 × 135	405	YAR050w (*FLO1*)	10 × 135	TM × 1	SHITT
212	I00220g (2)	4 × 177	708	YAR050w (*FLO1*)	10 × 135	TM × 1	SHITT degen.
213	I07293g (2)	16 × 135	2160				known (5)
214	I10147g (2)	32 × 300	9600	YHR211w (*FLO5*)	7 × 135; 3 × 21	PA14; β-lac.; NADP	SFFIT
215	I10246g (2)	5 × 300	1500			TM × 1	SFFIT
216	I10340g (2)	3 × 300	900	YAL063c (*FLO9*)	13 × 135	TM × 1	SFFIT
217	I10362g (2)	3 × 135	405			PA14; Invasin	SHITT
218		4 × 135	540				SHITT
219	J01800g (2)	27 × 12	324			TM × 6; Ribo	
220		4 × 135	540				unknown (4)
91	K13024g (2)	4 × 39	156	YIR019c (*FLO11*)	5 × 30; 5 × 36	TM × 8	
221		5 × 132	660				SHITT
222	L00157g (2)	11 × 141	1551	YAR050w (*FLO1*)		β-lac.; GLNA	SHITT
223		4 × 300	1200				SFFIT
224	L09911g (2)	5 × 300	1500				SFFIT
225	L13310g (2)	10 × 141	1410			PA14	SHITT
226	(*EPA11*)	7 × 300	2100				SFFIT
227	L13332g (2) (*EPA13*)	4 × 297	1188	YAL063c (*FLO9*)	13 × 135	TM × 1	SHITT-V
228	L10092g (2)	3 × 300	900			TM × 1; β-lac.	SFFIT
229		3 × 306	918				SHITT-V

C. g. gene names are abbreviated, only the chromosome letter and the five-digit number are given. Underlined *C. g.* names bear the signature of cell wall components or involved in cell wall metabolism (see text).
Please refer to Tables 2 and 3 for cues (1–3).
(4) No occurence of this motif was found in databases.
(5) Several occurences of genes containing this motif were found in *K. delphensis* (see text).

Table 5. Minisatellites in *C. glabrata* pseudogenes

MS#	Pseudogene Name	MS	Size (nt)	Coordinates (4)	Domain (1)	Motif (3)
230	A04873g (2)	4 × 300	1200	482 956–484 291	GLNA	SFFIT
114		20 × 141/10 × 309	2780/3090			SHITT/SHITT-V
231	B05093g (2)	11 × 135	1485	499 712–501 364	TM × 1	SHITT
92	C01067g (2)	17 × 12	204	104 419–106 401	TM × 9; PT × 2; ABC	
93		29 × 12	348			
232	E00143g (2)	6 × 141	846	4621–6420	EGF × 2	SHITT
233		4 × 300	1200			SFFIT
115	F00110g (2)	11 × 9/5 × 126	99/630	2275–2910	TM × 2; EGF	-/SHITT
234	H00132g (2)	4 × 129	516	4229–4837	TM × 2; Q6CXZ8	SHITT
235	I00110g (2)	3 × 141	423	2407–5280	TM × 2; α−β Hyd.	SHITT
236		9 × 300	2700			SFFIT
237	I10200g (2)	10 × 300	3000	992 434–998 401	PA14; TM × 1	SFFIT

C. g. gene names are abbreviated, only the chromosome letter and the five-digit number are given. Underlined *C. g.* names bear the signature of cell wall components or involved in cell wall metabolism (see text).
Please refer to Tables 2 and 3 for cues (1–3).
(4) Coordinates of beginning and end of the pseudogene, in nucleotides.

minisatellite found within the *EPA1* gene, varied from three to six (four repeat units are found in the CBS138 and in the BG2 strains), among a panel of 25 clinical isolates of *C. glabrata*. Additional experiments using PCR and Southern blot analyses to determine the size of minisatellites in EPA genes, confirmed that they exhibit size polymorphism among four different strains of *C. glabrata* (Muller,H. *et al.*, manuscript in preparation). We concluded that, like microsatellites and minisatellites in *S. cerevisiae*, several minisatellites exhibit size polymorphism in *C. glabrata*.

Proteins encoded by megasatellite-containing genes

There are 44 megasatellites, encoded by 33 different genes. Sixteen out of these 44 megasatellites share a common motif, that was called the SFFIT motif, conserved in all cases except two in which it is slightly degenerated (MS#203 and MS#110, Tables 3, 4, 5 and Figure 4). This 100 amino-acid SFFIT motif is conserved in 37 proteins in *Kluyveromyces delphensis*, a hemiascomycetous yeast closely related to *C. glabrata*. In these proteins, it is tandemly repeated, like in *C. glabrata*. This protein

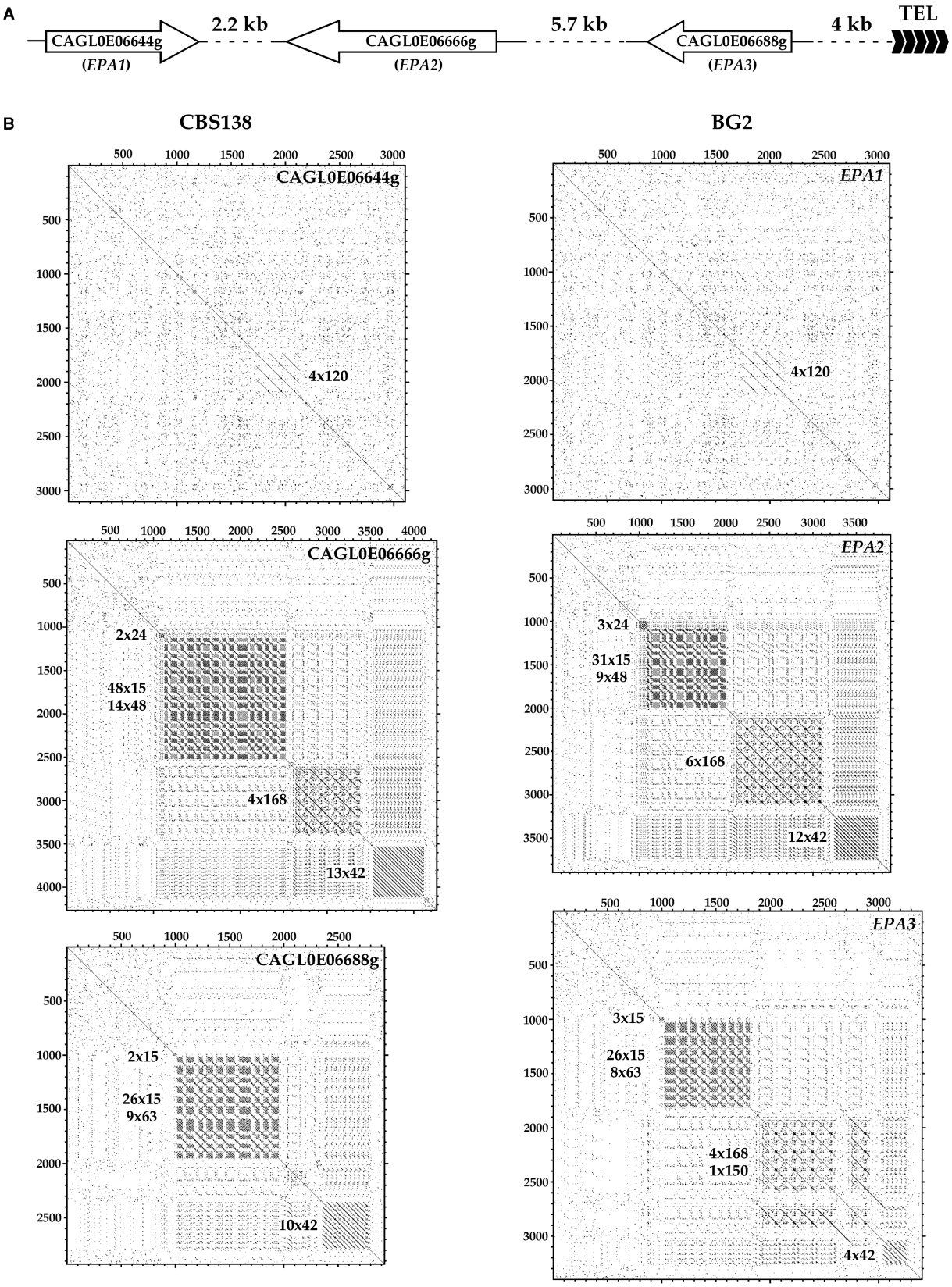

Figure 3. Comparison of minisatellites in EPA genes in two different *C. glabrata* strains. (**A**) Schematic representation of the *EPA1*, *EPA2* and *EPA3* genes, located on the right subtelomeric region of chromosome VI. Note that gene order and organization are identical in both the BG2 and the CBS138 strains. (**B**) Minisatellites in the three EPA genes show size polymorphism. DNA self-matrix of *EPA1*, *EPA2* and *EPA3* are shown for each of the two strains studied (BG2 and CBS138). Gene names are indicated in the right upper corner of each matrix. Number and size of each motif are shown next to each minisatellite. Note the additional compound minisatellite in *EPA3* in the BG2 strain. The smaller repeats (2×24 bp and 2×15 bp), not detected in the CBS138 strain due to the parameters chosen for the program (see Materials and methods section), are slightly expanded in the BG2 strain.

motif is threonine rich (20%), but also contains numerous serine (9%) and proline (8%) residues.

We identified a second long motif, contained in 16 megasatellites, that was called the SHITT motif. It is 45 amino acid long, rich in threonine (28%), serine (16%) and valine (14%) residues. In one megasatellite (MS#212), it is slightly degenerated, and in another one (MS#109b) it contains a small insertion of nine amino acids (Figure 4). Two variant forms of the SHITT motif are also found, one corresponds to an insertion of 60 amino acids in the middle of the motif and the other one to an insertion of 30 amino acids at the same position. The first one was called SHITT-V (three occurences: MS#114b, #227 and #229), and is rich in threonine (20%) and valine (16%), residues, and the second one was called SHITT-G (three occurences: MS#105b, #108b and #112c; Figure 4), and is rich in glycines (48%) and serines (23%). Seven occurences of the SHITT motif were detected in *K. delphensis*, tandemly repeated in proteins, but no other match was found in protein databases, proving that this motif, like the SFFIT motif, is specific to species closely related to *C. glabrata*.

Another long motif was found in two megasatellites (MS#201 and #204). It was called the TITTL motif, and is rich is threonine (26%), serine (16%) and aspartic acid, glycine and proline residues (9% each). It is found in two genes (*EPA2* and CAGL0K00170g) in the completely sequenced CBS138 strain of *C. glabrata*, and in addition, in two other genes (*EPA4* and *EPA5*), in the BG2 strain of *C. glabrata* (13,14). All these proteins are involved in cell–cell adhesion. Finally, three megasatellites contain a unique motif in *C. glabrata* (MS#210, #213, #220), one of them being homologous to a motif found in five proteins in *K. delphensis*, tandemly repeated form two to eight times (MS#213).

Proteins containing the SFFIT or any of the three SHITT motifs often show weak matches with the flocculin family of proteins in *S. cerevisiae* [FLO/STA superfamily; (16)], involved in flocculation and cell adhesion, but this similarity is restricted to the serine-rich repeated region. We have subsequently used the InterProScan software (see Materials and methods section) to look for possible conserved domains within genes that contain megasatellites. Several of the corresponding proteins contain putative transmembrane spans (TM in Tables 3, 4, 5), the signature of membrane-anchored proteins (17). Another frequent domain encountered is the PA14 domain found in bacterial toxins, glucosidases and adhesins (18), and shown to be involved in carbohydrate binding, both in *C. glabrata* (19) and in *S. cerevisiae* (20), making proteins containing this domain good candidates to interact with membrane glycoproteins. In *S. cerevisiae*, the PA14 domain is found in four proteins encoded by minisatellite-containing genes, belonging to the family of flocculins, directly involved in

Figure 4. Alignments of megasatellite motifs. The first motif of each megasatellite was aligned using ClustalW. The signature motif in each family (SFFIT, SHITT, TTITL) is shown in a light gray box to the left. The megasatellite number is indicated left to the sequence (MS#), followed by the number of repeat motifs within the megasatellite (in parentheses). The central part of the SHITT motif in which insertions occur (see text) is indicated by a dark gray box.

cell–cell adhesion: *FLO1*, *FLO5*, *FLO9* and *FLO10*. The PRich domain is found in highly glycosylated proline-rich cell wall proteins in plants, and is probably involved in interactions with cell wall carbohydrates (21). It is present in two proteins encoded by genes that contain compound minisatellites (CAGL0E06666g/*EPA2* and CAGL0I10098g; Table 3). *EPA2* encodes an adhesin responsible for cell–cell adhesion in *C. glabrata* (13,14). Altogether, 10 proteins encoded by megasatellite-containing genes, out of 33, show the signature of membrane proteins, many of them probably involved in interactions with glycoproteins.

InterProScan was also used to find structural domains in proteins that are not encoded by megasatellite-containing genes. Out of 79 such genes (Table 2), 12 contain at least three transmembrane spans and are therefore good candidates to be membrane proteins. Altogether, ~30–40% of minisatellite-containing genes are suspected to encode either cell wall components or proteins involved in cell wall formation, a much higher figure than expected if minisatellites were randomly distributed among *C. glabrata* genes.

SHITT motifs are well conserved among the different minisatellites in the N- and C-terminal parts of the motif, but insertions are found in the central region (Figure 4). The SHITT-G and SHITT-V motifs correspond to insertions of 90 and 180 nt (30 and 60 amino acids), respectively. At the DNA level, SHITT motifs are split into two parts, the sequence corresponding to the N-terminal region is very rich in cytosines (GC skew: −0.7) and adenosines (AT skew: +0.4), whereas the sequence encoding the C-terminal part does not show such biases. The central region, in which minisatellite insertions occur (MS#105b, 108b, 112c, 114b, 227 and 229), is also rich in cytosines and adenosines. This observation suggests that negative GC skews (and to a lesser extent positive AT skews) are a determinant favoring the insertion of new DNA sequences, a conclusion that was also reached for *S. cerevisiae* minisatellites (3). In comparison, the SFFIT motif does not exhibit any particular sequence bias, whereas the TTITL motif is almost as skewed as the SHITT motif (GC skew: −0.5, AT skew: +0.3).

The remaining 93 minisatellites contain shorter motifs (up to 120 nt), that do not belong to any of the families described above. The global amino acid composition of proteins encoded by these 93 minisatellites is given in Table 7. The most common amino acid found in such repeats is serine, followed by glycine, proline and asparagine. This is quite different from motif composition of *S. cerevisiae* minisatellites, in which, serine and threonine residues are the most frequent amino acids encountered, as in the *C. glabrata* SFFITT and SHITT megasatellites. In *S. cerevisiae*, serine- and threonine-rich repeats are thought to be the sites of *O*-glycosylations of cell wall proteins by the Pmt4 protein (22,23). It is therefore possible that in *C. glabrata*, proteins containing long-motif minisatellites are targets of similar posttranslational modifications and play a role at the cell wall surface, whereas short-motif minisatellites are involved in a variety of other cellular processes.

DISCUSSION

In the present work, we analyzed the distribution and composition of all minisatellites detected in the genome of the pathogenic yeast *C. glabrata*. Although similar in size to that of *S. cerevisiae*, the genome of *C. glabrata* exhibits a much larger number of minisatellites. The human genome was estimated to contain approximately 6000 minisatellites (≈2 minisatellites/Mb of sequences), whereas 6 and 7 minisatellites/Mb were found in *Arabidopsis thaliana* and in *Caenorhabditis elegans*, respectively (2,24). Similar figures were found in *S. cerevisiae* [9 minisatellites/Mb; (3)], but a larger number of minisatellites was found in the present study of the *C. glabrata* genome (19 minisatellites/Mb). Of particular interest are two new two types of minisatellites absent in *S. cerevisiae*: compound minisatellites, containing two different intermingled motifs, and megasatellites with long motifs (126–429 bp), that can be tandemly repeated up to 32 times. The latter are often encountered in genes whose products show signatures of cell wall proteins (Tables 3, 4, 5).

In contrast to microsatellites, that have been the subject of numerous studies in all sequenced organisms, there are very few reports in the litterature on minisatellite distribution in eukaryotic genomes. The genome of *Tetraodon nigroviridis*, extensively examined in search of such elements (25), revealed that minisatellites cover only 0.41% of the total sequence, compared to 0.7% in *C. glabrata*. In *T. negroviridis*, minisatellites are mainly located in two regions: a subtelocentric minisatellite (10 bp highly polymorphic motif) hybridizing on the short arm of 10 out of 11 subtelocentric chromosomes and a minisatellite with a 118 bp repeated motif, found at all centromeres. Except for these two minisatellites, found in very large arrays in the tetraodon genome, no minisatellite with a repeat motif size >200 bp was detected, nor any kind of tandem repeat resembling compound minisatellites or megasatellites.

Possible origin of *C. glabrata* minisatellites

One intriguing question is the origin of the numerous *C. glabrata* minisatellites. Are they *de novo* created, or are they propagated when the genes that contain them are duplicated? We classified minisatellites into families, based on their motif length and sequence. In total, 109 different motifs are found in 117 simple minisatellites (Table 6, top), showing that, most of the time, each motif is unique. Therefore, minisatellites do not propagate by duplicating a minisatellite-containing gene, but are probably *de novo* created in existing genes.

Megasatellites can be classified into defined families, even though their motif size exhibits some size variation (Table 6, bottom). We compared ten genes containing SFFIT megasatellites with each other, and found that only two of them (CAGL0L00157g and CAGL0E0 0231g), are similar in their 3′-end (35% identity at the nucleotidic level), and are therefore, most probably paralogues. The remaining genes do not show any significant similarity (besides the SFFIT motif itself), suggesting that these megasatellites are also, most of the time, *de novo* created in genes.

Table 6. Size distribution of *C. glabrata* minisatellites and megasatellites

	Minisatellites																				
Motif size	9	12	15	18	21	24	27	30	33	39	42	45	48	51	54	57	63	75	108	120	
Nb of occurrences	6	41	19	17	2	5	3	3	3	3	3	2	3	1	1	1	1	1	1	1	
Nb of families	5	38	18	15	2	5	3	3	3	3	2	2	3	1	1	1	1	1	1	1	
	Megasatellites																				
Motif size	126	129	132	135		138	141	168		177	240	243	258	270	297	300	306	309	429		
Nb of occurrences	1	1	1	9		1	5	3		1	1	1	1	1	1	14	1	1	1		
Nb of families	1	1	1	4	4	1	1	2	2	1	1	1	1	1	1	1	1	1	1		
Family (1)	H	H	H		H (2)	H	H		H (3)	H	G	G	G	S	V	S	V	V	S		

(1) Megasatellite family: H, SHITT; G, SHITT-G; V, SHITT-V; S, SFFIT; T, TTITL.
(2) Four families, including one SHITT (H) and three other unrelated familes
(3) Two families, including one SHITT (H) and one TTITL (T)

Table 7. Amino acids encoded by minisatellites and megasatellites

	Minisatellites			Megasatellites	
Amino acid	AA	*n* (%)	Amino acid	AA	*n* (%)
Serine	S	27.8	Threonine	T	20.5
Glycine	G	18.0	Serine	S	9.9
Proline	P	11.8	Aspartic acid	D	7.8
Asparagine	N	10.0	Valine	V	7.3
Alanine	A	6.8	Glycine	G	6.8
Threonine	T	3.8	Proline	P	6.7
Glutamic acid	E	3.8	Isoleucine	I	6.4
Valine	V	3.5	Glutamic acid	E	5.8
Aspartic acid	D	3.3	Alanine	A	4.7
Lysine	K	2.9	Asparagine	N	3.9
Glutamine	Q	2.6	Tyrosine	Y	3.6
Methionine	M	1.7	Leucine	L	3.4
Isoleucine	I	1.0	Phenylalanine	F	3.3
Leucine	L	<1.0	Lysine	K	3.3
Arginine	R	<1.0	Histidine	H	2.4
Histidine	H	<1.0	Arginine	R	1.6
Tyrosine	Y	<1.0	Tryptophane	W	1.1
Phenylalanine	F	<1.0	Glutamine	Q	<1.0
Cysteine	C	<1.0	Methionine	N	<1.0
Tryptophane	W	<1.0	Cysteine	C	<1.0

It was previously proposed that minisatellites result from replication slippage between two short DNA sequences located downstream and upstream of a central element (26). Almost all *S. cerevisiae* minisatellites exhibit such short repeated DNA sequences, consistent with this model (3). However, in *C. glabrata*, only half of the simple minisatellites show such short repeats, upstream and downstream of the minisatellite. When present, their mean size is 5 ± 0.8 nt, very similar to what was observed in *S. cerevisiae*. (5 ± 0.4 nt). The absence of such repeats in so many minisatellites in *C. glabrata* suggests that an additional mechanism may exist to create minisatellites in *C. glabrata*, or that these short repeats were subsequently erased by mutational decay in this yeast species.

Evolution of *C. glabrata* minisatellites

In the present study, only 15 of the 65 (23%) *S. cerevisiae* homologs to the *C. glabrata* genes containing simple minisatellites, also contain a minisatellite in *S. cerevisiae* (Table 2). It was previously reported (3), that out of 24 minisatellite-containing *S. cerevisiae* genes, only six of them (25%) also contain a minisatellite in *C. glabrata*, a similar proportion to what was found in the present study. Hence, minisatellites evolve faster than the genes containing them. It is interesting to note that among the 53 *S. cerevisiae* homologs that do not contain a minisatellite (Table 2), only six encode products that probably play a role in cell wall metabolism (YLR194c, *OSW2*, *SAG1*, *DSE1*, *BNI1* and *KRE1*). The others exhibit various functions, in all the known cellular compartments. This suggests that minisatellites in *C. glabrata* are found in a much wider variety of genes than in *S. cerevisiae*, in which they are mostly found in cell wall genes (3). This could be due to a higher flexibility of the *C. glabrata* genes to accomodate such tandem repeats, and underlines the fact that minisatellites may have a function in other genes besides cell wall genes.

The insertion of internal motifs into the SHITT motif itself (Figure 4), can be explained by two models, not mutually exclusive (Figure 5). A pre-existing gene already containing a minisatellite will be modified by the insertion of a second motif into one of the previous motifs, and subsequently either lost or propagated by intra-allelic gene conversion or replication slippage (Figure 5A) [for an in-depth review on gene conversion, see ref. (27)]. An

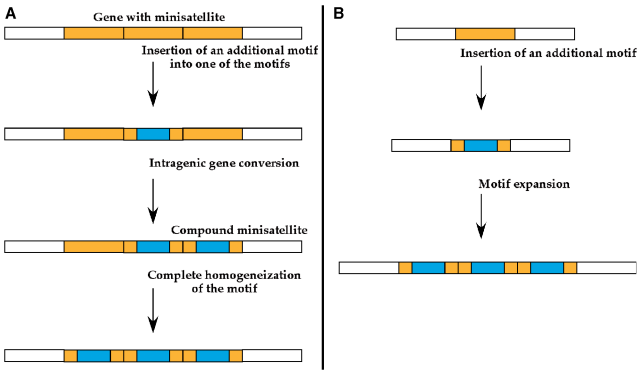

Figure 5. Insertion of a new motif within a minisatellite: two possible models. The motif may target a pre-existing minisatellite, and subsequently spread by intragenic gene conversion (**A**). Alternatively, the same motif may target a gene that does not contain a minisatellite, and is afterwards expanded in a minisatellite (**B**). Note that both models are not mutually exclusive, but only model A may lead to compound minisatellites.

alternative hypothesis supposes that one gene contains a nonrepeated motif. A new amino acid motif in inserted into it, and is subsequently amplified to give rise to a minisatellite (Figure 5B). Note that both hypotheses postulate that a short DNA sequence has the propensity to 'jump' into another DNA sequence, a property reminiscent of transposable elements (2,28). The same model may be used to explain the presence of the compound minisatellites, showing an irregular alternation of two different motifs (Figure 2), that would result from intermediate steps before complete homogeneization of the minisatellite (Figure 5A, bottom).

In *S. cerevisiae*, the *CUP1* locus is amplified under selection pressure, by intra-allelic gene conversion and unequal crossing-over between tandem repeats of the *CUP1* gene (29). Similarly, human minisatellites CEB1 and MS32 show high levels of inter- and intra-allelic gene conversions during meiosis and mitosis in *S. cerevisiae*, leading to complex reshuffling of repeat order and composition (30–32). Such mechanisms are also operating on human minisatellites, both during meiosis (33) and mitosis (34), and are probably also active in *C. glabrata*. Given that its genome contains significantly more unusual minisatellites than the *S. cerevisiae* genome, one can hypothesize that replication and/or recombination machineries have slightly different properties in each yeast species. *In silico* comparisons of the gene content of several hemiascomycetous yeasts showed that both replication and recombination machineries are very well conserved between *C. glabrata* and *S. cerevisiae*, exhibiting very few differences (35). However, the few differences found (like the presence of two copies of the *TOP1* gene and an extra truncated copy of the *SGS1* helicase in *C. glabrata*), might point to some specific properties of replication and/or recombination of the *C. glabrata* genome, that may explain the numerous peculiar minisatellites found there.

In a very recent analysis, Muller and colleagues (Muller,H. *et al.*, manuscript in preparation) showed that two deletions in two *C. glabrata* strains isolated from infected patients (F11017Blo1 and F15035Blo1), were located in close proximity to three megasatellites (MS#228/229 and MS#214, the largest megasatellite in the genome), suggesting that megasatellites may behave as fragile sites. Fragile sites are natural sites of chromosomal breakage in humans (36) and in yeast (37). It is therefore possible that due to the large repeated nature of megasatellites, spontaneous breakage occur during DNA replication at or near the megasatellite, giving rise to deletions around it (2).

ACKNOWLEDGEMENTS

We thank our colleagues of the Unité de Génétique Moléculaire des Levures for many fruitful discussions and C. Fairhead for careful reading of the article. We also thank the Génolevures consortium, particularly Tiphaine Martin for expert assistance with the Génolevures database. B.D. is a member of the Institut Universitaire de France.

FUNDING

Agence Nationale de la Recherche (ANR-05-BLAN-0331). Funding for open access charge: Agence Nationale de la Recherche.

Conflict of interest statement. None declared.

REFERENCES

1. Dujon,B. (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet.*, **22**, 375–387.
2. Richard,G.F., Kerrest,A. and Dujon,B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.*, In press.
3. Richard,G.-F. and Dujon,B. (2006) Molecular evolution of minisatellites in hemiascomycetous yeasts. *Mol. Biol. Evol.*, **23**, 189–202.
4. Bowen,S., Roberts,C. and Wheals,A.E. (2005) Patterns of polymorphism and divergence in stress-related yest proteins. *Yeast*, **22**, 659–668.
5. Verstrepen,K.J., Jansen,A., Lewitter,F. and Fink,G.R. (2005) Intragenic tandem repeats generate functional variability. *Nat. Genet.*, **37**, 986–990.
6. Fidalgo,M., Barrales,R.R., Ibeas,J.I. and Jimenez,J. (2006) Adaptive evolution by mutations in the FLO11 gene. *Proc. Natl Acad. Sci. USA*, **103**, 11228–11233.
7. Malpertuy,A., Dujon,B. and Richard,G.-F. (2003) Analysis of microsatellites in 13 hemiascomycetous yeast species: mechanisms involved in genome dynamics. *J. Mol. Evol.*, **56**, 730–741.
8. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., Marck,C., Neuveglise,C., Talla,E. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
9. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
10. Richard,G.-F. and Dujon,B. (1996) Distribution and variability of trinucleotide repeats in the genome of the yeast *Saccharomyces cerevisiae*. *Gene*, **174**, 165–174.
11. Cormack,B.P. and Falkow,S. (1999) Efficient homologous and illegitimate recombination in the opportunistic yeast pathogene *Candida glabrata*. *Genetics*, **151**, 979–987.
12. Marck,C. (1988) 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucleic Acids Res.*, **16**, 1829–1836.
13. Castano,I., Pan,S.-J., Zupancic,M., Hennequin,C., Dujon,B. and Cormack,B.P. (2005) Telomere length control and transcriptional

regulation of subtelomeric adhesins in Candida glabrata. *Mol. Microbiol.*, **55**, 1246–1258.
14. De Las Penas,A., Pan,S.J., Castano,I., Alder,J., Cregg,R. and Cormack,B.P. (2003) Virulence-related surface glycoproteins in the yeast pathogen Candida glabrata are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. *Genes Dev.*, **17**, 2245–2258.
15. Frieman,M.B., McCaffery,J.M. and Cormack,B.P. (2002) Modular domain structure in the *Candida glabrata* adhesin Epa1p, a β1,6 glucan-cross-linked cell wall protein. *Mol. Microbiol.*, **46**, 479–492.
16. Fabre,E., Muller,H., Therizols,P., Lafontaine,I., Dujon,B. and Fairhead,C. (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Mol. Biol. Evol.*, **22**, 856–873.
17. De Hertogh,B., Hancy,F., Goffeau,A. and Baret,P.V. (2006) Emergence of species-specific transporters during evolution of the hemiascomycete phylum. *Genetics*, **172**, 771–781.
18. Rigden,D., Mello,L.V. and Galperin,M.Y. (2004) The PA14 domain, a conserved all-ß domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends Biochem. Sci.*, **29**, 335–339.
19. Zupancic,M., Frieman,M.B., Smith,D., Alvarez,R.A., Cummings,R.D. and Cormack,B.P. (2008) Glycan microarray analysis of *Candida glabrata* adhesin ligand specificity. *Mol. Microbiol.*, **68**, 547–559.
20. Kobayashi,O., Hayashi,N., Kuroki,R. and Sone,H. (1998) Region of Flo1 proteins responsible for sugar recognition. *J. Bacteriol.*, **180**, 6503–6510.
21. Stiefel,V., Pérez-Grau,L., Albericio,F., Giralt,E., Ruiz-Avila,L., Ludevid,M.D. and Puigdomènech,P. (1988) Molecular cloning of cDNAs encoding a putative cell wall protein from Zea mays and immunological identification of related polypeptides. *Plant Mol. Biol.*, **11**, 483–493.
22. Ecker,M., Mrsa,V., Hagen,I., Deutzmann,R., Strahl,S. and Tanner,W. (2003) O-mannosylation precedes and potentially controls the N-glycosylation of a yeast cell wall glycoprotein. *EMBO Rep.*, **4**, 628–632.
23. Latgé,J.-P. and Calderone,R. (2005) In: Esser, K., and Fischer, R. (eds), *The Mycota XIII*. Springer, Berlin.
24. Vergnaud,G. and Denoeud,F. (2008) Minisatellites: mutability and genome architecture. *Genome Research*, **10**, 899–907.
25. Roest Crollius,H., Jaillon,O., Dasilva,C., Ozouf-Costaz,C., Fizames,C., Fischer,C., Bouneau,L., Billault,A., Quetier,F., Saurin,W. *et al.* (2000) Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.*, **10**, 939–949.
26. Haber,J.E. and Louis,E.J. (1998) Minisatellite origins in yeast and humans. *Genomics*, **48**, 132–135.
27. Pâques,F. and Haber,J.E. (1999) Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.*, **63**, 349–404.
28. Wicker,T., Sabot,F., Hua-Van,A., Bennetzen,J.L., Capy,P., Chalhoub,B., Flavell,A., Leroy,P., Morgante,M., Panaud,O. *et al.* (2007) A unified classification system for eukaryotic transposbale elments. *Nat. Rev. Genet.*, **8**, 973–982.
29. Welch,J.W., Maloney,D.H. and Fogel,S. (1990) Unequal crossing-over and gene conversion at the amplified *CUP1* locus of yeast. *Mol. Gen. Genet.*, **222**, 304–310.
30. Appelgren,H., Cederberg,H. and Rannug,U. (1997) Mutations at the human minisatellite MS32 integrated in yeast occur with high frequency in meiosis and involve complex recombination events. *Mol. Gen. Genet.*, **256**, 7–17.
31. Debrauwère,H., Buard,J., Tessier,J., Aubert,D., Vergnaud,G. and Nicolas,A. (1999) Meiotic instability of human minisatellite CEB1 in yeast requires double-strand breaks. *Nat. Genet.*, **23**, 367–371.
32. Lopes,J., Ribeyre,C. and Nicolas,A. (2006) Complex minisatellite rearrangements generated in the total or partial absence of Rad27/hFEN1 activity occur in a single generation and are Rad51 and Rad52 dependent. *Mol. Cell Biol.*, **26**, 6675–6689.
33. Jeffreys,A.J., Tamaki,K., McLeod,A., Monckton,D.G., Neil,D.L. and Armour,J.A.L. (1994) Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.*, **6**, 136–145.
34. Jeffreys,A.J. and Neumann,R. (1997) Somatic mutation processes at a human minisatellite. *Hum. Mol. Genet.*, **6**, 129–136.
35. Richard,G.-F., Kerrest,A., Lafontaine,I. and Dujon,B. (2005) Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. *Mol. Biol. Evol.*, **22**, 1011–1023.
36. Debacker,K. and Kooy,R.F. (2007) Fragile sites and human disease. *Hum. Mol. Genet.*, **16 (**Spec No. 2), R150–R158.
37. Zhang,H. and Freudenreich,C.H. (2007) An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in *S. cerevisiae*. *Mol. Cell*, **27**, 367–379.
38. Durrens,P. and Sherman,D.J. (2005) A systematic nomenclature of chromosomal elements for hemiascomycete yeasts. *Yeast*, **22**, 337–342.