DE NOVO GENOME ASSEMBLY AND ANALYSIS OF NON-ALLELIC RECOMBINATION IN PATHOGENIC YEAST *CANDIDA GLABRATA*

by

Zhuwei Xu

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

November 2019

© 2019 Zhuwei Xu

All Rights Reserved

Abstract

Candida glabrata is an opportunistic pathogen in humans, responsible for approximately 20% of disseminated candidiasis. *C. glabrata*'s ability to adhere to host tissue is mediated by GPI-anchored cell wall proteins (GPI-CWPs); the corresponding genes contain long tandem repeat regions and form large gene families. These tandem repeats cause mis-assemblies of GPI-CWP genes in *C. glabrata* genome. Subtelomeres of *C. glabrata* are particularly rich in GPI-CWP genes and share homology with each other. Consequently, the subtelomeres are mis-assembled in genome sequences assembled from short sequencing reads.

In this thesis, we used the long single-molecule real time (SMRT) reads and performed *de novo* genome assembly of the *C. glabrata* genome to establish the correct structure of GPI-CWP genes and the subtelomeres. We assembled the genome of six *C. glabrata* strains: the type strain, CBS138; our lab strain, BG2; four serial clinical isolates, BG3993-96 to assess genome changes during infection. With high quality sequences in hand, we then assess recombinational exchange between GPI-CWP genes by non-allelic mitotic recombination. This question is difficult to address with normal aligners, and we developed a k-mer based method to identify recombination.

Our assembly established the correct subtelomere structure of *Candida glabrata* and provides correct structure of the GPI-CWP gene families. Our analysis of the clinical isolates showed a very modest level of genetic change during the period of infection. Two of the four isolates are hyperadherent, and we identified a mutation in the gene encoding the transcription factor Yap6 as a likely candidate resulting in this phenotypic change. Our k-mer based method was applied genome wide to identify non allelic mitotic recombination events including in complex repeat regions. We documented a higher apparent recombination rate between subtelomeric genes and overall between GPI-CWP genes, independent of their location. In addition, we could document mitotic exchange between non-subtelomeric, non-GPI-CWP genes.

Thesis Advisor	Brendan Cormack
Thesis Reader	Michael Schatz

Acknowledgement

I am grateful to my parents, Zhongyuan Xu and Xiuhua Yang for their assistance and support during my graduate career. I am grateful to Jef Boeke in whose lab I worked for the first part of my graduate career, and for the opportunity to do research with him on synthesis of yeast chromosomes. I thank Brendan for his tremendous help and guidance. He welcomed me in my transition from benchwork to computational biology. He introduced me to the computational analysis of *C. glabrata* genome, and gave me an interesting and complicated problem to study - mitotic recombination events in the complex *C. glabrata* GPI-CWP genes. I would like to thank Cindy Rogers for all her help during my time in the Cormack lab.

I also would like to thank my thesis committee members Jeff Corden, Sarah Wheelan, and Michael Schatz for their guidance. Michael and Sarah are also co-mentors for my thesis, and made many invaluable intellectual contributions to my thesis work. I also thank Michael for reading my thesis, and I am grateful for his expeditious review to finish my thesis in time.

Finally, I would like to thank the BCMB program for giving me the chance to do research at Hopkins. I am particularly grateful to Carolyn Machamer for her help during and after my transition to the Cormack Lab. Lastly, I want to thank Arhonda Gogos, who is the heart of the program, for all her help during my time at Hopkins.

Table of Contents

Abstract ii

Acknowledgement iv

Table of Contents v

List of Tables vii

List of Figures viii

Chapter 1 Background and Significance 1

- Cell wall structure 3
- Structure and function of adhesin genes 4
- Adhesins in *Candida albicans* *5*
- Adhesins in *Saccharomyces cerevisiae* 8
- Adhesins in *Candida glabrata* 9
- Subtelomeres in *Candida glabrata* 11
- Subtelomeric silencing in *Saccharomycese cerevisiae* 12
- Subtelomeric silencing in *Candida glabrata* 13
- Genome assembly of *Candida glabrata* 13
- *De novo* genome assembly using long sequencing reads 15
- Reference 17

Chapter 2 *De novo* assembly of the *Candida glabrata* type strain reveals cell wall protein complement and structure of dispersed tandem repeat arrays 26

- Introduction 27
- Experimental Procedures 30
- Results 40
- Reference 56

Chapter 3 *De novo* assembly of *C. glabrata* BG2 strain 62

- Introduction 63
- Experimental Procedures 65
- Results 72
- Discussion 89
- Reference 92

Chapter 4 *De novo* assembly of four clinical serial isolates 96

- Introduction 97
- Experimental Procedures 99

Results 103

Discussion 114

Reference 120

Chapter 5 Non-allelic mitotic recombination in *Candida glabrata* 122

Introduction 123

Experimental Procedures 125

Results 139

Discussion 171

Reference 172

Conclusion 173

De novo assembly of the type strain CBS138 174

De novo assembly of strain BG2 176

De novo assembly of serial clinical isolates, BG3993-96 177

Analysis of non-allelic mitotic recombination in *C. glabrata* 178

Future directions 178

Reference 180

Appendices 181

Supplementary Tables 181

Supplementary Figures 183

Curriculum Vitae 185

List of Tables

Table 1 Metrics of Pacbio sequencing 116

Table 2 Chromosome assembly metrics 117

List of Figures

Figure 1	Illumina Read Coverage over the assembled genome	42
Figure 2	Structure of the *C. glabrata* subtelomeres	47
Figure 3	Gene Length Difference between two assemblies	51
Figure 4	Phylogenetic tree of GPI-anchored adhesins	55
Figure 5	Chromosome rearrangements	74
Figure 6	Illumina read coverage	76
Figure 7	Subtelomere structure	80
Figure 8	Gene comparison between BG2 and CBS138	83
Figure 9	Phylogenetic tree of CWP-GPIs	88
Figure 10	Chromosome rearrangements between BG3993 and BG2	105
Figure 11	Subtelomere structure of BG3993	107
Figure 12	Phylogenetic tree of GPI-CWPs	111
Figure 13	Diagram of k-mer based recombination analysis	126
Figure 14	Example of one conflict recombination event	133
Figure 15	Optimization of k-mer size in the first step	141
Figure 16	Optimization of k-mer size in the second step	142
Figure 17	Recombination of BG2 CAGL0A01284g (EPA10) relative to CBS138	144
Figure 18	Multiple sequence alignment of one recombination event	147
Figure 19	Structure of CAGL0B05061g ORF sequence in strain BG2 and CBS138	150
Figure 20	Recombination of BG2 CAGL0B05061g and BG2 CAGL0F09251g	151
Figure 21	Megablast alignment of CBS138 CAGL0B05061g to BG2 ORFs	154

Figure 22	Recombination of BG2 CAGL0J11968g (EPA15) and BG2 CAGL0M00132g (EPA12)	157
Figure 23	Genome-wide recombination map in BG2 relative to CBS138	161
Figure 24	Histogram of weighted average of bitscores	166
Figure 25	Ratio of recombined pairs to homologous pairs	169
Figure 26	Influence of distance between gene pairs in recombination	170

Chapter 1 Background and Significance

Candida glabrata is an opportunistic pathogen of humans that causes both superficial mucosal infection and severe disseminated infection (Gonçalves et al., 2016; Pappas et al., 2018). *C. glabrata* accounts for up to 30% of *Candida* bloodstream infections in hospitalized patients in Europe and the United States (Andes et al., 2016; Astvad et al., 2018; Chapman et al., 2017; Cleveland et al., 2015). The contributing factors of *C. glabrata* virulence are not completely understood, but secretion of hydrolytic enzymes, ability to evade phagocytic killing and adherence to host tissue are all thought to contribute to the virulence of *C. glabrata* (Kumar et al., 2019). Cell adherence is mediated by some of the *C. glabrata* surface proteins (Timmermans et al., 2018). The GPI-anchored cell wall proteins (GPI-CWPs) are the main class of surface proteins and the main class of proteins mediating adherence. The GPI-CWPs form large gene families, and the encoded ORFs are characterized by the presence of long tandem repeats. In *C. glabrata*, the repetitive nature of the GPI-CWPs, and the homology between members of GPI-CWP families, complicates genome assembly and has resulted in gross mis-assemblies of *C. glabrata* genome sequences. Since these mis-assemblies correspond to the GPI-CWP class of gene, a major impact of these errors is to complicate genetic investigation of the GPI-CWPs.

This thesis provides an approach to generate high-quality genome assembly of *C. glabrata* that permits further study of GPI-CWPs with the application of *de novo* genome assembly using long Single-Molecule Real-Time (SMRT) sequencing reads. In addition, using very high quality genome sequences generated for *C. glabrata*, this thesis studies whether the GPI-CWP encoding genes exchange information through mitotic recombination, and investigates general mitotic recombination events between *C. glabrata* strains. In this introductory chapter, I review the current knowledge of cell adherence, subtelomeric localization of the GPI-CWP encoding genes, transcriptional regulation of subtelomeric GPI-CWP genes, and the current status of the *C. glabrata* genome assembly.

Cell wall structure

The fungal cell wall is composed of proteins and polysaccharides. The cell wall has a general structure with an alkali-insoluble layer close to the plasma membrane, an alkali-soluble layer close to the outer face, and an outermost layer where most cell wall proteins are located. The alkali-insoluble layer is constituted of a complex of β1,3-glucans and chitin: β1,3-glucans are cross-linked to chitin, forming a complex to bind to other polysaccharides. (Latgé and Beauvais, 2014). In contrast, the alkali-soluble layer is composed of amorphous polysaccharides (Latgé and Beauvais, 2014). The polysaccharide components of the two layers vary amongst pathogens. For instance, β1,3- and β1,4-glucans and galactomannans bind to the glucan-chitin complex in *Aspergillus fumigatus*, while, β1-6 glucans are the primary component in *Candida albicans* (Latgé, 2010).

Cell wall proteins (CWPs) constitute the outermost layer of the cell wall. The GPI-anchored CWPs are covalently bound to β1-6 glucans through a remnant of the GPI-anchor; In addition, some cell wall proteins are proposed to be linked directly to β1,3-glucans, including Pir (proteins with internal repeats) proteins in *S. cerevisiae* and ASL-CWPs (alkali-sensitive-linkage cell wall proteins) in *C. albicans* (De Groot et al., 2005); Lastly, some CWPs are probably linked by disulfide bonds to other proteins covalently bound to the cell well because these CWPs are released using treatment with reducing agents (Klis et al., 2002).

C. glabrata encodes approximately 100a large number of GPI-CWPs, and the GPI-CWPs have been shown to mediate adherence to epithelial and endothelial cells of *C. glabrata*. (Cormack et al., 1999; Desai et al., 2011; Domergue et al., 2005; Frieman et al., 2002; Kaur et al., 2007). The major class of adhesins, encoded by the *EPA* (**ep**ithelial **a**dhesin) genes, are GPI-CWPs with a proposed important role in virulence. However, the correct number of *EPA* genes encoded in the genome, as well as the number and structure of other families of GPI-CWPs is unclear due to mis-assembly of these repetitive genes in current genome assemblies.

Structure and function of adhesin genes

I will start with a brief review of adhesins in higher eukaryotes. Canonical adhesins in metazoans maintain confluent layers of cells in the living tissue. They are usually located at cell junctions. Adhesins can interact with adhesins on the same cell, with cognate molecules on other cells, or with disparate ligands. The most common adhesins are cadherins. Cadherins participates in cell migration, differentiation and pathogenesis (Beavon, 2000). The adhesins form a large family of transmembrane glycoproteins which are widely expressed in multiple cell types among multiple tissues. One of the most widely conserved adhesins is E-cadherin. E-cadherin is expressed in epithelial cells, and it encodes five tandem repeat regions, EC1-EC5 (Beavon, 2000). N-cadherin is named by its function in neural tissue, and it is expressed in multiple tissues. N-cadherin has a similar structure with the EC1-5 tandem repeats (Langer et al., 2012). All the EC1-5 repeat domains are essential for cadherin mediated adhesion, and each EC domain encodes two calcium binding sites conserved during evolution. These cadherins can participate as the *cis*-acting element on the cell surface, and as the *trans*-acting element through binding (Beavon, 2000). In various types of tumor cells, alteration of binding patterns are associated with hypo- or hyper-glycosylation patterns on EC domains (Langer et al., 2012). The cadherins also modulate actin interactions and cytoskeletal remodeling through binding with α, β and/or γ-catenin. The cadherin-catenin interactions further regulate cell proliferation or internalization of cadherin-bound cells (Beavon, 2000).

Human pathogens can hijack the cadherin-catenin interactions using their adhesins. For instance, *Fusobacterium nucleatum* is an opportunistic pathogen associated with colorectal carcinoma. *F.*

nucleatum is reported to invade epithelial and endothelial cells. *F. nucleatum* utilizes the adhesin, FadA to modulate the E-cadherin/β-catenin signaling in colorectal cancer cells. The hijacking of cadherin-catenin interaction results in tumor proliferation (Rubinstein et al., 2013). Fungal pathogens also interact with N-cadherin, which contributes to their virulence. The *Candida albicans* hyphae is reported to interact with N-cadherin through the adhesin, Als3p. This interaction leads to endocytosis of *C. albicans* by endothelial cells, which is proposed to contribute to tissue invasion (Phan et al., 2005, 2007).

Fungal adhesins have a general structure of three domains: The N-terminal function domain faces the exterior of the cell and functions as a ligand binding region; the large central domain encodes highly glycosylated serine and/or threonine repeats, and it has a structural function to locate the N-terminal function domain outside the cell wall; the C-terminal domain encodes a GPI-anchor addition signal (Verstrepen and Klis, 2006). Fungal pathogens encode different classes of adhesins, probably for adaptation to various environmental conditions, and in some species encode large adhesin families. The exact complement of adhesins varies between species and can also vary between strains within a species. Since diploid species can encode additional allelic variations of adhesin genes, allelic differences also can in some cases contribute to the adhesin profile for strains within a species (Hoyer et al., 2008).

Adhesins in *Candida albicans*

Candida albicans is one of the most prevalent fungal pathogens. The best characterized family of adhesins in *C. albicans* are the agglutinin-like sequence (*ALS*) family. *ALS* genes are also encoded in *Candida dubliniensis, Candida tropicalis,* and *Candida parapsilosis* (Hoyer, 2001). *C. albicans* expresses at least 8 *ALS* genes, each encoding a large glycoprotein (de Groot et al., 2013). The *ALS* genes are located in multiple loci in the diploid yeast, which led to mis-assembly and mis-annotation in early studies. For instance, *ALS8* was re-annotated as an *ALS3* allele; *ALS10, ALS11,* and *ALS12* were mis-assembled hybrid chimeras of the real ORFs (Hoyer et al., 2008).

Als adhesins have the general fungal adhesin structure described in the previous section. In the corresponding ORF sequences, about 1300 nt corresponds to the N-terminal functional domain. The nucleotide sequence identity in the N-terminal domains ranges from 55% to 90% (Hoyer, 2001). Particularly, *ALS1, ALS5,* and *ALS3/8* share 85% identity, while *ALS7* shares 55% identity with other *ALS* genes. The Als N-terminal domain encodes an immunoglobulin-like region, which is important for host cell-ligand interactions (de Groot et al., 2013).

The central regions of *ALS* genes are made up of tandem repeats. The tandem repeats are constituted of 108 nt motifs. The *ALS* genes are classified into subcategories based on the tandem repeats. The tandem repeats of *ALS1-4* cross hybridize with *ALS1*; The tandem repeats of *ALS5-7* cross hybridize with *ALS5*; *ALS9* encodes a distinctive tandem repeat region (Hoyer, 2001).

ALS genes encode a variety of C-terminal domains. In common, all the C-terminal regions are serine/threonine rich. In addition, some *ALS* genes share similar C-terminal regions: *ALS2 and ALS4* share >95% sequence identity, and *ALS5* and *ALS9* are 93% identical. Notably, *ALS7* encodes a unique tandem repeat region, which contains the Val-Ala-Ser-Glu-Ser (VASES) motif (Hoyer, 2001). The *ALS* genes are located on multiple chromosomes, and the regulation of *ALS* genes is not reported to be related to genomic location. *ALS1, 2, 4, 5, 9* are located on chromosome 6; *ALS6* and *ALS7* are located on chromosome 3; *ALS3* and *ALS8* are located on chromosome R (Hoyer, 2001). Interestingly, *ALS* genes are relatively close to chromosome ends:. *ALS2* and *ALS4* are located within 100 kb from the telomere ends of chromosome 6, and other *ALS* genes on chromosome 6 are located within 250 kb from the ends. *ALS6* and *ALS7* are located within 400 kb from the end of chromosome 3. *ALS3/8* locus is centromere proximal on chromosome R (according to the *C. albicans* reference genome Candida Genome Database (http://www.candidagenome.org)).

The *ALS* genes are reported to be regulated by nutrient conditions, morphological form, and growth stage in *C. albicans*. *ALS* genes are transcribed constitutively in *C. dubliniensis*, suggesting different regulation of the *ALS* genes between *Candida* species. Orthologs of the *ALS* genes between *C. albicans* strains have sequence variation. For instance, the ORF length of the *ALS1* gene changes between strains due to changes in the copy number of tandem repeats; sequence polymorphisms are also reported in *ALS5* ORF between strains (Hoyer, 2001).

ALS genes have various functions in cell adherence. Als1-4p*,* Als9p are crucial in binding to endothelial cells. In addition, Als1p and Als3p can regulate the adherence to epithelial cells (de Groot et al., 2013). Remarkably, deletion mutants of Als5-7p are hyper-adherent to epithelial and endothelial cells. According to the transcription data, subsets of *ALS* genes have concurrent up- and down- regulation in response to specific extracellular signals (Hoyer et al., 2008). Als1p, Als3p, Als5p and Als9p bind to extracellular matrix components. All *ALS* genes impact the binding to abiotic surfaces, such as glass and plastics (de Groot et al., 2013).

Hyphal wall proteins (Hwps) are the second major family of adhesins in *C. albicans*. These are mannoproteins which are only expressed during germ tube and hyphal stages (de Groot et al., 2013; Staab et al., 1999). These two morphologies are important factors of *C. albicans* virulence. Hwp1p is the substrate of host cell-expressed transglutaminase. The transglutaminase catalyzes the covalent bond between the glutamines in the glutamine-rich region of Hwp1p and the substrate on the host buccal epithelium (Staab et al., 1999). The ECM proteins bound to the epithelial cells are consequently covalently bound to Hwp1p (de Groot et al., 2013; Ponniah et al., 2007).

EAP1 is an *HWP1* related gene, classified as such based on a conserved 42 amino acid domain in the putative effector region. This domain influences formation of amyloid-like patches, which is speculated to reflect adhesin oligomerization (de Groot et al., 2013). *Eap1* regulates cell adherence to yeast cells, epithelial cells and polystyrene as well as invasive growth on agar (Li and Palecek, 2008). *HWP* genes also mediate adherence to bacteria. For example, both Hwp1p and Eap1p mediate adherence to *Streptococcus gordonii*, which colonizes the oral cavity (de Groot et al., 2013; Nobbs et al., 2010).

Adhesins in *Saccharomyces cerevisiae*

Candida glabrata, even though it is classified as a *Candida* species, is more closely related to *S. cerevisiae* rather than to other pathogenic Candida species, such as *Candida albicans* or *Candida tropicalis*. Therefore, adhesins in *S. cerevisiae* also provide insights of the structure and function of adhesins in *C. glabrata*. The largest adhesin family in *S. cerevisiae* is the *FLO* family. The *FLO* genes regulate cell adherence and flocculation which is relevant in industrial fermentation. The *FLO* genes are essential for yeasts to form visible “flocs”, which are aggregates of thousands of cells. Differences in flocs formation between lager and ale strains exist such that lager strains form flocs that sediment to the bottom of the fermentation vats while flocs of ale strains float to the surface (Verstrepen and Klis, 2006). These differences depend on differences in *FLO* gene adhesins complement and expression. There are five formally classified *FLO* genes in *S. cerevisiae*, and the *FLO genes* are highly similar in sequence. *FLO1, 5, 9, 10* mediate cell-cell adhesion, or flocculation; *FLO11* binds to other substrates involved in cell-cell interaction during pseudohyphal growth. There are two additional members of the *FLO* family: *FIG2 and AGA1. FIG2* and *AGA1* share the same structure as *FLO11*. They carry out their adhesion function in mating type switching. Aga1p is covalently bound to soluble Aga2p on the surface of MATa cells, which forms a complex interacting with Sag1p on MATα cells to facilitate adherence (Zhao et al., 2001). Fig2p is the paralog of Aga1p, and it is essential for mating cell integrity during the mating process. It localizes to the region of the mating projection, and probably modulates and organizes local cell wall structure (Zhang et al., 2002).

Adhesins in *Candida glabrata*

The first class of adhesins described in *C. glabrata* were encoded by the *EPA* gene family. *EPA1* was the first discovered member, identified through a forward genetic screen based on adherence to an epithelial cell line (Cormack et al., 1999). Epa1 is a lectin, and requires calcium for adherence. The ligands of *EPA1* are N-acetyllactosamine or N-acetyllactosamine-containing glycoconjugates. (Cormack et al., 1999). Many additional *EPA* genes were discovered in subsequent studies. The *EPA* genes form clusters in the subtelomeric regions of *C. glabrata*: *EPA1, EPA2,* and *EPA3* are located in a 24 kb cluster close to the telomere; *EPA4* and *EPA5* are located together, 4 kb from a different telomere (De Las Peñas et al., 2003). The subtelomeric location of these genes make these *EPA* genes subject to subtelomeric silencing. Deletion of either *EPA* cluster results in 3-5 fold modest decrease in kidney colonization in the murine infection model. (De Las Peñas et al., 2003).

The genome assembly of the *C. glabrata* type strain, ATCC2001 (CBS138) was first carried out by the Dujon lab *(Dujon et al., 2004)*. There were additional *EPA1*-like sequences in the assembled genome; however, whether all these sequences were full length ORFs of *EPA* genes was unclear in the assembly (Castaño et al., 2005). Two additional *EPA1*-like sequences, *EPA6* and *EPA7, (Castaño et al., 2005; Iraqui et al., 2005)* were identified in a forward genetic screen for hyperadherent mutants (Castaño et al., 2005). *EPA6 and EPA7* are also located in the subtelomere and are subject to subtelomeric silencing together with *EPA1-5*. Epa6p was also classified as the major adhesin important in biofilm formation (Iraqui et al., 2005).

In addition to the *EPA* or *EPA*-like sequences identified in CBS138, there were 23 *EPA* or *EPA*-like sequences identified by sequence homology in another *C. glabrata* clinical isolate, BG2. If these 23 sequences are classified using the sequence homology of the non-repeat containing N-terminal domains, 17 of the 23 sequences are shared with the CBS138 strain, and 6 sequences are specific to BG2 relative to CBS138 (Kaur et al., 2005).

Adhesin ligand specificity in *Candida glabrata*

Epa1p, Epa6p and Epa7p bind to different glycan moieties (Zupancic et al., 2008). The function of the three *EPA* genes was studied by expression in *S. cerevisiae* due to the high redundancy of the *EPA* genes in *C. glabrata*. *S. cerevisiae* is a non-adherent yeast, therefore adherence conferred by *EPA* genes can be monitored by expression in *S. cerevisiae*. The N-terminal lectin domains of the three proteins were expressed in the *S. cerevisiae* cell wall. These three Epa lectins were shown to prefer particular terminal disaccharides, with the preferred disaccharides differing in terminal sugar identity and linkage to the penultimate residue. Epa6p has the broadest binding spectrum recognizing galactose bound through α or β1-3 and β1-4 glycosidic linkages with glucose, or N-acetyl glucosamine. Epa1p and Epa7p do not bind to α-linked moieties. Epa7p has a preferentially binding to Gal β1-4-Glc (lactose), Gal β1-3-Gal, and N-acetylated derivatives (Zupancic et al., 2008).

Epa6p and Epa7p bind to different in glycan moieties; however, the two genes are highly similar with 92% sequence identity in protein sequence. Domain swapping and sugar inhibition experiments identified two hypervariable regions in the two genes resulting in the change in binding specificities (Zupancic et al., 2008). The *EPA* genes are members of a larger family of lectins - PA14 containing genes. The PA14 domain was initially described in the anthrax toxin protective antigen, but is a widely distributed domain found in fungal and bacterial proteins, including the *FLO* genes in *S. cerevisiae* (Rigden et al., 2004).

The second adhesin family in *C. glabrata* is the PA14 domain-containing wall proteins (*PWPs*). There are 7 *PWP* genes in *C. glabrata*, related to the *EPA* genes, but forming a distinct grouping. They are initially identified from the *in silico* screen for GPI-anchored cell wall proteins (Weig et al., 2004). The *PWP*s form a distinct family of adhesin from the phylogenetic tree of the N-terminus functional domains (de Groot et al., 2008). The deletion of *PWP7* is reported to modestly decrease adherence to human umbilical vein endothelial cells (Desai et al., 2011).

The third proposed adhesin family includes the Adhesin-like wall protein (AWPs). This family was first identified by four family members from a tandem mass spectrometry experiment (de Groot et al., 2008). There are three additional members discovered in later studies (Kraneveld et al., 2011). Further sequence analysis separates the *AWPs* into two subfamilies. *AWP1* and *AWP3* are in one subfamily, while *AWP2* and *AWP4* are in other. Both subfamilies contain additional members in subsequent studies (de Groot et al., 2008, 2013). *AWP5,* with the alias of *AED1*, has a similar function with *PWP7*, and deletion reportedly leads to decreased adherence to human umbilical vein endothelial cells (Desai et al., 2011). Awp2p, 4p, 5p, 6p are expressed in biofilms. (Kraneveld et al., 2011).

Subtelomeres in *Candida glabrata*

Prior to this thesis, there are no high-quality subtelomeric genome sequences published for *C. glabrata*. The subtelomeres of *C. glabrata* encodes many GPI-CWPs, including the *EPA* and *AWP* genes introduced in previous sections (Candida Genome Database (http://www.candidagenome.org). However, the subtelomeres are highly repetitive due to the tandem-repeats encoded by the GPI-CWPs; these repeat regions are highly similar and shared between different proteins. Consequently, in existing genome sequences of *C. glabrata*, there are many mis-assemblies of the subtelomeres leading to mis-assembled ORFs corresponding to GPI-CWPs. One major contribution of this thesis is that we generated a high-quality subtelomeric structure with high-quality sequences of the ORFs to document the complete number and structure of the GPI-CWPs in *C. glabrata*.

Subtelomeric silencing in *Saccharomycese cerevisiae*

In *S. cerevisiae,* Rap1 binds to consensus telomeric repeats to initiate silencing (Rusche et al., 2003). The silent information regulator (Sir) proteins are recruited by Rap1. The Sir proteins spread from the telomeric repeats into the subtelomere and repress transcription. Sir2 is a histone deacetylase and it regulates the deacetylation of the N-termini of histone 3 and histone 4. The deacetylation permits the binding of Sir3 and Sir4 to the histones. The Rap1 and Sir proteins form a complex which inhibits transcription. This silencing mechanism also regulates the silencing of mating loci and a similar mechanism involving Sir2 inhibits transcription of the rDNA arrays (Rusche et al., 2003). Rap1 interacts with two additional proteins, Rif1 and Rif2. The two proteins are negative regulators of telomere length and impact subtelomeric silencing; the deletion mutants of Rif1 and Rif2 have increased telomeric repeats and increased subtelomeric silencing in *S. cerevisiae* (Rusche et al., 2003). yKu70 and yKu80 form a heterodimer that binds to telomeric ends and is required to maintain telomere length. Mutants have short telomeres and reduced subtelomeric silencing. (Mishra and Shore, 1999).

Subtelomeric silencing in *Candida glabrata*

C. glabrata employs a subtelomeric silencing mechanism similar to that in *S. cerevisiae.*. The *SIR* genes as well as the *RAP1*, *RIF1,* and the *yKu* genes are required for subtelomeric silencing, and the *SIR* complex can extend > 20 kb from telomere towards the centromere (De Las Peñas et al., 2003; Domergue et al., 2005; Rosas-Hernández et al., 2008). Different precise sets of proteins are required for subtelomeric silencing in different subtelomeres. For instance, *yKU70* and *yKU*80 are apparently not required for silencing in the ChrE right subtelomere, which encodes the *EPA1, EPA2* and *EPA3* cluster, possibly because there is a *cis*-acting protosilencer (Sil2126) between *EPA3* and the telomere repeats, which carries out *yKu70* and *yKu*80 independent silencing (Juárez-Reyes et al., 2012). Since silencing requires the NAD+ dependent histone deacetylase Sir2, cellular NAD+ concentrations impact silencing of the *EPA* genes. *C. glabrata* is an NAD+ auxotroph and requires vitamin precursors of NAD+ to grow. Limitation of niacin results in transcriptional derepression of the normally repressed *EPA* genes in the subtelomeres; NAD+ limitation occurs during urinary tract infection (Domergue et al., 2005) with the resulting derepression of *EPA* gene transcription..

Genome assembly of *Candida glabrata*

The type strain of *C. glabrata* is ATCC2001 (CBS138). The CBS138 reference genome was assembled using shotgun sequencing from 3-5 kb plasmid libraries with Sanger sequencing, and finished by Bacterial Artificial Chromosome (BAC) end sequencing (Dujon et al., 2004). Dujon *et al.* generated a high-quality reference genome with overall high accuracy. However, shotgun sequencing cannot correctly assemble the long tandem-repeats, and the long tandem repeats also lead to systematic mis-assemblies. Dujon *et.al* used Sanger sequencing for the original C. glabrata genome assembly. The effective length of sequences is about 1kb. Many tandem repeat regions, composed of repeats with very few or no SNVs are much longer than this read length, and consequently, assembly cannot "span" the repeat region. This results in two problems. First, while the repeat unit is known, the number of units in the array cannot be discovered. Second, since different GPI-CWP genes share tandem repeat arrays of the same sequence, the unique sequences flanking tandem repeat arrays can be mis-assembled to the wrong array. Consequently, the GPI-CWPs, which encode the long tandem repeats, and the subtelomeres, which encode many GPI-CWPs, are both systematically mis-assembled in the reference genome. More recent genome sequencing of clinical isolates has generally used short read sequencing libraries from PCR-based whole genome amplification (WGA) (Barber et al., 2019; Carreté et al., 2019; Guo et al., 2019). PCR-based WGA can result in sequencing error or bias (Beerenwinkel et al., 2012), and lead to an overestimation of SNPs. Furthermore, it is difficult to detect large structural variants or to obtain the correct structure of long tandem repeat regions using short read sequencing.

Amplification-free long read single molecule sequencing technologies are therefore one solution to obtain genome assemblies with high accuracy and correct genome structure. Single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) is one of the most widely used long read approaches (Eid et al., 2009). SMRT sequencing monitors the incorporation events of dNTPs using a nanophotonic structure, the zero-mode waveguide (ZMW) which can reduce the volume of observation by more than three orders of magnitude relative to confocal fluorescence microscopy (Levene et al., 2003). Therefore, the incorporating dNTP-fluorophore can be detected against the bulk solution. Each ZMW is only occupied by ~0.01 to 1 molecule on average, which permits single-molecule sequencing (Foquet et al., 2008; Levene et al., 2003). The ZMW employes special surface chemistry that the polymerase is only immobilized to the bottom of the

ZMW rather than the side walls. The dNTPs are phospholinked to the fluorophores at their terminal phosphate moiety to generate natural DNA molecule after incorporation and to permit consecutive real-time sequencing (Eid et al., 2009).

During SMRT sequencing, the DNA sequence is determined by detection of fluorophores. ZMW provides excitation confinement in the zeptoliter (10^{-21} liter) regime, permitting observations of single dNTP incorporations. The template DNA is a single-stranded circular DNA, and each template is consecutively sequenced multiple times from rolling circle amplification, generating long polymerase sequencing reads (Eid et al., 2009). Concatenated reads are separated into reads which contains a single read of the template DNA which are called as subreads. The consensus sequence of the subreads from the same template DNA molecule are called circular consensus reads. SMRT sequencing generates long reads without context bias (Chaisson et al., 2015; Quail et al., 2012; Ross et al., 2013) and allows efficient and highly accurate *de novo* assembly, including assembly of the long tandem repeat regions. Vale-Silva et al. have applied this technology to assemble genome sequence of two serial clinical *C. glabrata* isolates (Vale-Silva et al., 2017).

De novo **genome assembly using long sequencing reads**

Long SMRT sequencing reads have low context bias, and they can address long tandem repeats that cannot be resolved by short read sequencing. However, they have relatively high error rates compared to short read sequencing. The assembly of the single-molecule long sequencing reads require sensitive alignment methods of the reads to purge sequencing errors while maintaining the discrimination of divergent alleles and nonexact repeats. There are three approaches to assemble the long reads: 1) Hybrid methods use both long and short read sequencing together. Long reads reconstruct the long-range structure, and short reads are applied for accurate base calling (Hackl et al., 2014; Koren et al., 2012; Lee et al., 2014; Salmela and Rivals, 2014; Ye et al., 2016). 2) Hierarchy methods only use the long read sequencing. These methods improve the quality of the single-molecule sequencing reads prior to assembly using multiple rounds of read sequencing alignment and correction (Chin et al., 2013; Koren et al., 2013, 2017). 3) Direct methods assemble the single-molecule long sequencing reads using overlaps without any *a priori* corrections (Kolmogorov et al., 2019; Li, 2016; Tørresen et al., 2017).

Canu is a *de novo* sequence assembler designed for the noisy single-molecule long sequencing reads (Koren et al., 2017). It is a successor of the Celera Assembler (Miller et al., 2008; Myers et al., 2000). Canu employs a hierarchy method. The assembly is performed in three stages: correction, trimming, and assembly. In all the three stages, the assembler performs an all-to-all overlap of the reads. The correction stage applies the best overlaps to correct sequencing errors; The trimming stage identifies regions which are not supported by overlap with other reads, and subsequently trims or separates reads to generate high quality reads, which are fully supported by overlap with other reads; the assembly stage performs a final correction for sequencing errors, and generates the assembled contigs. The overlaps between the trimmed reads are used for the final correction of the trimmed reads. An assembly graph is established based on the overlaps and the reads are assembled into contigs based on this graph.

We show in this thesis that PacBio generated long sequencing reads are highly useful in generating complete genome assemblies for *Candida glabrata*, and can resolve structure for even the longest tandem repeat arrays in *C. glabrata*, thereby providing the correct ORF sequences of these complex adhesins for the first time.

Reference

Andes, D.R., Safdar, N., Baddley, J.W., Alexander, B., Brumble, L., Freifeld, A., Hadley, S., Herwaldt, L., Kauffman, C., Lyon, G.M., et al. (2016). The epidemiology and outcomes of invasive Candida infections among organ transplant recipients in the United States: results of the Transplant-Associated Infection Surveillance Network (TRANSNET). Transpl. Infect. Dis. *18*, 921–931.

Astvad, K.M.T., Johansen, H.K., Røder, B.L., Rosenvinge, F.S., Knudsen, J.D., Lemming, L., Schønheyder, H.C., Hare, R.K., Kristensen, L., Nielsen, L., et al. (2018). Update from a 12-Year Nationwide Fungemia Surveillance: Increasing Intrinsic and Acquired Resistance Causes Concern. J. Clin. Microbiol. *56*.

Barber, A.E., Weber, M., Kaerger, K., Linde, J., Gölz, H., Duerschmied, D., Markert, A., Guthke, R., Walther, G., and Kurzai, O. (2019). Comparative Genomics of Serial Candida glabrata Isolates and the Rapid Acquisition of Echinocandin Resistance during Therapy. Antimicrob. Agents Chemother. *63*.

Beavon, I.R. (2000). The E-cadherin-catenin complex in tumour metastasis: structure, function and regulation. Eur. J. Cancer *36*, 1607–1620.

Beerenwinkel, N., Günthard, H.F., Roth, V., and Metzner, K.J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. *3*, 329.

Carreté, L., Ksiezopolska, E., Gómez-Molero, E., Angoulvant, A., Bader, O., Fairhead, C., and Gabaldón, T. (2019). Genome Comparisons of Candida glabrata Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations. Front. Microbiol. *10*, 112.

Castaño, 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.

Chaisson, M.J.P., Wilson, R.K., and Eichler, E.E. (2015). Genetic variation and the de novo assembly of human genomes. Nat. Rev. Genet. *16*, 627–640.

Chapman, B., Slavin, M., Marriott, D., Halliday, C., Kidd, S., Arthur, I., Bak, N., Heath, C.H., Kennedy, K., Morrissey, C.O., et al. (2017). Changing epidemiology of candidaemia in Australia. J. Antimicrob. Chemother. *72*, 1270.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569.

Cleveland, A.A., Harrison, L.H., Farley, M.M., Hollick, R., Stein, B., Chiller, T.M., Lockhart, S.R., and Park, B.J. (2015). Declining incidence of candidemia and the shifting epidemiology of Candida resistance in two US metropolitan areas, 2008-2013: results from population-based surveillance. PLoS ONE *10*, e0120452.

Cormack, B.P., Ghori, N., and Falkow, S. (1999). An adhesin of the yeast pathogen Candida glabrata mediating adherence to human epithelial cells. Science *285*, 578–582.

De Las Peñas, A., Pan, S.-J., Castaño, 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.

Desai, C., Mavrianos, J., and Chauhan, N. (2011). Candida glabrata Pwp7p and Aed1p are required for adherence to human endothelial cells. FEMS Yeast Res. *11*, 595–601.

De Groot, P.W.J., Ram, A.F., and Klis, F.M. (2005). Features and functions of covalently linked proteins in fungal cell walls. Fungal Genet. Biol. *42*, 657–675.

Domergue, R., Castaño, I., De Las Peñas, A., Zupancic, M., Lockatell, V., Hebel, J.R., Johnson, D., and Cormack, B.P. (2005). Nicotinic acid limitation regulates silencing of Candida adhesins during UTI. Science *308*, 866–870.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., et al. (2004). Genome evolution in yeasts. Nature *430*, 35–44.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Foquet, M., Samiee, K.T., Kong, X., Chauduri, B.P., Lundquist, P.M., Turner, S.W., Freudenthal, J., and Roitman, D.B. (2008). Improved fabrication of zero-mode waveguides for single-molecule detection. J. Appl. Phys. *103*, 034301.

Frieman, M.B., McCaffery, J.M., and Cormack, B.P. (2002). Modular domain structure in the Candida glabrata adhesin Epa1p, a beta1,6 glucan-cross-linked cell wall protein. Mol. Microbiol. *46*, 479–492.

Gonçalves, B., Ferreira, C., Alves, C.T., Henriques, M., Azeredo, J., and Silva, S. (2016). Vulvovaginal candidiasis: Epidemiology, microbiology and risk factors. Crit. Rev. Microbiol. *42*, 905–927.

de Groot, P.W.J., Kraneveld, E.A., Yin, Q.Y., Dekker, H.L., Gross, U., Crielaard, W., de Koster, C.G., Bader, O., Klis, F.M., and Weig, M. (2008). The cell wall of the human pathogen Candida glabrata: differential incorporation of novel adhesin-like wall proteins. Eukaryotic Cell *7*, 1951–1964.

de Groot, P.W.J., Bader, O., de Boer, A.D., Weig, M., and Chauhan, N. (2013). Adhesins in human fungal pathogens: glue with plenty of stick. Eukaryotic Cell *12*, 470–481.

Guo, X., Zhang, R., Li, Y., Wang, Z., Ishchuk, O.P., Ahmad, K.M., Wee, J., Piskur, J., Shapiro, J.A., and Gu, Z. (2019). Understand the genomic diversity and evolution of fungal pathogen Candida glabrata by genome-wide analysis of genetic variations. Methods.

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics *30*, 3004–3011.

Hoyer, L.L. (2001). The ALS gene family of Candida albicans. Trends Microbiol. *9*, 176–180.

Hoyer, L.L., Green, C.B., Oh, S.-H., and Zhao, X. (2008). Discovering the secrets of the Candida albicans agglutinin-like sequence (ALS) gene family--a sticky pursuit. Med. Mycol. *46*, 1–15.

Iraqui, I., Garcia-Sanchez, S., Aubert, S., Dromer, F., Ghigo, J.-M., d'Enfert, C., and Janbon, G. (2005). The Yak1p kinase controls expression of adhesins and biofilm formation in Candida glabrata in a Sir4p-dependent pathway. Mol. Microbiol. *55*, 1259–1271.

Juárez-Reyes, A., Ramírez-Zavaleta, C.Y., Medina-Sánchez, L., De Las Peñas, A., and Castaño, I. (2012). A protosilencer of subtelomeric gene expression in Candida glabrata with unique properties. Genetics *190*, 101–111.

Kaur, R., Domergue, R., Zupancic, M.L., and Cormack, B.P. (2005). A yeast by any other name: Candida glabrata and its interaction with the host. Curr. Opin. Microbiol. *8*, 378–384.

Kaur, R., Ma, B., and Cormack, B.P. (2007). A family of glycosylphosphatidylinositol-linked aspartyl proteases is required for virulence of Candida glabrata. Proc Natl Acad Sci USA *104*, 7628–7633.

Klis, F.M., Mol, P., Hellingwerf, K., and Brul, S. (2002). Dynamics of cell wall structure in

Saccharomyces cerevisiae. FEMS Microbiol. Rev. *26*, 239–256.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. *37*, 540–546.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. *30*, 693–700.

Koren, S., Harhay, G.P., Smith, T.P.L., Bono, J.L., Harhay, D.M., Mcvey, S.D., Radune, D., Bergman, N.H., and Phillippy, A.M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. *14*, R101.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Kraneveld, E.A., de Soet, J.J., Deng, D.M., Dekker, H.L., de Koster, C.G., Klis, F.M., Crielaard, W., and de Groot, P.W.J. (2011). Identification and differential gene expression of adhesin-like wall proteins in Candida glabrata biofilms. Mycopathologia *172*, 415–427.

Kumar, K., Askari, F., Sahu, M.S., and Kaur, R. (2019). Candida glabrata: A Lot More Than Meets the Eye. Microorganisms *7*.

Langer, M.D., Guo, H., Shashikanth, N., Pierce, J.M., and Leckband, D.E. (2012). N-glycosylation alters cadherin-mediated intercellular binding kinetics. J. Cell Sci. *125*, 2478–2485.

Latgé, J.-P. (2010). Tasting the fungal cell wall. Cell. Microbiol. *12*, 863–872.

Latgé, J.-P., and Beauvais, A. (2014). Functional duality of the cell wall. Curr. Opin. Microbiol. *20*, 111–117.

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W.R., and Schatz, M. (2014). Error correction and assembly complexity of single molecule sequencing reads. BioRxiv.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., and Webb, W.W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science *299*, 682–686.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics *32*, 2103–2110.

Li, F., and Palecek, S.P. (2008). Distinct domains of the Candida albicans adhesin Eap1p mediate cell-cell and cell-substrate interactions. Microbiology (Reading, Engl) *154*, 1193–1203.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics *24*, 2818–2824.

Mishra, K., and Shore, D. (1999). Yeast Ku protein plays a direct role in telomeric silencing and counteracts inhibition by rif proteins. Curr. Biol. *9*, 1123–1126.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al. (2000). A whole-genome assembly of Drosophila. Science *287*, 2196–2204.

Nobbs, A.H., Vickerman, M.M., and Jenkinson, H.F. (2010). Heterologous expression of Candida albicans cell wall-associated adhesins in Saccharomyces cerevisiae Reveals differential specificities in adherence and biofilm formation and in binding oral Streptococcus gordonii. Eukaryotic Cell *9*, 1622–1634.

Pappas, P.G., Lionakis, M.S., Arendrup, M.C., Ostrosky-Zeichner, L., and Kullberg, B.J. (2018). Invasive candidiasis. Nat. Rev. Dis. Primers *4*, 18026.

Phan, Q.T., Fratti, R.A., Prasadarao, N.V., Edwards, J.E., and Filler, S.G. (2005). N-cadherin mediates endocytosis of Candida albicans by endothelial cells. J. Biol. Chem. *280*, 10455–10461.

Phan, Q.T., Myers, C.L., Fu, Y., Sheppard, D.C., Yeaman, M.R., Welch, W.H., Ibrahim, A.S., Edwards, J.E., and Filler, S.G. (2007). Als3 is a Candida albicans invasin that binds to cadherins and induces endocytosis by host cells. PLoS Biol. *5*, e64.

Ponniah, G., Rollenhagen, C., Bahn, Y.-S., Staab, J.F., and Sundstrom, P. (2007). State of differentiation defines buccal epithelial cell affinity for cross-linking to Candida albicans Hwp1. J. Oral Pathol. Med. *36*, 456–467.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*, 341.

Rigden, D.J., Mello, L.V., and Galperin, M.Y. (2004). The PA14 domain, a conserved all-beta domain in bacterial toxins, enzymes, adhesins and signaling molecules. Trends Biochem. Sci. *29*, 335–339.

Rosas-Hernández, L.L., Juárez-Reyes, A., Arroyo-Helguera, O.E., De Las Peñas, A., Pan, S.-J., Cormack, B.P., and Castaño, I. (2008). yKu70/yKu80 and Rif1 regulate silencing differentially at telomeres in Candida glabrata. Eukaryotic Cell *7*, 2168–2178.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol. *14*, R51.

Rubinstein, M.R., Wang, X., Liu, W., Hao, Y., Cai, G., and Han, Y.W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell Host Microbe *14*, 195–206.

Rusche, L.N., Kirchmaier, A.L., and Rine, J. (2003). The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae. Annu. Rev. Biochem. *72*, 481–516.

Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. Bioinformatics *30*, 3506–3514.

Staab, J.F., Bradway, S.D., Fidel, P.L., and Sundstrom, P. (1999). Adhesive and mammalian transglutaminase substrate properties of Candida albicans Hwp1. Science *283*, 1535–1538.

Timmermans, B., De Las Peñas, A., Castaño, I., and Van Dijck, P. (2018). Adhesins in Candida glabrata. J. Fungi (Basel) *4*.

Tørresen, O.K., Star, B., Jentoft, S., Reinar, W.B., Grove, H., Miller, J.R., Walenz, B.P., Knight, J., Ekholm, J.M., Peluso, P., et al. (2017). An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics *18*, 95.

Vale-Silva, L., Beaudoing, E., Tran, V.D.T., and Sanglard, D. (2017). Comparative Genomics of Two Sequential Candida glabrata Clinical Isolates. G3 (Bethesda) *7*, 2413–2426.

Verstrepen, K.J., and Klis, F.M. (2006). Flocculation, adhesion and biofilm formation in yeasts. Mol. Microbiol. *60*, 5–15.

Weig, M., Jänsch, L., Gross, U., De Koster, C.G., Klis, F.M., and De Groot, P.W.J. (2004). Systematic identification in silico of covalently bound cell wall proteins and analysis of protein-polysaccharide linkages of the human pathogen Candida glabrata. Microbiology (Reading, Engl) *150*, 3129–3144.

Ye, C., Hill, C.M., Wu, S., Ruan, J., and Ma, Z.S. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci. Rep. *6*, 31900.

Zhang, M., Bennett, D., and Erdman, S.E. (2002). Maintenance of mating cell integrity requires the adhesin Fig2p. Eukaryotic Cell *1*, 811–822.

Zhao, H., Shen, Z.M., Kahn, P.C., and Lipke, P.N. (2001). Interaction of alpha-agglutinin and a-agglutinin, Saccharomyces cerevisiae sexual cell adhesion molecules. J. Bacteriol. *183*, 2874–2880.

Zupancic, M.L., Frieman, M., 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.

Chapter 2 *De novo* assembly of the *Candida glabrata* type strain reveals cell wall protein complement and structure of dispersed tandem repeat arrays

Introduction

In this chapter, I performed *de novo* genome assembly of the *C. glabrata* type strain, CBS138 and all the analysis.

The subcloning of *C. glabrata* subtelomeres into fosmids, and Sanger sequencing of the subcloned fosmids was carried out by a previous lab member, Brian Green. Brendan Cormack prepared the genome DNA of *C. glabrata,* and the PacBio sequencing was carried out by Haiping Hao at the Transcriptomes and Deep Sequencing Core in Hopkins. The manuscript of this chapter is under revision at Molecular Microbiology (Xu Z, Green B, Benoit N, Schatz M, Wheelan S, Cormack B, *De novo* genome assembly of *Candida glabrata* reveals cell wall protein complement and structure of dispersed tandem repeat arrays, Molecular Microbiology).

Candida glabrata is a major fungal pathogen in humans, causing both superficial mucosal infection and serious disseminated infection (Gonçalves et al., 2016; Pappas et al., 2018). In Europe and the United States, *C. glabrata* is responsible for up to 30% of Candida bloodstream infections in hospitalized patients (Andes et al., 2016; Astvad et al., 2018; Chapman et al., 2017; Cleveland et al., 2015). The factors contributing to the virulence of *C. glabrata* are incompletely understood but include secretion of hydrolytic enzymes, ability to evade phagocytic killing and adherence to host tissue (Kumar et al., 2019). In terms of adherence, *C. glabrata* is known to encode a large repertoire of surface proteins, some of which directly mediate adherence to mammalian cells (Timmermans et al., 2018). The *C. glabrata* genome encodes a family of adhesins encoded by the *EPA* genes, which mediate adherence to host glycans (Castaño et al., 2005; De Las Peñas et al., 2003; Maestre-Reyna et al., 2012; Zupancic et al., 2008) as well as additional cell wall proteins, including the Awp, Aed, and Pwp proteins, that have been implicated in *C. glabrata* adherence (Desai et al., 2011; Timmermans et al., 2018). The exact complement of predicted cell wall proteins in *C. glabrata* is unknown because of limitations of current genome assemblies for *C. glabrata*. These limitations are related to the nature of cell wall proteins in *C. glabrata*. The major cell wall proteins are GPI-anchored cell wall proteins (GPI-CWPs) which are covalently anchored to the cell wall through a remnant GPI anchor (present at the C-terminus of the protein) and have large, low complexity spacer regions that act to project the N-terminal domains of these proteins away from the site of cell wall attachment. Importantly, these repeat regions include tandem repeat sequences, some of which are large (termed megasatellites (Thierry et al., 2008)) that confound genome assembly efforts. About half of all cell wall proteins are encoded in the sub-telomeric regions of *C. glabrata*, and these regions are particularly difficult to assemble from short read sequences because multiple sub-telomeres share tandem repeat sequences (including megasatellites). Genome sequence reports of clinical isolates suggest substantial variation in the complement of cell wall protein genes relative to the reference assembly (Barber et al., 2019; Carreté et al., 2019; Guo et al., 2019; Vale-Silva et al., 2017) , though potential errors in assembly in the reference sequence as well as in new genome sequences make such claims uncertain. A high-quality genome sequence is therefore needed to generate first a more accurate map of the genome, and also to provide a reference for the total repertoire of adhesin and adhesin-like proteins, which are likely to play important roles in *C. glabrata* virulence.

The reference sequence of *C. glabrata* was assembled using shotgun sequencing from 3-5 kb plasmid libraries with Sanger sequencing, and finished by Bacterial artificial chromosome (BAC) end sequencing (Dujon et al., 2004). This resulted in a high-quality genome assembly and accurate overall reference sequence. However, difficulties in assembling long tandem-repeat regions by shotgun sequencing resulted in systematic mis-assemblies within the tandem-repeat regions, and corresponding misassembly of the cell wall protein genes containing those tandem repeat regions. More recent genome sequencing of clinical isolates has generally used short read sequencing libraries from PCR-based whole genome amplification (WGA) (Barber et al., 2019; Carreté et al., 2019; Guo et al., 2019). PCR-based WGA can result in sequencing error or bias (Beerenwinkel et al., 2012), and lead to an overestimation of SNPs. Furthermore, it is difficult to detect large structural variants or to obtain the correct structure of long tandem repeat regions using short read sequencing.

Amplification-free long read single molecule sequencing technologies are therefore one solution to obtain genome assemblies with high accuracy and correct genome structure. Single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) is one of the most widely used long read approaches (Eid et al., 2009). SMRT sequencing generates long reads without context bias (Chaisson et al., 2015; Quail et al., 2012; Ross et al., 2013) that allows efficient and highly accurate de novo assembly, including assembly of the long tandem repeat regions. Vale-Silva et al. have applied this technology to assemble genome sequence of serial clinical C. glabrata isolates (Vale-Silva et al., 2017). We have applied the SMRT sequencing for de novo assembly of the CBS138 *C. glabrata* reference strain. Our sequence corrects multiple assembly errors in the current reference, resulting in elimination of 60 ORFs and identification of 31 new ORFs. A major unexpected finding is the substantial length of tandem repeat regions contained within the coding region of several GPI-CWP genes, suggesting that the *C. glabrata* genome encodes several extremely long cell surface proteins.

Experimental Procedures

Genomic DNA and PacBio library preparation

Genomic DNA of strain ATCC2001 (gift of Cecile Fairhead, Pasteur Institute) was made by preparation of spheroplasts, followed by lysis and spooling of high molecular weight genomic DNA after ethanol precipitation. The sequencing library was prepared using DNA Template Prep kit v.2 (3-10kb) following the PacBio shared protocol guidelines for preparing size-selected ~20kb SMARTbell templates. Briefly, 7.5 µg of genomic DNA was diluted to 150 µl and sheared using Covaris G-tube by centrifugation in an Eppendorf 5424 microcentrifuge at 4600 rpm for 60 seconds. Sheared DNA was then purified using AMPure beads and quality controlled for concentration and size. 5 µg of sheared DNA was then used for DNA damage repair and end repair. End repaired DNA was purified using AMPure beads and ligated to SMRTbell adapter via blunt end ligation and exonuclease III/VII treatment. The exonuclease treated SMRTbell template was purified again using AMPure beads and quality controlled for yield and library size.

The purified SMRTbell library was size selected on Blue Pippin with a size cut off of 5kb. The size selected library was then sequenced on PacBio RS II using PacBio P4/C2 chemistry.

De novo Assembly of CBS138 genome

The de novo assembly of the genome using the SMRT sequencing reads was performed using CANU 1.5 (Koren et al., 2017) on the cluster of the Maryland Advanced Research Computing Center (MARCC). The default CANU protocol was applied with the following parameter adjustments: 1) corOutCoverage=300; 2) genomeSize=12.3m. The draft contigs were polished by Quiver from the GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus) with the same PacBio reads. We performed whole genome sequence alignment by NUCMER from MUMMER3 (http://mummer.sourceforge.net/) with the CBS138_s02-m07-r41 reference genome (http://www.candidagenome.org/) (Dujon et al., 2004) to assign contigs to chromosomes, and used the telomere seed sequence (GGGGTCTGGGTGCTG) to locate the telomere repeats in each contig. We obtained 17 contigs from the CANU assembly (see Results). 10 contigs were the telomere to telomere assembly of ChrA, ChrB, ChrD – ChrK. 1 contig was assigned as the mitochondrial genome, and 1 contig was an array of rDNA repeats. Two contigs were assigned to the left and right part of ChrC, overlapping a repetitive region around 80 kb– 119 kb in the reference ChrC. We inserted 1 kb of N to join the two ChrC contigs to generate a single contig for ChrC. The two major contigs of ChrL and ChrM were telomere-to-rDNA assemblies, with both contigs ending in one rDNA repeat. One contig had 4 rDNA repeats with a 10 kb downstream region and terminating in telomere repeats, which we speculated to correspond to the ends of both ChrL and ChrM. We aligned the 6 rDNA repeats using ClustalW2 (Larkin et al., 2007). We generated the ~ 11 kb consensus rDNA sequence using consamig from the EMBOSS package (Rice et al., 2000), and manually corrected ambiguous nucleotides in the consensus sequence.

The rDNA-to-telomere fragment contained three complete rDNA repeats followed by a truncated rDNA repeat and the 10 kb downstream region. We fused this region to the end of Chr L and M to generate single contigs for Chr L and M. We verified the rDNA consensus with 50 bp paired Illumina reads using BWA-backtrack (Li and Durbin, 2009) with default settings. Illumina reads supported our rDNA consensus (data not shown), interestingly, we found two polymorphisms within the rDNA non-coding region at a frequency of 21% and 35%. We manually added each polymorphism separately in the second-to-last rDNA repeats in ChrL and ChrM (Table S1)

Sequence qualification and further correction

We aligned the Illumina sequencing reads to our assembly as well as the reference genome by Bowtie2 (Langmead and Salzberg, 2012) with default settings and filtered for properly paired reads by samtools (Li et al., 2009). The per base read coverage was counted by BEDTools (Quinlan and Hall, 2010), and we calculated the mean read coverage in 100 bp windows in both assemblies. We also calculated the read average in the ORF regions of single-exon genes in both assemblies to compare the sequence quality in repeat regions. We compared both assemblies by counting 20-mers, and masked structural variant regions with tandem-repeat differences (Table S1). We performed the whole genome alignment by NUCMER in non-subtelomeric genes to discover the SNPs and INDELs. After masking the manually defined structural variant regions (where called variants are due to mis-assembled repeats), we discovered 40 SNPs and 102 INDELs in our assembly compared with the reference genome (Table S1). To validate these variants, we polished both our assembly and the reference genome using the same Illumina sequencing reads with Pilon 1.22 (Walker et al., 2014). If the polishing supported the variant in either the reference or our assembly, it was classified as Illumina-supported. 17 SNPs were supported, and all correspond to the variant in our assembly. 19 additional SNPs were spurious, generated by NUCMER due to misalignment. 4 SNPs were unconfirmed by Illumina, and these have been kept in the final genome sequence. Of the 102 INDELs, 55 were supported by Illumina, and 46 of these correspond to sequence in the reference genome; 35 of these 46 INDELs were in homopolymer (>=5N) regions. 47 INDELs were unsupported by Illumina: 43 of these were in homopolymer (>=5N) regions. We used PCR to amplify and sequence 5 variant regions (INDELs and SNPs) (Primers in Table S2) - all of these verified the Illumina-supported sequence (data not shown). We adjusted the Illumina-supported INDELs (46) as well as the unsupported INDELs (47) in our assembly to the variants the reference genome. Our final assembly, then, has 9 INDELs and 21 SNPs relative to the reference genome.

For structural variation, we used Assemblytics (Nattestad and Schatz, 2016) to extract the structural variants in addition to manually defining the structural variant regions, and verified that all the structural variants are expansions, contractions, or other changes within repeat regions (Table S1). We also visualized the change at chromosome level by dotplot (window=30 nt) (Figure S2). We illustrated the change in subtelomeres by NUCMER alignment (--no-extend, --mum) of our subtelomeres to the reference subtelomeres and visualized the alignments (Figure S4). The tandem-repeat regions were identified by Tandem Repeat Finder (TRF) (Benson, 1999). The alignments of one subtelomere in our assembly to other subtelomere(s) in reference in non-tandem-repeat regions indicate possible mis-assemblies. To further qualify the structure of the subtelomere structure, we compared the subtelomeric regions in our assembly against the subtelomeres subcloned into fosmids (see below). The comparison of our subtelomeres to subcloned fosmids were illustrated by dotplots (Figure S5).

Targeting fosmid construction and integration

To ensure that clones originated from well assembled regions of the genome, the terminal ORF of the most terminal block of three syntenic ORFs between C. glabrata and S. cerevisiae for each telomere end was selected to be the site of fosmid integration. In cases where targeting fosmids or integrants could not be recovered, new targeting sequence was selected that was centromere proximal to the terminal syntenic ORF. Fusion PCR was used to generate, for each targeting fosmid, an approximately 1,000 bp MluI to SacII fragment with BG1182 (CBS138) as the template. The first round of PCRs yielded approximately 500 bp fragments using the "left AS" and "left S" and the "right AS" and "right S" pairs (Table S3). A PCR purification kit was used to isolate those two fragments, which were combined and used as a template with the appropriate "left S" and "right AS" oligos. The resulting fragment, containing an internal PpiI site, was ligated into MluI to KpnI and KpnI to SacII fragments of pBAC-NAT (Green et al. 2012). These targeting plasmids were digested with PpiI and used to transform BG1182 (CBS138) and BG2, followed by selection on plates contained clonNAT. Correct integrants were identified by PCR.

Telomere cloning in Fosmids, Sequencing and Assembly

To prepare genomic DNA, the cell pellet from 1.5 mL of YPD stationary phase culture of each integrant was resuspended in 250 uL zymolyase buffer (1.2 M sorbitol, 10 mM tris pH 8, 10 mM CaCl2, 1% beta-mercaptoethanol, 0.7 mg/ml zymolyase). After approximately 30 minutes at 37 C, 200 uL of lysis buffer (50 mM tris pH 8, 50 mM EDTA, 1.2% SDS) was added to each tube and samples were inverted to mix. Next, 100 uL of 3M NaAc pH 5.2 was added, followed by inverting the tube to mix and centrifugation in a microfuge at full speed for 10 minutes. The supernatant was transferred to a new tube, centrifuged for another 5 minutes, and then transferred again. Isopropanol precipitated DNA was resuspended in 300 uL of TE plus 1.5 μL 10 mg/ml RNAse, and incubated at 37 C for 30 to 60 minutes. After another isopropanol precipitation, the DNA was resuspended in 50 μL TE.

To clone the subtelomeric region, the DNA from the integrated fosmid to the telomere was liberated and circularized. First, AscI was used to cleave inside the fosmid and release the end, after which the AscI was heat killed. The ends were blunted with T4 polymerase, which was then heat killed. After a three-fold dilution, the fosmid was circularized with T4 ligase, precipitated, and transformed into MegaX DH10B cells. Colony PCR was used to check for fosmid to telomere junction suggestive of full-length clones with ON3629 and ON3654 (Table S3).

We sequenced fosmids by generating transposon insertions using Tn7 transposition was used as described (Castano et al., 2003). Tn7 was chosen because it has a less pronounced sequence bias than other transposons (Green et al., 2012). After transposition and transformation into MegaX DH10B cells, colonies were picked and arrayed into 96 well plates, initially 2 per telomere rescue. Fosmids were then sequenced using Sanger sequencing with paired reads out of both ends of the transposon with ON661 and ON662 (Table S3). A standard phred/phrap/consed (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998) assembly pipeline was used with default parameters to assemble the sequencing reads. Some repeat regions were longer than could be resolved with standard Sanger sequencing, and those fosmid assemblies contain truncated tandem repeat arrays.

Genome annotation and gene comparison

For annotation of our draft genome assembly, we treated the body of the chromosomes and the subtelomeres differently. For the body of the chromosomes, we did not carry out de novo ORF calling. Rather, we re-annotated our assembly by homology and synteny information. Our reannotation was very conservative, preserving as much as possible the systematic names in the reference. All multi-exon genes in the reference genome were aligned to our draft assembly by BLASTN in BLAST 2.6.0+ (Camacho et al., 2009) and directly annotated. Single-exon genes were first aligned to our draft assembly by BLASTN, and, as necessary, the homologous regions were extended to generate a list of all ORFs. We next performed BLASTN comparing our ORFs to annotated genes in the reference. If an ORF was a reciprocal best hit with an annotated gene, we assigned that systematic name. For all other ORFs (in multigene families) with multiple BLASTN hits, we used synteny information to assign systematic names, *i.e.*, we assigned the ORF to the unassigned reference gene from after the previous step that shared homology if it was in synteny with the reference genes assigned to neighbor ORFs.

Because the subtelomeric regions in the reference genome are broadly misassembled, these regions were were re-annotated separately, as follows. First, to operationally define the subtelomeric regions, we obtained synteny information from Yeast Gene Order Browser (Byrne and Wolfe, 2005) for the terminal genes in the reference CBS138 genome. We define sub-telomeric regions as the regions telomeric to the coding regions of the last pair of genes syntenic between *C. glabrata* and *S. cerevisiae* or the putative ancestor. We also chose to exclude from the subtelomeric regions genes associated with the MTL (Mating type like) loci and MLT-related translocations. We identified all predicted ORFs with length > 200 nt, and aligned them to the annotated genes in the reference (Table S4). All ORFs with the same N-terminus and C-terminus as a currently annotated gene in the reference genome were assigned that systematic name. For novel ORFs >600 nt, we assigned a new systematic name, generated by addition or subtraction of the multiple of 11 to the adjacent preserved gene. For novel ORFs < 600 nt, we examined RNASeq data (from strain BG2), and assigned a new systematic name if the RNAseq data supported the existence of a transcription unit corresponding to that gene. To assess whether we had missed any subtelomeric genes in the reference, we aligned all unassigned subtelomeric genes in the reference to our assembly by BLASTN, which showed that all correspond to misassembled ORFs: this is visualized by a comparison of the protein sequence of the misassembled gene to the correct overlapping gene in our assembly by dotplot with window size=10 (Figure S7). These misassembled genes have been removed from the new annotation. Apart from the subtelomeres, we performed the same workflow with the ChrC region between the CAGL0C00759g and the CAGL0C01155g, which is also largely misassembled in the reference.

To capture potentially important genome differences between the reference and our assembly, we compared the sequences of all single-exon genes present in the two genomes. Genes with a length difference > 50 nt were classified as structural variants, and for these, we compared the protein sequence by dotplot with window size=10. 4 genes (CAGL0A04851g, CAGL0I10246g, CAGL0J05159g and CAGL0L13299g) had a frameshift in the tandem repeat region, due to the presence of one or more mutated repeat units. Solely for the purpose of ORF definition and downstream ORF analysis, we manually corrected the frame by replacing tandem repeats with adjacent repeats (i.e. the nucleotide sequence for these genes was not changed in the assembly) (Table S6).

Analysis of the GPI-anchored genes

Since most of the genes corrected in the new assembly correspond to GPI-anchored cell wall proteins (GPI-CWPs), we systematically classified all genes encoding GPI-anchored proteins and additionally annotated all GPI-anchored adhesins. We first identified all putative GPI anchored proteins using the PredGPI GPI anchor predictor (Pierleoni et al., 2008) for all the ORFs in our assembly. The genes with ORFs having a PredGPI FDR < 0.005 were classified as putative GPI-protein encoding genes. Additionally, for all genes with FDR >= 0.005, we identified homologs (BLASTP e-value < 1E–9) of annotated GPI-anchor protein encoding genes in the reference; among these, ORFs with boundary PredGPI scores (FDR < 0.03) were also classified as putative GPI-anchored genes. Lastly, we also included any genes annotated as encoding GPI-anchored proteins in the reference. To identify the subset of GPI proteins encoding GPI-anchored adhesins or adhesin like proteins, we identified those tandem repeat regions using dotplots of the protein sequences with window size=5 (Figure S8). Seven GPI-anchored proteins did not contain tandem repeats, but were homologous to GPI adhesins (BLASTP e-value < 1 e –9) and these were also classified as GPI anchored putative adhesins. Finally, we included three additional genes (with no internal repeats and no homology to established adhesins) but which have been annotated as GPI-CWP adhesin-like genes in the reference.

We analyzed the adhesin like genes and assigned them into clusters by a neighbor-joining phylogenetic tree with bootstrap values (1000 bootstraps) of the N-terminal domains of the GPI adhesin like proteins (Figure 4). GPI-anchored adhesin like genes have an N terminal region followed by a repeat containing region. Operationally, we defined the N-terminal regions of the GPI-anchored adhesin-like proteins as the region preceding repeated sequence by analyzing the occurrence of 5-mers in protein sequences (Figure S8). We counted the occurrence of the 5-mers from the beginning of the gene, and the first occurrence of the first 5-mer that had 3 occurrences in the sequence were defined as the start site of the repeat region. Accordingly, the N-terminal region extends from the beginning of the ORF to the amino acid preceding the repeat start site. The neighbor-joining phylogenetic tree for N terminal regions was generated by ClustalW2 with default setting from Clustal 2.1 (Larkin et al., 2007) with seed=111 and 1000 bootstrap trials. All the branches with bootstrap values > 500 were first collapsed together to form clusters, and we compared our clusters with the current adhesin clusters defined by de Groot et al. (de Groot et al., 2008). As our clusters correspond well to the current adhesin clusters, we preserve that nomenclature.

Results

Telomere to telomere assembly of CBS138 genome

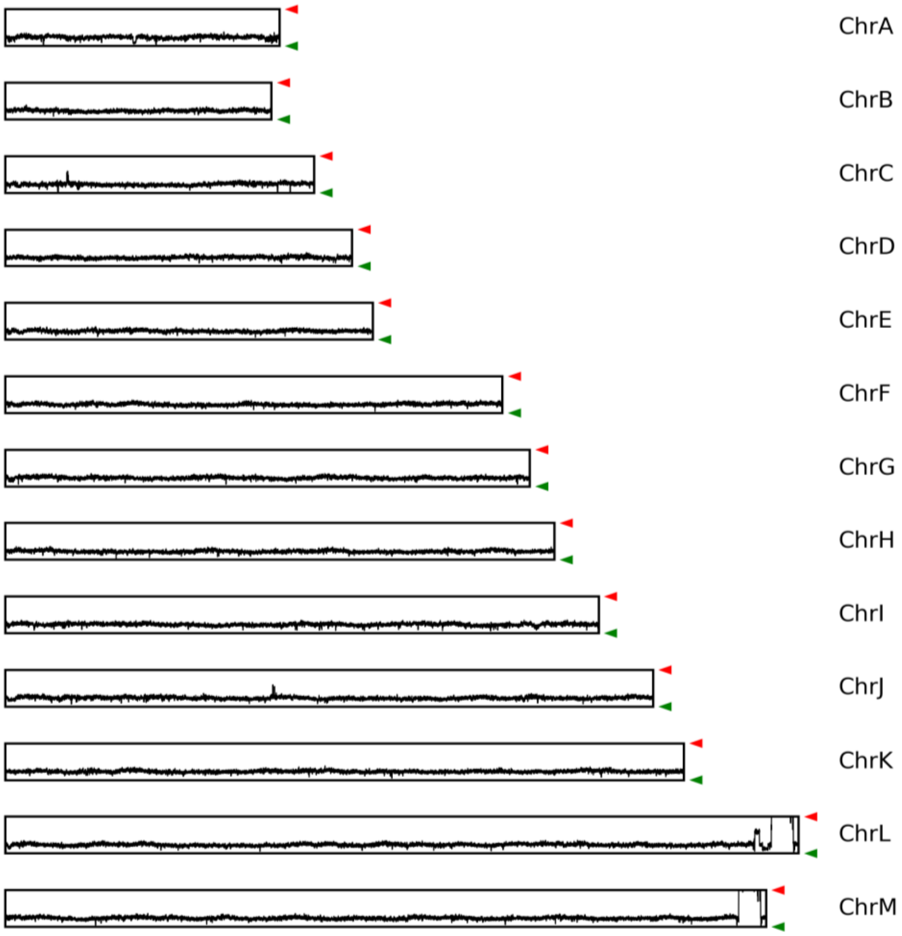
We sequenced C. glabrata strain ATCC2001 (CBS138) using SMRT long read sequencing (Table S5). We obtained 203,355 reads (approximately 102-fold coverage) with a subread N50=9522 (Figure S1). We then assembled the contigs using the CANU 1.5 assembler (Koren et al., 2017). Draft contigs were polished with the same PacBio reads using Quiver from PacBio GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus). We annotated our contigs using the CBS138 reference genome (s02-m07-r41) (Dujon et al. 2004) in the Candida Genome Database (http://www.candidagenome.org/). Our draft assembly consisted of just 17 contigs: the mitochondrial genome; a contig which contained rDNA sequence only; a contig consisting of rDNA repeats and an additional 10 kb region; 10 contigs which were telomere-to-telomere assembly of all the chromosomes except ChrC, ChrL and ChrM; two contigs corresponding to ChrC; two contigs corresponding to telomere-to-rDNA repeat array of ChrL and ChrM. Consequently, with minimal manual changes to these three chromosomes (described below) we generated a draft genome sequence with 13 telomere-to-telomere chromosome assemblies (Table S6).

Chromosome structure

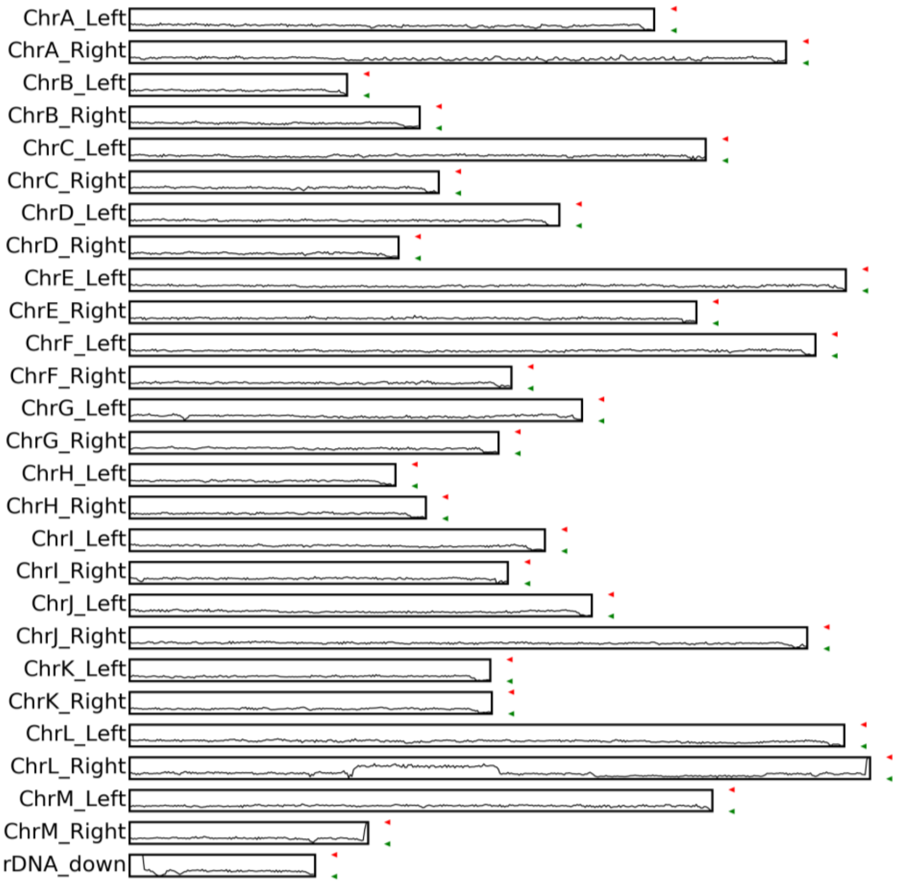
Most contigs in the draft assembly, with the exception of Chr L, M, and C were whole chromosome telomere to telomere assemblies, in overall agreement with the reference genome (Figure S2). ChrL and ChrM terminated in rDNA repeats in our assembly. Our assembly included a small contig which contained three complete and one truncated copy of the rDNA repeat linked to a 10 kb region, terminating in telomere repeats. This 10 kb region contains one ORF, EPA14.

We verified using PCR that the rDNA array is in fact linked to this novel 10 kb region, and verified that our unassembled CANU unitigs with rDNA sequence did not contain adjacent novel regions (data not shown). Illumina reads corresponding to the 10 kb region are present at twice the depth as for single copy regions of the genome consistent with the 10 kb region being present twice in the genome. We conclude therefore that ChrL and ChrM both terminate in the rDNA array followed by identical 10 kb EPA14 containing regions. We name these duplicated genes EPA14a and EPA14b.

Chromosome C was assembled in two contigs. We could not resolve the exact structure of the region in ChrC corresponding to the breakpoint, and added 1 kb of Ns to join the two contigs. The ChrC region between CAGL0C00759g and CAGL0C01155g (80 kb – 119 kb genomic location in the reference CBS138 genome) is highly repetitive, containing multiple copies of several ORFs: there is a single copy of the EPA8 gene and several putative GPI-CWP encoding genes, homologous to CAGL0C00968g, as well as multiple copies of small genes homologous to CAGL0C00781g and CAGL0C00946g. This region in the reference genome had significant mis-assemblies as assessed by depth of Illumina coverage (Figure S3); these mis-assemblies were mostly but not completely resolved in our assembly, leaving only a small peak in the Illumina coverage (Figure 1).


ChrA
ChrB
ChrC
ChrD
ChrE
ChrF
ChrG
ChrH
ChrI
ChrJ
ChrK
ChrL
ChrM

(A)



(B)

Figure 1 Illumina Read Coverage over the assembled genome.

The 50-bp pair-end Illumina reads were aligned to our assembled genome by bowtie2, and the properly paired reads were filtered by samtools. The per-base read coverage was calculated by BEDTools, and we drew the average read coverage in 100 bp windows. A) Read coverage of the entire assembly. The green triangle indicates 0 coverage, and the red triangle indicates coverage greater than or equal to 200. The Illumina reads were evenly distributed in the genome, indicating the correct structure of tandem-repeat regions in our assembly. Only four regions had an odd read coverage peak: 1) the dynamic ChrC region; 2) the region in ChrJ where the CAGL0J05159g gene is located; 3) the region where the CAGL0L013299g (EPA11) gene and CAGL0L013332g (EPA13) gene are located; 4) rDNA regions B) The read coverage of sub-telomeric regions. The green triangle indicates 0 coverage, and the red triangle indicates coverage greater than or equal to 200. Left and Right subtelomeres are both drawn in a centromere to telomere orientation.

Comparison and validation relative to reference genome

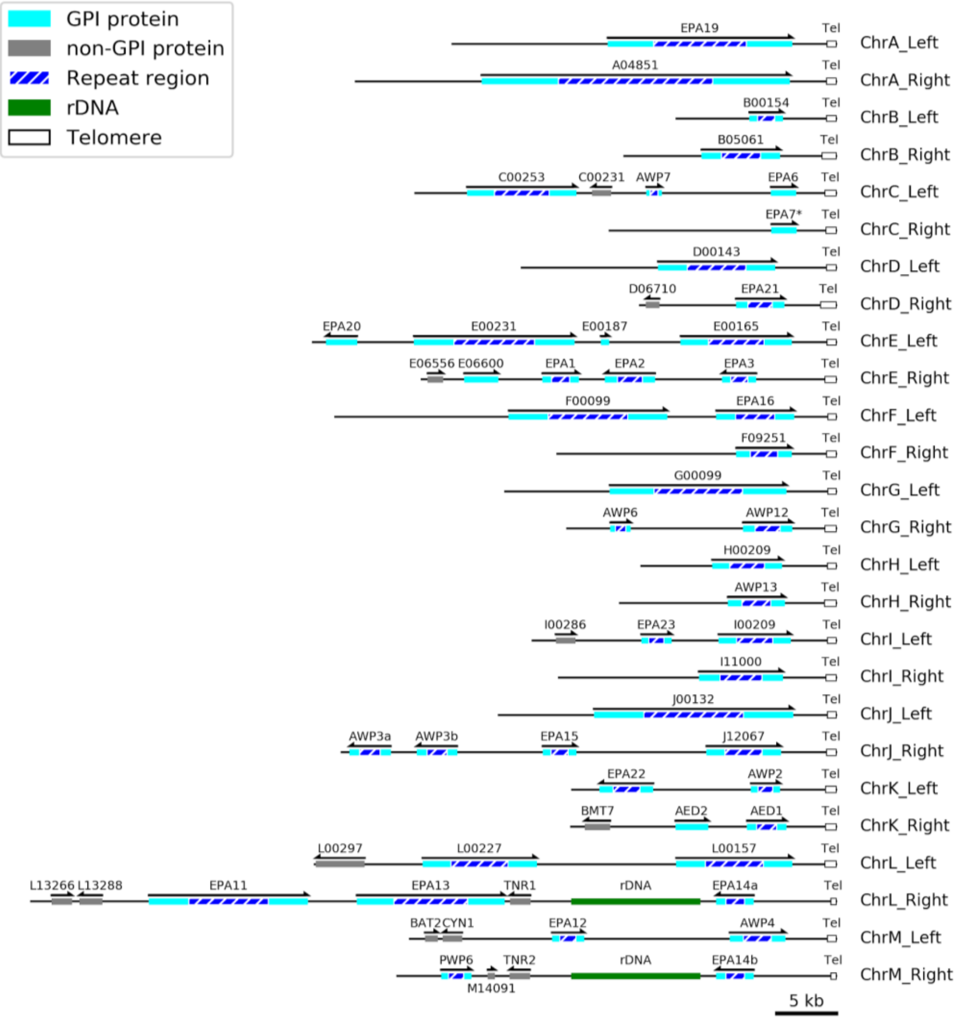
Excluding the sub-telomeric regions, and structural variant regions (defined as length differences of greater than 50 bp, see below) we found only 30 SNPs and INDELs (Table S1) between our assembly and the reference genome. In addition, we discovered two polymorphisms in the rDNA repeat region, with the frequency of 21% and 35% in the non-coding region (Table S1), suggesting heterogeneity in rDNA repeats. We added the two polymorphisms in the second-to-last repeat in ChrL and ChrM. While the chromosome structure across the body of all chromosomes was in agreement with the reference genome (Figure S2), we identified significant mis-assemblies in the subtelomeres in the reference genome. We define the subtelomeric regions as telomeric to the last gene pair that is syntenic with the ascomycete ancestor (see Methods). There were dozens of small translocations and one large translocation resulting from the mis-assembly of ChrM-Left and ChrJ-Right subtelomeres (Figure S4). In addition to these translocations, the major assembly errors we identified were structural variants resulting in length differences of greater than 50 nt. Using Assemblytics (Nattestad and Schatz, 2016), we identified 29 structural variants between our assembly and the reference genome, all of which were due to a change in repeat number within tandem repeat arrays, and accounting for a total of 188.34 kb of length differences in our assembly relative to the reference genome (Table S1).

To validate the accuracy of our genome specifically in these tandem repeat regions, we verified our assembly in several additional ways. First, we assessed read depth for Illumina short read sequencing when mapped against our draft assembly or the published reference. The short reads were aligned by Bowtie2 with default settings, and we filtered for properly paired reads (Langmead and Salzberg, 2012). Multi-mapping reads (corresponding mostly to large tandem-repeat arrays) were randomly distributed across potential target sites. The sequencing depth coverage relative to the reference genome (Figure S3) showed multiple high coverage peaks indicating regions where the tandem repeat array has been truncated in the reference genome. In contrast, the read coverage of our assembly shows a much more even distribution over all chromosomes (Figure 1). We compared the mean read coverage in 100 bp windows in non-rDNA and non-telomeric regions for quantitative verification. The average read coverage was 47.5 for the published genome, and 46.5 for our assembly. If tandem repeat regions are correctly assembled in terms of the number of repeats, the depth of coverage for these regions should match that for the genome as a whole. Specifically, the read coverage distribution should fit a Poisson distribution, with a prediction that less than 0.19% (20.3 kb) of the genome is expected to have read coverage > 68 (3 standard deviation above the mean). The published CBS138 genome had 1038 windows (103.8 kb) with coverage > 68, and the standard deviation of read coverage was 22.2. In our draft assembly, by contrast, only 254 windows (25.4 kb) had a coverage > 68, and the standard deviation was 10.7. These 254 windows correspond primarily to three regions in our assembly with aberrant read depth based on the Illumina read coverage: 1) the ChrC region discussed above; 2) the gene CAGL0J05159g; 3) the genes CAGL0L13299g and CAGL0L13332g.

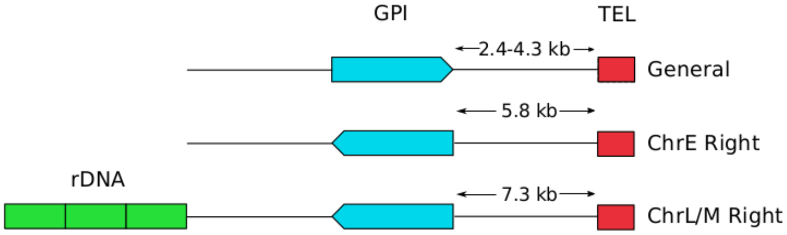
Because the subtelomeres share substantial homology between different subtelomeres in multiple chromosomes, and are broadly misassembled in the reference genome, we wished to confirm that the overall assembly of these regions was correct in our draft assembly. We cloned each individual subtelomere into fosmids, and sequenced these using Sanger sequencing. Comparison of the fosmid sequences to the subtelomeric regions in our draft genome (Figure S5) show perfect alignment for all non-repeat regions, demonstrating that overall assembly of these regions in our draft assembly is correct (see Methods)

Subtelomere Structure

Most of the subtelomeres in the reference genome were mis-assembled compared to our assembly, requiring a re-annotation of all the subtelomeric regions (Figure 2, Table S4). We corrected the tandem-repeat length of 9 genes (Figure S6). Several ORFs in the reference genome were mis-assembled and scrambled, and we removed 53 genes from the reference genome. Within the subtelomeres, we annotated 21 new ORFs (length > 600 nt), with a total size of 177 kb. In addition, we annotated 1 novel subtelomeric gene of 480 nt, CAGL0G00143g with no homology with the genes in the reference genome. This is an intact ORF in strain CBS138, and in an additional strain (BG2), RNAseq data showed that it was transcribed (data not shown). After the re-annotation of the CBS138 subtelomeres, we discovered a common overall structure to the subtelomeres (fig_22). The terminal gene on each subtelomere is a GPI-anchored adhesin gene, with almost all of these being transcribed towards the telomere, followed by a 2.4-4.3 kb non-coding region. There are two exceptions: the terminal GPI-CWP gene of ChrE-Right (EPA3), and EPA14a/b (downstream of the rDNA arrays on ChrL/M-right) are transcribed towards the centromere.


GPI protein
non-GPI protein
Repeat region
rDNA
Telomere
EPA19
Tel
ChrA_Left
A04851
ChrA_Right
B00154
ChrB_Left
B05061
ChrB_Right
C00253
C00231
AWP7
EPA6
ChrC_Left
EPA7*
ChrC_Right
D00143
ChrD_Left
D06710
EPA21
ChrD_Right
EPA20
E00231
E00187
E00165
ChrE_Left
E06556
E06600
EPA1
EPA2
EPA3
ChrE_Right
F00099
EPA16
ChrF_Left
F09251
ChrF_Right
G00099
ChrG_Left
AWP6
AWP12
ChrG_Right
H00209
ChrH_Left
AWP13
ChrH_Right
I00286
EPA23
I00209
ChrI_Left
I11000
ChrI_Right
J00132
ChrJ_Left
AWP3a
AWP3b
EPA15
J12067
ChrJ_Right
EPA22
AWP2
ChrK_Left
BMT7
AED2
AED1
ChrK_Right
L00297
L00227
L00157
ChrL_Left
L13266
L13288
EPA11
EPA13
TNR1
rDNA
EPA14a
ChrL_Right
BAT2
CYN1
EPA12
AWP4
ChrM_Left
PWP6
TNR2
EPA14b
M14091
ChrM_Right
5 kb

(A)



(B)

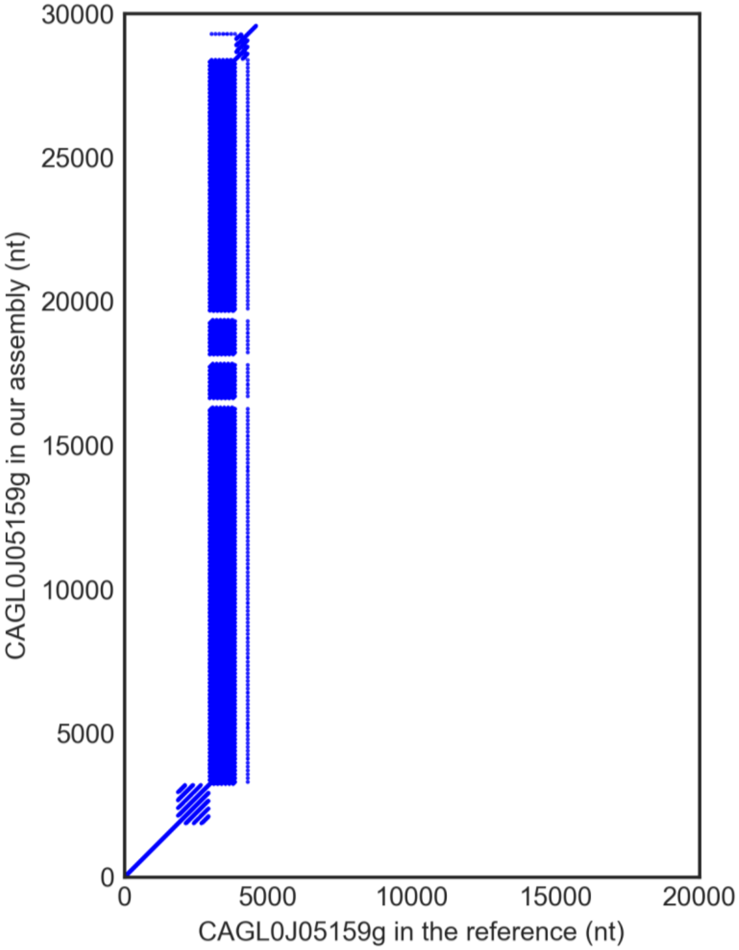
Figure 2 Structure of the C. glabrata subtelomeres.

A) Location, size and orientation of ORFs in the subtelomeric regions of the CBS138 strain. All ORF lengths are indicated to scale. ORFs for predicted GPI-anchored adhesins are colored in light blue, and all other genes are colored in grey. Tandem repeats are indicated with a shadowed dark blue box (not to scale). One copy of the rDNA repeat (11 kb) is indicated to represent the rDNA cluster. Telomere repeats are indicated by a black box. B) The general structure of the subtelomeres. The terminal gene for all subtelomeres encodes a GPI-anchored adhesin-like protein. The terminal GPI genes are transcribed towards the telomere end with a terminal intergenic region of 2.4 - 4.3 kb for all subtelomeres except for those of ChrE right, and the two rDNA containing chromosomes, ChrL right and ChrM right.

Gene comparison

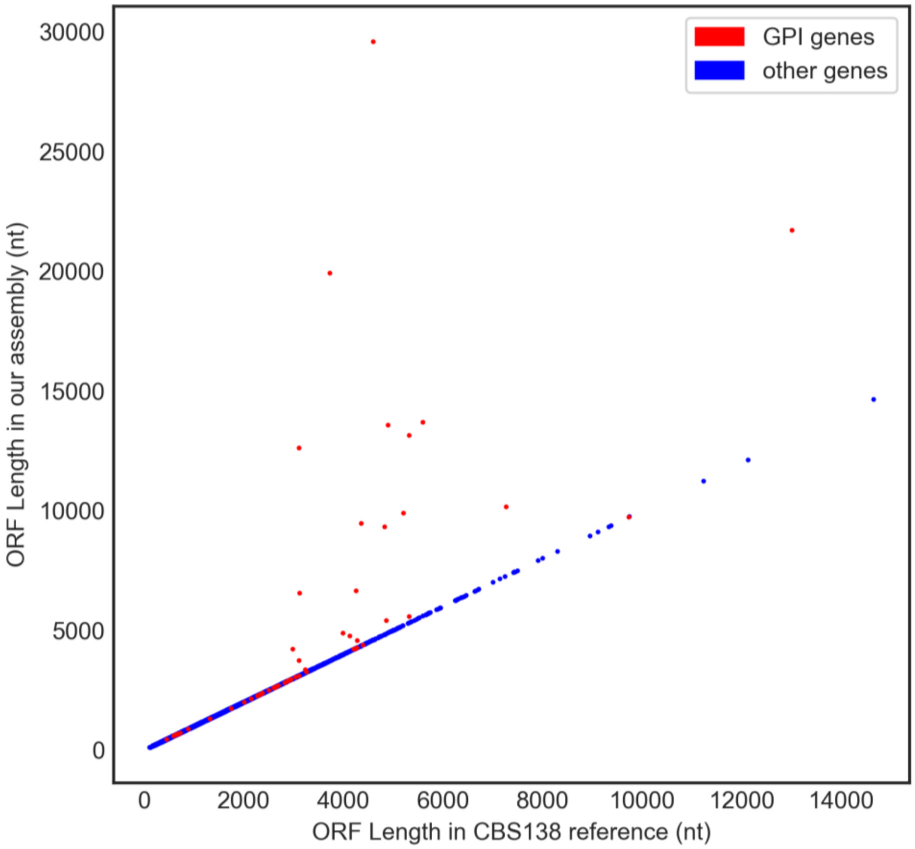
Most ORFs in our assembly are identical with those in the reference genome: There were 5269 ORFs annotated in our assembly. 5209 genes were identical in protein sequence to the reference genome. 31 genes were novel. 8 genes contained small variants in protein sequence. 21 genes contained structural variants correcting 111.7 kb in tandem repeat regions (Figure 3, Figure S6, Table S7). Compared to the 5300 annotated ORFs in reference, 62 genes were removed. We validated our correction in tandem repeat structure using read depth for Illumina sequence coverage. We calculated the mean read depth of each coding gene (Table S7). In the published reference genome, many genes had much higher mean read depth (>3 expected standard deviations) consistent with truncation of the tandem-repeat array in the reference genome. The mean read depth for all but one of the tandem-repeat containing genes in our assembly were within 3 deviations, consistent with correct assembly. EPA11(CAGL0L13299g) had a read depth of 5.7 times the average genome coverage in the published genome, and approximately twice the coverage in our assembly, suggesting that our assembly of EPA11 still truncates the actual tandem repeat array. We annotated 31 new coding genes with a total length of ~191 kb. 22 were subtelomeric; seven novel genes were additional homologs of CAGL0C00781g, CAGL0C00946g and CAGL0C00968g in the repetitive ChrC region; we annotate one extra copy of the duplicated MT-II gene (CAGL0H04290g); CAGL0I02838g in the reference genome contains a stop codon that we corrected, resulting in a fusion with the CAGL0I02816g gene, and we name this new ORF CAGL0I02827g. 62 genes were removed of which 53 were located in the subtelomeres (Table S5). 45 of the 62 were partial or scrambled ORFs due to mis-assembly (Figure S7). Six genes were short subtelomeric genes (length < 400 bp) in the reference genome, which are not ORFs in our assembly, or for which RNAseq data (in strain BG2) did not show gene-associated transcription (data not shown). One multi-exon gene (CAGL0K13013g) overlapping with the

CAGL0K13002g in the reference did not have expression with the RNAseq data (in strain BG2), and it was also removed. CAGL0A00132g was one multi-exon gene (212 nt ORF) with long intron (739 nt), and it was not supported by RNAseq data and removed. CAGL0H00110g is located in the terminal 5 kb region of Chr D. This putative ORF shares low-level homology with the C-termini of many terminal GPI-CWP genes and likely represents the remnants of a deletion event. RNAseq data (in strain BG2) did not show gene-associated transcription for CAGL0H00110g (data not shown), and we removed it from the annotation. Three small spurious ORFs in rDNA regions were removed. Three were in the repetitive ChrC region and removed because of re-annotation. Two were the CAGL0I02816g and CAGL0I02838g that were merged as the new gene, CAGL0I02827g.



30000
25000
20000
15000
10000
5000
0
0
5000
10000
15000
20000
CAGL0J05159g in our assembly (nt)
CAGL0J05159g in the reference (nt)

(A)



(B)

Figure 3 Gene Length Difference between two assemblies

A) Comparison of CAGL0J05159g gene between the reference genome and our assembly. The dotplot is drawn with window size = 15 nt. This illustrates an example of sequence correction within tandem-repeat regions. Gaps inside the tandem-repeat regions were linkers lost in the reference genome. B) Gene length comparison between the reference and our assembly. We compared the nucleotide lengths of ORFs of all the single-exon genes we kept in our assembly against those in the reference. The GPI-anchored adhesins are colored in red, and all the other genes are colored in blue.

Classification of the GPI genes

We systematically identified the GPI proteins in our genome (Table S8). Using PredGPI predictor (Pierleoni et al., 2008), we identified 127 genes with a false-positive rate (FDR) < 0.005. We also searched for homologs to known GPI -anchored Cell Wall Proteins (GPI-CWPs) in the reference by BLASTP and found eight genes with E-value < 10-9 and with boundary FDR (< 0.03) that were classified as predicted GPI genes as well. Fungal adhesins form as subset of GPI-CWPs, and typically these contain substantial internal tandem repeat regions. For the 135 GPI-CWP genes, we generated a protein sequence self-dotplot, identifying 71 genes with internal tandem repeat regions (Figure S8). These we classify as putative adhesin-like proteins. In addition, seven GPI proteins homologous to known adhesin like proteins (E-value < 10^{-9}) were also annotated as putative adhesins. Finally, we included three additional genes (with no internal repeats and no homology to established adhesins) but which have been annotated as GPI-CWP adhesin-like genes in the reference. Of these 81 GPI-CWP adhesin-like proteins, 34 are identical in both assemblies, two have small sequence variation (length difference < 50 nt), 21 are structural variants compared with the reference genome (length difference > 50 nt), and 24 are newly annotated.

A striking finding of our assembly is the corrected length of many genes encoding GPI-CWPs. In the reference genome, the annotated length of genes that correctly assembled except repeat regions in the reference genome ranges from 0.4 kb to 13 kb, while in our assembly, the range is 0.4 kb to 29.5 kb, corresponding to tandem repeat arrays that are more than twice as long as previously appreciated. 19 genes are predicted to encode proteins longer than 2500 aa, and the coding region of the longest corresponds to a protein of 9860 amino acids. This has implications, discussed below, on the organization of the cell envelope in Candida glabrata and the potential functions of these long GPI-CWPs.

We used this set of adhesin-like proteins to redefine the adhesin clusters in the C. glabrata genome, generating a bootstrap phylogenetic tree for the N-terminal regions of the proteins (Figure 4). We defined adhesin clusters by branches with bootstrap > 500 (1000 trials for bootstrap). Published adhesin clusters were defined by the N-terminal 300 amino acids (de Groot et al. 2008). We compared our clusters with the current adhesin clusters defined by de Groot et al. (de Groot et al. 2008) for the genes found in both assemblies. We found that all proteins in our clustering were assigned to the same cluster as the published adhesin clusters except CAGL0E00187g. CAGL0E00187g is annotated in adhesin cluster IV in the reference. This gene only shared homology with the C-terminus of CAGL0C00209g in cluster IV. Based on N-terminal sequence, we have assigned CAGL0E00187g as a singleton. The numbers of genes in each adhesin cluster are: 20 in cluster I; 7 in cluster II; 13 in cluster III; 2 in cluster IV; 13 in cluster V; 8 in cluster VI; 6 in cluster VII; 12 singletons.

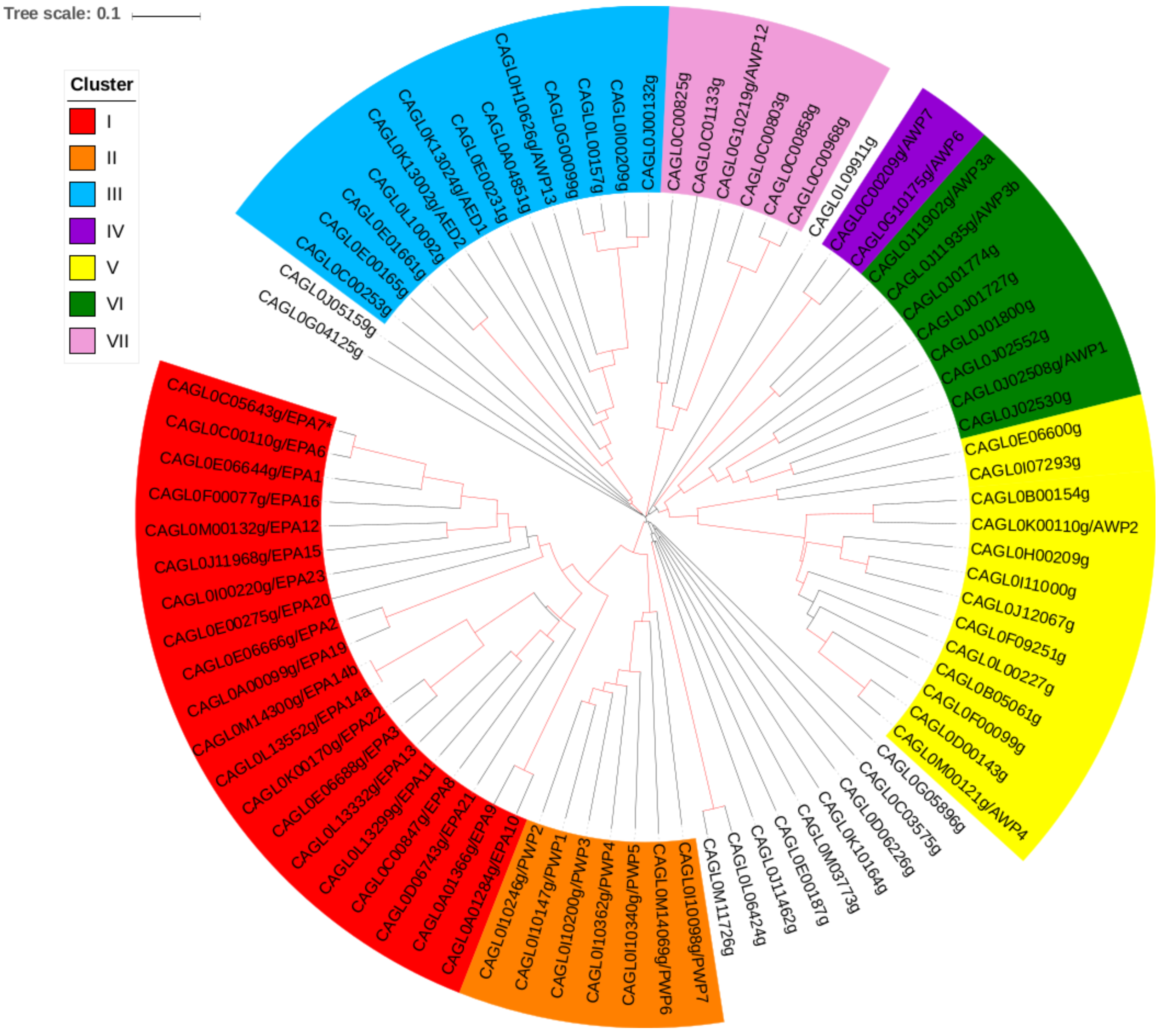


Figure 4 Phylogenetic tree of GPI-anchored adhesins.

The adhesin clusters are indicated by color, and the common name, if any, is shown in the tree as well. We extracted the N-terminal regions of the GPI-anchored adhesins as described in Methods, and used these regions to generate a bootstrap phylogenetic tree was with ClustalW2 with seed=111 and 1000 bootstrap trials. The branches with bootstrap number > 500 were colored in red. All the branches with bootstrap > 500 were first collapsed together to form clusters. Cluster I and II, V and VI were further classified based on current annotation. Our clusters are generally the same as with the current adhesin nomenclature (de Groot et al. 2008), and we named our clusters according to the current annotation. CAGL0L06424g and CAGL0M11726g are two small genes that form a cluster not previously annotated.

Reference

Andes, D.R., Safdar, N., Baddley, J.W., Alexander, B., Brumble, L., Freifeld, A., Hadley, S., Herwaldt, L., Kauffman, C., Lyon, G.M., et al. (2016). The epidemiology and outcomes of invasive Candida infections among organ transplant recipients in the United States: results of the Transplant-Associated Infection Surveillance Network (TRANSNET). Transpl. Infect. Dis. *18*, 921–931.

Astvad, K.M.T., Johansen, H.K., Røder, B.L., Rosenvinge, F.S., Knudsen, J.D., Lemming, L., Schønheyder, H.C., Hare, R.K., Kristensen, L., Nielsen, L., et al. (2018). Update from a 12-Year Nationwide Fungemia Surveillance: Increasing Intrinsic and Acquired Resistance Causes Concern. J. Clin. Microbiol. *56*.

Barber, A.E., Weber, M., Kaerger, K., Linde, J., Gölz, H., Duerschmied, D., Markert, A., Guthke, R., Walther, G., and Kurzai, O. (2019). Comparative Genomics of Serial Candida glabrata Isolates and the Rapid Acquisition of Echinocandin Resistance during Therapy. Antimicrob. Agents Chemother. *63*.

Beerenwinkel, N., Günthard, H.F., Roth, V., and Metzner, K.J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. *3*, 329.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. *15*, 1456–1461.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Carreté, L., Ksiezopolska, E., Gómez-Molero, E., Angoulvant, A., Bader, O., Fairhead, C., and Gabaldón, T. (2019). Genome Comparisons of Candida glabrata Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations. Front. Microbiol. *10*, 112.

Castano, I., Kaur, R., Pan, S., Cregg, R., Penas, A.D.L., Guo, N., Biery, M.C., Craig, N.L., and Cormack, B.P. (2003). Tn7-based genome-wide random insertional mutagenesis of Candida glabrata. Genome Res. *13*, 905–915.

Castaño, 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.

Chaisson, M.J.P., Wilson, R.K., and Eichler, E.E. (2015). Genetic variation and the de novo assembly of human genomes. Nat. Rev. Genet. *16*, 627–640.

Chapman, B., Slavin, M., Marriott, D., Halliday, C., Kidd, S., Arthur, I., Bak, N., Heath, C.H., Kennedy, K., Morrissey, C.O., et al. (2017). Changing epidemiology of candidaemia in Australia. J. Antimicrob. Chemother. *72*, 1270.

Cleveland, A.A., Harrison, L.H., Farley, M.M., Hollick, R., Stein, B., Chiller, T.M., Lockhart, S.R., and Park, B.J. (2015). Declining incidence of candidemia and the shifting epidemiology of Candida resistance in two US metropolitan areas, 2008-2013: results from population-based surveillance. PLoS ONE *10*, e0120452.

De Las Peñas, A., Pan, S.-J., Castaño, 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.

Desai, C., Mavrianos, J., and Chauhan, N. (2011). Candida glabrata Pwp7p and Aed1p are required for adherence to human endothelial cells. FEMS Yeast Res. *11*, 595–601.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., et al. (2004). Genome evolution in yeasts. Nature *430*, 35–44.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.

Gonçalves, B., Ferreira, C., Alves, C.T., Henriques, M., Azeredo, J., and Silva, S. (2016). Vulvovaginal candidiasis: Epidemiology, microbiology and risk factors. Crit. Rev. Microbiol. *42*, 905–927.

Gordon, D., Abajian, C., and Green, P. (1998). *Consed:* A Graphical Tool for Sequence Finishing. Genome Res. *8*, 195–202.

Green, B., Bouchier, C., Fairhead, C., Craig, N.L., and Cormack, B.P. (2012). Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob. DNA *3*, 3.

de Groot, P.W.J., Kraneveld, E.A., Yin, Q.Y., Dekker, H.L., Gross, U., Crielaard, W., de Koster, C.G., Bader, O., Klis, F.M., and Weig, M. (2008). The cell wall of the human pathogen Candida glabrata: differential incorporation of novel adhesin-like wall proteins. Eukaryotic Cell *7*, 1951–1964.

Guo, X., Zhang, R., Li, Y., Wang, Z., Ishchuk, O.P., Ahmad, K.M., Wee, J., Piskur, J., Shapiro, J.A., and Gu, Z. (2019). Understand the genomic diversity and evolution of fungal pathogen Candida glabrata by genome-wide analysis of genetic variations. Methods.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Kumar, K., Askari, F., Sahu, M.S., and Kaur, R. (2019). Candida glabrata: A Lot More Than Meets the Eye. Microorganisms *7*.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Maestre-Reyna, M., Diderrich, R., Veelders, M.S., Eulenburg, G., Kalugin, V., Brückner, S., Keller, P., Rupp, S., Mösch, H.-U., and Essen, L.-O. (2012). Structural basis for promiscuity and specificity during Candida glabrata invasion of host epithelia. Proc Natl Acad Sci USA *109*, 16864–16869.

Nattestad, M., and Schatz, M.C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics *32*, 3021–3023.

Pappas, P.G., Lionakis, M.S., Arendrup, M.C., Ostrosky-Zeichner, L., and Kullberg, B.J. (2018). Invasive candidiasis. Nat. Rev. Dis. Primers *4*, 18026.

Pierleoni, A., Martelli, P.L., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. BMC Bioinformatics *9*, 392.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics *13*, 341.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. Trends Genet. *16*, 276–277.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol. *14*, R51.

Thierry, A., Bouchier, C., Dujon, B., and Richard, G.-F. (2008). Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in Candida glabrata. Nucleic Acids Res. *36*, 5970–5982.

Timmermans, B., De Las Peñas, A., Castaño, I., and Van Dijck, P. (2018). Adhesins in Candida glabrata. J. Fungi (Basel) *4*.

Vale-Silva, L., Beaudoing, E., Tran, V.D.T., and Sanglard, D. (2017). Comparative Genomics of Two Sequential Candida glabrata Clinical Isolates. G3 (Bethesda) *7*, 2413–2426.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE *9*, e112963.

Zupancic, M.L., Frieman, M., 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.

Chapter 3 *De novo* assembly of *C. glabrata* BG2 strain

Introduction

In this chapter, I performed *de novo* genome assembly of the *C. glabrata* strain BG2 and all analysis. The subcloning of *C. glabrata* subtelomeres into fosmids, and Sanger sequencing of the subcloned fosmids was carried out by a previous lab member, Brian Green. The RNAseq experiment was performed by a former lab member, Shi-jung Pan. Brendan Cormack prepared the genome DNA of *C. glabrata,* and the PacBio sequencing is done by Haiping Hao at the Transcriptomes and Deep Sequencing Core in Hopkins.

The *C. glabrata* BG2 strain is a clinical isolate and a major strain used for genetic studies (Castano et al., 2003; Castaño et al., 2005; De Las Peñas et al., 2003; Domergue et al., 2005; Frieman et al., 2002; Kaur et al., 2005, 2007; Zupancic et al., 2008). BG2 strain encodes strain specific sets of GPI-CWPs relative to the type strain ATCC2001 (CBS138). For instance, there are reportedly 23 *EPA* or *EPA*-like sequences in the *C. glabrata* clinical isolate BG2; 6 of the 23 sequences are specific to BG2 (Kaur et al., 2005). In Chapter 2, we performed *de novo* genome assembly of strain CBS138 using long SMRT sequencing reads. We obtained a high quality genome with the correct sequences of GPI-CWPs, and corrected the mis-assemblies in the CBS138 genome. In this chapter, we apply the long SMRT sequencing reads of BG2 strain to generate a high-quality genome of BG2.

We employed a similar *de novo* genome assembly pipeline discussed in Chapter 2, and obtained a high-quality genome sequence of BG2. We identified the complete and accurate sequence of GPI-CWPs encoding genes in BG2, which will benefit further studies of the GPI-CWPs in this strain. In addition, we discovered significant sequence variation in the structures of GPI-CWP genes relative to the orthologue in CBS138. These sequence variations can result from non-allelic mitotic recombination events, which will be further investigated in Chapter 5. We also analyzed chromosome rearrangement events between BG2 and CBS138 and found translocation events both in the chromosome body and in subtelomeres. We identified two retrotransposons and the associated LTRs across the genome. The retrotransposons encode ORFs homologous to the *TKP5* genes in DSY562 strain (Vale-Silva et al., 2017); the CBS138 genome does not encode any complete retrotransposons. The genome also has many copies of the corresponding LTRs and these are located at similar locations in BG2, DSY562 and CBS138..

Experimental Procedures

PacBio library preparation

Genomic DNA of strain BG2 was made by preparation of spheroplasts, followed by lysis and spooling of high molecular weight genomic DNA after ethanol precipitation. The genomic DNA of the four isolates were sheared to 30 kb with MegaRupter . Samples were diluted to under 50 ng/µl per the MegaRupter protocol with TE. Samples were then purified using a 1x AMPure bead cleanup and elutied in 62 µl of EB. Libraries were prepared using the standard PacBio 30kb protocol. Briefly, after fragmentation DNA end repaired and A-tailed followed but SMRTbell hairpin ligation. Upon completion of the library, samples were then size selected on the Blue Pippin to the range of 20-50kb. The size selected library was then sequenced on PacBio RS II using PacBio P6/C4 chemistry.

De novo genome assembly of strain BG2

The de novo assembly of the genome using the SMRT sequencing reads was performed using CANU 1.5 (Koren et al., 2017) on the cluster of the Maryland Advanced Research Computing Center (MARCC). The default CANU protocol was applied with the following parameter adjustments: 1) corOutCoverage=300; 2) genomeSize=12.3m. The draft contigs were polished by Arrow from the GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus) with the same PacBio reads. We performed whole genome sequence alignment by NUCMER from MUMMER3 (http://mummer.sourceforge.net/) with the CBS138 genome (Dujon et al., 2004)Xu et al in print(Dujon et al., 2004) to assign contigs to chromosomes which the largest alignment, and used the telomere seed sequence (GGGGTCTGGGTGCTG) to locate the telomere repeats in each contig.

Sequence qualification and further correction

We aligned the Illumina sequencing reads to our assembly by Bowtie2 (Langmead and Salzberg, 2012) with default settings. We polished our assembly with the aligned reads using Pilon 1.22 (Walker et al., 2014). The aligned reads are further filtered for properly paired reads by samtools (Li et al., 2009) to calculate the per base read coverage, which was counted by BEDTools (Quinlan and Hall, 2010). We calculated the mean read coverage in 100 bp windows of the 13 chromosomes to detect mis-assemblies in repeat regions. In addition, we verified SNPs in rRNA genes using Illumina reads. We aligned the Illumina sequencing reads to the first rDNA copy on ChrL using Bowtie2 (identified by BLASTN alignment of the rDNA repeat in CBS138 genome), and filtered for the properly paired reads in the same way. The rRNA genes were identified using BLASTN alignment of the rRNA genes in the CBS138 genome. The frequencies of SNPs were calculated from the aligned reads using IGV (Robinson et al., 2017) (Table S1). To further validate the structure of the highly repetitive subtelomere structure, we compared the subtelomeric regions in our assembly against the subtelomeres subcloned into fosmids (see below). For the comparison of our subtelomeres to subcloned fosmids we used dotplots with kmer size = 15 nt (Figure S2).

To assess structural variation, we used Assemblytics (Nattestad and Schatz, 2016) to extract the structural variants (Table S2). Chromosome rearrangements were annotated after annotation of the BG2 genome (see below). Non-syntenic genes define the boundaries of the rearrangements. We extracted intergenic regions of non-syntenic genes and used BLASTN to align these regions to the CBS138 genome, and found the boundaries of rearrangements. The rearrangement events are visualized by CIRCOS (Krzywinski et al., 2009).

Targeting fosmid construction and integration

To clone subtelomeric regions of the chromosome, we targeted a fosmid integration to the telomeric ORF within the most telomeric block of two ORFs that are syntenic between *C. glabrata* and *S. cerevisiae* for each telomere. Fusion PCR was used to generate, for each targeting sequence, an approximately 1,000 bp MluI to SacII fragment using BG2 genomic DNA as the template. The first round of PCRs yielded approximately 500 bp fragments using the "left AS" and "left S" and the "right AS" and "right S" pairs (Table S3). A PCR purification kit was used to isolate those two fragments, which were combined and used as a template with the appropriate "left S" and "right AS" oligos. The resulting fragment, containing an internal *Ppi*I site, was ligated into *Mlu*I to *Kpn*I and *Kpn*I to *Sac*II fragments of pBAC-NAT (Green et al., 2012). These targeting plasmids were digested with *Ppi*I and used to transform BG2, followed by selection on plates contained clonNAT. Correct integrants were identified by PCR.

Telomere cloning in Fosmids, Sequencing and Assembly

To prepare genomic DNA, the cell pellet from 1.5 mL of YPD stationary phase culture of each integrant was resuspended in 250 µL zymolyase buffer (1.2 M sorbitol, 10 mM tris pH 8, 10 mM CaCl2, 1% beta-mercaptoethanol, 0.7 mg/ml zymolyase). After approximately 30 minutes at 37 C, 200 µL of lysis buffer (50 mM tris pH 8, 50 mM EDTA, 1.2% SDS) was added to each tube and samples were inverted to mix. Next, 100 µL of 3M NaAc pH 5.2 was added, followed by inverting the tube to mix and centrifugation in a microfuge at full speed for 10 minutes. The supernatant was transferred to a new tube, centrifuged for another 5 minutes, and then transferred again. Isopropanol precipitated DNA was resuspended in 300 µL of TE plus 1.5 µL 10 mg/ml

RNAse, and incubated at 37 C for 30 to 60 minutes. After another isopropanol precipitation, the DNA was resuspended in 50 μL TE.

To clone the subtelomeric region, the DNA from the integrated fosmid to the telomere was liberated and circularized. First, AscI was used to cleave inside the fosmid and release the end, after which the AscI was heat killed. The ends were blunted with T4 polymerase, which was then heat killed. After a three-fold dilution, the fosmid was circularized with T4 ligase, precipitated, and transformed into MegaX DH10B cells. Colony PCR was used to check for fosmid to telomere junction suggestive of full-length clones with ON3629 and ON3654 (Table S3).

We sequenced fosmids by generating transposon insertions using Tn7 transposition was used as described (Castano et al., 2003). Tn7 was chosen because it has a less pronounced sequence bias than other transposons (Green et al., 2012). After transposition and transformation into MegaX DH10B cells, colonies were picked and arrayed into 96 well plates, initially 2 per telomere rescue. Fosmids were then sequenced using Sanger sequencing with paired reads out of both ends of the transposon with ON661 and ON662 (Table S4). A standard phred/phrap/consed (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998) assembly pipeline was used with default parameters to assemble the sequencing reads. Some repeat regions were longer than could be resolved with standard Sanger sequencing, and those fosmid assemblies contain truncated tandem repeat arrays.

RNAseq analysis of the subtelomere-activated BG2 strain

The single-end RNA sequencing reads from two biological replicates were aligned together to our BG2 assembly using Bowtie2 with default settings. The per base coverage was calculated by BEDTools. We also used IGV to manually check the expression pattern.

Genome annotation and gene comparison

Our reannotation was very conservative, preserving as much as possible the systematic names in the CBS138 genome. All multi-exon genes in the reference genome were aligned to our draft assembly by BLASTN in BLAST 2.6.0+ (Camacho et al. 2009) and directly annotated. For single-CDS genes, we performed *de novo* ORF calling for single-CDS genes (all possible ORFs in all six frames of BG2 genome), and aligned all the predicted ORFs with the single-CDS ORFs in CBS138 using BLASTN, and kept all predicted ORFs that have in-frame alignments. If an ORF was a reciprocal best hit with an annotated gene, we assigned that systematic name. For all other ORFs (in multigene families) with multiple BLASTN hits, we used synteny information to assign systematic names, i.e., we assigned the ORF to the unassigned reference gene that shared homology if it was in synteny with the reference genes assigned to neighboring ORFs. We further checked ORFs with alternative start codons by RNA seq data (Figure S6). Single-CDS genes with frame-shift generate multiple ORFs homologous to the same gene in CBS138, and we manually merged these ORFs, and note the presence of a frame shift in the ORF annotation. After the assignment of BG2 ORFs to CBS138 ORFs, we manually annotated the BG2 specific ORFs that shared homology with CBS138 ORFs with new systematic names. For ORFs with no homology to annotated CBS138 ORFs, we checked the RNAseq coverage using IGV, and annotated genes supported by RNAseq evidence. We re-aligned the CBS138 ORFs, which are not assigned to BG2 ORFs (and therefore deleted in the BG2 annotation), to our BG2 assembly, to verify the deletion of those ORFs in the BG2 genome.

We compared the genome ORF lengths of single-CDS genes between BG2 and CBS138, and classified genes with structural variants as ORF length difference > 50 nt. The genes with structural variants were visualized by dotplot with kmer size=15 nt (Figure S5). We verified transcription of ORFs by RNAseq data (Figure S6). For ORFs with length difference less than or equal to 50 nt, we compared their ORF genome sequence as well as protein sequence by BLASTN and BLASTP, respectively, and calculated the percentage sequence identity of these ORFs.

Annotation of transposons and LTRs

We aligned the genome sequence of our novel ORFs, CAGL0K02618g and CAGL0H01952g, with protein sequences of the *TKP5-1-TKP5-9* retrotransposons by BLASTX. The flanking regions of both genes were extracted, and we defined the direct repeats in the flanking regions as LTRs associated with the two retrotransposon genes. The LTRs were aligned to the BG2 assembly by BLASTN, and we identified the LTR regions with e-value $< 10^{-9}$ (Table S5). They were also aligned to CBS138 and DSY562 strain to compare the LTR locations (data not shown).

Analysis of the GPI-CWPs

We first classified the BG2 genes assigned to putative GPI-CWPs in CBS138 as GPI-CWPs as well. Secondly, we identified all putative GPI anchored proteins using the PredGPI GPI anchor Predictor (Pierleoni et al., 2008) for all the ORFs in our assembly. In addition, we used BLASTP to align the BG2 specific ORFs to CBS138 GPI-CWPs, and filtered by e-value $< 10^{-9}$ and classified these as additional potential GPI-CWPs (Table S4). We analyzed adhesin like genes and assigned them into clusters by a neighbor-joining phylogenetic tree with bootstrap values (1000 bootstraps) of the N-terminal domains of the GPI adhesin like proteins (Figure 8, Figure S7). We identified these N-terminal domains as follows: GPI-anchored adhesin like genes have an N-terminal region followed by a repeat containing region. Operationally, we defined the N-terminal regions of the GPI-anchored adhesin-like proteins as the region preceding repeat containing sequence. To identify repeat sequences, we analyzed the occurrence of 15-mers (5 aa) in DNA sequences (Figure S8), and counted the occurrence of 15-mers from the beginning of the gene; the first occurrence of the first 15-mer that had 3 occurrences in the sequence were defined as the start site of the repeat region. Accordingly, the N-terminal region extends from the beginning of the ORF to the amino acid preceding the repeat start site. The neighbor-joining phylogenetic tree for N-terminal regions was generated by ClustalW2 with default setting from Clustal 2.1 (Larkin et al., 2007) with seed=111 and 1000 bootstrap trials. All the branches with bootstrap values > 500 were highlighted to indicate clusters. We assigned the BG2 specific GPI-CWPs to adhesin clusters based on the phylogenetic tree and named the 6 BG2 specific *EPA* genes as *EPA4, EPA5, EPA17, EPA24, EPA25, EPA26*.

Results

Telomere to telomere assembly of BG2 genome

We sequenced *C. glabrata* strain BG2 using SMRT long read sequencing (Table S6). We obtained 44,649 reads (approximately 40-fold coverage) with a subread N50=21529 (Figure S1). We then assembled the contigs using the CANU 1.5 assembler (Koren et al., 2017). Draft contigs were polished with the same PacBio reads using Arrow from the PacBio GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus). We annotated our draft genome with reference to the ATCC2001 (CBS138) genome sequence (Dujon et al., 2004). Our draft assembly consisted of 14 contigs: the mitochondrial genome and 13 contigs which were telomere-to-telomere assemblies of all the chromosomes.

Chromosome structure

We obtained telomere-to-telomere assembly for all 13 chromosomes (Table S7). In general, the BG2 genome has the same genome structure as the type strain CBS138. However there are 11 chromosome-level rearrangements (Figure 5, Table S2): one large inversion of 631 kb in ChrL; a large reciprocal translocation in the body of ChrL and ChrI; four non-reciprocal translocations in subtelomeric regions in which the terminal gene is translocated; one reciprocal translocation between the CBS138 ChrL and ChrD left subtelomeres; one *Ty3* transposition; one inversion of the two *TIR1* paralogs, CAGL0H09592g and CAGL0H09614g; one intragenic inversion resulting from recombination within the ORF region of *PWP1* (CAGL0I10147g) and *PWP3* (CAGL0I0200g).

Sequence comparison to type strain CBS138

BG2 and CBS138 have approximately 99.06% identity in the genome sequence. We documented 79,289 SNPs and 27,272 indels between BG2 and CBS138. We describe gross chromosome rearrangements above. We also documented major structural variation events (genome sequence length difference > 50 nt) most of which are located within repeat regions (Table S2): 42 insertions and deletions, accounting for 30,725 bp change in genome sequence; 29 expansions and contractions within tandem-repeat regions, accounting for 67,448 bp change in genome sequence; 27 expansions and contractions in other repeat regions that make a total change of 65,531 bp.

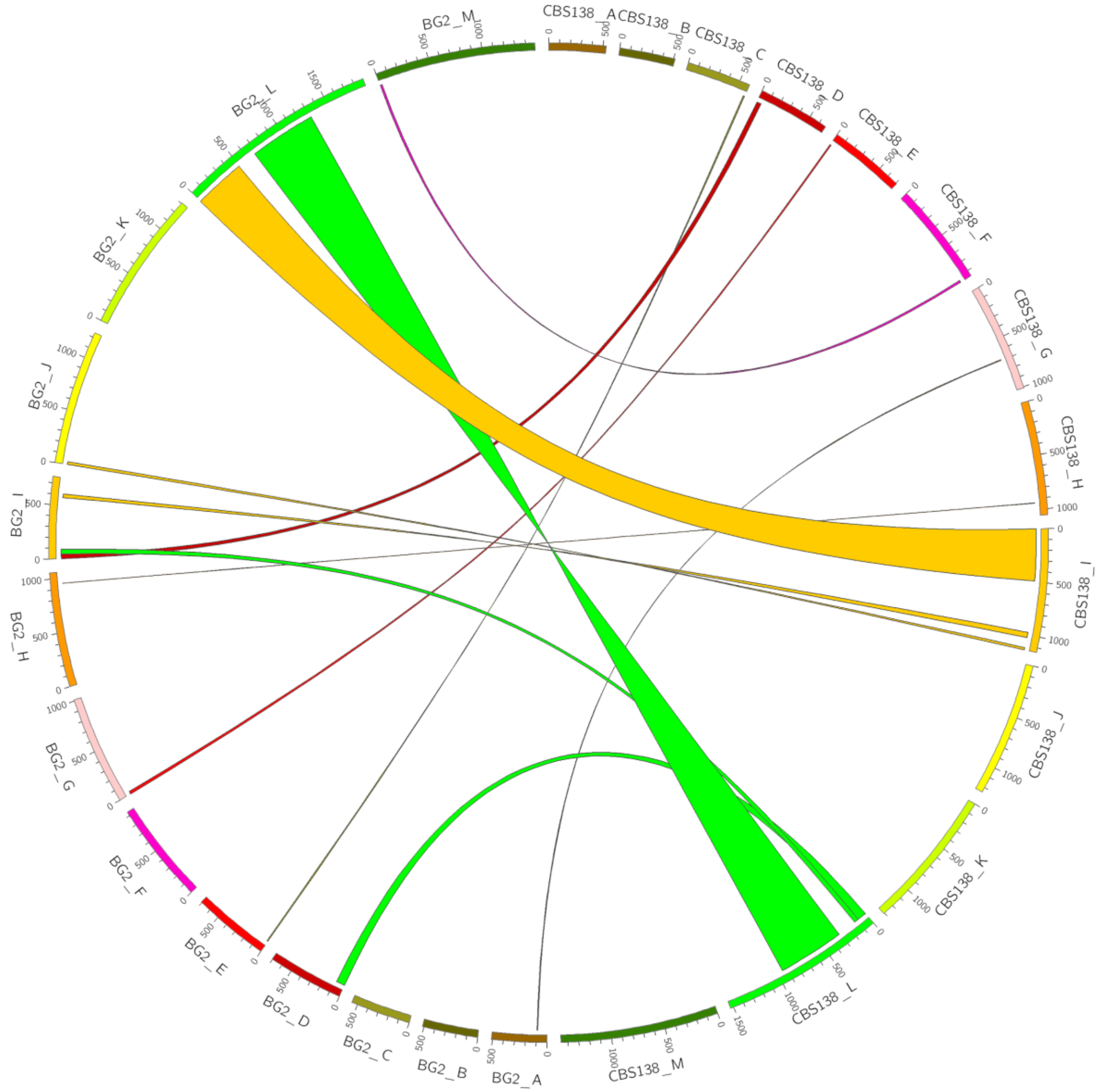


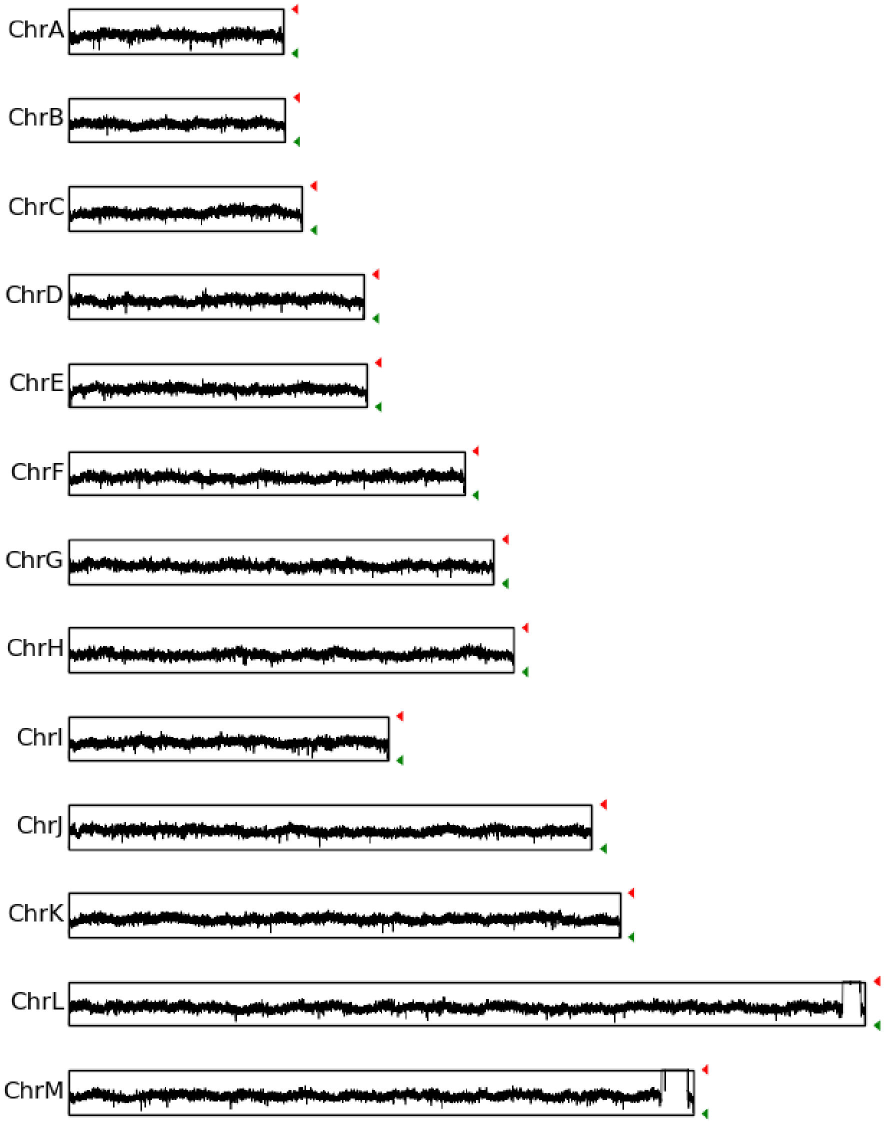
Figure 5 Chromosome rearrangements

The Chromosome rearrangements in our BG2 assembly relative to CBS138 genome are illustrated using CIRCOS (Krzywinski et al., 2009). The rearrangement events are colored using the color of the CBS138 chromosome. The 11 chromosome rearrangement events are identified using translocated genes in BG2 relative to CBS138, and the translocation sites are identified using BLASTN alignment of the intergenic regions around the translocated genes (Table S2).

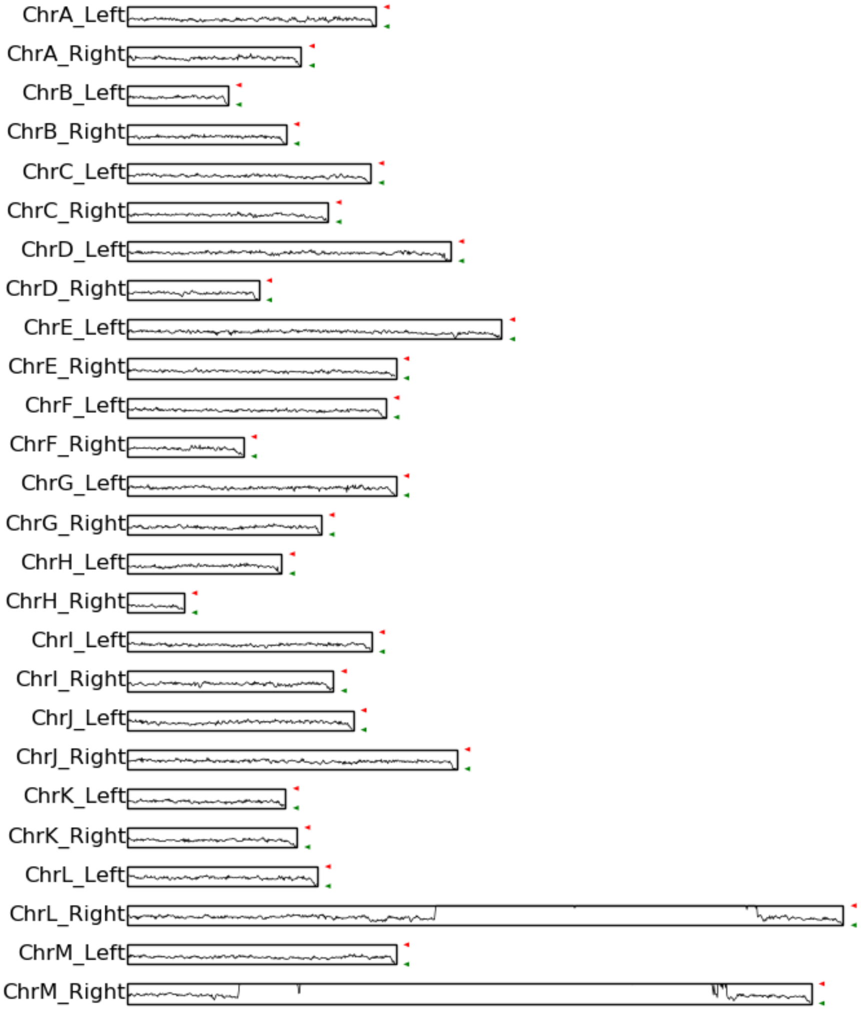
Validation of BG2 genome assembly

To verify the quality of our genome assembly particularly in the repeat regions, we assessed read depth for Illumina pair-end short sequencing reads. The short reads were aligned by Bowtie2 with default settings, and we filtered for properly paired reads (Langmead and Salzberg, 2012). Multi-mapping reads (mostly aligned to large tandem-repeat arrays) were randomly distributed across potential target sites. The read coverage of our assembly shows an even distribution over all chromosomes (Figure 6). To identify mis-assembled repeat regions, we compared the mean read coverage in 100 bp windows in non-rDNA regions. The average Illumina read coverage is 40.03, and there are 160 windows (16 kb in genome length) that have a mean read coverage > 59 (> 3 standard deviations) which is less than the expected 237 windows (23.7 kb) for 12.6 Mb genome. The windows with high read coverage are small and distributed in the genome, demonstrating that our assembly has a high fidelity sequence across the repeat regions.

The subtelomeric regions are highly repetitive, and they share substantial homology with different subtelomeres in multiple chromosomes. Therefore, we confirmed the structure of our assembled subtelomeres by cloning and individually sequencing each subtelomere, using Sanger sequencing. The near exact alignment for all non-repetitive regions between our assembled subtelomeres and the cloned fosmids demonstrates that our draft assembly contains the correct structure of the subtelomeric regions (Figure S2).


ChrA
ChrB
ChrC
ChrD
ChrE
ChrF
ChrG
ChrH
ChrI
ChrJ
ChrK
ChrL
ChrM

(A)



(B)

Figure 6 Illumina read coverage

The 50-bp pair-end Illumina reads were aligned to our assembled genome by Bowtie2, and the properly paired reads were filtered by samtools. The per-base read coverage was calculated by BEDTools, and we drew the average read coverage in 100 bp windows. A) Read coverage of the entire assembly. The green triangle indicates 0 coverage, and the red triangle indicates coverage greater than or equal to 200. The Illumina reads were evenly distributed in the genome, indicating the correct structure of tandem-repeat regions in our assembly. Only the rDNA regions had an odd read coverage peak. B) The read coverage of sub-telomeric regions. The green triangle indicates 0 coverage, and the red triangle indicates coverage greater than or equal to 200. Left and Right subtelomeres are both drawn in a centromere to telomere orientation.

Subtelomere Structure

Subtelomeric regions contain half of the highly repetitive GPI-CWPs, and these regions differ between strains. For characterizing these regions, we use an operational definition for the subtelomere - namely, the chromosome region telomeric to the coding regions of the last pair of genes syntenic between *C. glabrata* and *S. cerevisiae* (or the putative ancestor) from the Yeast Gene Order Browser (Byrne and Wolfe, 2005). We obtained the correct structure of the BG2 subtelomeric regions and annotated them (Figure 7 , Table S8). There are 72 genes located in the subtelomeres, 50 of which encode GPI-CWPs. Seven subtelomeric genes are specific to BG2 relative to the CBS138 type strain, and all of them are GPI-CWPs. The subtelomeres in the BG2 strain in general have the same structure with that of the CBS138 genome (Figure S3), with four translocations relative to the CBS138 strain, all of which are associated with the terminal GPI-CWPs (Figure S4): CAGL0I11000g translocated from CBS138 ChrI_Right subtelomere to BG2 ChrJ_Left subtelomere; CAGL0F09251g translocated from CBS138 ChrF_Right subtelomere to BG2 ChrM_Left subtelomere; *EPA7* (CAGL0C5643g) translocated from CBS138 ChrC_Right subtelomere to BG2 ChrE_Left subtelomere; CAGL0E00165g at CBS138 ChrE_Left subtelomere translocated to BG2 ChrG_Left subtelomere. All the translocations are non-reciprocal; as a consequence, there are five BG2 specific GPI-CWPs in the corresponding BG2 loci involved in the translocations, and three CBS138 specific GPI-CWPs deleted at the translocation target loci. In addition, the BG2 specific *EPA24* (CAGL0A00143g) gene is inserted between CAGL0A00165g and *EPA19* (CAGL0A00099g) and the BG2 specific *EPA17*(CAGL0H10648g) gene replaced the CBS138 specific *AWP13* (CAGL0H10626g) gene.

The *C.glabrata* ChrL and ChrM left subtelomeres contain the rDNA arrays with a terminal rDNA downstream region. Consistent with our findings in the assembly of CBS138 assembly (Xu *et al.* in revision), we found a region downstream of the rDNA arrays, encoding the *EPA14* gene; the two *EPA14* copies are identical to each other, but relative to CBS138, both BG2 copies of *EPA14*

ORF have significant changes in tandem-repeat structure relative to the CBS138 copies of *EPA14* (Figure S5).

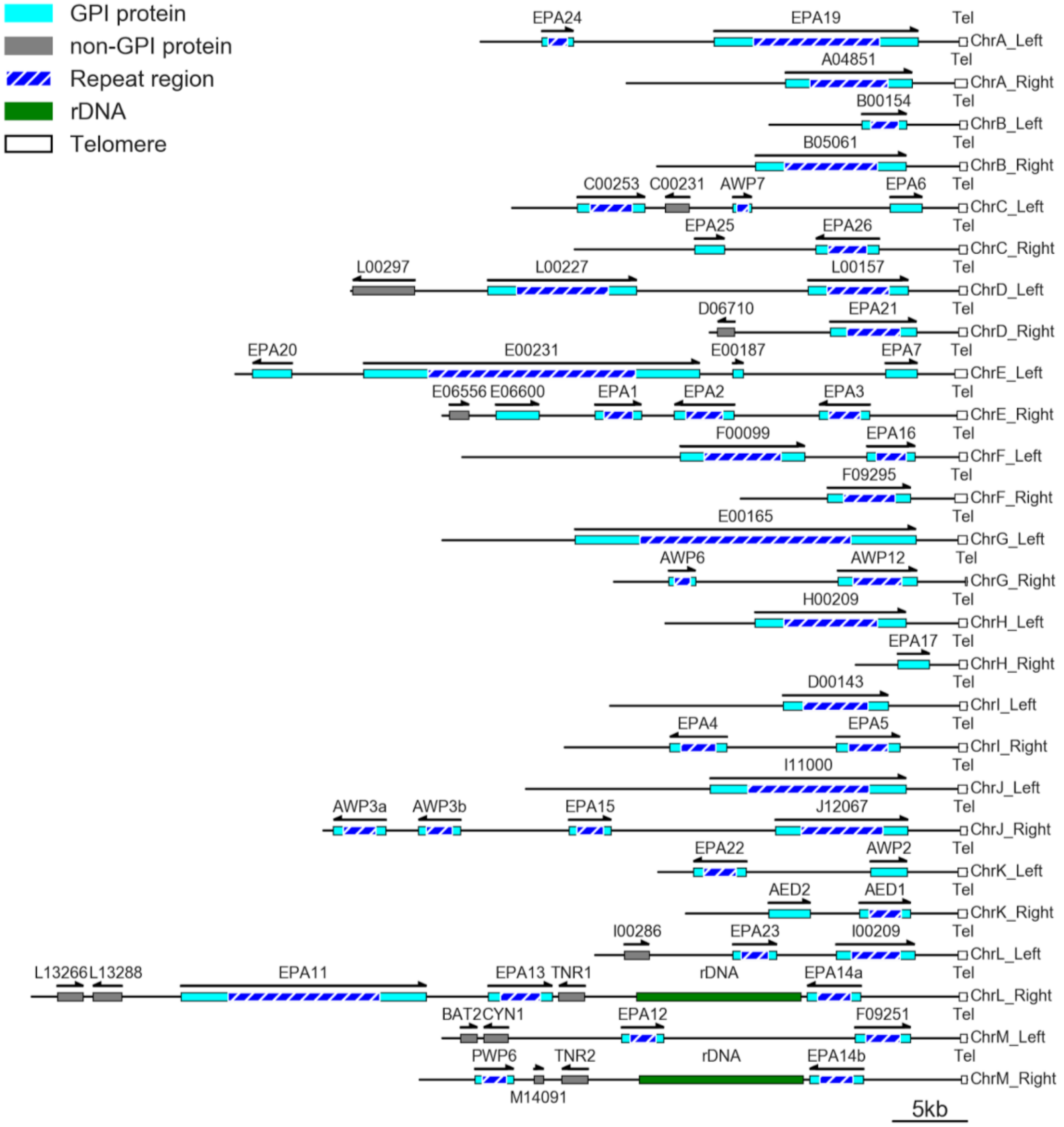


Figure 7 Subtelomere structure

Location, size and orientation of ORFs in the subtelomeric regions of the BG2 assembly. All ORF lengths are indicated to scale. ORFs for predicted GPI-anchored adhesins are colored in light blue, and all other genes are colored in grey. Tandem repeats are indicated with a shadowed dark blue box (not to scale). One copy of the rDNA repeat (11 kb) is indicated to represent the rDNA cluster. Telomere repeats are indicated by a black box.

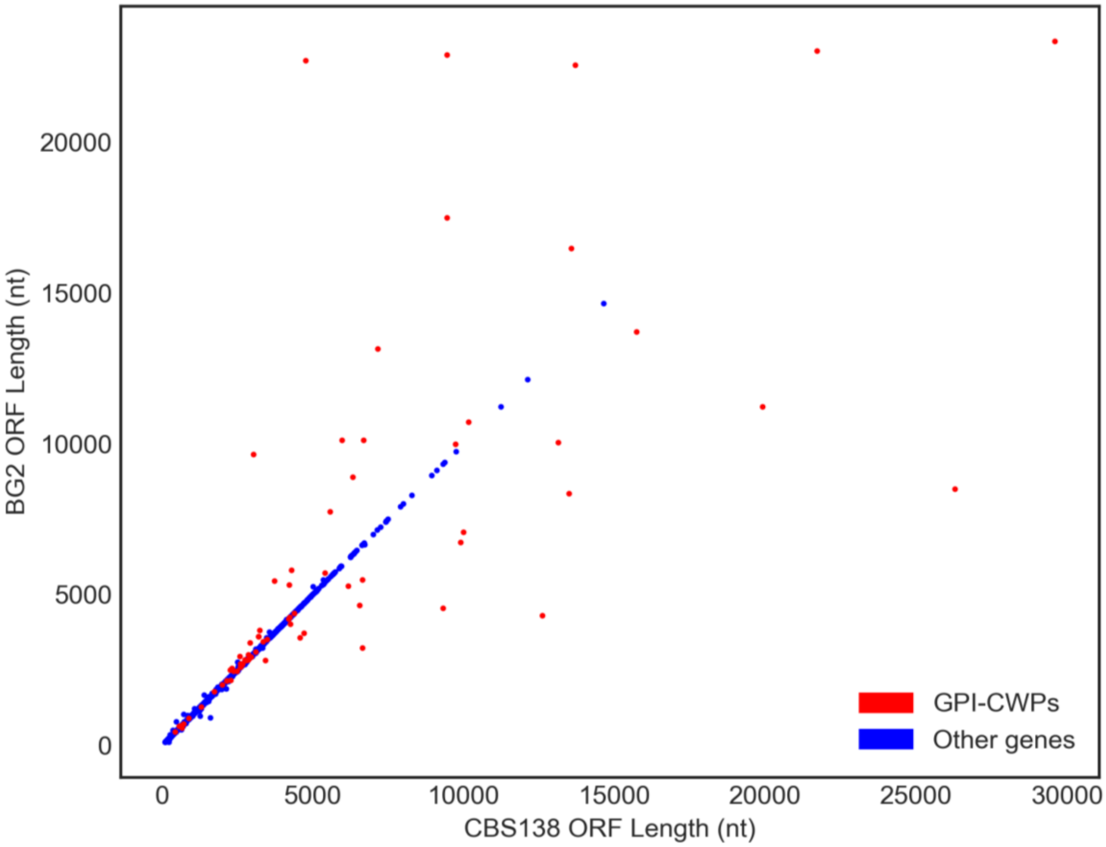
Gene comparison

The genes in BG2 are in general highly homologous to their orthologs in CBS138; there are, however, significant changes in many genes, especially in the GPI-CWPs (Figure 8). There are 5235 coding genes in our BG2 assembly. 138 are multi-exon genes with introns located within the ORF and all these genes are shared with the CBS138 strain and identical in length between the two strains.. Of the 5097 single-exon genes (single-exon genes as well as multi-exon genes with intron(s) only in the 5' or 3' UTR), 1865 encode proteins with the same sequence, 3123 encode proteins with small variations (amino acid substitutions or ORF genome length difference less than or equal to 50 nt), and 91 genes contain structural variations (ORF genome length difference > 50 nt between BG2 and CBS138). Finally, there are 18 BG2 specific genes (Table S9). The average ORF genome sequence identity of genes with the same protein sequence is 99.7%, and that of the genes with small variations is 99.3%. The average protein sequence identity of genes with small variants is 99.2%.

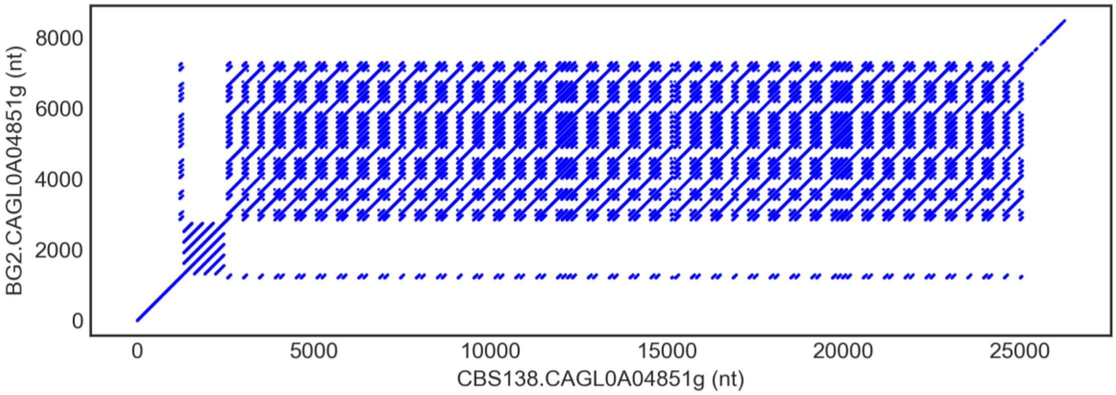
For the 91 structural variant genes (Figure S5, Figure S6): 64 have changes in repeat regions; 19 have an alternative start codon relative to the CBS138 annotated ORF. For these genes we used RNAseq data to show that the start codon falls appropriately within the transcriptional unit (Figure S6). 51 of the 64 genes with change in repeat regions encode GPI-CWPs, which are also the genes with the largest change in ORF length (Figure 8), indicating that the GPI-CWPs are the group of genes with the largest variation between strains BG2 and CBS138. The total absolute ORF length differences between shared genes encoding GPI-CWPs is 158.9 kb, and the total ORF length of shared GPI-CWPs of BG2 and CBS138 is 452.6 kb and 439.4 kb, respectively. Thus, the variant regions constitute in total approximately one third of the total ORF length of GPI-CWPs genes.

In addition to the changes in GPI-CWPs, we identified copy number variation within the *UBI4* homolog CAGL0D06226g. *UBI4* is the poly-ubiquitin precursor that encodes multiple copies of ubiquitin as tandem-repeats. There are seven copies of ubiquitin in CAGL0D06226g in the CBS138 strain, but only four copies in the BG2 strain. Notably, for CAGL0D06226g, although all the protein sequences of each tandem-repeat are identical, there is significant variation between the corresponding nucleotide sequences. Considering the CAGL0D06226g gene in both CBS138 and BG2 strain, there are in total 11 ubiquitin repeats, all of which are different from each other. The pairwise sequence identity ranges from 92.1% to 99.1% with an average of 95.7% (corresponding to approximately 11 SNPs in each 228 nt repeat unit.

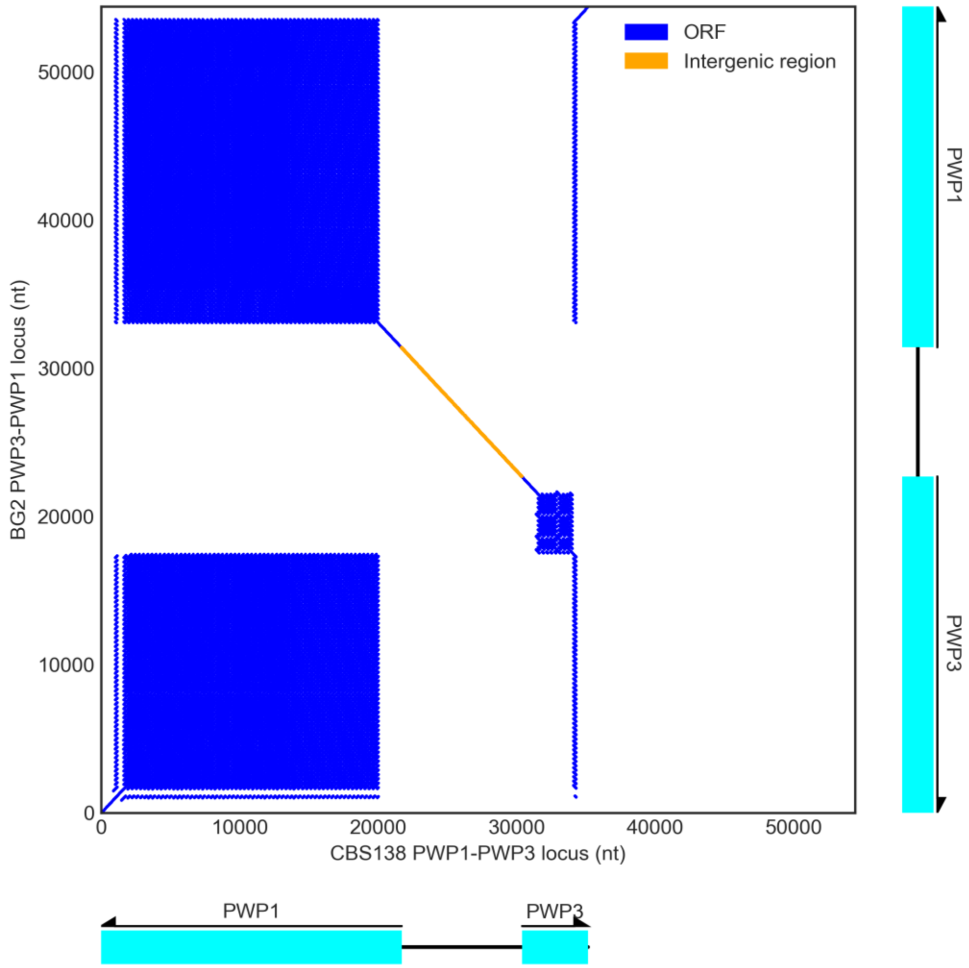
For the 18 BG2 specific genes, seven are GPI-CWPs, six of which are genes in the *EPA* family (*EPA4, EPA5, EPA17, EPA24, EPA25, EPA26*); CAGL0B02975g is the duplication of *MLT1*a (CAGL-E00341g); four are merged ORFs resulting from variations (CAGL0B02975g is the merged ORF of CAGL0B02981g and CAGL0B03014g; CAGL0K07353g is the merged ORF of CAGL0K07348g and CAGL0K07381g; CAGL0C01848g is the merged ORF of two annotated multi-exon genes CAGL0C01837g and CAGL0C01859g; CAGL0E03426g is the merged ORF of CAGL0E03421g (CDA2 homolog) and CAGL0E03432g (CDA1 homolog)); two are ORFs within retrotransposons, CAGL0H01952g and CAGL0K02618g; four are additional copies of metallothionein (*MT-II*) genes.


BG2 ORF Length (nt)
20000
15000
10000
5000
0
0
5000
10000
15000
20000
25000
30000
CBS138 ORF Length (nt)
🟥 GPI-CWPs
🟦 Other genes

(A)


BG2.CAGL0A04851g (nt)
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
CBS138.CAGL0A04851g (nt)

(B)



(C)

Figure 8 Gene comparison between BG2 and CBS138

A) Gene length comparison between the reference and our assembly. We compared the nucleotide lengths of ORFs of all the single-CDS genes we kept in our BG2 assembly against those in the CBS138 genome. The GPI-CWPs are colored in red, and all the other genes are colored in blue B) Comparison of ORF genome sequence of CAGL0A04851g in BG2 assembly and CBS138 genome using dotplot. Each blue dot represents one exact match of 15-mer. C) Illustration of intragenic recombination of *PWP1* and *PWP3* in BG2 relative to CBS138 using dotplot. Each dot represents one exact match of 15-mer. Blue dots are matches in ORFs, and the orange dots are matches in the intergenic region. The recombination of PWP1 and PWP3 results in the inversion of the N-termini of *PWP1*, *PWP3*, the *PWP3* specific repeat, and the intergenic region. The large repeat of BG2 *PWP3* that shared with *PWP1* gene is deleted in CBS138, probably during the recombination.

There are 53 CBS138 genes deleted in strain BG2: 19 gene deletions; 8 genes merged into novel ORFs (as mentioned above); 8 genes have nonsense mutations resulting in no ORF in BG2; 18 genes have nonsense mutations resulting in short ORFs (<300 nt). For these, we used RNAseq evidence to show there is no appropriate transcriptional unit corresponding to the short ORF, and therefore eliminated these from our list of annotated putative ORFs.

We identified multiple SNPs and indels in the rRNA genes relative to CBS138 genome (Table S1), consistent with substantial phylogenetic distance between strains BG2 and CBS138. The frequencies of the variants in our assembly is roughly the same as those calculated using aligned Illumina reads. All 10 *RDN25* copies in our assembly have 586T>C. One *RDN*25 gene has 1514T>C, and other two *RDN25* genes have 3070InsT. Seven *RDN18* genes have 644C>T, one of them has additional 1112delG. One *RDN18* gene only has 1112delG. The other two short rRNA genes, *RDN5* and *RDN58* do not contain variant sequence relative to strain CBS138.

Retrotransposon and LTRs

We identified two retrotransposons that include the CAGL0H01952g and CAGL0K02618g ORFs and associated LTRs consistent with these being retrotransposons. CAGL0H01952g contains an internal frameshift while CAGL0K02618g has an intact ORF. The two genes are homologs of the *TKP5* genes. These are retrotransposon associated ORFs, present in 9 copies (*TKP51-TKP59*) in the DSY562 strain (Vale-Silva et al., 2017). *TKP51-59* fall into three groups with partial homology. In BG2, CAGL0H01952g is the homolog of the *TKP53 and TKP54* retrotransposons (99% identity protein sequence) and CAGL0K02618g is the homolog of *TKP52* and *TKP55-59* retrotransposons (99% identity in protein sequence). The associated LTRs correspond to LTRs defined in DSY562. We obtained the full distribution of LTR sequences in the BG2 strain (Table

S5) Most of these share similar chromosome locations in strain BG2, DSY562 and CBS138 strains. However, the intact retrotransposons are at different chromosome locations in strains BG2 and DSY562, consistent with recent mobilization. The CBS138 genome does not encode any *Tkp5* homologs and therefore lacks any intact retrotransposons. Notably, one LTR colocalizes with a previously characterized negative transcriptional regulatory element downstream of the *EPA1* gene (Gallegos-García et al., 2012), suggesting that LTR sequences can contribute to gene regulation.

Classification of the GPI-CWPs

We identified 81 GPI-CWPs in the BG2 assembly (Figure 9, Figure S7, Figure S8). 74 are shared with CBS138 genome. Seven GPI-CWPs are specific to the BG2 strain. Seven GPI-CWPs in the CBS138 genome are deleted in BG2. All the GPI-CWPs can be classified into the adhesin clusters of de Groot *et. al (de Groot et al., 2008).* Six of the seven BG2 specific genes are *EPA* genes, *EPA4*, *EPA5*, *EPA17*, *EPA24*, *EPA25*, *EPA26*; the other one, CAGL0F09295g, is in adhesin cluster III. For the seven CBS138 specific GPI-CWPs: three of them (CAGL0C00803g, CAGL0C00825g, CAGL0C00858g) are located in the repetitive ChrC region discussed below; four genes, *AWP13* (CAGL0H10626g), CAGL0G00099g, *AWP4* (CAGL0M00121g), CAGL0J00132g, are deleted due to the non-reciprocal translocations of four other GPI-CWPs, *EPA17*, CAGL0E00165g, CAGL0F09251g, and CAGL0I11000g, respectively. 50 of 81 GPI-CWPs are located in the subtelomeric regions. There is a complex ChrC region that encodes the *EPA8* gene, and additional GPI-CWP genes. In this region, there are five additional GPI-CWP genes in the CBS138 strain and two additional GPI-CWP in BG2 assembly, all of which are cluster VII (Figure S7). The cluster VII genes in the ChrC region are closely related to each other, resulting in a highly repetitive region.

Consistent with what we observed in the sequence of strain CBS138, GPI-CWPs are strikingly large in strain BG2. The ORF length of GPI-CWPs ranges from 0.4 kb to 23.3 kb. The largest GPI-CWP gene, CAGL0J05159g with a predicted ORF length of 23.3 kb is also the largest gene in CBS138 strain, and there are 14 GPI-CWPs with ORF length > 10 kb. Notably, some of the longest genes in BG2 are much shorter in CBS138, and some long GPI-CWP genes in CBS138 GPI-CWPs are much shorter in BG2. For example, the CBS138 CAGL0A04851g with an ORF length of 26.3 kb undergoes large tandem-repeat contraction in strain BG2, resulting in an 8.5 kb ORF (Figure 8). In BG2 *PWP3* has an ORF length of 22.7 kb, while in CBS138, the ORF length is 4.8 kb ORF (Figure 8).

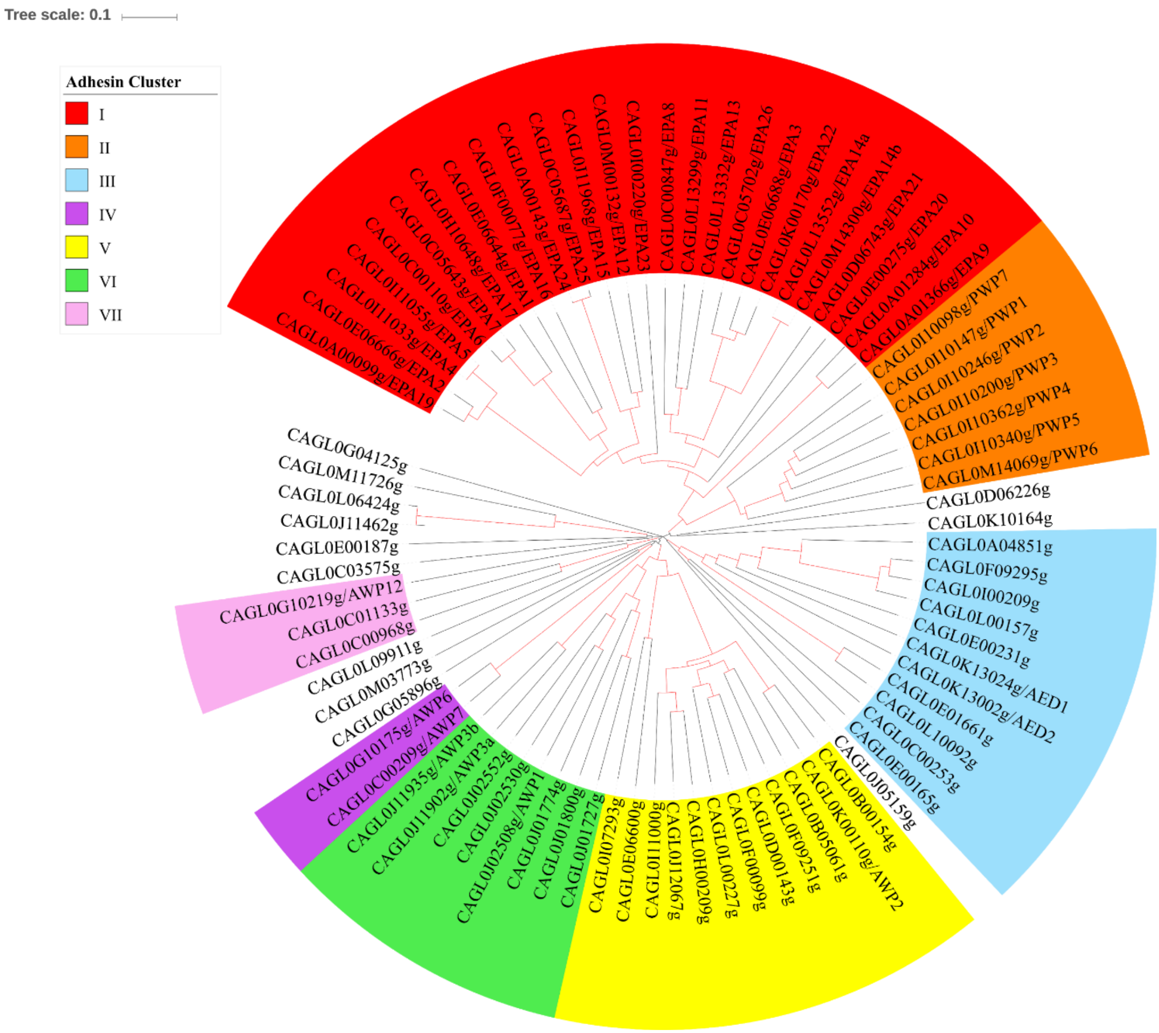


Figure 9 Phylogenetic tree of CWP-GPIs

The adhesin clusters are indicated by color, and the common name, if any, is shown in the tree as well. We extracted the N-terminal regions of the GPI-anchored adhesins as described in Methods, and used these regions to generate a bootstrap phylogenetic tree using ClustalW2 with seed=111 and 1000 bootstrap trials. The branches with bootstrap number > 500 were colored in red.

Discussion

We generated a high-quality *de novo* genome assembly of *C. glabrata* BG2 strain containing telomere-to-telomere contigs of all 13 chromosomes. Our assembly demonstrated that the 13 Mb *C. glabrata* genome can be assembled *de novo* using only the CANU assembler and P6C4 SMRT sequencing reads (with a subread length of approximately 20 kb) and with approximately 40-fold genome coverage. We verified that our assembly is reliable in highly repetitive regions. The subtelomeric regions, are highly repetitive, and also share homology between chromosomes. Even so, we show by comparison with Sanger sequencing of cloned subtelomeres that our assembly is correct. This sequence provides a high-confidence reference genome for genetic analyses in the BG2 strain.

Relative to the CBS138 genome, we identified 11 chromosome rearrangements, seven of which are translocations (Figure 5). The four non-reciprocal translocations are all in the subtelomeric regions, and all the translocated genes are the terminal GPI-CWPs. Although the translocation events are non-reciprocal, they do not break one key feature of *C. glabrata* subtelomere structure - namely that the *C. glabrata* subtelomeres end in a GPI-CWP gene, generally transcribed towards the telomere. The BG2 and CBS138 strains have strain-specific GPI-CWP genes, and these are generally the terminal gene on the respective subtelomeres, involved in translocation events. For example, in strain CBS138, the terminal ChrM right gene is CAGL0F09251g. In BG2, the CAGL0F09251g gene has translocated to the terminal position on ChrM left, and the terminal gene on ChrF right in BG2 is the BG2 specific gene CAGL0F09295g which is highly homologous to another terminal GPI-CWP, CAGL0I00209g.

The BG2 strain has a nearly identical complement of genes relative to the CBS138 strain. 5307 of the 5325 coding genes are shared between BG2 and CBS138. For the 5097 single-exon genes, 1685 genes have the same protein sequence with genome ORF sequence identity of 99.7%; 3123 have small variations (ORF genome length difference less than or equal to 50 nt) with genome ORF sequence identity of 99.3%. However, 90 genes have structural variation between BG2 and CBS138 (ORF genome length difference > 50 nt); of these, 64 are genes with changes within repeat regions, and 51 of these 64 genes are GPI-CWPs, encoding cell surface proteins. (Figure 8).

BG2 encodes 81 GPI-CWPs, seven of them are specific to BG2 relative to CSB138, of which six are *EPA* genes. All the BG2 specific GPI-CWPs are located in the subtelomeres. Apart from *EPA24*, all the BG2 specific GPI-CWPs are located at terminal of the subtelomeres (Figure S4) as discussed above. Of the seven CBS138 specific GPI-CWPs, three resulting from the gene copy number variation of the repetitive ChrC region which encodes a cluster of GPI-CWPs in adhesin cluster VII; four are terminal GPI-CWPs resulting from translocation or replacement of BG2 specific GPI-CWPs. In addition to the conserved general structure of terminal GPI-CWPs, GPI-CWPs encodes the largest ORFs in BG2. There are 14 GPI-CWPs with ORF genome length > 10 kb. The largest ORF in BG2 is the GPI-CWP gene CAGL0J05159g, encoding approximately 8,000 amino acids, which is also the largest gene in CBS138, encoding approximately 10,000 amino acids. The long GPI-CWPs in both strains support our previous hypothesis that the GPI-CWPs, which are located to the outer layer of the cell wall, encode long tandem-repeats that may permit interaction with ligands at some distance from the cell.

The GPI-CWPs are the genes with the most significant changes in BG2 relative to CBS138 (Figure 8). The total absolute ORF length difference in GPI-CWPs is 185.9 kb, which is approximately a third of the total ORF length. Interestingly, the total ORF length of GPI-CWPs between BG2 and CBS138 is similar, which is 452.6 kb and 439.3 kb, respectively. In other words, some GPI-CWP genes are longer in one strain and shorter in the other and vice versa. 51

of the 74 shared GPI-CWPs have structural variations in ORF genome sequences compared to CBS138. Many of them undergo tandem-repeat contractions and expansions, however, some GPI-CWPs have more complex changes (Figure S5). The GPI-CWPs share identical or highly similar tandem-repeats, and they are probable target sites for non-allelic recombination. Therefore, the complex changes in repeat regions may result from recombination events, and we will investigate the recombinations in future research.

Additionally, we identified tandem-repeat copy number variations of the *UBI4* homolog, CAGL0D06226g. CAGL0D06226g encodes four copies of ubiquitin in BG2 and seven copies in CBS138. Interestingly, although all the 11 copies of ubiquitin have the same protein sequence, none of them have identical genome sequence. The average pairwise sequence identity is only 95.7%. However, most of the SNPs are shared between strains are have multiple copies in each strain, indicating intragenic recombination events within the CAGL0D06226g gene.

We identified five SNPs and indels in *RDN25* and *RDN18* rRNA relative to CBS138 genome. All the CBS138 rRNA genes have the same sequence, however, different copies of BG2 rRNAs have SNPs and indels relative to each other as well. Only one SNP, 586T>C in *RDN25* is shared for all the copies. In our assembly, we have included three variants of *RDN25* gene and four variants of the *RDN18* gene resulting from SNPs and indels.

There are two *Tkp5* retrotransposons in BG2, and we identified the two different LTRs in the flanking regions of the two transposons. The *Tkp5* transposons in BG2 are highly similar to the transposons in DSY562 strain (Vale-Silva et al., 2017), and the two LTRs are the same with the LTRs annotated in DSY562 strain. BG2 and DSY562 have different locations and copy numbers of the transposon genes, and CBS138 does not encode *Tkp5* transposon. Nevertheless, most of the LTR loci (expect the flanking LTRs with the transposon genes) are shared between the three strains, demonstrating that these LTRs are derived from ancestral insertion events. Finally, one LTR maps with a negative regulatory element characterized downstream of the *EPA1* gene (Gallegos-García et al., 2012), indicating a potential role for these LTRs in gene regulation.

Reference

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. *15*, 1456–1461.

Castano, I., Kaur, R., Pan, S., Cregg, R., Penas, A.D.L., Guo, N., Biery, M.C., Craig, N.L., and Cormack, B.P. (2003). Tn7-based genome-wide random insertional mutagenesis of Candida glabrata. Genome Res. *13*, 905–915.

Castaño, 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.

De Las Peñas, A., Pan, S.-J., Castaño, 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.

Domergue, R., Castaño, I., De Las Peñas, A., Zupancic, M., Lockatell, V., Hebel, J.R., Johnson, D., and Cormack, B.P. (2005). Nicotinic acid limitation regulates silencing of Candida adhesins during UTI. Science *308*, 866–870.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., et al. (2004). Genome evolution in yeasts. Nature *430*, 35–44.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.

Frieman, M.B., McCaffery, J.M., and Cormack, B.P. (2002). Modular domain structure in the Candida glabrata adhesin Epa1p, a beta1,6 glucan-cross-linked cell wall protein. Mol. Microbiol. *46*, 479–492.

Gallegos-García, V., Pan, S.-J., Juárez-Cepeda, J., Ramírez-Zavaleta, C.Y., Martin-del-Campo, M.B., Martínez-Jiménez, V., Castaño, I., Cormack, B., and De Las Peñas, A. (2012). A novel downstream regulatory element cooperates with the silencing machinery to repress EPA1 expression in Candida glabrata. Genetics *190*, 1285–1297.

Gordon, D., Abajian, C., and Green, P. (1998). *Consed:* A Graphical Tool for Sequence Finishing. Genome Res. *8*, 195–202.

Green, B., Bouchier, C., Fairhead, C., Craig, N.L., and Cormack, B.P. (2012). Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob. DNA *3*, 3.

de Groot, P.W.J., Kraneveld, E.A., Yin, Q.Y., Dekker, H.L., Gross, U., Crielaard, W., de Koster, C.G., Bader, O., Klis, F.M., and Weig, M. (2008). The cell wall of the human pathogen Candida glabrata: differential incorporation of novel adhesin-like wall proteins. Eukaryotic Cell *7*, 1951–1964.

Kaur, R., Domergue, R., Zupancic, M.L., and Cormack, B.P. (2005). A yeast by any other name: Candida glabrata and its interaction with the host. Curr. Opin. Microbiol. *8*, 378–384.

Kaur, R., Ma, B., and Cormack, B.P. (2007). A family of glycosylphosphatidylinositol-linked aspartyl proteases is required for virulence of Candida glabrata. Proc Natl Acad Sci USA *104*, 7628–7633.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Nattestad, M., and Schatz, M.C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics *32*, 3021–3023.

Pierleoni, A., Martelli, P.L., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. BMC Bioinformatics *9*, 392.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A., and Mesirov, J.P. (2017). Variant

Review with the Integrative Genomics Viewer. Cancer Res. *77*, e31–e34.

Vale-Silva, L., Beaudoing, E., Tran, V.D.T., and Sanglard, D. (2017). Comparative Genomics of Two Sequential Candida glabrata Clinical Isolates. G3 (Bethesda) *7*, 2413–2426.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE *9*, e112963.

Zupancic, M.L., Frieman, M., 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.

Chapter 4 *De novo* assembly of four clinical serial isolates

Introduction

In this chapter, I performed *de novo* genome assembly of four serial clinical isolates of *C. glabrata* and all analysis. The serial clinical isolates are received from Jack Sobel in Wayne State University Medical School. Brendan Cormack prepared the genomic DNA of *C. glabrata,* and the PacBio sequencing was done by Sara Goodwin at Cold Spring Harbor Laboratory. The re-sequencing of the four strains was done by Haiping Hao at the Transcriptomes and Deep Sequencing Core in Hopkins.

There are multiple factors reported to influence the virulence of *C. glabrata,* including secretion of hydrolytic enzymes, ability to evade phagocytic killing and adherence to host tissues (Kumar et al., 2019). We previously obtained four serial clinical isolates from the same patient, BG3993-BG3996. The four isolates were vaginal isolates collected during office visits for a patient over a 21-month period. BG3995 and BG3996 which were isolated from later visits have a higher cell adherence than BG3993 and BG3994. We were interested in understanding the microevolution that accounts for this change in adherence during infection. We employed the *de novo* genome assembly described in chapters 2 and 3 to generate high quality genomes of the four isolates to further study the micro-evolution during infection. The genomes of the four strains are highly similar, with only 220 SNPs and indels in total between BG3994, BG3995 and BG3996 with BG3993. We identified one candidate gene, *YAP6* that likely is responsible for the change in adherence. In addition, we document novel GPI-CWPs in the serial isolates, and changes of the GPI-CWP reservoir in the clinical isolates relative to BG2 and CBS138. This indicates that the *C. glabrata* complement of GPI-CWPs can vary between strains, possibly as an adaptation to different environments.

Experimental Procedures

Genomic DNA preparation

Genomic DNA of strain BG3993-BG3996 was made by preparation of spheroplasts, followed by lysis and spooling of high molecular weight genomic DNA after ethanol precipitation.

PacBio library preparation

The genomic DNA of the four isolates were sheared to 30 kb with MegaRupter in 350 volume. Samples were diluted to under 50 ng/µl per MegaRupter protocol with TE. Samples then went through a 1x AMPure bead cleanup and eluting in 62 µl of EB. Libraries were then prepared using the standard PacBio 30kb protocol. Briefly, after fragmentation, DNA was end repaired and A-tailed followed but SMRTbell hairpin ligation. Samples were then size selected on the Blue Pippin to the range of 20-50kb. The size selected library was then sequenced on PacBio RS II using PacBio P6/C4 chemistry.

To increase the quality of our final assemblies, we sequenced the four strains a second time, and for this, the DNA was prepared differently. The sequencing library was prepared using DNA Template Prep kit v.2 (3-10kb) following the PacBio shared protocol guidelines for preparing size-selected ~20kb SMARTbell templates. Briefly, 7.5 µg of genomic DNA was diluted to 150 µl and sheared using Covaris G-tube by centrifugation in an Eppendorf 5424 microcentrifuge at 4600 rpm for 60 seconds. Sheared DNA was then purified using AMPure beads and quality controlled for concentration and size. 5 µg of sheared DNA was then used for DNA damage repair and end repair. End repaired DNA was purified using AMPure beads and ligated to SMRTbell adapter via blunt end ligation and exonuclease III/VII treatment. The exonuclease treated SMRTbell template was purified again using AMPure beads and quality controlled for yield and library size. The purified SMRTbell library was size selected on Blue Pippin with a size cut off of 5kb. The size selected library was then sequenced on PacBio RS II using PacBio P4/C2 chemistry.

De novo genome assembly

The de novo assembly was performed by CANU 1.5 (Koren et al., 2017) on the cluster of the Maryland Advanced Research Computing Center (MARCC) with the Pacbio sequencing reads. The default Canu protocol was applied with the following parameters: 1) corOutCoverage=300; 2) genomeSize=12.3m. The draft contigs were polished by arrow 2.3.2 from the GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus) with the same Pacbio reads. We performed whole genome sequence alignment by NUCMER from the MUMMER3 package (http://mummer.sourceforge.net/) with our BG2 assembly to identify contigs. We used the telomere seed sequence (GGGGTCTGGGTGCTG) to locate the telomere repeats in each contig. We made two manual corrections in the BG3994 strain: the ChrL of BG3994 was assembled in two contigs, and the ChrG chromosome lost the first gene in the assembly. To correct ChrL, we 1) used the two 20 kb flanking regions around the ChrL break point, and aligned these two regions to the BG3993, BG3995, BG3996 assemblies by BLASTN (Camacho et al., 2009); 2) the corresponding regions were aligned by MUSCLE using default parameters (Edgar, 2004). 3) The consensus sequence was generated by Consambig from EMBOSS (Rice et al., 2000); 4) we joined the BG3994 ChrL contigs with this consensus. To correct the terminal gene in ChrG, left subtelomere, we aligned the corresponding region in BG3993, BG3995, BG3996 and corrected the BG3994 ChrG in the same way as for the Chr L correction. We re-sequenced the BG3993-3996 strain using SMRT long read sequencing, and we performed *de novo* assembly with CANU 1.5 and polished with arrow using the same protocol. Procedurally, to correct sequencing errors, we started by identifying all SNPs and indels between all the 8 assemblies (each of the four isolates have 2 assemblies from two libraries of SMRT reads) using NUCMER. We corrected the probable sequencing errors (SNPs and indels between the assembly of the same strain) by comparison with the other strains: ie SNPs or indels between the two assemblies was corrected if and only if only one variant was shared with the other strains, making it likely that other variant was a sequencing error. Finally, with final draft sequences, SNPs and indels in BG3994, BG3995 and BG3996 were identified relative to the BG3993 strain using NUCMER. The SNPs and indels between BG3993 and BG2 are also identified using NUCMER. Structural variants between BG3993 and BG3994, BG3995, BG3996 as well as those between BG2 and BG3993 are identified using Assemblytics (Nattestad and Schatz, 2016).

Genome annotation and gene comparison

We first annotated the BG3993 genome. We used the genes from our BG2 assembly as the reference. The multi-exon genes were directly annotated using BLASTN. We applied the same annotation pipeline to annotate single-exon genes as that to annotate the BG2 genome in Chapter 3. Novel genes in BG3993 were aligned to the CBS138 to identify shared genes in strain CBS138. Annotated genes in BG3993 were compared with those in BG2, and all the chromosome rearrangements were identified from translocated genes. We extracted the flanking regions of the translocated genes and locate the breakpoints for rearrangements. The rearrangements are visualized using CIRCOS (Krzywinski et al., 2009). The genome of BG3994, BG3995, BG3996 was annotated using the annotation of BG3993 strain with the same annotation pipeline. The ORF genome sequences of BG3993-BG3996 strain were extracted and used to identify all genes with sequencing variants between the four strains.

Analysis of the GPI-anchored cell wall proteins (GPI-CWPs)

We classified the novel genes in BG3993 by PredGPI GPI anchor predictor (Pierleoni et al., 2008) and assigned genes with FDR < 0.005 as novel GPI-CWPs. All the discovered novel GPI genes were homologous to annotated GPI-adhesins. We extracted the N-termini of all the BG3993 GPI-CWPs to establish the phylogenetic tree, together with GPI-CWPs in BG2 and CBS138. The N-termini of the GPI-anchored adhesins were defined by counting 15 mers in genome ORF sequences with the same protocol in Chapter 3. We counted the occurrence of the 15-mers from the beginning of the gene in its genome ORF sequence, and the first occurrence of a 15-mer that had 3 occurrences in the sequence was defined as the start of the repeat region. Accordingly, the N-terminal region extends from the beginning of the ORF to the nucleotide preceding the repeat start site. The first half of the gene was identified as the “N-terminus” if that gene was not repetitive, i.e., the gene didn’t have any 15 mer that repeated three times. The bootstrap phylogenetic tree was generated by ClustalW2 with seed=111 and 1000 bootstrap trials (Larkin et al., 2007). All novel genes were additional genes within current clusters.

Results

Telomere to telomere assembly of the four clinical isolates

We sequenced four *C. glabrata* serial isolates, BG3993-BG3996 using SMRT long read sequencing (Table 1). We obtained subreads of approximately 40-fold coverage of the 13 Mb genome with N50=20 kb (Figure S1). We then assembled the contigs using the CANU 1.5 assembler (Koren et al., 2017). Draft contigs were polished with the same PacBio reads using Arrow from PacBio GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus). We further made minimum corrections to generate the four assemblies (see Experimental Procedures).

We obtained telomere-to-telomere assembly for most of the chromosomes in all four isolates (Table 2). There were two exceptions: 1) the rDNA downstream region. *C. glabrata* ChrL and ChrM encode a rDNA array at end right end. Both ChrL and ChrM share an identical terminal rDNA downstream region that encodes the *EPA14* gene (see chapter 2). The rDNA arrays on both chromosomes for strains BG3993 and 3996 were generally identical. Our BG3993 and BG3996 assemblies had one copy of the rDNA downstream region in ChrL, but not for ChrM. For BG3994 and BG3995 assembly, the ChrL and ChrM contigs did not include the rDNA downstream region , but these regions were found in the unassembled unitigs of the Canu assembly (data not shown). 2) BG3995 did not assemble the terminal 3 kb for ChrG left.

The four genomes are highly similar to each other (Figure S2). There are no chromosome rearrangements between the four strains. In addition, there are only approximately 220 SNPs and indels in BG3994, BG3995, BG3996 individually relative to BG3993 (Table S1). We used the BG3993 genome to compare genome sequences of the BG2 and CBS138 *C. glabrata* strains.

The clinical isolates share a similar chromosome structure with the BG2 strain. There are only 6 chromosome rearrangements compared to BG2 (Figure 10, Table S2). Notably, 5 of the 6 rearrangements are subtelomeric and associated with the translocation of the terminal ORFs. The genome sequence identity between BG2 and BG3993 is 98.9%. There are 89422 SNPs and 31205 indels between BG2 and BG3993. The major structural variants (genome sequence length difference > 50 nt) between BG2 and BG3993 are located in repeat regions (Table S1): 45 insertions and deletions, accounting for 35,242 bp change in genome sequence; 32 expansions and contractions in repeat regions, adding up to 126,162 bp change in genome sequence. Strikingly, there are only three structure variants between BG3993, BG3995, BG3996 and BG3993, all of which are in coding regions: BG3993 have one tandem-repeat contraction in the PWP2 gene (CAGl0I10246g); BG3994 and BG3995 have two different repeat changes in CAGL0J05159g relative to BG3993 and BG3996 (Table S1).

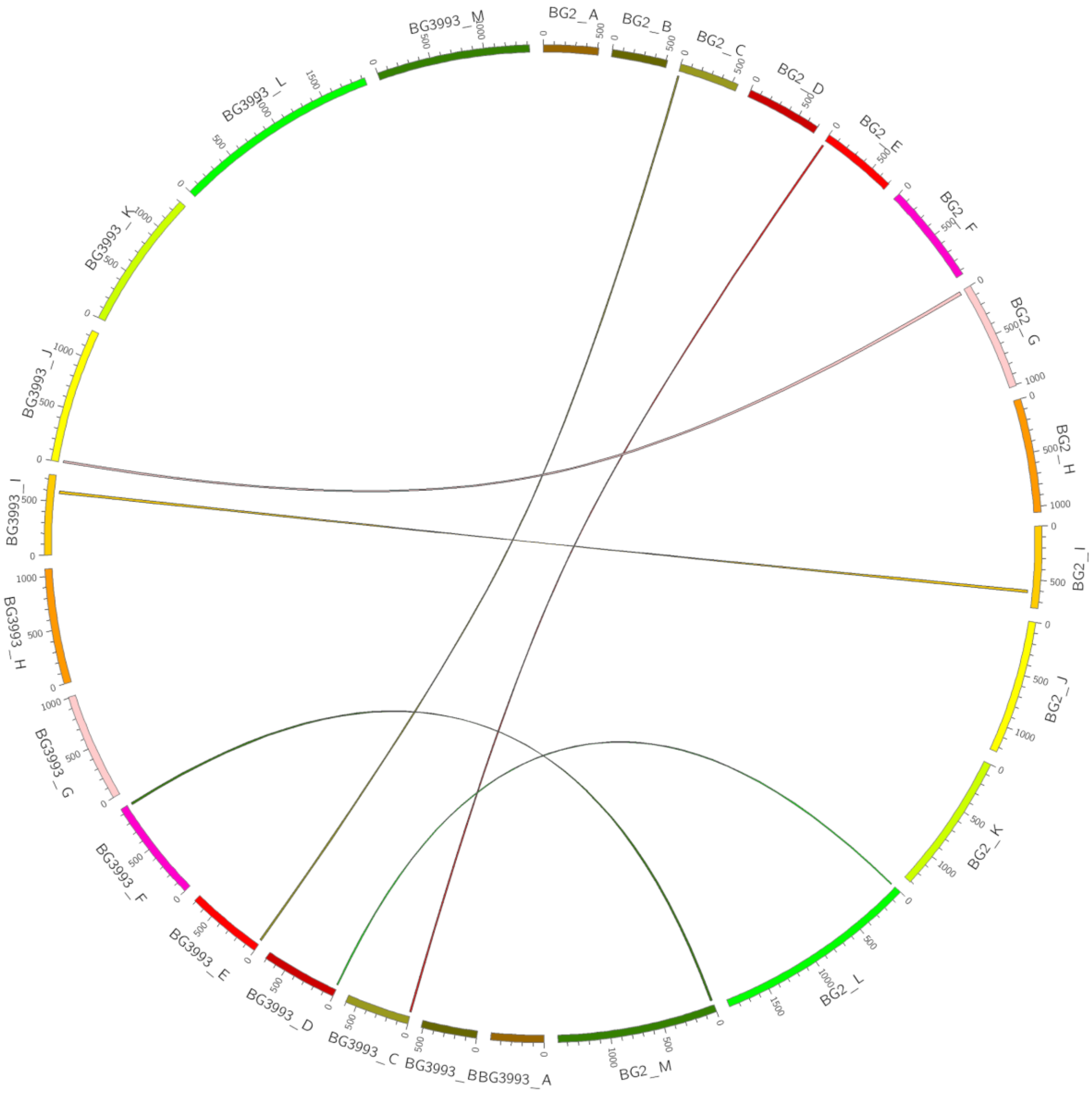


Figure 10 Chromosome rearrangements between BG3993 and BG2.

The Chromosome rearrangements in our BG3993 assembly relative to BG2 genome are illustrated using CIRCOS (Krzywinski et al., 2009). The rearrangement events are colored using the color of the CBS138 chromosome. The 6 chromosome rearrangement events are identified using translocated genes in BG3993 relative to BG2, and the translocation sites are identified using BLASTN alignment of the intergenic regions around the translocated genes (Table S2).

Subtelomere structure

The subtelomeres of *C. glabrata* are highly repetitive, and share homology with each other across different chromosomes. The subtelomeric regions encode half of the GPI-CWPs and have significant variation between strains. Therefore we analyzed the subtelomere structure separately. We annotated the subtelomeric regions and obtained the high-fidelity structure of the subtelomeres (Figure 11, Table S3). All four isolates have the same subtelomere structure, and all the changes in subtelomere structure in BG3993 relative to BG2 discussed below are shared changes with BG3994-BG3996 relative to BG2. There are 71 subtelomeric genes in BG3993 strain. 64 of the 71 genes are shared with BG2 strain. There are 2 terminal GPI-CWP genes, CAGL0G00099g and CAGL0M00121g (*AWP4*) that are encoded in BG3993 and CBS138 strain, not in BG2. BG3993 encodes 5 novel genes relative to BG2 and CBS138. CAGL0J00161g and CAGL0J00163g are genes related with Ty3 retrotransposon, and they are inserted into the BG3993 ChrJ Left subtelomere. CAGL0A00088g, CAGL0H10670g and CAGL0I00198g are three novel GPI-CWPs in BG3993 relative to both BG2 and CBS138. In addition, there are five terminal GPI-CWPs are translocated to a different subtelomere in BG3993 relative to BG2 (Table S2). The CAGL0A00143g (*EPA24*) and CAGL0E00187g genes are degenerate in all four clinical isolates. Compared to the subtelomeres, the structure of the chromosome body is more conserved between BG2 and BG3993 (discussed in the following section). This indicates that subtelomeres may undergo higher rate of translocation during evolution. Although there are multiple changes in the subtelomere structure between BG2 and BG3993, there are no changes between the four clinical isolates. This indicates that the subtelomeres of *C. glabrata* are relatively stable in a short period (21 months for our clinical isolates).

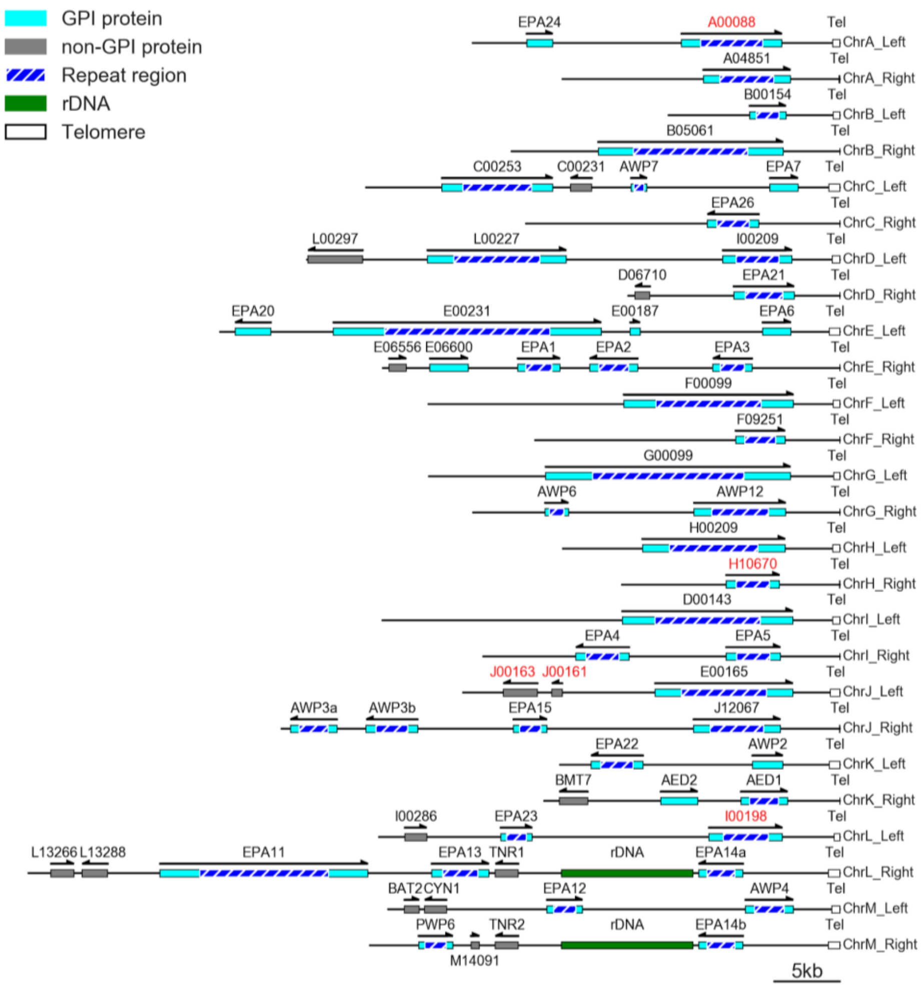


Figure 11 Subtelomere structure of BG3993.

Location, size and orientation of ORFs in the subtelomeric regions of the BG3993 assembly. All ORF lengths are indicated to scale. ORFs for predicted GPI-anchored adhesins are colored in light blue, and all other genes are colored in grey. Tandem repeats are indicated with a shadowed dark blue box (not to scale). One copy of the rDNA repeat (11 kb) is indicated to represent the rDNA cluster. Telomere repeats are indicated by a black box. The genes with their names highlighted in red are the novel genes in the subtelomeres relative to BG2 and CBS138 strain. The rDNA downstream region are identical in ChrL and ChrM as illustrated in the diagram, however, we only have the downstream region assembled in ChrL of the BG3993 assembly. The downstream region of ChrM is drawn to show the inferred structure. Dubious ORFs and ORFs which are not specific to *C. glabrata* subtelomeres are not drawn in this diagram.

Gene comparison

The four clinical isolates shared identical ORFs in general. All four clinical isolates encode 5252 ORFs. 5236 ORFs are identical in genome sequence in the four strains. Only 16 ORFs have any sequence variation whatsoever between the four strains (Table S4). 6 of the 16 ORFs have tandem-repeat contractions or extensions, and all of these 6 ORFs encode GPI-CWPs. The *EPA14* gene contains repeat expansion and changes in the expanded repeat in BG3996 relative to BG3993 (*EPA14* was not assembled in BG3994 and BG3995 assembly as discussed before). 9 ORFs only have 1-4 SNPs or indels.

In terms of genome alterations that might account for the increase in adherence of BG3995 and BG3996 relative to BG3993 and BG3994, we found only one candidate variant - and this was within the *YAP6* coding sequence. The *YAP6* gene in BG3995 and BG3996 has an insertion of four bases AACC at nt 313 of the ORF, relative to the sequence in strains BG3993 and BG3994. The insertion in BG3995 and BG3996 *YAP6* there resulting in a frameshift mutation in the *YAP6* gene. Therefore, loss of *YAP6* function is a possible candidate responsible for the increase of cell adherence in BG3994 and BG3995 relative to BG3993 and BG3994.

While extremely similar to each other, the four clinical isolates have significant changes in ORFs compared to BG2. There are 5086 shared single-exon ORFs in BG3993 and BG2. Only 1839 ORFs share identical protein sequence; 3099 ORFs contain substitutions in protein sequence or changes in ORF length ≤ 50 nt. 148 ORFs contains structural variants which have genome ORF length difference > 50 nt. There are 11 BG2 ORFs deleted in BG3993: 3 simple gene deletions; One GPI-CWP, CAGL0F09295g, is replaced by the GPI-CWP, CAGL0M00121g (*AWP4*)In terms of what genome alteration might account for the increase in adherence of BG3995 and BG3996 relative to BG3993 and BG3994, we found only one candidate. YAP6 encodes a shared variation found in BG3995 and BG3996 relative to BG3993 and BG3994. In BG3995 and

BG3996 YAP6 there is an insertion of four bases AACC at nt 313 of the ORF, resulting in a frameshift mutation in the YAP6 gene. Therefore, loss of YAP6 function is a possible candidate responsible for the increase of cell adherence in BG3994 and BG3995 relative to BG3993 and BG3994.

The clinical isolates have significant changes in ORFs compared to BG2. There are 5086 shared single-exon ORFs in BG3993 and BG2. Only 1839 ORFs share identical protein sequence; 3099 ORFs contain substitutions in protein sequence or changes in ORF length ≤ 50 nt. 148 ORFs contains structural variants which have genome ORF length difference > 50 nt. There are 11 BG2 ORFs deleted in BG3993: 3 simple gene deletions; One GPI-CWP, CAGL0F09295g, is replaced by the GPI-CWP, CAGL0M00121g (*AWP4*), found in the CBS138 strain; One GPI-CWP, CAGL0L00157g, is deleted due to non-reciprocal translocation; One gene, CAGL0F00077g (*EPA16*), recombined with CAGL0F00099g, resulting in a chimeric gene (CAGL0F00099g with the *EPA16* C-terminus) and the deletion of *EPA16*; 3 GPI-CWPs are replaced by novel GPI-CWPs (specific to the BG3993-BG3996 relative to BG2 and CBS138); Two *TKP5* homologs in BG2, CAGL0K02618g and CAGL0H01952g are deleted.

BG3993 encodes 11 genes which are specific to CBS138 relative to BG2 (Table S4). 2 are shared gene duplications in CBS138 relative to BG2. Two are GPI-CWP encoding genes specific to CBS138 relative to BG2. 7 are shared copy number variations of genes within the complex ChrC region. As discussed in chapter 2 and chapter 3, there is a complex ChrC region which encodes one copy of CALG0C00847g (*EPA8*) and various copies of GPI-CWP encoding genes from adhesin cluster VII (5 in CBS138, 2 in BG2) (de Groot et al., 2008). The cluster VII GPI-CWPs are highly similar with each other, and also share homology in flanking regions and flanking short ORFs. 3 of the 7 genes encode GPI-CWPs, and 4 are short duplicated ORFs in this region.

BG3993 encodes 20 novel ORFs relative to both BG2 and CBS138. 1) 4 ORFs are encoded in the complex ChrC region: one of these ORFs is the duplicated GPI-CWP in cluster VII, and the other three are the duplicated short ORFs. 2) 6 ORFs represent copy number variation of the metallothionein (*MT-II*) genes. Although there are various copies of *MT-II* genes in BG2, CBS138, and BG3993, all the four isolates encode the same number of *MT-II* genes. 3) There are 4 novel Ty3 retrotransposon related ORFs in BG3993 resulting from transposon insertion events. 4) There are two *TKP5* homologs, CAGL0F03278g and CALG0E05841g; 5) CAGL0C04774g is a homolog of CBS138 CAGL0C01859g (97.3% sequence identity). CAGL0C04774g might be the result of a translocation of CAGL01859g to the CAGL0C04774g locus, followed by sequence diversification to generate CAGL04774g. 6) Finally, BG3993 has 3 novel GPI-CWP encoding genes, probably resulting from gene duplication (see the next section).

Classification of the GPI-CWPs

We identified 82 GPI-CWPs in BG3993 and we classified them into adhesin clusters defined by de Groot *et. al* (de Groot et al., 2008) (Figure 12, Table S5). All the four isolates encode the same GPI-CWPs with nearly identical sequences. Only 10 GPI-CWPs contain sequence variations and all of the variations are simple SNPs and indels or repeat extensions or contractions (Table S4). There are 4 novel GPI-CWPs relative to BG2 and CBS138. Three of the four novel GPI-CWPs are terminal GPI-CWPs (CAGL0A00088g, CAGL0I00198g and CAGL0H10670g, all in adhesin cluster V), and one of them is in the complex ChrC discussed previously (Table S4). There are 2 GPI-CWPs shared with CBS138, but not encoded in strain BG2.

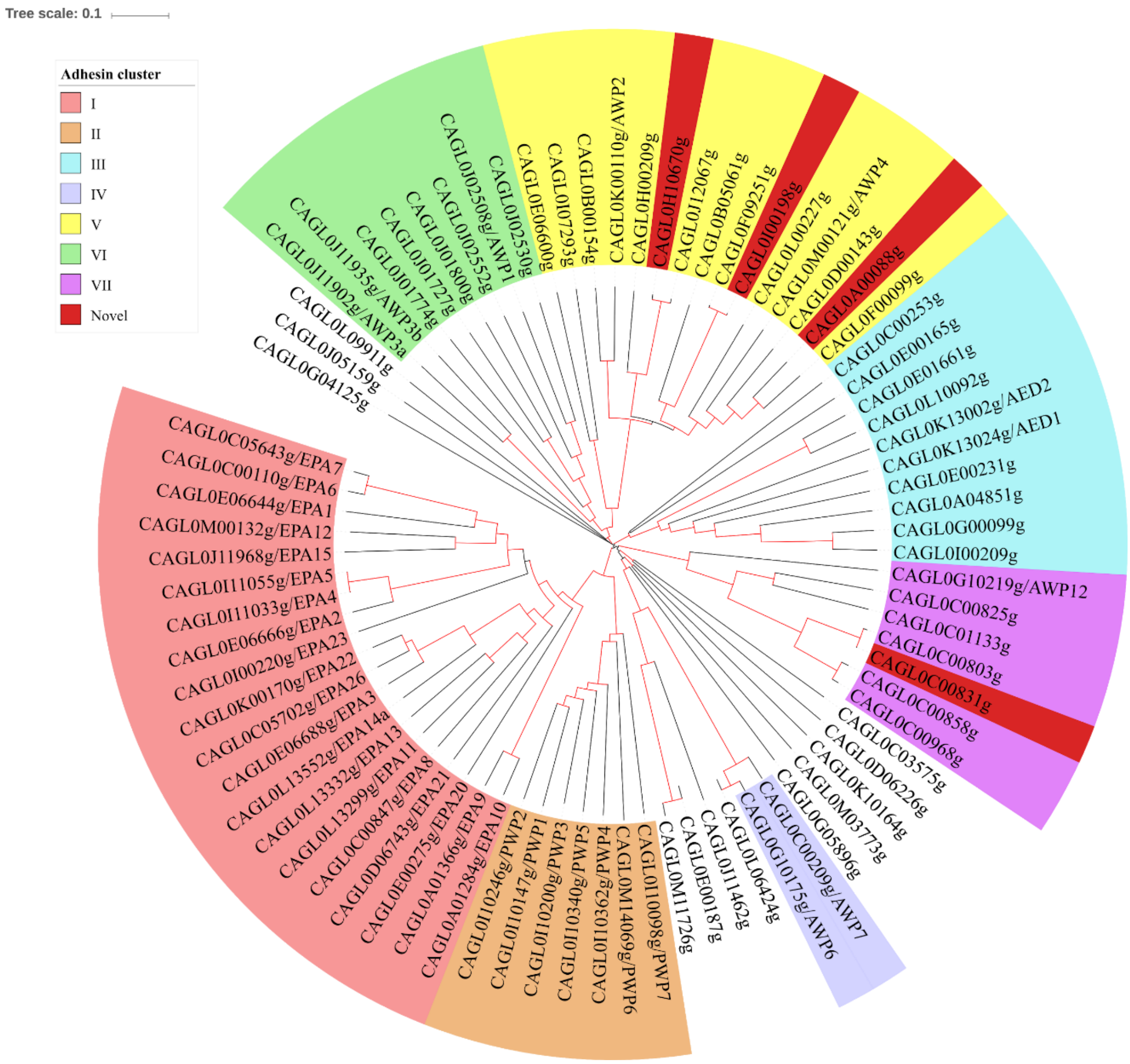


Figure 12 Phylogenetic tree of GPI-CWPs

The adhesin clusters are indicated by color, and the common name, if any, is shown in the tree as well. In addition, we highlighted the novel GPI-CWPs in red. We extracted the N-terminal regions of the GPI-anchored adhesins as described in Experimental Procedures, and used these regions to generate a bootstrap phylogenetic tree using ClustalW2 with seed=111 and 1000 bootstrap trials. The branches with bootstrap number > 500 were colored in red. The adhesin clusters are annotated based on published annotations (de Groot et al., 2008). CAGL0A00143g (*EPA24*) is degraded and the N-terminus of the degraded gene is not located in the *EPA* cluster, which is not drawn in this diagram.

The subtelomeres of the clinical isolates have the general subtelomere structure in which the terminal gene is a GPI-CWP transcribes towards the telomere. As discussed in chapter 3, strain BG2 and strain CBS138 diverge in chromosome structure due to multiple translocation events. Strain BG3993 is closely related in terms of chromosome structure to BG2. There are translocation events associated with the terminal GPI-CWPs between the four isolates and strain BG2. All the six chromosome rearrangement events are associated with GPI-CWPs, and five of six events are translocations associated with the terminal GPI-CWP genes. *EPA6* and *EPA7* undergo reciprocal recombination switching their location; CAGL0I00209g is translocated from ChrL left terminal locus to ChrD left terminal locus relative to BG2. The ChrL left terminal locus encodes a novel GPI-CWP, CAGL0I00198g. CAGL0I00198g is highly similar with another terminal GPI-CWP, CAGL0F09251g. In addition, the two genes share similar upstream intergenic region (98% sequence identity in 12 kb intergenic region); CAGL0F09251g is translocated from ChrM left terminal locus to ChrF right terminal locus relative to BG2, the ChrM Left terminal locus encodes CAGL0M00121g (*AWP4*) a GPI-CWP found at this locus in CBS138 rather than in BG2; CAGL0E00165g is translocated from ChrG left terminal position to ChrJ left terminal position, and the ChrG left terminal position encodes the GPI-CWP, CAGL0G00099g, also a GPI-CWP found at this locus in CBS138 rather than BG2.

In addition to changes in terminal genes, related to translocation events, there are two GPI-CWP genes that are directly replaced by novel GPI-CWPs. CAGL0H10670g replaces the CAGL0H10647g (*EPA17*) found in BG2 at the ChrH right terminal location and CAGL0A00088g replaces the CAGL0A00099g (*EPA19*) found in strain BG2 at the ChrA left terminal location. CAGL0H10670g is highly similar with CAGL0J12067g, which is the terminal GPI-CWP at the ChrJ right position, and the two genes share 93% sequence identity in 7 kb flanking region. CAGL0A00088g is a novel homolog of CAGL0F00099g (83% protein sequence identity in N-terminus) and the flanking sequence shared with CAGL0F00099g also has modest sequence homology (86% genome sequence identity in 7 kb flanking region). This suggests that CAGL0A00088g represents a relatively older duplication of CAGL0F00099g than, for example, the duplication generating the CAGL0H10670g/ CAGL0J12067g gene pair (98%-99% protein sequence identity in N-terminus). CAGL0F00099g is a terminal gene only in BG3993-96, where is a chimera with CAGL0F00077g (a terminal gene in CBS138 and BG2). *C. glabrata* encodes strikingly long GPI-CWPs in strain CBS138 and BG2 (see chapter 2, 3). In our clinical isolates, we also observed long GPI-CWPs. The ORF lengths of GPI-CWPs ranges from 0.4 kb to 33.8 kb. Strikingly, for the four isolates the ORFs for those long genes are nearly identical (Table S5). There are, for example, 15 GPI-CWPs with ORF length > 10 kb. 11 of these 15 “long” GPI-CWPs have identical ORF lengths. 3 of the other 4 GPI-CWPs have tandem-repeat extension/contraction only in one isolate. Even for CAGL0J05159g, the longest GPI-CWP ORF, is identical in length in BG3993 and BG3995, while the ORF length in BG3994 and BG3996 shows small changes in the tandem-repeat region. Therefore, the four assemblies from independent sequencing and *de novo* genome assembly show high reproducibility even for the highly repetitive long ORFs, validating our assembly, but more importantly indicating the relative stability of these regions in strains colonizing a patient across many months.

We observed significant changes in ORF length of the GPI-CWPs relative to BG2 and CBS138. For instance, the ORF length of the largest ORF, CAGL0J05159g is 33.8 kb in BG3993, 23.3 kb in BG2, and 29.6 kb in CBS138. The ORF length of CAGL0E00231g is 20.4 kb in BG3993, 22.5 kb in BG2, and 13.7 kb in CBS138. These changes could result from repeat extensions or contractions. However, they could also result from non-allelic mitotic recombination events between GPI-CWPs encoding genes. We take up this question of the non-allelic mitotic recombination in the next chapter.

Discussion

The four clinical isolates are highly similar to each other. There were only 220 SNPs and indels in the 13 Mb genome between the four isolates. The four isolates encodes the same 5252 genes with no strain specific genes. 5236 of the 5252 encoded ORFs had identical genomic ORF sequence across all four isolates, and only 16 ORFs contain sequence variations between the four isolates (Table S4). The ORFs with identical sequence include 11 ORFs with ORF length > 10 kb. Since the four assemblies were obtained from independent SMRT sequencing and independent *de novo* genome assemblies, this demonstrates that our *de novo* assembly is highly reproducible, and increases confidence in the assemblies of even the most complex tandem repeat rich sequences.

One aim of assembling the genomes of the four isolates was to identify candidates responsible for the increase in cell adherence in BG3995, BG3996 relative to BG3993, BG3994. There is only one 1 gene, *YAP6,* for which we found a shared sequence variation (c.313_314insAACC) in BG3995, BG3996 relative to BG3993, BG3994, and this insertion is predicted to cause a frameshift mutation. *YAP6* has not been studied in *C. glabrata*, but investigations of the *YAP6* orthologue in *S. cerevisiae* implicates it in stress response (Rodrigues-Pousada et al., 2019). *YAP6* binds to subtelomeric genes and regulates chromatin-remodeling during stress response in *S. cerevisiae* (Mak et al., 2009). The subtelomeres of *C. glabrata* has many GPI-CWP adhesin genes, and *YAP6* may regulate cell adherence through the regulation of these genes. Further experiments are needed to analyze the function of *YAP6* in *C. glabrata.*

The subtelomere structure of the four isolates supports the general subtelomere structure proposed in chapter 2: the *C. glabrata* subtelomeres encode a terminal GPI-CWP that is almost always transcribed towards the telomere. We observed that the terminal GPI-CWPs are often involved in translocations between strains: 5 of 6 chromosome-rearrangement events between BG2 and BG3993 are translocations of the terminal GPI-CWP gene. Interestingly, the non-reciprocal translocation of terminal GPI-CWPs does not result in a deletion. That is, at the “donor” locus for the translocation, there is no deletion (resulting in the penultimate gene now being terminal), but rather, a new strain specific terminal GPI-CWP gene is now found at this locus. Here, we observed this in comparing strains 3993 and BG2, but also observed the phenomenon comparing BG2 and CBS138 in chapter 3. This suggests that *C. glabrata* subtelomere diversification involves the reciprocal exchange of terminal GPI-CWPs. In other words, even where translocations apparent in comparing two strains appear to be non-reciprocal, it is possible that the underlying historic translocation was a reciprocal exchange of terminal GPI-CWPs.

We documented repeat extensions or contractions in only 6 GPI-CWPs, again suggesting a surprisingly modest dynamic change in these genes across the period of infection (20 months). The dramatic changes in GPI-CWP structure between more distant strains may reflect environmental adaptation or genome dynamics operating on longer evolutionary time scales.

Table 1 Metrics of PacBio Sequencing

Sample	BG3993-1	BG3994-1	BG3995-1	BG3996-1
Number of Reads	54369	45708	49564	41748
Read N50	20750	19687	19570	21149
Total Number of Bases	597265766	476772217	474786630	498052558
Sample	**BG3993-2**	**BG3994-2**	**BG3995-2**	**BG3996-2**
Number of Reads	259559	170632	144123	290457
Read N50	8810	9372	8964	8371
Total Number of Bases	1576233091	1016386802	768015191	1714267514

Metrics of the PacBio SMRT sequencing subreads. Sample BG3993-1 – BG3996-1 are SMRT sequencing reads for *de novo* assembly. Sample BG3993-2 – BG3996-2 are re-sequencing of the BG3993-BG3996, which were used to correct sequencing errors (see Experimental procedures).

Table 2 Chromosome assembly metrics

Strain	Chrom	Chrom_Length	Left_Telomere	Right_Telomere
BG3993	ChrA	494217	577	592
BG3993	ChrB	518902	528	645
BG3993	ChrC	604662	862	804
BG3993	ChrD	707303	617	590
BG3993	ChrE	712011	812	804
BG3993	ChrF	951544	527	606
BG3993	ChrG	1009067	489	822
BG3993	ChrH	1069838	565	803
BG3993	ChrI	743751	573	653
BG3993	ChrJ	1254702	736	699
BG3993	ChrK	1306201	808	814
BG3993	ChrL	1924090	471	889
BG3993	ChrM	1420520	643	N/A
BG3994	ChrA	494132	482	621
BG3994	ChrB	518750	536	408
BG3994	ChrC	604297	703	597
BG3994	ChrD	707238	554	590
BG3994	ChrE	712092	950	746
BG3994	ChrF	951489	479	584
BG3994	ChrG	1007930	390	683
BG3994	ChrH	1069813	695	652
BG3994	ChrI	744358	686	852
BG3994	ChrJ	1249708	683	572
BG3994	ChrK	1306212	771	865

Strain	Chrom	Chrom_Length	Left_Telomere	Right_Telomere
BG3994	ChrL	1884164	381	N/A
BG3994	ChrM	1437530	797	N/A
BG3995	ChrA	494470	838	528
BG3995	ChrB	518998	589	701
BG3995	ChrC	604363	655	711
BG3995	ChrD	707182	593	494
BG3995	ChrE	712403	827	1181
BG3995	ChrF	951592	577	600
BG3995	ChrG	1005504	N/A	677
BG3995	ChrH	1069683	641	577
BG3995	ChrI	744102	702	560
BG3995	ChrJ	1254125	90	638
BG3995	ChrK	1306374	807	1007
BG3995	ChrL	1873931	687	N/A
BG3995	ChrM	1429365	567	N/A
BG3996	ChrA	494207	564	550
BG3996	ChrB	518882	584	569
BG3996	ChrC	604353	754	604
BG3996	ChrD	707585	805	686
BG3996	ChrE	712044	753	892
BG3996	ChrF	951614	659	546
BG3996	ChrG	1008911	614	544
BG3996	ChrH	1069827	671	637
BG3996	ChrI	744204	652	677
BG3996	ChrJ	1254637	840	550
BG3996	ChrK	1305987	557	838

Strain	Chrom	Chrom_Length	Left_Telomere	Right_Telomere
BG3996	ChrL	1901730	498	729
BG3996	ChrM	1435065	658	N/A

The length of each chromosome and the length of the telomere repeats on each chromosome. ChrL and ChrM with N/A in right telomere repeats are chromosomes that ended in the rDNA region. They are not full assembled chromosomes, lacking the complete rDNA array, and missing the 10 kb rDNA downstream region that encodes the *EPA14* gene. BG3995 with N/A ChrG Left telomere has a ChrG chromosome lacking the telomeric repeats and the terminal intergenic region (approximately 3 kb in length) assembled in the contig.

Reference

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

de Groot, P.W.J., Kraneveld, E.A., Yin, Q.Y., Dekker, H.L., Gross, U., Crielaard, W., de Koster, C.G., Bader, O., Klis, F.M., and Weig, M. (2008). The cell wall of the human pathogen Candida glabrata: differential incorporation of novel adhesin-like wall proteins. Eukaryotic Cell *7*, 1951–1964.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Kumar, K., Askari, F., Sahu, M.S., and Kaur, R. (2019). Candida glabrata: A Lot More Than Meets the Eye. Microorganisms *7*.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Mak, H.C., Pillus, L., and Ideker, T. (2009). Dynamic reprogramming of transcription factors to and from the subtelomere. Genome Res. *19*, 1014–1025.

Nattestad, M., and Schatz, M.C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics *32*, 3021–3023.

Pierleoni, A., Martelli, P.L., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. BMC Bioinformatics *9*, 392.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. Trends Genet. *16*, 276–277.

Rodrigues-Pousada, C., Devaux, F., Caetano, S.M., Pimentel, C., da Silva, S., Cordeiro, A.C., and Amaral, C. (2019). Yeast AP-1 like transcription factors (Yap) and stress response: a current overview. Microb. Cell *6*, 267–285.

Chapter 5 Non-allelic mitotic recombination in *Candida glabrata*

Introduction

The *Candida glabrata* genome encodes large families of GPI-CWP genes. The GPI-CWP genes mediate cell adherence which is an important step for the virulence of *C. glabrata*. We obtained high-quality ORF sequences of 6 *C. glabrata* strains and these high quality genome sequences permit further analysis to study the sequence variations between strains. We observed significant changes in the complement of GPI-CWP genes between different *C. glabrata* strains and significant changes in gene structure for those GPI-CWP genes that are conserved. For instance, 51 of 74 shared GPI-CWP genes between BG2 and CBS138 have genome ORF length difference > 50 nt, resulting in a total change of 158.9 kb (approximately 35% of the total ORF length of GPI-CWP) (see chapter 3). We also observed significant changes between BG2 and the serial isolates, BG3993-96 (detailed in chapter 4). By contrast, in comparing the genome sequence of the four clinical isolates, we found evidence of only modest changes in GPI-CWP genes gene structure, in spite of the fact that the strains were taken from an infected patient over a 21-month period (see chapter 4). Therefore, dynamic change within GPI-CWP genes may pertain primarily to long-term evolution of *C. glabrata* rather than short-term microevolution during infection.

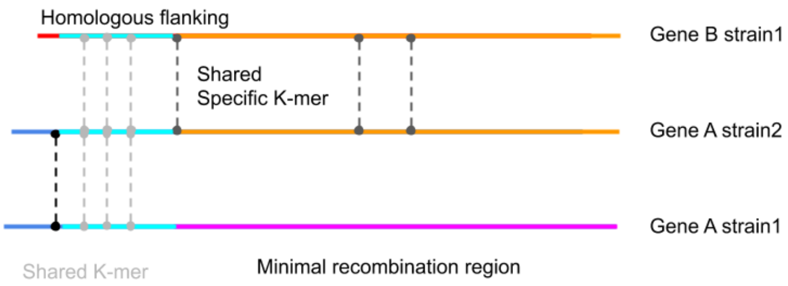
GPI-CWP genes undergo intragenic recombination, resulting in tandem-repeat extensions and contractions between strains. We are interested in whether ORFs in large gene families also exchange information with each other through intergenic recombination. *C. glabrata* is an asexual haploid yeast, and therefore, the intergenic recombination events are all non-allelic mitotic recombination events. The GPI-CWP genes include long tandem-repeat regions which share homology between different genes. It is challenging to access recombination events in those repeats using general sequence aligners, for instance, BLAST (see Results). Therefore, we developed a k-mer based method to identify recombination events in repeat regions, and use this to carry out a genome-wide analysis of non-allelic mitotic recombination events.

Our method identifies the sequence variation for a given gene (or genomic region) in two strains, identifies other loci containing the variant sequence, and then models the minimum number of recombination events that can account for the migration of SNVs or indels between the relevant homologous regions. To avoid mischaracterizing mutational events as recombination events, the minimal recombination region between the "recipient" ortholog of one recombinant gene and the "donor" paralog must include at least two independent SNVs/indels specific to the recombined paralog. The repeat expansions and contractions are not identified in our analysis because these events do not alter the encoded k-mers. An important advantage of our method is that it simplifies the identification and analysis of recombination events within even complex repeat regions. Using this algorithm, we successfully documented substantial recombination between both tandem-repeat and non-tandem repeat regions in GPI-CWP genes in the subtelomeric regions. We performed genome-wide recombination analysis of ORFs as well and documented many examples of non-allelic recombination within the body of the chromosomes. Our analysis demonstrates that subtelomeric regions and GPI-CWP genes have more apparent mitotic recombination than the chromosome body, although there are also many recombination events between parologous genes in non-subtelomeric regions.

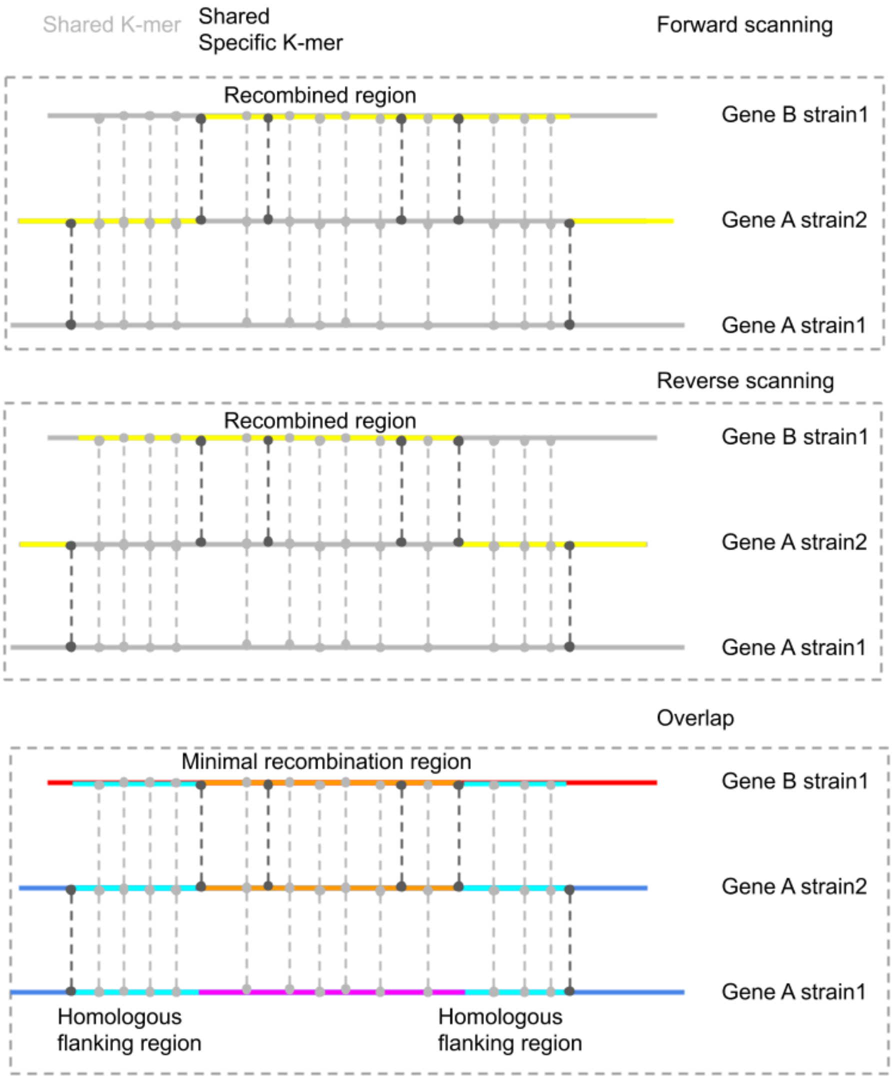
Experimental Procedures

Non-allelic mitotic recombination analysis

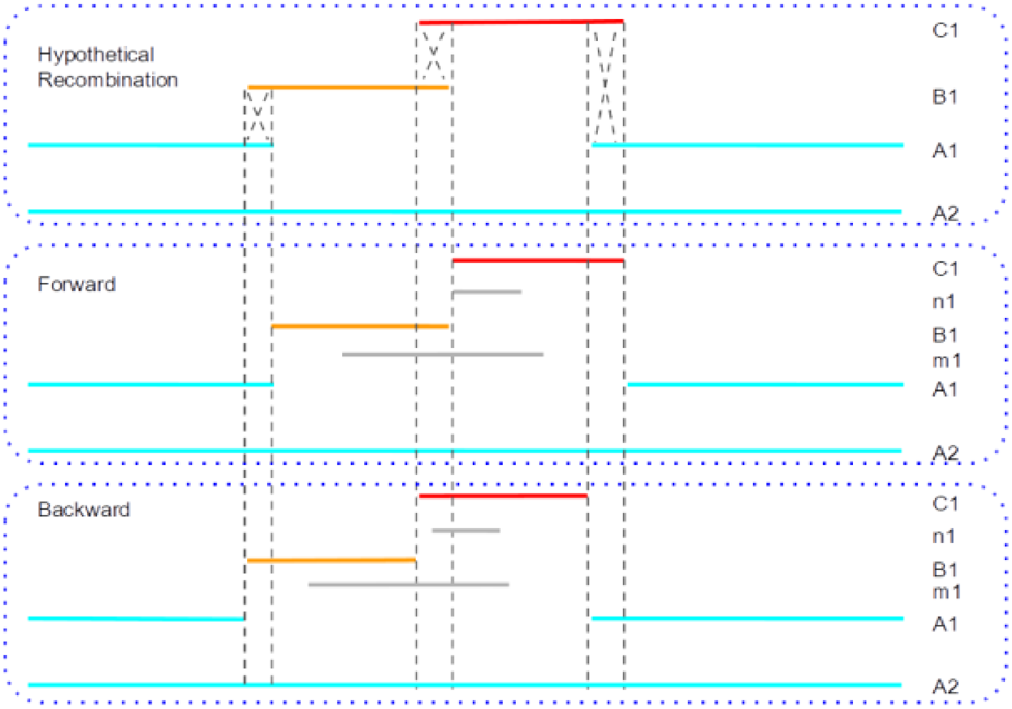
I developed a pipeline to analyze non-allelic mitotic recombination events between different *C. glabrata* strains (https://github.com/zhuweix/recombination_analysis.git/) (See Figure 13 for an overview of the algorithm). This pipeline identifies sequence variation between a “query gene” in one strain (the “query strain”) and its orthologue (“reference gene”) in a second strain (the “reference strain”), and models putative recombination events between genes (the reference gene and its paralogues) in the reference genome that account for sequence variation of the query gene in the query genome. Shared k-mers are used for identification of recombination events (Salzberg et al., 2014). The basic assumption of the pipeline is that all kmers that a query gene shares with its paralog in the reference strain but not with its ortholog in the reference strain, indicate historic recombination events between the ortholog and paralog. The algorithm identifies the maximum size of recombination regions corresponding to a minimum number of recombination events to account for a particular SNV pattern.


Homologous flanking
Gene B strain1
Shared
Specific K-mer
Gene A strain2
Gene A strain1
Shared K-mer
Minimal recombination region

(A)


Shared K-mer
Shared
Specific K-mer
Forward scanning
Recombined region
Gene B strain1
Gene A strain2
Gene A strain1
Reverse scanning
Recombined region
Gene B strain1
Gene A strain2
Gene A strain1
Overlap
Minimal recombination region
Gene B strain1
Gene A strain2
Gene A strain1
Homologous
flanking region
Homologous
flanking region

(B)



(C)

Figure 13 Diagram of k-mer based recombination analysis.

(A) Definition of minimum recombination region. The recombination of gene A in strain 2 relative to stain 1 is assessed in this diagram. There is a hypothetical recombination between gene A and gene B. The minimal recombination regions are regions with SNVs specific to the recombined gene. In this example, there is one minimal recombined region assigned to gene B in strain 1 (orange) which correlates the SNVs in the purple region between gene A in strain 1 and strain 2. The region assigned to gene A itself, indicating a non-recombined region, is indicated in dark blue. ortholog specific k-mers, representing specific SNVs are indicated in black and the shared k-mers are indicated in grey. The boundaries of minimal recombination regions are defined by specific k-mers. Flanking the minimal recombination region is the homologous flanking region. In this region, all shared k-mers are shared between gene A in strains 1 and 2 and gene B in strain 1.

(B) Diagram of k-mer scanning and region overlapping. The recombination of gene A in strain 2 relative to stain 1 is assessed in this diagram. There is a hypothetical recombination between gene A and gene B. The specific k-mers are indicated in black, while the k-mers shared with gene A strain 1, gene A strain 2, and gene B strain 1 are indicated in grey. The recombined region identified in forward scanning is indicated in yellow. In forward scanning, we read k-mers in the ORF sequence of gene A strain 2 from 5' to 3'. The recombined region starts at the first specific k-mer shared with gene B strain 1 and gene A strain 2, and not with gene A strain 1. The region is extended until there is a specific k-mer with gene A in strain 1 and strain 2, not with gene B in strain 1. In forward scanning, the left end of the recombined region contains the identifying SNV between gene A and gene B, and it is the left end of the minimal recombination within this recombined region. In reverse scanning, we read k-mers in the ORF sequence of gene A strain 2 from 3' to 5'. Similarly, we obtain the recombined regions of which the right end is the right end of the minimal recombination region within this region. In the overlapping step, the forward recombined regions and reverse recombined regions are overlapped to define minimal recombination regions as well as the homologous flanking regions.

(C) Diagram of identification of hypothetical mitotic recombination with multiple paralogs. Here we also assess the recombination of gene A in strain 1 and stain 2. For simplicity, gene A in strain 1 and strain 2 are named as gene A1, A2. The homolog of A2 in strain 1, gene B, C, m ,n are named as gene B1, C1, m1, n1. There are two hypothetical mitotic recombination between gene A1, B1, C1, resulting in sequence variations in gene A2. As illustrated in (a), the k-mer based approach scanned the sequence of gene A2 in forward (5'-3') direction, and determined the left end of minimal recombination regions. Similarly, gene A2 is then scanned in reverse (3'-5') direction to determine the right end of minimal recombination regions. The results of forward and reverse scanning are overlapped to generate the minimal recombination region and the homologous flanking region. Our approach ensures that all the recombined genes (B1, C1) have at least one k-mer in the query sequence which is specific to these genes. Thus, the homologs with no specific k-mers (m1, n1) are filtered in this process. Finally, we obtained the minimal recombination region with the high-fidelity list of recombined genes from this approach.

The first step of our analysis is to define recombination events. Our analysis documents all the sequence variation that can be explained by recombination between reference genes. Variation specific to the query gene (i.e. not found in any homologs in the reference strain) is considered to be due to mutation. Each query gene is initially assigned an ortholog in the reference genome to serve as a baseline for recombination events. We then model the minimum number of sequential recombination events that can account for a given pattern of variation between the query gene and its reference orthologue. This is done sequentially, so that at each recombination site (between the orthologue A and some homolog B), the "reference" gene for assessment of subsequent recombination events changes to homolog B(the "current reference gene"). Thus, sequence variation is always assessed between the query gene and the current reference gene, not necessarily the original ortholog of the query gene. -ie the reference gene changes at each recombination site.

The search for recombination partners is done in both directions across the query gene length. The beginning of the recombination region is assigned as a function of the first SNV that is shared with other reference gene(s), but not with the reference orthologue. We then use a "greedy" strategy that will expand this recombination region for as long as possible. This recombination region is extended maximally such that there is at least one reference gene for which any sequence variation (in this recombination region) is specific to the query, *i.e,* not found anywhere in the reference genome; functionally, this means that recombination regions can be assigned even when they have diverged by mutation. Reference genes that fulfil these conditions are "assigned" to this region, and the set of assigned genes for one region is called the "assigned set". The recombination region terminates at the position where a sequence variation between the query and the assigned set is shared with genes outside the assigned set (i.e. can be explained by recombination with an additional reference gene). This marks the beginning of another recombination region. We continue the sequential assignment and expansion of recombination regions until the end of the query sequence. After scanning the query sequence in the forward direction, we scan the query sequence in the reverse direction. The overlap of forward and reverse regions defines "homologous flanking" regions and "minimal recombination" regions (see below). This process identifies the minimum number of sequential recombination events to generate the query sequence.

Homologous Flanking Region and Minimal Recombination Region

Each recombination region is defined by a left homologous flanking region, a minimal recombination region, and a right homologous flanking region. In the left homologous flanking region, all sequences are shared between the query gene, the assigned set of reference genes for this region, and the assigned set or reference genes for the previous recombination region. The left homologous flanking region, therefore, can be thought of as the homologous region in which recombination occurs between the assigned set of reference gene(s) assigned to the previous recombination region and the assigned set of reference gene(s) assigned to the current region. The right homologous flanking region, similarly, can serve as the homologous region between the assigned set of the current region and the assigned set of the subsequent region.

The minimal recombination region is the recombination region that excludes homologous flanking regions. *i.e.*, this region encodes sequences in the query strain which are only shared with the current assigned set rather than the assigned set of the preceding or subsequent recombination regions. Forward scanning of the query gene identifies the left end of the minimal recombination region, and expands the right end of the recombination region for as long as possible. Similarly, reverse scanning of the query gene identifies the right end of the minimal recombination region, and expands the left end of each recombination region for as long as possible. The overlap of the forward and reverse recombination regions set the boundaries for the minimal recombination regions as well as the homologous flanking regions. The minimal recombination regions also minimize the number of sequential recombination events between the reference genes to generate the query sequence.

Implementation of our algorithm

Our algorithm identifies sequence variation based on k-mers. Shared k-mers indicate shared sequences, and the sequence variation at one specific location between the query gene and the reference gene is indicated by the k-mers in the query gene which cannot be found in the reference gene. Query gene sequence variation are classified as mutations if the corresponding k-mer is not found in any reference gene. Copy number and the location of the k-mer is not considered for the comparison; the important consequence of this is that intragenic recombination events which result in copy number variation or rearrangement of k-mers within the same gene are not identified as sequence variation in our method.

We first identify all the k-mers in the reference genes. For every k-mer we store whether this k-mer is encoded in each reference gene into a binary array. The k-mer and the corresponding binary array is stored in a hash map (dictionary in python), and any k-mer in the query that cannot be found in the hash map indicates a mutation. In forward scanning, we search reference genes for k-mers shared with the query region, starting the first query region with the first k-mer shared with reference genes. All reference genes that share a k-mer are initially "assigned" to this region. As the forward scanning proceeds, we perform intersections of the set of reference genes sharing the current k-mer and the set of genes assigned the current region to update the assigned set of the region using bitwise operations of the binary arrays. This is to ensure that all the assigned genes to the current regions encode all the shared k-mers with the query. The region terminates when the intersection is empty, indicating that there would be no reference gene sharing all the shared k-mers in the query region if this region were extended by one extra nucleotide. The assigned set of the terminated region is the set of reference genes that encode all the non-mutational k-mers in the query. A new region is initiated with the specific k-mer at its left end. As the scanning proceeds, we identify all the left ends of minimal recombination regions. In reverse scanning, we use a similar procedure to identify the right ends of minimal recombination regions. The regions between the minimal recombination regions are homologous flanking regions. The homology between the query sequence and the assigned reference gene in each region is accessed by the number of shared k-mer between the query gene and the reference gene.

If the overlapping regions from forward and reverse scanning share at least one reference gene, this overlap of the two regions is the minimal recombination region. However, there are rare cases when there is no shared reference gene in the overlap. For instance, reference gene A may have longer right homologous region, and reference gene B may have longer left homologous region (Figure 14). Therefore, forward scanning identifies a recombination region assigned to gene A, and reverse scanning identifies a recombination region assigned to gene B because of maximum extension of homologous flanking regions, and these overlaps are re-classified as minimal recombination regions (see the following section of genome-wide mitotic recombination analysis).

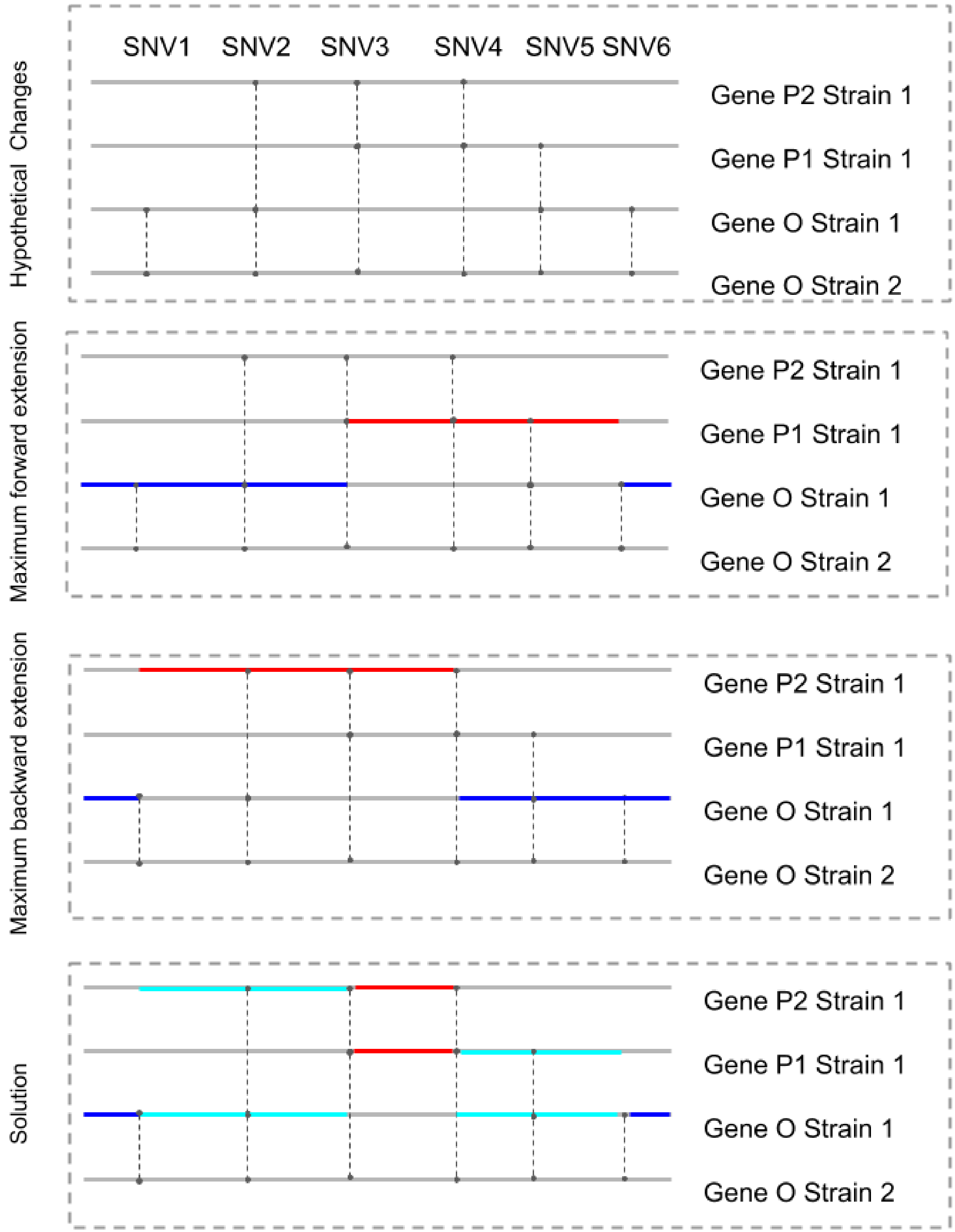


Figure 14 Example of one conflict recombination event

This is a hypothetical recombination event that the forward scanning and reverse scanning generate conflicting recombination events. This hypothetical recombination event moves SNV3 and SNV4 between gene O in strain 1 and strain 2, that can be the result of recombination with the paralog P1 or P2. The four sequences are identical in general, however, there are 6 SNVs that gene O in strain 2 only shares with subsets of gene O, P1, P2 in strain 1 (the genes in strain 1 containing the SNV are indicated in the parenthesis): SNV1(O), SNV2(O and P2), SNV3(P1 and P2), SNV4(P1 and P2), SNV5(O and P1), SNV6 (O). Therefore, recombination with P1 allows the maximum left extension (longest right flanking region) that the recombination region is extended to SNV6 (recombination with P2 only extend the region to SNV5); however, recombination with P2 allows maximum right extension (longest left flanking region) and the recombination region is extended to SNV1 (recombination with P1 only extend the region to SNV2). The algorithm defines the minimal recombination with either P1 or P2 between SNV3 and SNV4. However, the left flanking is the flanking region of recombination between O and P2, ranging from SNV1 to SNV3; the right flanking is the flanking region of recombination between O and P1, ranging from SNV3 to SNV6.

Visualization of recombination events

The raw recombination events are visualized using the assignment graph and dotplots. The assignment graph illustrates each minimal recombination region and the corresponding assigned genes in solid horizontal lines; the homologous regions and the corresponding assigned genes are horizontal dashed lines; the break points are indicated as vertical lines. The color of the horizontal line represents the ratio of shared k-mers to total k-mers in each query region.

The sequence comparison is illustrated by dotpots. The self-dotplot of the query gene is plotted to illustrate the identified recombination regions and homologous flanking regions, together with the recombined kmers which define the minimal recombination region. For each query, up to 3 recombined reference paralogues (sorted by length of recombination regions in the query) are compared with the query gene. We draw three plots for each query/paralogue pair. a) The dotplot of query-to-reference, highlighting recombination regions as well as the recombined k-mers within the recombination regions. b) The dotplot comparison between query and the ortholog, illustrating sequence variations. c) The dotplot of the reference-to-reference, highlighting the recombination regions as well as the recombined k-mers within the recombination regions in the reference structure.

Optimization of the recombination analysis

The choice of k-mer size is crucial to the analysis. A smaller k-mer size will identify random mutations that by chance share homology with other regions of the genome, while a larger k-mer size will mis-identify recombinant variants as random mutations. For instance, if multiple SNVs derived from multiple genes are located within one k-mer, none of the internal SNVs can make an exact match with the reference gene. We applied the genome-wide recombination analysis between BG2 and BG3993 for optimization and, based on this analysis, adopted a two-step recombination analysis for analysis of the comparison of BG3993 and CBS138.

We first aligned the ORF sequences of BG3993 to BG2 using BLASTN (multi-CDS genes were excluded) (Camacho et al., 2009). Genes in BG2 with an alignment with e-value < 1e-6 to a query BG3993 gene were defined as BG2 gene references. To determine optimal k-mer size for the analysis, we identified recombination events using the k-mer size of 1-60 (increment size = 1); 70-290 (increment size = 10); 400, 500, 600, 1000. The sequences of the minimal recombination regions were extracted, and we counted the unique (counted once) recombinant k-mers (k-mers shared with recombinant references but not with orthologs) and the unique mutated k-mers (k-mers in the query sequence that are not in any reference sequence). We also counted the unique recombination events (recombination regions share the same sequences are counted only once) for evaluation. Optimal k-mer size was the size that minimized false mutated k-mers and while maximizing called recombination events. The optimized k-mer size is 24 (see Results). For the second step, the analysis was performed using the k-mer size of 1-30 (increment size = 1) with the query and reference genes identified in the first step. Unique recombinant and mutated k-mers were counted in the same way. In addition, the number of reference genes (including the ortholog) per query gene was counted. The optimized k-mer size, which minimized the number of paralogues in the assigned set of reference genes was k-mer =14 (see Results).

Genome-wide analysis of mitotic recombination events between BG2 and CBS138

As discussed in the optimization procedure, candidate reference genes for each query gene were first selected by BLASTN with e-value < 1e-6; then we performed a two-step recombination analysis with k-mer size=24 and 14. The first step is to identify the minimal recombination regions and homologous recombination regions. In rare cases when the forward and reverse scanning identify different minimal recombination events assigned to non-overlapping sets of reference genes. This occurs because our approach is conservative and maximizes the extension of each region in forward/reverse direction to minimize called recombination events, and genes in the same minimal recombination region can have different maximized forward/reverse extension. One simplified hypothetical example is illustrated in Figure 14. The signature of these regions is that although these regions have different assigned sets in forward and reverse scanning, they do not have an adjacent minimal recombination region, and they are corrected to be minimal recombination regions, with an assigned set that is the union of the two conflicting assigned sets.

Secondly, the minimal recombination regions from the query genes were used to identify corresponding regions in the reference genome (using BLASTN). The best Blast alignment regions (indicated by largest bitscore) with the assigned reference genes from the analysis are combined together as the potential reference regions of the query sequence. Query sequences that are assigned to the same reference regions are further combined. Hence, only minimal recombination events are counted, and the effect of repeat expansions and contractions is minimized. The best query-to-reference alignment is compared with the best query-to-ortholog alignment to further evaluate the sequence variation generated by recombination events. Potentially false positive recombination events involving low-complexity k-mers (for example, k-mers with homopolymers) are excluded as they are filtered during BLASTN alignment by default.

Third, the genome locations of the genes with recombination regions are extracted. To permit classification based on genome location - subtelomere or chromosome body. The distribution of the genome-wide recombination events are visualized using CIRCOS (Krzywinski et al., 2009).

To analyze the apparent rates of recombination, we identified all possible homologous gene pairs (pairs that in principle are substrates for recombination). Homology of gene pairs was estimated using the weighted average of the bitscores of the BLASTN alignments. For each gene pair with multiple alignment regions (each of which has an e-value < 1 e-6 from BLASTN alignment), the average bitscore is calculated using the ratio of individual alignment length to total alignment length as the weight for each individual alignment bitscore. For gene pairs with single alignment regions, the weighted average is therefore its alignment bitscore. We compared the distribution of the average bitscores of all the gene pairs with at least one alignment of e-value < 1 e-6 against those of all the recombined gene pairs, and found the lower bitscore threshold to be <70 for strains BG2 and BG3993. We filtered all the homolog gene pairs, genome wide, with bitscore < 70 and used this set for comparisons to gene pairs that have actually undergone recombination. We classified gene pairs based on whether the two genes are GPI-CWP genes or not (all the recombined gene pairs are either both GWP-CWPs or both non GPI-CWP genes); location in subtelomere or chromosome body; whether the gene pair is adjacent with each other (distance < 50 kb within the same chromosome). We compared the ratios of recombined gene pairs to homologous (bitscore<70) gene pairs to estimate relative recombination rates in those groups. To assess the influence of adjacency in intrachromosomal recombination in non-subtelomeric regions, we calculated the cumulative distribution of number of recombined and the homologous gene pair relative to the distance between two genes in the gene pair, using the cumulative distribution of all possible gene pairs as the background.

Results

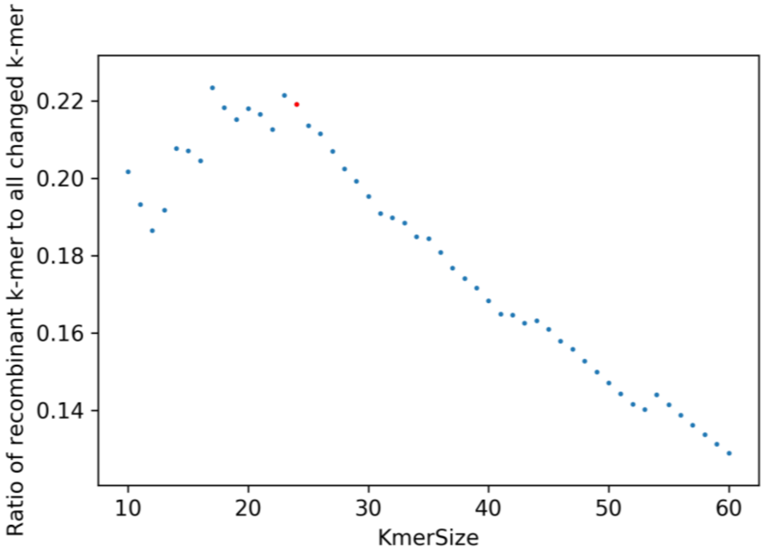
Optimization of identification of non-allelic mitotic recombination events

Our algorithm identifies sequence variation based on k-mer analysis. Shared k-mers indicate shared sequences, and the sequence variation at one specific location between the query gene and the reference gene is indicated by the k-mer in the query gene which cannot be found in the reference gene. Shared k-mers identify sequence variations resulting from recombination while unique k-mers indicate mutation. A critical parameter of the analysis is the k-mer size. Short k-mer size results in mutations being mis-characterized as recombination events due to random sharing of the mutated k-mer. So, increase in k-mer size increases specificity and reduces false positive recombination calls. However, increase in k-mer size also leads to lower sensitivity. Short recombination regions may not be long enough to have one k-mer exactly matched to the recombined paralog. In addition, if the k-mer size is too large, recombinant SNVs may be omitted because they are contained only within k-mers that also have adjacent mutations.

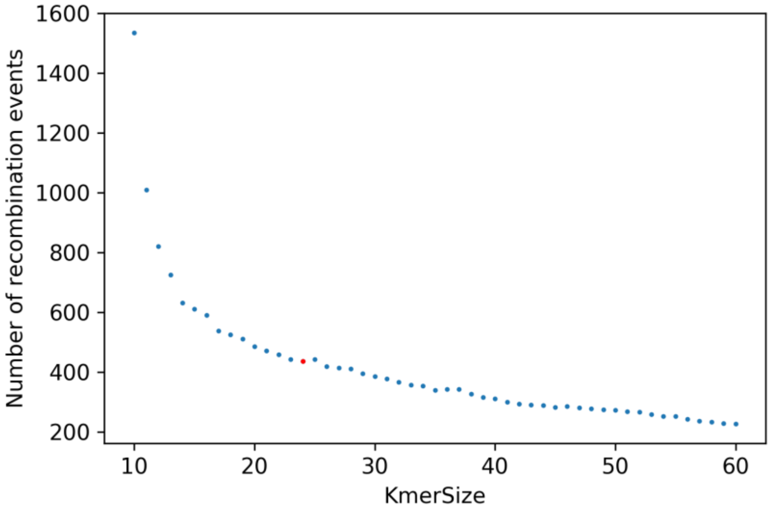
Our method uses a two-step identification procedure. First, we used a larger k-mer size to eliminate paralogs with no recombination events. Second, we used a smaller k-mer size to refine the definition of recombination regions. We used the genome-wide recombination between BG2 and BG3993 to search for optimal k-mer size. We do not have estimates of recombination and mutational events to directly assess sensitivity and specificity for each k-mer size. Instead, for each k-mer size, we counted the number of changed k-mers between a recombinant gene, and its ortholog, and further classified them as recombinant k-mers if they are shared with a paralog, or mutational k-mers if not. The optimized k-mer query size is expected to have maximum ratio of recombinant k-mer to all change k-mer to identify as many recombinant k-mers as possible. We also counted the number of recombination events to quantify decreased sensitivity with larger k-mer sizes.

We performed recombination analysis with all the k-mer sizes from 10 to 60, and compared the change in the ratio of recombinant k-mer to all changed k-mers and change in recombination events (Figure 15). We found that the ratio of recombinant k-mer to all changed k-mers form a peak around k-mer size = 20, and the ratio is almost linearly decreased after k-mer size=23. Interestingly, we found that the number of recombination events and k-mer size also forms a linear relationship after k-mer size=18. We selected k-mer size=24 as the proximal terminal point in the linear trend.

In the second step, we perform our k-mer based recombination analysis using only the ortholog and the identified paralogs from the first step. A smaller k-mer size is used to increase sensitivity. In addition, we found that the smaller k-mer size reduced predicted recombination events with suboptimal paralogs (paralogs that have fewer recombination-defining SNVs but longer homologous regions). We chose k-mer size based on reduction of suboptimal paralogs using the average recombined paralog per recombinant gene (Figure 16). For K-mer size < 11, recombinant k-mers are mis-identified as k-mers shared with orthologs because of short motifs randomly shared within the gene, leading to fewer predicted recombined genes, with more paralogs per recombined gene. The optimal k-mer size of the second step is 14 because analysis using 14-mers minimizes the average number of paralogs per recombined gene.



(A)



(B)

Figure 15 Optimization of k-mer size in the first step

The k-mer size is chosen based on ratio of recombinant k-mer to all changed k-mer as well as number of recombination events. (A) Ratio of recombinant k-mer to all changed k-mer at k-mer size from 10 to 60. The data of k-mer size = 24 is highlighted in red. (B) Number of recombination events at k-mer size from 10 to 60. The data of k-mer size = 24 is highlighted in red.

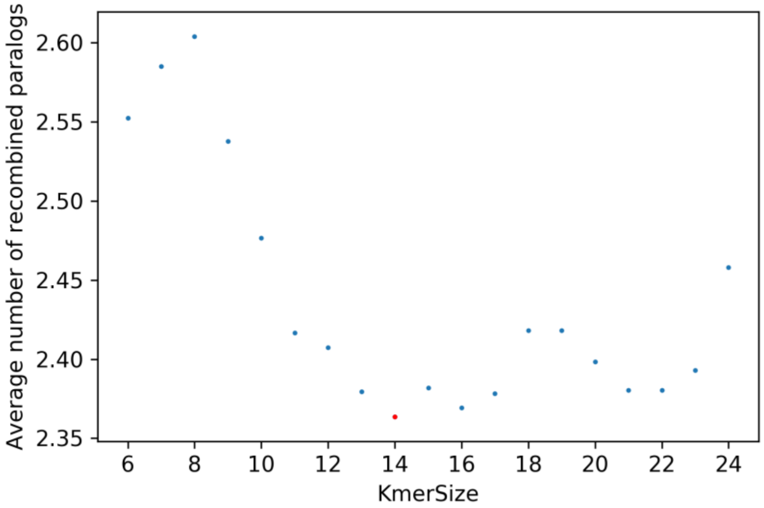
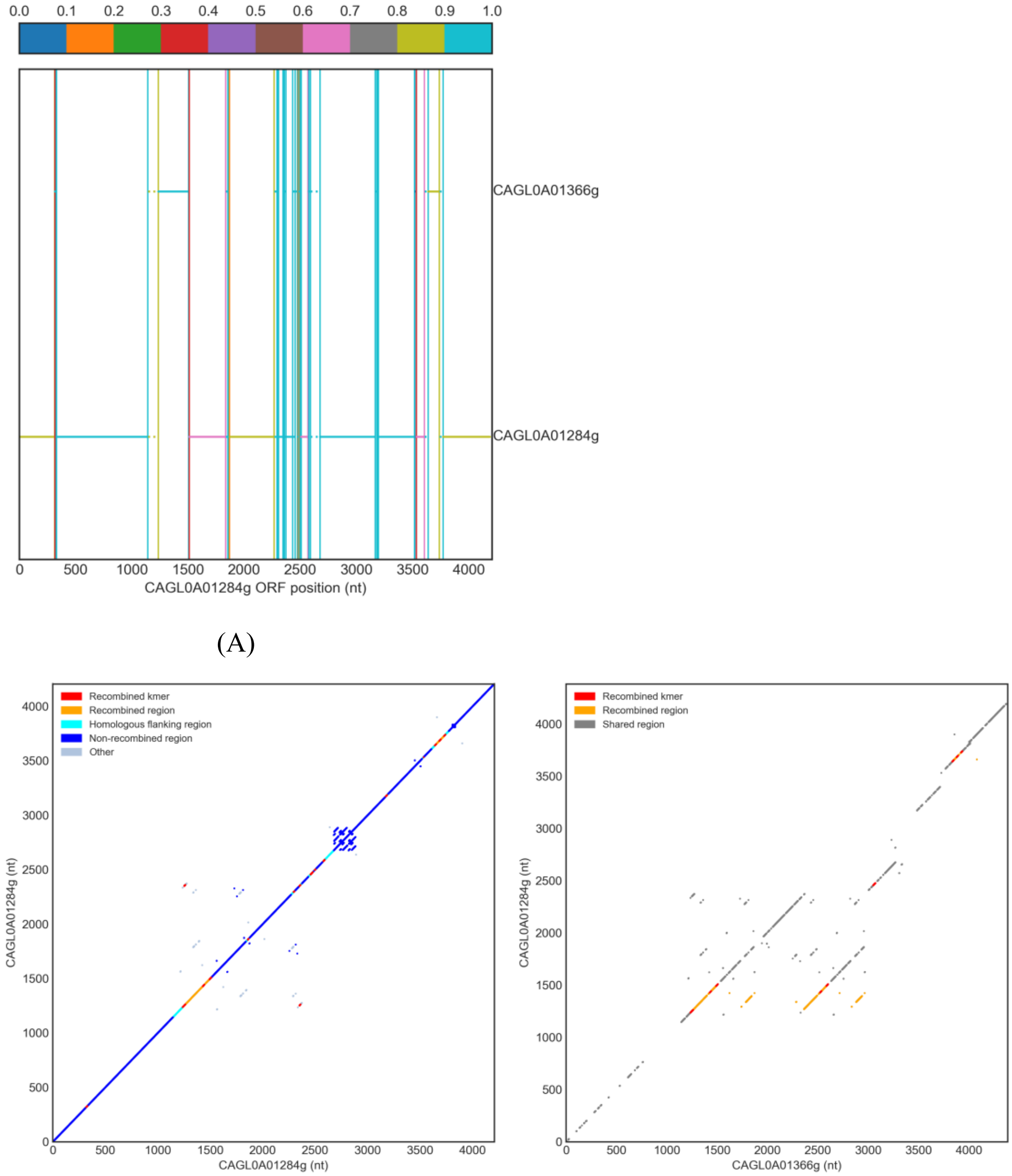


Figure 16 Optimization of k-mer size in the second step

The average number of recombined paralogs per gene using k-mer size from 6 to 24. We observed an average higher than that in the first step (k-mer size=24) because we lose all the recombination events for some genes with small k-mer sizes, resulting in fewer recombined gene to calculate the average. The higher average indicates that those remnant recombined genes have more recombined paralogs than the original set of recombined genes. The data of k-mer size=14 is highlighted in red.

Identification of non-allelic mitotic recombination events

We used our algorithm to identify various classes of recombination events between *C. glabrata* strains. One simple class of recombination is recombination within non-repetitive regions. Recombination events in BG2 CALG0A01284g (*EPA10*) gene relative to CBS138 genes is illustrated as an example of our analysis (Figure 17). BG2 *EPA10* underwent recombination events with BG2 CAGL0A01366g (*EPA9*), observable as sequence variation in CBS138 *EPA10*. In the map of recombined paralogs, there is one minimal recombination region between 1100-1500 nt in the ORF of CBS138 *EPA10*, resulting from recombination with BG2 CAGL0A01366g (*EPA9*). Over 90% of the k-mers in this region is shared with the recombinant. Furthermore, there is a long left homologous region with 80%-90% shared k-mers. According to the structure of CBS138 *EPA10*, the minimal recombination region between 1100-1500 nt is a non-repetitive region encoding multiple recombined k-mers (k-mers shared with the paralog but not with the ortholog). This minimal recombination region is derived from the region between nt 1100-1500 of the recombined paralog, BG2 *EPA9*.



(A)

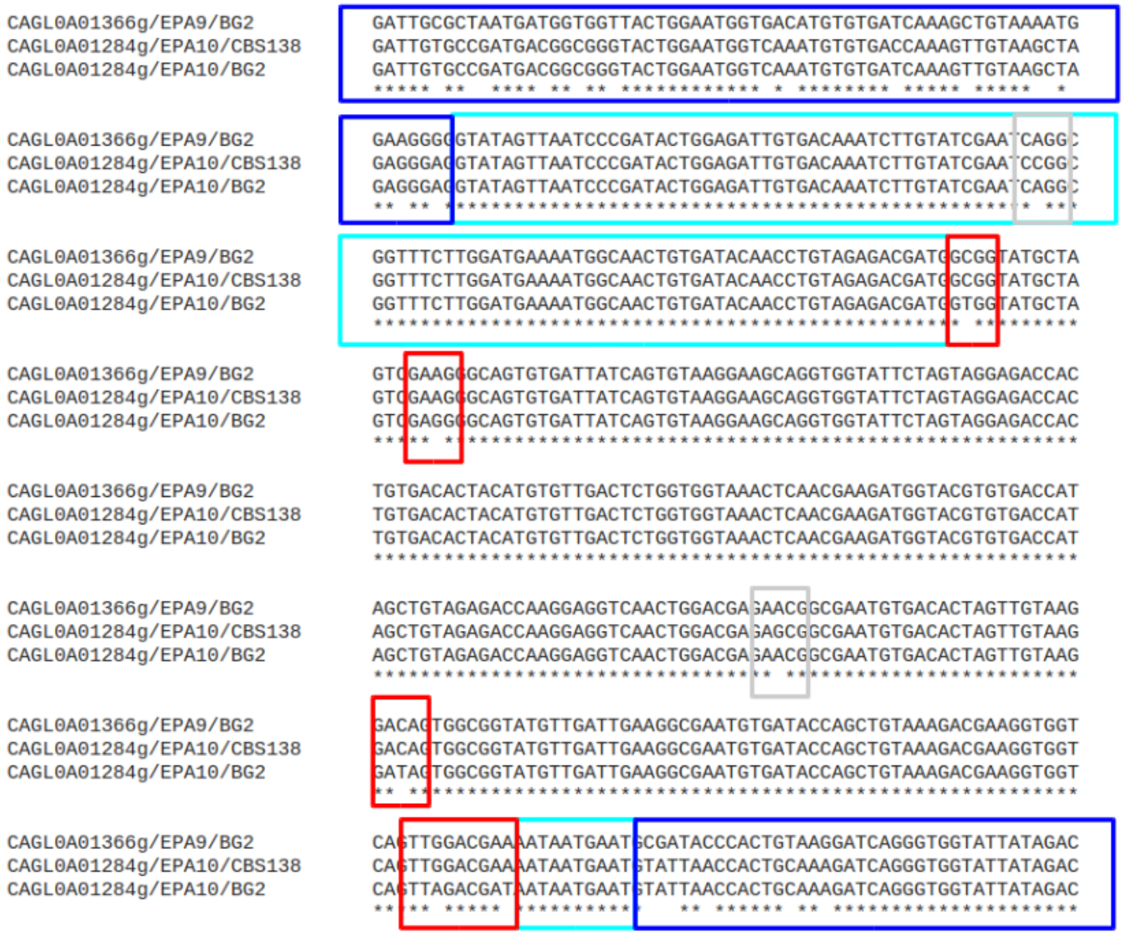
(B)

(C)

Figure 17 Recombination of BG2 CAGL0A01284g (*EPA10*) relative to CBS138

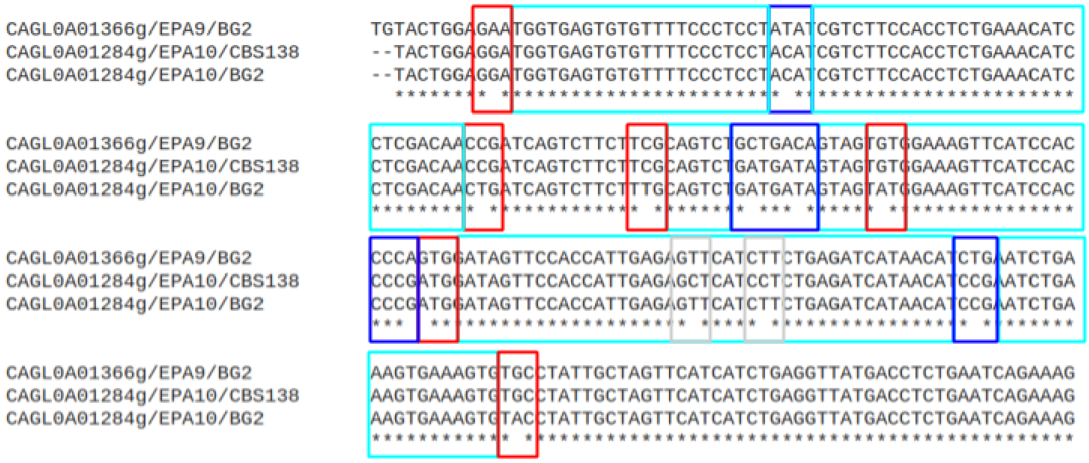
(A) The identification of recombined paralogs of *EPA10*. The location of the recombination regions in CBS138 *EPA10* ORF is shown in the x-axis, and the ortholog and recombined paralogs in BG2 are shown in the y-axis. Each solid horizontal line is a minimal recombination region with the paralog or a region with no recombination relative to the ortholog; each dashed horizontal line is the homologous flanking region. Vertical lines are boundaries of recombination regions. The ratio of shared k-mers between the paralog/ortholog to the total number of k-mers are illustrated in colors shown in the color bar. Therefore, the color of horizontal lines and the proceeding vertical lines indicates the homology between the recombined gene and the corresponding ortholog/paralog. A poor shared ratio indicates a hypermutated region, therefore, indicates a low-confidence recombination event. On the contrary, a high shared ratio indicates a high-confidence recombination event. (B) Illustration of recombination events relative to the gene structure. The dotplot of the CBS138 *EPA10* ORF genome sequence is drawn with window size=kmer size for analysis (window size=14 nt). Non-recombined regions are colored in dark blue; homologous flanking regions are colored in cyan and the minimal recombination regions are colored in orange. The internal homology between non-recombined regions and recombination regions, if any, are indicated in light steel blue. In addition, the recombined k-mers shared with the paralog but not with the ortholog within the minimal recombination regions are colored in red. (C) Comparison of the structure of CBS138 *EPA10* with the recombined paralog, BG2 CAGL0A01366g (*EPA9*) highlighting recombination regions and specific k-mers. The shared 14-mers in the minimal recombination regions between *EPA9* and *EPA10* are colored in red. The regions not participating in recombination are colored in grey. The recombined k-mers are colored in red.

The corresponding sequences from BG2 *EPA9*, BG2 *EPA10,* and CBS138 *EPA10* of the recombination between 1100-1500 was extracted, and displayed as a multiple sequence alignment to illustrate the basic concept of our method using MUSCLE (Figure 18A) (Edgar, 2004). The minimal recombination region with BG2 *EPA9* is identified by the recombined SNVs shared with CBS138 *EPA10* and BG2 *EPA9*, but not with the ortholog, BG2 *EPA10*. Notably, if we ascribed all variation in this region to mutation, we would estimate 7 SNPs in the approximately 400 bp region, amounting to 98.25% sequence identity. We might conclude that this region is hypermutated given the overall ORF sequence identity between BG2 and CBS138 , genome wide, is 99.5%. In fact, there are only 2 SNPs derived from mutation, which is in agreement with the 0.5% sequence divergence in ORF sequences; the other 5 variants are due to recombination. In a second region around 2500 nt in CBS138 *EPA10*, we document many small recombination events clustered together, indicating a region in which SNVs shared only between the two *EPA10* genes alternate with SNVs shared only between CBS138 *EPA10* and BG2 *EPA9* within this small region (Figure 18B).



```
CAGL0A01366g/EPA9/BG2         GATTGCGCTAATGATGGTGGTTACTGGAATGGTGACATGTGTGATCAAAGCTGTAAAATG
CAGL0A01284g/EPA10/CBS138     GATTGTGCCGATGACGGCGGGTACTGGAATGGTCAAATGTGTGACCAAAGTTGTAAGCTA
CAGL0A01284g/EPA10/BG2        GATTGTGCCGATGACGGCGGGTACTGGAATGGTCAAATGTGTGATCAAAGTTGTAAGCTA
                              ***** **  **** ** ** ************ * ******** ***** *****  * 

CAGL0A01366g/EPA9/BG2         GAAGGGGGTATAGTTAATCCCGATACTGGAGATTGTGACAAATCTTGTATCGAATCAGGC
CAGL0A01284g/EPA10/CBS138     GAGGGAGGTATAGTTAATCCCGATACTGGAGATTGTGACAAATCTTGTATCGAATCCGGC
CAGL0A01284g/EPA10/BG2        GAGGGAGGTATAGTTAATCCCGATACTGGAGATTGTGACAAATCTTGTATCGAATCAGGC
                              ** ** ************************************************** ***

CAGL0A01366g/EPA9/BG2         GGTTTCTTGGATGAAAATGGCAACTGTGATACAACCTGTAGAGACGATGGCGGTATGCTA
CAGL0A01284g/EPA10/CBS138     GGTTTCTTGGATGAAAATGGCAACTGTGATACAACCTGTAGAGACGATGGCGGTATGCTA
CAGL0A01284g/EPA10/BG2        GGTTTCTTGGATGAAAATGGCAACTGTGATACAACCTGTAGAGACGATGGTGGTATGCTA
                              ************************************************* **********

CAGL0A01366g/EPA9/BG2         GTCGAAGGGCAGTGTGATTATCAGTGTAAGGAAGCAGGTGGTATTCTAGTAGGAGACCAC
CAGL0A01284g/EPA10/CBS138     GTCGAAGGGCAGTGTGATTATCAGTGTAAGGAAGCAGGTGGTATTCTAGTAGGAGACCAC
CAGL0A01284g/EPA10/BG2        GTCGAGGGGCAGTGTGATTATCAGTGTAAGGAAGCAGGTGGTATTCTAGTAGGAGACCAC
                              ***** ******************************************************

CAGL0A01366g/EPA9/BG2         TGTGACACTACATGTGTTGACTCTGGTGGTAAACTCAACGAAGATGGTACGTGTGACCAT
CAGL0A01284g/EPA10/CBS138     TGTGACACTACATGTGTTGACTCTGGTGGTAAACTCAACGAAGATGGTACGTGTGACCAT
CAGL0A01284g/EPA10/BG2        TGTGACACTACATGTGTTGACTCTGGTGGTAAACTCAACGAAGATGGTACGTGTGACCAT
                              ************************************************************

CAGL0A01366g/EPA9/BG2         AGCTGTAGAGACCAAGGAGGTCAACTGGACGAGAACGGCGAATGTGACACTAGTTGTAAG
CAGL0A01284g/EPA10/CBS138     AGCTGTAGAGACCAAGGAGGTCAACTGGACGAGAGCGGCGAATGTGACACTAGTTGTAAG
CAGL0A01284g/EPA10/BG2        AGCTGTAGAGACCAAGGAGGTCAACTGGACGAGAACGGCGAATGTGACACTAGTTGTAAG
                              ********************************** *************************

CAGL0A01366g/EPA9/BG2         GACAGTGGCGGTATGTTGATTGAAGGCGAATGTGATACCAGCTGTAAAGACGAAGGTGGT
CAGL0A01284g/EPA10/CBS138     GACAGTGGCGGTATGTTGATTGAAGGCGAATGTGATACCAGCTGTAAAGACGAAGGTGGT
CAGL0A01284g/EPA10/BG2        GATAGTGGCGGTATGTTGATTGAAGGCGAATGTGATACCAGCTGTAAAGACGAAGGTGGT
                              ** *********************************************************

CAGL0A01366g/EPA9/BG2         CAGTTGGACGAAAATAATGAATGCGATACCCACTGTAAGGATCAGGGTGGTATTATAGAC
CAGL0A01284g/EPA10/CBS138     CAGTTGGACGAAAATAATGAATGTATTAACCACTGCAAAGATCAGGGTGGTATTATAGAC
CAGL0A01284g/EPA10/BG2        CAGTTAGACGATAATAATGAATGTATTAACCACTGCAAAGATCAGGGTGGTATTATAGAC
                              ***** ***** ***********   ** ****** ** *********************
```

(A)



```
CAGL0A01366g/EPA9/BG2         TGTACTGGAGAATGGTGAGTGTGTTTTCCCTCCTATATCGTCTTCCACCTCTGAAACATC
CAGL0A01284g/EPA10/CBS138     --TACTGGAGGATGGTGAGTGTGTTTTCCCTCCTACATCGTCTTCCACCTCTGAAACATC
CAGL0A01284g/EPA10/BG2        --TACTGGAGGATGGTGAGTGTGTTTTCCCTCCTACATCGTCTTCCACCTCTGAAACATC
                                ******** ************************ ************************

CAGL0A01366g/EPA9/BG2         CTCGACAACCGATCAGTCTTCTTCGCAGTCTGCTGACAGTAGTGTGGAAAGTTCATCCAC
CAGL0A01284g/EPA10/CBS138     CTCGACAACCGATCAGTCTTCTTCGCAGTCTGATGATAGTAGTGTGGAAAGTTCATCCAC
CAGL0A01284g/EPA10/BG2        CTCGACAACTGATCAGTCTTCTTTGCAGTCTGATGATAGTAGTATGGAAAGTTCATCCAC
                              ********* ************* ******** *** ****** ****************

CAGL0A01366g/EPA9/BG2         CCCAGTGGATAGTTCCACCATTGAGAGTTCATCTTCTGAGATCATAACATCTGAATCTGA
CAGL0A01284g/EPA10/CBS138     CCCGATGGATAGTTCCACCATTGAGAGCTCATCCTCTGAGATCATAACATCCGAATCTGA
CAGL0A01284g/EPA10/BG2        CCCGATGGATAGTTCCACCATTGAGAGTTCATCTTCTGAGATCATAACATCCGAATCTGA
                              ***  ********************** ***** ***************** ********

CAGL0A01366g/EPA9/BG2         AAGTGAAAGTGTGCCTATTGCTAGTTCATCATCTGAGGTTATGACCTCTGAATCAGAAAG
CAGL0A01284g/EPA10/CBS138     AAGTGAAAGTGTGCCTATTGCTAGTTCATCATCTGAGGTTATGACCTCTGAATCAGAAAG
CAGL0A01284g/EPA10/BG2        AAGTGAAAGTGTACCTATTGCTAGTTCATCATCTGAGGTTATGACCTCTGAATCAGAAAG
                              ************ ***********************************************
```

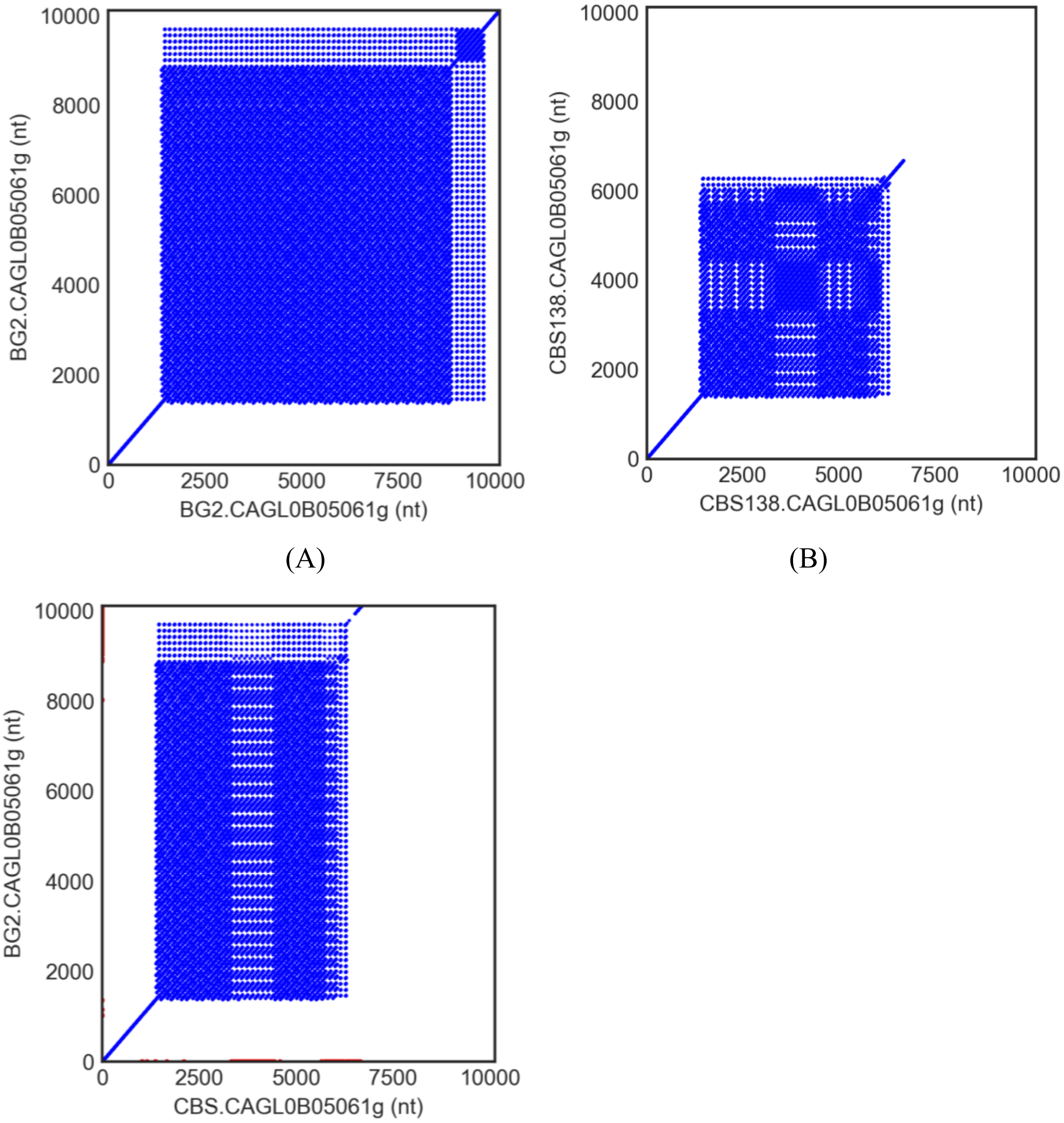
(B)

Figure 18 Multiple sequence alignment of one recombination event.

Multiple sequence alignment of the recombined gene, CBS138 *EPA10*, the ortholog BG2 *EPA10* and the recombined paralog BG2 *EPA9*. The blue boxes annotate the region with no recombination that there are SNVs specific for the two *EPA10* genes. The homologous flanks region are annotated in cyan boxes, where all the sequences, in spite of SNVs derived from mutations, are shared between the three genes. The SNVs shared only between CBS138 *EPA10* and the paralog, BG2 *EPA9* are highlighted in red boxes. These SNVs are derived from one recombination event, and the region between two cyan boxes is the minimal recombination region with recombined SNVs at both ends. SNVs derived from mutations, which only occur in the CBS138 *EPA10*, are highlighted in grey boxes. (A) Multiple sequence alignment of CBS138 *EPA10* 1100-1500 nt ORF region with the corresponding region from BG2 *EPA10* and BG2 *EPA9*. (B) Multiple sequence alignment of CBS138 *EPA10* 2400-2600 nt ORF region with the corresponding region from BG2 *EPA10* and BG2 *EPA9*.

A major advantage of our recombination analysis is that our method simplifies the identification of recombination between repeat regions. For example, the CAGL0B05061g undergoes significant changes in repeat region between BG2 and CBS138 (Figure 19). This gene encodes two tandem-repeat regions in BG2, while the repeat region in CBS138 is more complicated. There are two tandem repeats that share homology with each other, and they alternate in the overall repeat region. When we compared the ORFs between strains, we found that only one tandem repeat is shared. The tandem repeat region close to C-terminus in BG2 is not present in CBS138 (indicated by the specific 15-mers close to the C-terminus); one of the two tandem repeats in CBS138 is not encoded in BG2 either (indicated by the specific 15-mers at the corresponding tandem repeat region(approx 3-4000 nt).

We found that the tandem repeats in CBS138 which are not shared with it ortholog in BG2 result from recombination with the paralog CAGL0F09251g in BG2 (Figure 20). These tandem-repeats are located in two minimal recombination regions between CAGL0B05061g and CAGL0F09251g. The recombined k-mers derive from the tandem-repeat region of BG2 CAGL0F09251g (shown in red in Figure 20B) and are distributed across the tandem repeat regions of CBS138 CAGL0B05061g (shown in red in Figure 20C). Therefore, our algorithm can identify recombination events within tandem repeats.

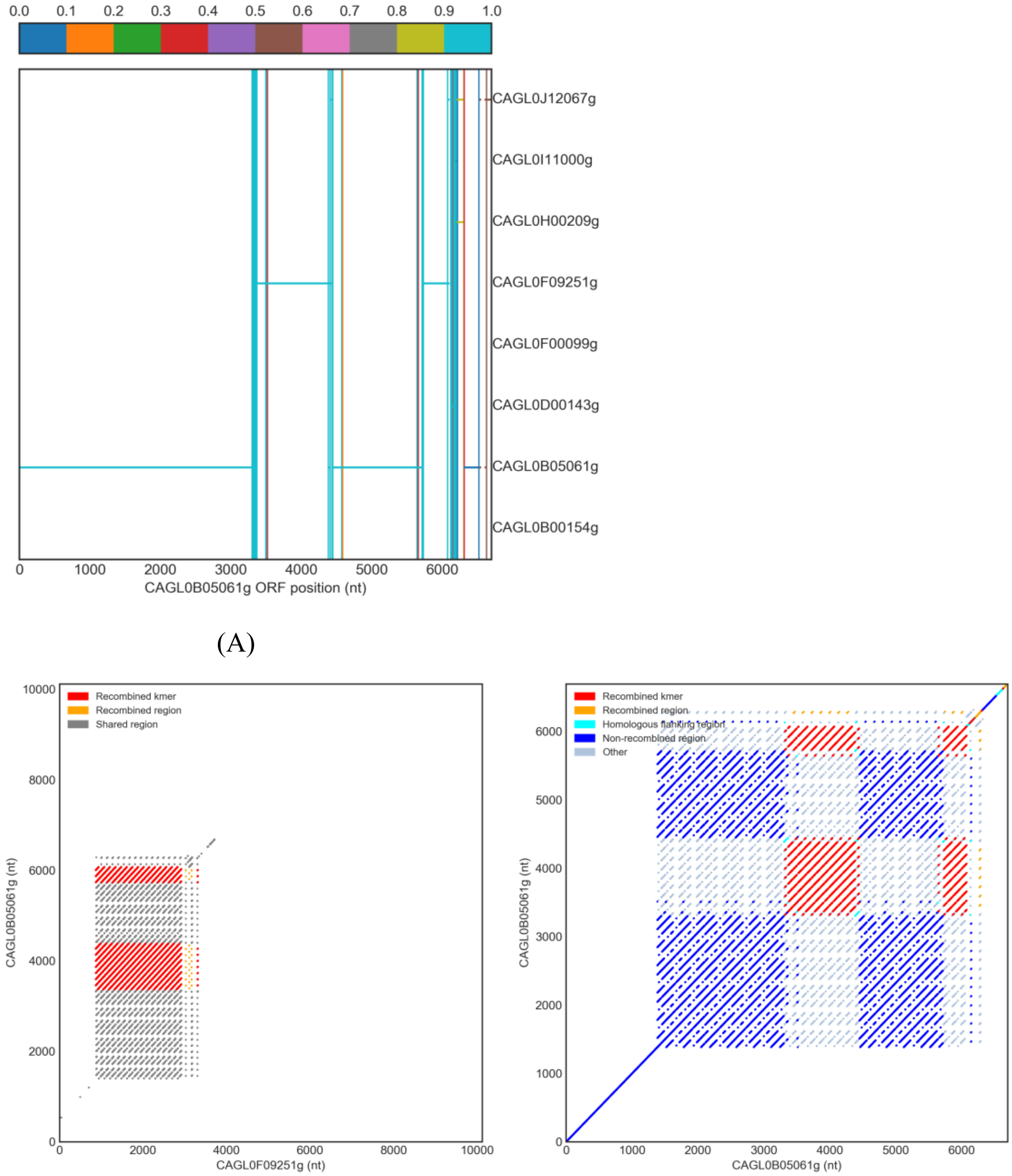


(A) (B)

(C)

Figure 19 Structure of CAGL0B05061g ORF sequence in strain BG2 and CBS138.

All the dotplots are illustrated with window size=15 nt. Each blue dot indicates an exact match of 15 nt. (A) Structure of BG2 CAGL0B05061g. (B) Structure of CBS138 CAGL0B05061g. (C) Comparison of the two ORF sequences. The location of specific 15-mers in BG2 or CBS138 are indicated in red dots along the corresponding axis.



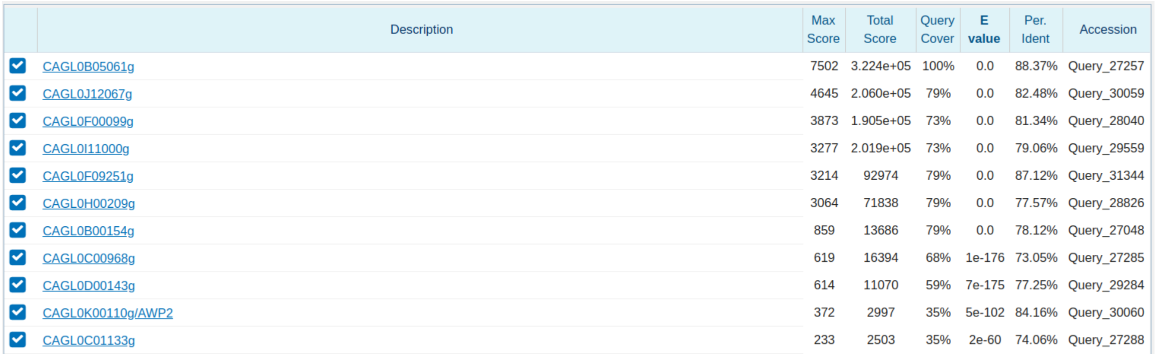
(A)

(B) (C)

Figure 20 Recombination of BG2 CAGL0B05061g and BG2 CAGL0F09251g.

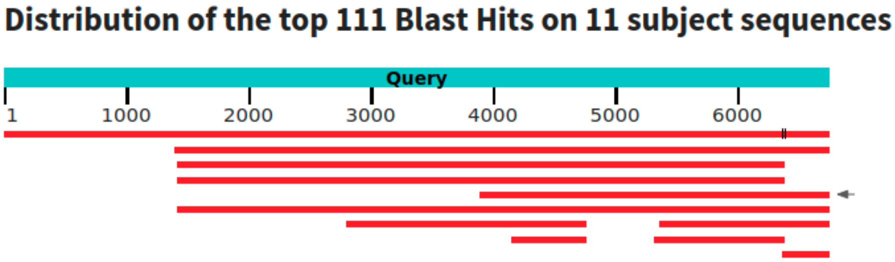
The tandem repeats in CBS138 CAGL0B05061g, which are not shared with the BG2 CAGL0B05061g, result from recombination with BG2 CAGL0F09251g. (A) The identification of recombined paralogs of CAGL0B05061g. The location of the recombination regions in CBS138 CAGL0B05061g ORF is shown in the x-axis, and the ortholog and recombined paralogs in BG2 are shown in the y-axis. Each solid horizontal line is a minimal recombination region with the paralog or a region with no recombination relative to the ortholog; each dashed horizontal line is the homologous flanking region. Vertical lines are boundaries of recombination regions. The ratio of shared k-mers between the paralog/ortholog to the total number of k-mers are illustrated in colors shown in the color bar. Therefore, the color of horizontal lines and the vertical lines indicates the homology between the recombined gene and the corresponding ortholog/paralog. A poor shared ratio indicates a hypermutated region, therefore, indicates a low-confident recombination event. On the contrary, a high shared ratio indicates a high-confident recombination event. (B) Comparison of the structure of CBS138 CAGL0B05061g with the recombined paralog, BG2 CAGL0F09251g highlighting recombination regions and specific k-mers. The shared 14-mers in the minimal recombination regions between CAGL0B05061g and CAGL0F09251g are colored in red. The regions not participating in recombination are colored in grey. The recombined k-mers are colored in red. (C) Illustration of recombination events relative to the gene structure. The dotplot of the CBS138 CAGL0B05061g ORF genome sequence is drawn with window size=kmer size for analysis (window size=14 nt). Non-recombined regions are colored in dark blue; homologous flanking regions are colored in cyan and the minimal recombination regions are colored in orange. The internal homology between non-recombined regions and recombination regions, if any, are indicated in light steel blue. In addition, the recombined k-mers shared with the paralog but not with the ortholog within the minimal recombination regions are colored in red.

It would be challenging to assess recombination using general sequence alignment because the tandem repeats are shared within large gene families. For instance, for the example just discussed, we aligned the genome ORF sequence of CBS138 CAGL0B05061g to all the single-exon ORFs in BG2 using megablast (https://blast.ncbi.nlm.nih.gov/) (Figure 21). The paralog involved in recombination, CAGL0F09251g, is only the 5-th hit, and we cannot easily observe recombination events from the graphical summary of the megablast alignment. Identification of the recombination region would require onerous manual inspection of the multiple sequence alignments (up to all 11 homologous genes, and the 111 Blast hits). Our method significantly simplifies the analysis of recombination events.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	CAGL0B05061g	7502	3.224e+05	100%	0.0	88.37%	Query_27257
☑	CAGL0J12067g	4645	2.060e+05	79%	0.0	82.48%	Query_30059
☑	CAGL0F00099g	3873	1.905e+05	73%	0.0	81.34%	Query_28040
☑	CAGL0I11000g	3277	2.019e+05	73%	0.0	79.06%	Query_29559
☑	CAGL0F09251g	3214	92974	79%	0.0	87.12%	Query_31344
☑	CAGL0H00209g	3064	71838	79%	0.0	77.57%	Query_28826
☑	CAGL0B00154g	859	13686	79%	0.0	78.12%	Query_27048
☑	CAGL0C00968g	619	16394	68%	1e-176	73.05%	Query_27285
☑	CAGL0D00143g	614	11070	59%	7e-175	77.25%	Query_29284
☑	CAGL0K00110g/AWP2	372	2997	35%	5e-102	84.16%	Query_30060
☑	CAGL0C01133g	233	2503	35%	2e-60	74.06%	Query_27288

(A)



(B)

Figure 21 Megablast alignment of CBS138 CAGL0B05061g to BG2 ORFs

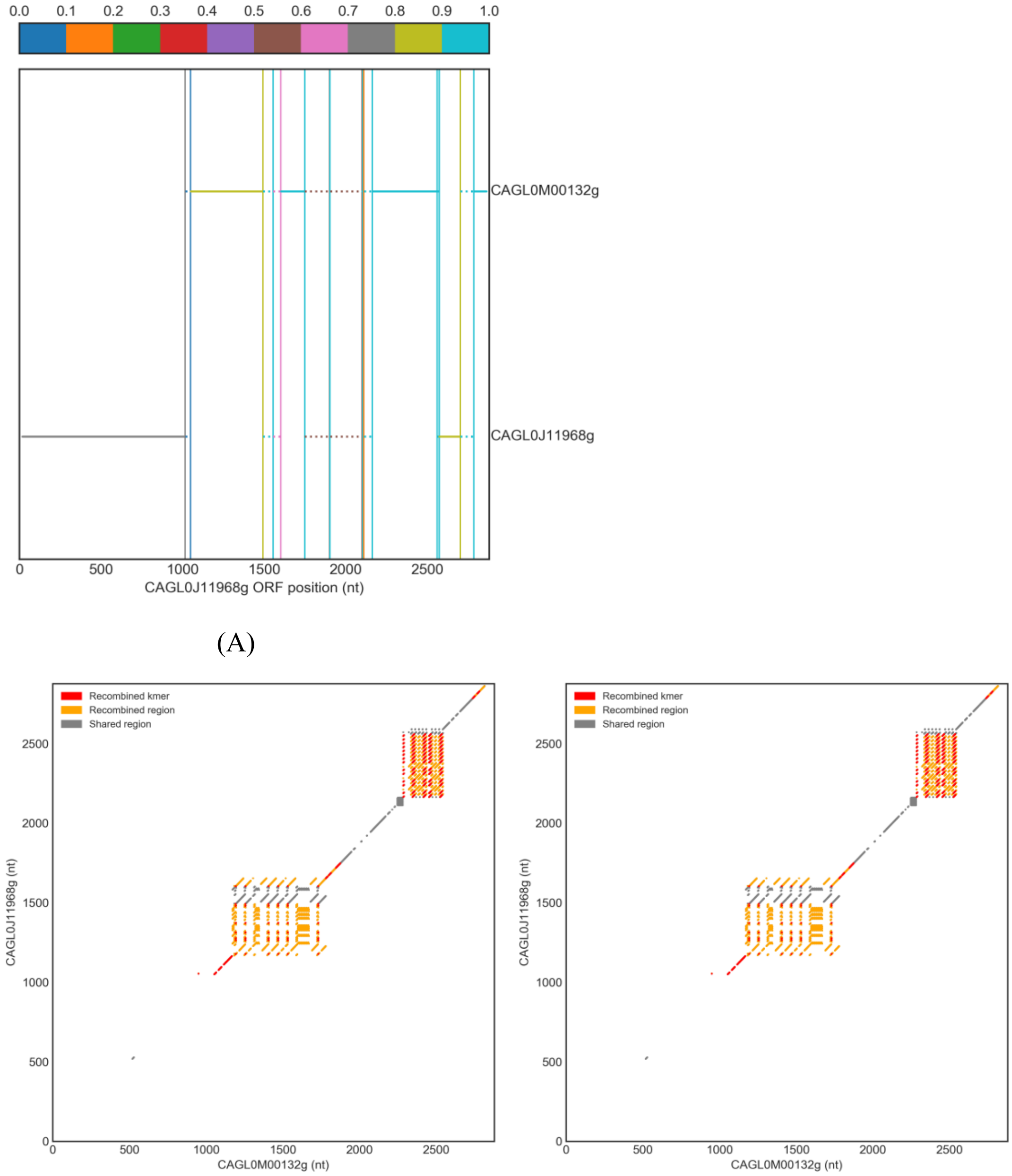
The genome ORF sequence of CBS138 CAGL0B05061g is aligned to all the single-exon ORFs in BG2 using megablast. (A) The ortholog BG2 CAGL0B05061g is the best hit of alignment, while the recombined paralog, CAGL0F09251g, is only the 5-th alignment. (B) Graphic summary of megablast. The alignment to the recombined CAGL0F09251g is indicated by the arrow.

Intergenic recombination in GPI-CWP genes

The GPI-CWP genes are present in large homologous families that are clearly potential substrates for recombination. We were particularly interested, therefore, to assess recombination within this group of genes. There are 92 GPI-CWP genes in BG2, BG3993 and CBS138. CBS138 and BG2 orthologous 74 GPI-CWP genes; BG3993 and BG2 shared 73 GPI-CWP genes. (Since repeat structure within GPI-CWP genes can vary between strains, orthologues are defined by the conserved non repeat containing N-terminal regions in different strains.) According to our genome-wide recombination analysis, 36 of the shared 74 GPI-CWP genes between BG2 and CBS138 undergo intergenic recombination; 46 of the 74 shared GPI-CWP genes between BG2 and CBS138 undergo intergenic recombination (Table S1, Table S2). There are more minimal recombination events (defined in the following section) in GPI-CWP genes that other genes. On average, there are approximately 6.6 and 6.3 recombination events per gene in recombined GPI-CWP genes in CBS138 and BG3993 relative to BG2, respectively. The average recombination events for other genes is only 1.5 and 1.8 in CBS138 and in BG3993, respectively. In *S. cerevisiae,* Zhao *et al.* reported an increased mitotic recombination rate in the tandemly repeated *CUP1* gene clusters (Zhao et al., 2017). However, recombination events in GPI-CWPs also take place between different subtelomeres, and we will characterize the events in GPI-CWPs in further research.

There are various classes of recombination events in the GPI-CWP genes. In the previous section, we have illustrated recombination in the middle of the ORFs in GPI-CWP genes (*EPA9* and *EPA10*) and recombination in tandem-repeat regions (CAGL0B05061g and CAGL0F09251g). The GPI-CWP genes can also have recombination affecting onl the C-terminus. For example, CAGL0J11968g (*EPA15*) undergoes recombination with CALG0M00132g (*EPA12*) (Figure

22)to generate a chimeric gene in which the C-terminal region of BG2 *EPA12* is derived from the C-terminal region of CBS138 *EPA15*.



(A)

(B)

(C)

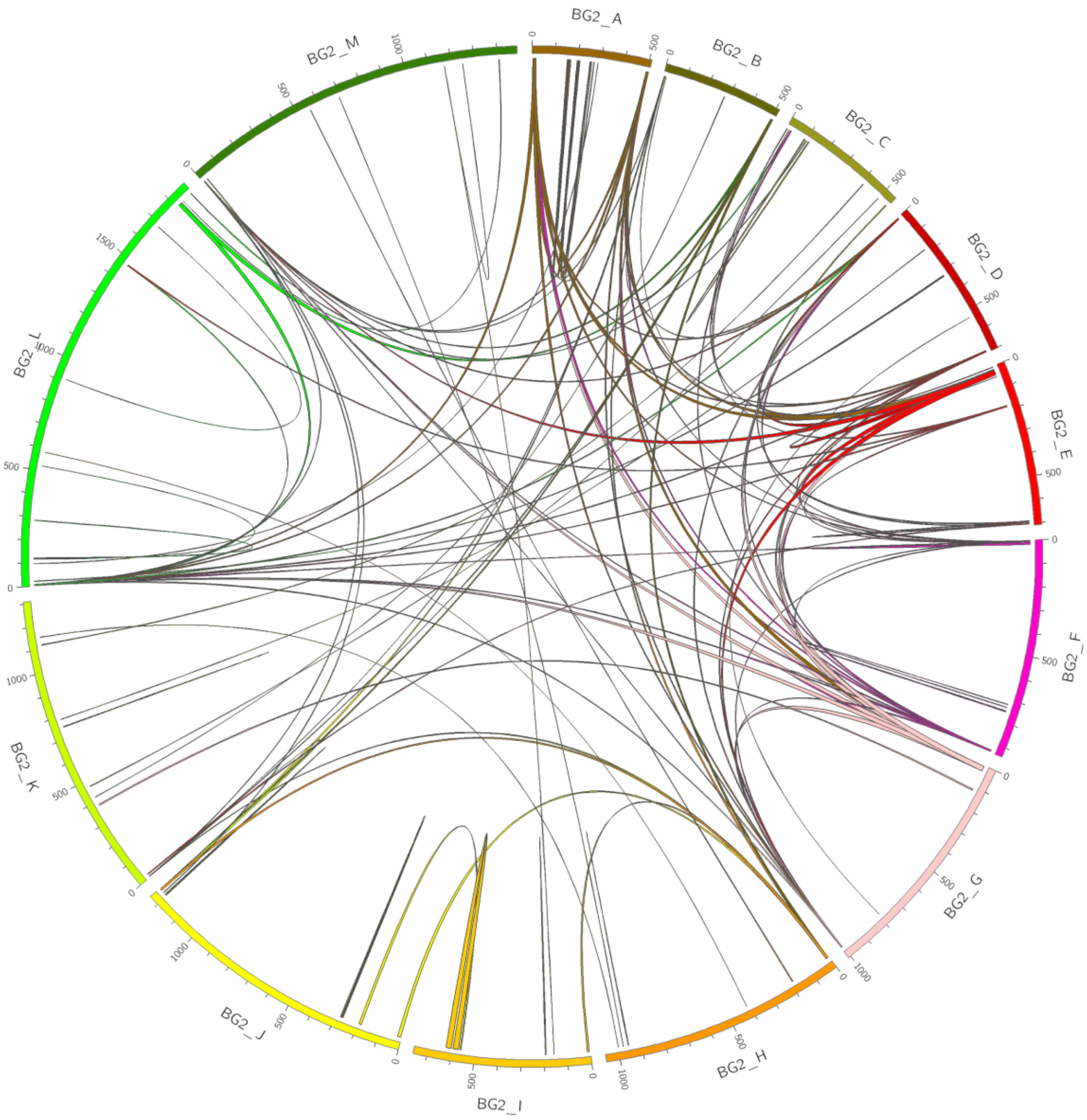
Figure 22 Recombination of BG2 CAGL0J11968g (*EPA15*) and BG2 CAGL0M00132g (*EPA12*).

BG2 *EPA12* recombines with BG2 *EPA15*, resulting in change in C-terminus of the CBS138 *EPA15*. BG2 *EPA12* and *EPA15* also undergo recombination in tandem repeats. (A) The identification of recombined paralogs of *EPA15*. The location of the recombination regions in CBS138 *EPA15* ORF is shown in the x-axis, and the ortholog and recombined paralogs in BG2 are shown in the y-axis. Each solid horizontal line is a minimal recombination region with the paralog or a region with no recombination relative to the ortholog; each dashed horizontal line is the homologous flanking region. Vertical lines are boundaries of recombination regions. The ratio of shared k-mers between the paralog/ortholog to the total number of k-mers are illustrated in colors shown in the color bar. Therefore, the color of horizontal lines and the vertical lines indicates the homology between the recombined gene and the corresponding ortholog/paralog. A poor shared ratio indicates a hypermutated region, therefore, indicates a low-confidence recombination event. On the contrary, a high shared ratio indicates a high-confidence recombination event. (B) Illustration of recombination events relative to the gene structure. The dotplot of the CBS138 *EPA15* ORF genome sequence is drawn with window size=kmer size for analysis (window size=14 nt). Non-recombined regions are colored in dark blue; homologous flanking regions are colored in cyan and the minimal recombination regions are colored in orange. The internal homology between non-recombined regions and recombination regions, if any, are indicated in light steel blue. In addition, the recombined k-mers shared with the paralog but not with the ortholog within the minimal recombination regions are colored in red. (C) Comparison of the structure of CBS138 *EPA15* with the recombined paralog, BG2 *EPA12* highlighting recombination regions and specific k-mers. The shared 14-mers in the minimal recombination regions between *EPA15* and *EPA12* are colored in red. The regions not participating in recombination are colored in grey. The recombined k-mers are colored in red.

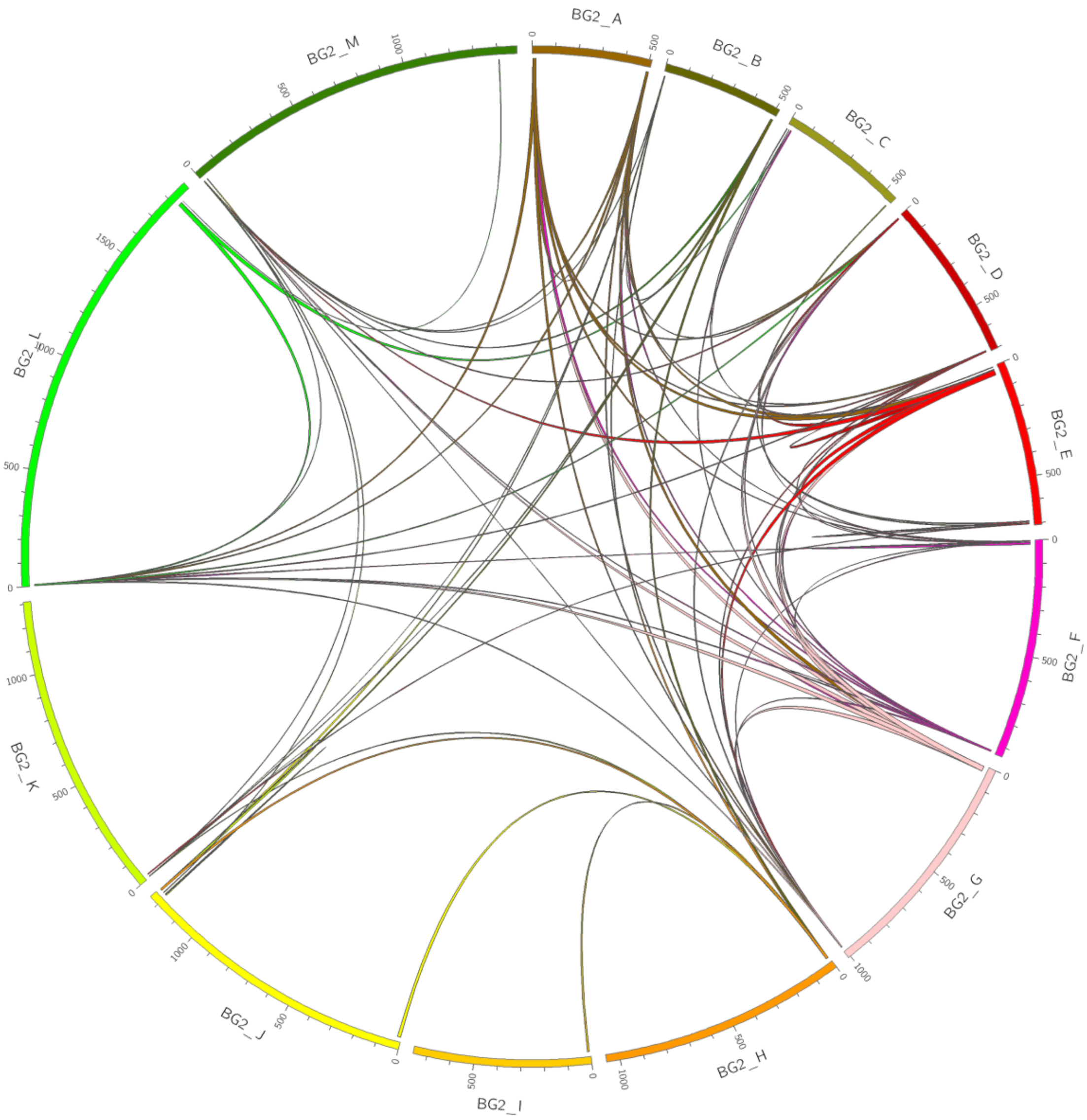
Genome-wide non-allelic recombination analysis

We performed genome-wide recombination analysis with the two-step identification for BG2 genes relative to CBS138 genes, *i.e,* we identified recombination events between BG2 ORF sequences resulting in sequence variations ORFs in strain CBS138. Recombination events within tandem repeats can generate multiple apparent recombination events because the recombinant SNVs can be separated by SNVs specific for the ortholog genes. These SNVs result in arrays of called recombination events within a repeat region. Such distribution of recombinant SNVs could clearly result from a single recombination between paralogs followed by intragenic events that re-order the recombinant and non-recombinant SNVs. To be conservative in calling intergenic recombination events, we collapse tandem repeat recombination events: we extracted all the minimal recombination regions from the recombinant ORFs in strain CBS138, aligned these regions to strain BG2 using BLASTN. Regions of BG2 that are aligned to the same CBS138 recombination region were collapsed; minimal recombination regions within the same gene in CBS138 that aligned to the same region (or collapsed region) in BG2 were also collapsed. Thus, we count identical called recombination events within repeat regions only once in our analysis, and obtain a conservative accounting of recombination events, minimizing the influence of intragenic repeat extensions, contractions and rearrangements.

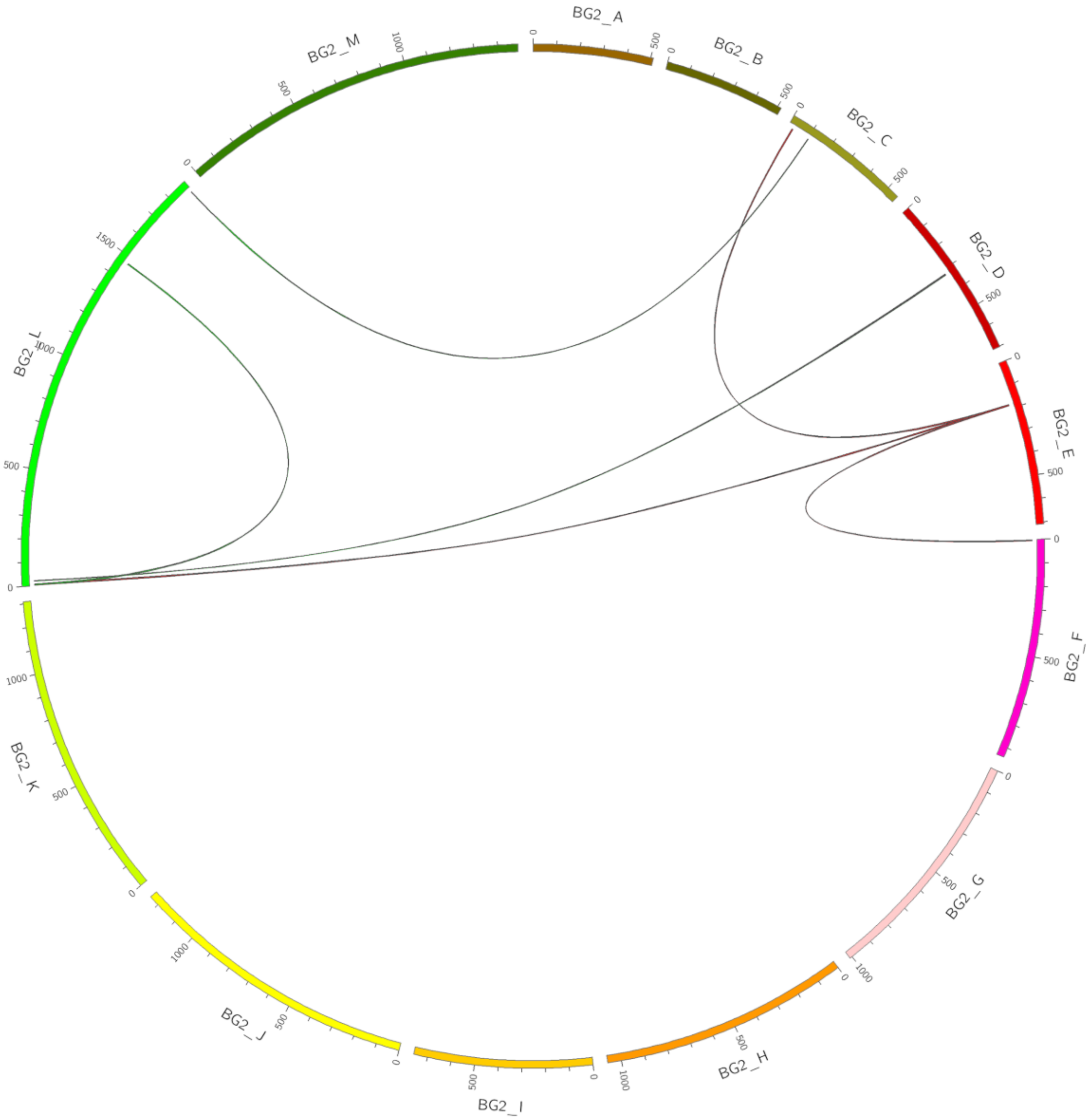
We identified 293 minimal recombination events after collapsing all the events in repeat regions (Table S1). The recombination events are derived from 158 pairs of recombined genes in BG2 (the “recipient” ortholog and “donor” paralog for the chimeric recombined gene in CBS138) (Table S2). We found those gene pairs are distributed across the chromosome (Figure 23). The majority of the recombined gene pairs are related to subtelomeres. 102 of the 159 pairs undergo recombination between subtelomeres, with 100 of the 102 recombined pairs being genes in different subtelomeres. The GPI-CWP genes contribute almost all the recombination in subtelomeres, 100 of the 102 pairs are pairs of GPI-CWP genes. Second, there are seven gene pairs indicating recombination between subtelomeric regions and non-subtelomeric regions, and all seven pairs are pairs of GPI-CWP genes. Lastly, the mitotic recombination can also take place in non-subtelomeric regions as well as in non-GPI-CWP genes. There are 49 recombined pairs in the non-subtelomeric regions. 12 of the 49 pairs are pairs of GPI-CWP genes, and 37 of the 49 pairs are pairs of non-GPI-CWP genes. Of these, 3 of the 12 pairs of GPI-CWP genes are interchromosomal (the genes in the gene pair are on different chromosomes) and 16 of the 37 pairs of non-GPI-CWP genes are interchromosomal.


BG2_A
BG2_B
BG2_C
BG2_D
BG2_E
BG2_F
BG2_G
BG2_H
BG2_I
BG2_J
BG2_K
BG2_L
BG2_M
0
500
1000
1500

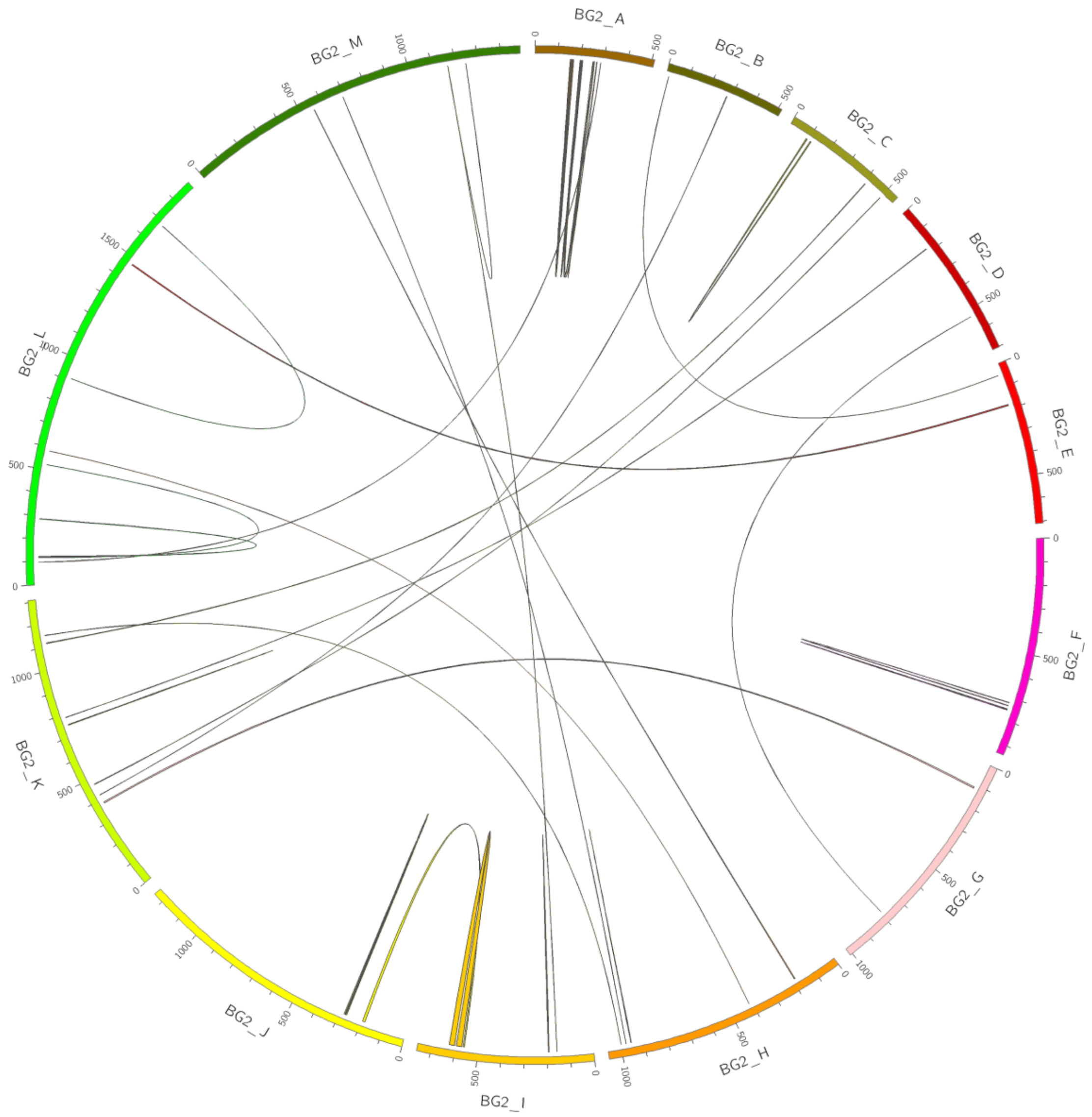
(A)


BG2_A
BG2_B
BG2_C
BG2_D
BG2_E
BG2_F
BG2_G
BG2_H
BG2_I
BG2_J
BG2_K
BG2_L
BG2_M
0
500
1000
1500

(B)


BG2_A
BG2_B
BG2_C
BG2_D
BG2_E
BG2_F
BG2_G
BG2_H
BG2_I
BG2_J
BG2_K
BG2_L
BG2_M
0
500
1000
1500

(C)



(D)

Figure 23 Genome-wide recombination map in BG2 relative to CBS138

Recombination events are identified from orthologs between strain BG2 and CBS138 that the sequence variations between the two orthologs are derived from recombination event with a paralog in BG2. The "recipient" ortholog encoding the sequence variation and the "donor" paralog that contributed the corresponding sequence are recombined gene pairs in BG2 relative to CBS138. The genome-wide recombination map illustrated the genome location of these recombined gene pairs in strain BG2 using CIRCOS (Krzywinski et al., 2009). (A) All the recombined gene pairs. (B) recombined gene pairs in subtelomeres. (C) recombined gene pairs between subtelomeres and non-subtelomeric regions. (D) recombined gene pairs in non-subtelomeric regions.

Distribution of Recombination Gene Pairs

Overall, most of the recombined gene pairs are pairs of subtelomeric genes, and pairs of GPI-CWP genes. GPI-CWP genes form large families, and half of the GPI-CWPgenes are located within subtelomeres. The larger number of recombined gene pairs at subtelomeres, or between GPI-CWP pairs may simply reflect more potential homologous gene pairs, rather than any preference for recombination between GPI-CWP genes or subtelomeric genes.. To address whether subtelomeres and GPI-CWP genes have higher rates of recombination, we need to know the number of homologous gene pairs (that are substrates for recombination in the first place). To define all homologous gene pairs, we performed pairwise alignment of all single-exon BG2 ORFs using BLASTN, and evaluated the homology between two genes that have alignment with e-value < 1e-6 using a weighted average bitscore of all the alignments with e-value <1e-6 between two ORFs (see Experimental Procedures). We found that there is a lower limit for recombined gene pairs (Figure 24), and operationally define the homology pairs are gene pairs with a weighted average above this limit (bitscore > 70).

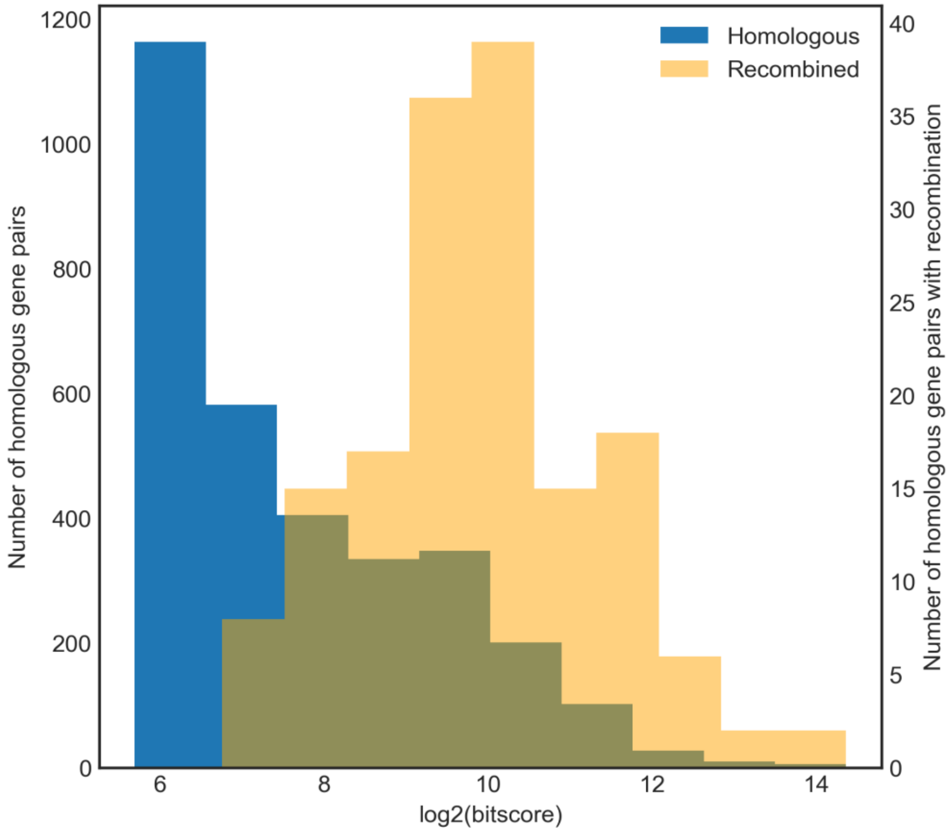


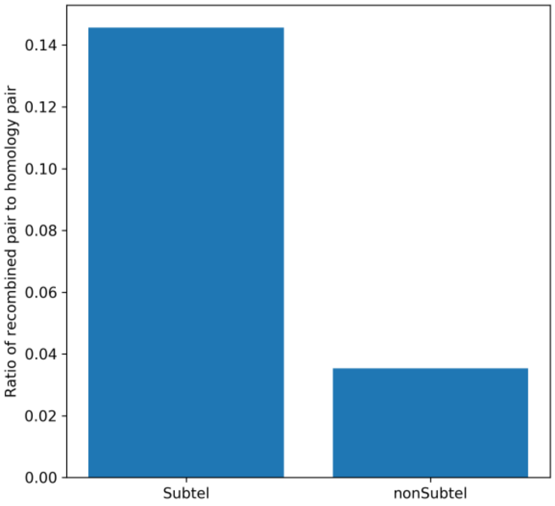
Figure 24 Histogram of weighted average of bitscores.

The histogram of weighted average of bitscores (see Experimental Procedures) of all homologous gene pairs (BLASTN e-value < 1e-6) is indicated in blue, and that of recombined gene pairs is indicated in yellow.

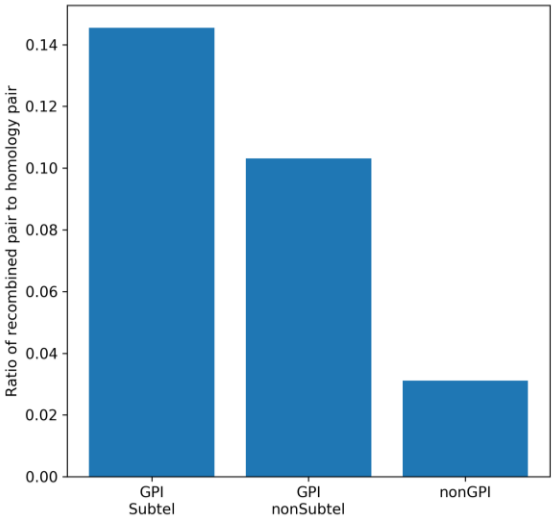
We estimated the recombination rate of any class of gene pairs using the ratio of number of recombined pairs to the number of homologous gene pairs within that class. We first compared the recombination rate for gene pairs in subtelomeres or non-subtelomeres (Figure 25), and observed higher rate of recombination as expected. To assess recombination rates within GPI-CWP genes relative to non-GPI-CWP genes, we wanted to control for the effect of sub-telomeric location, since half of the GPI-CWP genes are present in subtelomeres. We therefore classified the GPI-CWP genes into subtelomeric GPI-CWP genes genes and non-subtelomeric GPI-CWP genes, and we compared each of these subclasses to non-GPI-CWP genes (Figure 25). We found that subtelomeric GPI-CWP gene pairs have a higher rate of recombination than non-subtelomeric GPI-CWP genes. Nevertheless, non-subtelomeric GPI-CWP gene pairs also have higher recombination rates than non-GPI-CWP pairs. We conclude, therefore, that GPI-CWP genes exhibit higher probability of mitotic exchange than non-GPI-CWP pairs.

Subtelomeres in *S. cerevisiae* are clustered via Sir3 protein mediated mechanisms, and this is likely true in *C. glabrata* with the subtelomeres colocalizing in a small number of foci in the nucleus (Cormack Lab, data not shown). This mechanism may place subtelomeric genes physically close to one another, a prerequisite for recombination. Adjacent genes in the non-subtelomeric region may also have a higher rate of recombination because they are close to each other. What about genes in the body of the chromosome? We observed that the majority of recombination events in non-subtelomeric regions occur between adjacent genes (Figure 23), and there is a slight preference of recombination between adjacent gene pairs compared to non-adjacent gene pairs (Figure 26). Even this slight preference is due to recombination between clustered non-subtelomeric GPI-CWP genes, and may simply reflect the higher rates of recombination of this class of genes. Thus, if we exclude GPI-CWP genes, there is no difference in apparent recombination rates between adjacent (<50 kb) and non-adjacent non-GPI-CWP genes (distance < 50 kb) (Figure 26). Therefore, while most of the non-subtelomeric recombination that we document occurs between adjacent genes, surprisingly, we do not have evidence that recombination probability is distance-dependent.

To conclude, we observed a higher apparent rate of recombination among subtelomeric genes and separately, among GPI-CWP genes, but recombination is not limited to the two groups of genes, since we document recombination events between non-GPI-CWP genes in non-subtelomeric regions.



(A)



(B)

Figure 25 Ratio of recombined pairs to homologous pairs

We estimate the apparent recombination rate using the ratio of recombined gene pairs to homologous gene pairs. (A) Comparison of recombination rate between subtelomeric genes and non-subtelomeric regions. (B) Comparison of the recombination rate amongst subtelomeric GPI-CWP, non-subtelomeric GPI-CWP genes and non-GPI-GWP genes.

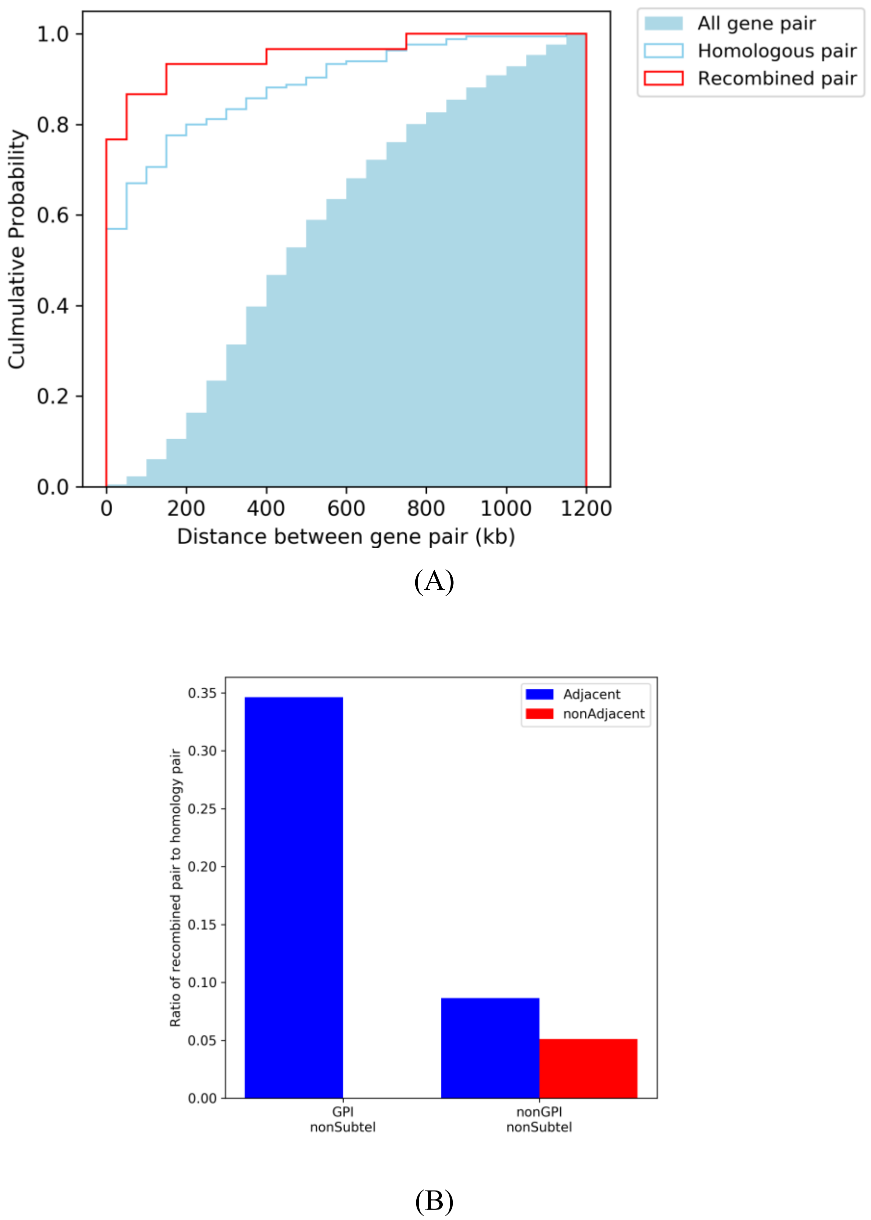


Figure 26 Influence of distance between gene pairs in recombination

(A) The cumulative distribution of distance between non-subtelomeric recombined pairs (in the same chromosome) are illustrated in the red line, and that of the homologous gene pairs are indicated in blue line. The blue area is the distribution for all possible gene pairs in non-subtelomeric regions in the same chromosome. (B) Ratio of recombined gene pairs to homologous adjacent (distance < 50 kb) and non-adjacent gene pairs.

Discussion

Our k-mer based recombination analysis pipeline successfully identified complicated recombination events between non allelic homologous genes. The events we identify are actually gene conversion events in which flanking regions of the genome are not recombined. These events do alter gene structure and the predicted structure of the encoded proteins. For example, we identified non-allelic mitotic recombination events between GPI-CWP genes, leading to exchange of simple SNPs between homologs, exchange of tandem-repeat regions, and exchange of the entire C-terminal regions. The functional consequence of these events is not clear, but since they alter the cell surface proteome, may alter the fitness of *C. glabrata* to different environments. In the future, it will be of interest to understand if recombination plays a role in adaptation to different host sites, which may be addressable by experimental adaptation of strains to different host niches, or by sequencing and analyzing multiple clinical isolates from different sites (blood, organs, mucosa).

Our analysis suggests that subtelomeric regions are somewhat more dynamic for mitotic recombination than the body of the chromosomes. However, recombination was not limited to subtelomeric region, and surprisingly occurred between distant and adjacent gene pairs, with similar apparent rates. A major caveat of our work is that the events we are monitoring are historic events, the imprint of which we see in these relatively distantly related strains. In data not shown, our analysis of recombination in the four clinical isolates from chapter 4 shows minimal intragenic recombination and no examples of non-allelic recombination of the kind analyzed in this chapter. It remains for future work to understand whether non-allelic recombination plays a role in microevolutionary adaptation over the short time periods of infection or carriage in a single individual.

Reference

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Salzberg, S.L., Pertea, M., Fahrner, J.A., and Sobreira, N. (2014). DIAMUND: direct comparison of genomes to detect mutations. Hum. Mutat. *35*, 283–288.

Zhao, Y., Dominska, M., Petrova, A., Bagshaw, H., Kokoska, R.J., and Petes, T.D. (2017). Properties of Mitotic and Meiotic Recombination in the Tandemly-Repeated CUP1 Gene Cluster in the Yeast Saccharomyces cerevisiae. Genetics *206*, 785–800.

Conclusion

In this thesis, we first applied long read SMRT sequencing reads to perform *de novo* genome assemblies of six *Candida glabrata* strains, CBS138, BG2 and BG3993-96 (chapter 2-4). We successfully obtained telomere to telomere assemblies of the six strains using the Canu assembler (Koren et al., 2017), with minimal manual corrections. CBS138 is the type strain of *C. glabrata*, and BG2 is the lab strain with which our lab uses for genetic analysis. The serial clinical isolates, BG3993-96 are vaginal isolates collected from a single patient during physician visits over 21 months, and they were sequenced to assess microevolution during infection. For all strains, we obtained correct structures of the subtelomeric regions, as well as the correct structures of GPI-CWP encoding genes. Within the clinical isolates, we identified one candidate gene *YAP6*, which is mutated in two of the four strains and likely accounts for these two strains being hyper-adherent relative to the other two less adherent strains in the same series.

GPI-CWP genes encode complex tandem-repeat regions and with our assemblies, we could document significant sequence variation between strains within the GPI-CWP genes. We are interested in how these GPI-CWP genes exchange information through non-allelic mitotic recombination. In chapter 5, we describe a k-mer counting based method to identify (historic) recombination events between GPI-CWP genes in different strains. This method was applied genome-wide to carry out a novel analysis of recombination events that explains some sequence differences between strain BG2 and strain CBS138, and to analyze factors affecting non-allelic mitotic recombination in *C. glabrata*.

***De novo* assembly of the type strain CBS138**

Strain CBS138 is the type strain of *C. glabrata*, and our assembly will be used to update the reference genome of *C. glabrata*. Our assembly is in overall agreement with the CBS138_s02-m07-r41 reference genome (http://www.candidagenome.org/) (Dujon et al., 2004). The Canu assembly with polishing of the same SMRT sequencing reads using Quiver from the PacBio GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus) had only 142 SNPs and indels relative to the current reference. After correction of Illumina sequencing reads, our CBS138 assembly contained 30 SNPs and indels relative to the reference genome, all of them supported by Illumina reads. The general agreement or our assembly with the reference genome, notwithstanding, our assembly corrected several errors in the current assembly.

C. glabrata encodes approximately 80 GPI-CWPs. About 70 of the GPI-CWPs in *C. glabrata* encode repeat regions, including long tandem repeats which cannot be accessed with short read sequencing. The subtelomeres of *C. glabrata* encode half of the GPI-CWPs. They are highly repetitive and share homology with each other on different chromosomes, and therefore, they are largely mis-assembled in the reference CBS138 genome. Our assembly resolves the correct structure of the CBS138 subtelomeres. We verified the structure of tandem repeats with the Illumina short sequencing reads. We further verified there is no translocations due to mis-assembled repeats using Sanger-sequenced cloned subtelomere sequences. Our assembly, as expected, have the same sequence in non-repetitive regions, but have different number of repeats relative to the fosmid sequences. Therefore, one major contribution of this thesis is that we, at the first time, obtained the correct structure of the subtelomeres for *C. glabrata*.

C. glabrata encodes rDNA arrays at the right end of ChrL and ChrM. In the reference genome, only ChrL ended with one rDNA repeat, and ChrM ended before the rDNA array. We identified a novel 10 kb rDNA downstream region, which is identical between ChrL and ChrM. We discovered a novel GPI-CWP, *EPA14* in the rDNA downstream region. We observe a general structure of the subtelomeres. All the subtelomeres of *C. glabrata* encodes a terminal gene which encodes a GPI-CWP. 23 of the 26 subtelomeres encode a terminal GPI-CWP transcribed towards the telomere. Only the terminal gene, *EPA3* on ChrE right and the terminal gene *EPA14a/b* on ChrL and ChrM right are transcribed towards the centromere.

The second major contribution of the CBS138 assembly is that we obtained the correct structures of the GPI-CWPs in *C. glabrata*. Compared to the reference genome, 21 GPI-CWPs have corrections resulting changes in ORF length > 50 nt. In fact, we corrected 111.7 kb (1.3% of total coding region in *C. glabrata*, 28% of the total ORF length of GPI-CWPs in strain CBS138). Furthermore, we annotated 21 novel GPI-CWPs because they were mis-assembled or scrambled in the reference genome.

From the correct structure of GPI-CWPs, we observed strikingly large ORFs of these genes. For instance, the largest ORF of GPI-CWP, CAGL0J05159g, which is also the largest ORF in fungi, encodes a hypothetical protein of approximately 10,000 amino acids. If this protein were in an alpha-helix conformation, it would be approximately 1.5 μm in length, which is at the same scale with the size of *C. glabrata* (1-2 μm).

In addition to long tandem repeats, the GPI-CWPs encode complicated repeat structure. They can encode multiple tandem repeat regions with the same repeat and the same linker between the repeat, therefore, the tandem repeat and the linker constitute a higher level of repeat unit. One extreme case of this structure is that the N-terminus region of *PWP4* serves as the linker between the encoded tandem repeats. GPI-CWPs can also encode multiple repeat units within the tandem repeat region, *i.e.*, the tandem repeats are alternatively distributed within the repeat region.

***De novo* assembly of strain BG2**

We perform all genetics in our lab using strain BG2 and it is known that strain BG2 encodes different GPI-CWPs from strain CBS138. We therefore sequenced BG2 to obtain the correct structure of the BG2 specific GPI-CWPs.

The BG2 genome has a similar structure with CBS138 genome, with 11 chromosome rearrangements. 8 of the 11 rearrangements are related to the subtelomeres, and 4 of them are located entirely within the subtelomeres. There are 18 BG2 specific ORFs relative to strain CBS138, and 7 of them are GPI-CWPs. Interestingly, all the 7 BG2 specific GPI-CWPs are located in the subtelomeres.

We obtained the correct structures of GPI-CWPs in BG2. BG2 encodes 81 GPI-CWPs, 74 of them are shared with strain CBS138. However, we observed significant sequence variations in the shared GPI-CWPs. In fact, there are only 91 ORFs with ORF length difference > 50 nt between BG2 and CBS138. 51 of the 91 ORFs are GPI-CWPs. The GPI-CWPs not only have most changed members than other genes, but the variations are also the most significant. The ORF length difference between the shared ORFs is 158.9 kb (approximately 35% of total ORF length of GPI-CWPs). However, the total size of encoded GPI-CWPs between strain BG2 and CBS138 is similar, with only 13.2 kb difference in total ORF length.

GPI-CWPs undergo various changes more than tandem repeat extensions and expansions. As discussed in the previous section, GPI-CWPs encode multiple tandem repeats with the same repeat, as well as with the same linker between the tandem repeats. The number of linkers can also vary between strains. For instance, *PWP4* encodes 5 tandem repeat regions with 4 linkers in CBS138, while in BG2, it encodes 4 tandem repeat regions with 3 linkers. The GPI-CWPs also have change in repeat structure and change in C-terminal region, a result of intergenic recombination, which is further analyzed in chapter 5.

De novo assembly of serial clinical isolates, BG3993-96

The serial clinical isolates, BG3993-96 are isolated from the same patient (received from Jack Sobel in Wayne State University Medical School) during a 21-month period. We aimed to assess microevolution of *C. glabrata* during infection with the assemblies of the four strains. One important change in phenotype in the 21-month period is that the latter isolates, BG3995-96 are hyper adherent to epithelial cells. Therefore, we want to identify candidate for change in adherence with our assemblies.

We obtained high-quality genomes of the four strains. Our initial hypothesis is that there are changes in GPI-CWPs, resulting in change of cell adherence. However, we found that the *C. glabrata* genome is highly stable even for the GPI-CWPs with long tandem repeats during infection. There are only 220 SNPs and indels between BG3993-96. Only 16 of the 5252 ORFs contain sequence variation in genome ORF sequences, and there are only 16 SNPs and indels in the non-repetitive regions in the changed ORFs. The GPI-CWPs only have trivial variations, and none of the changes is shared between the hyperadherent BG3995-96 strain. Instead, we identified a transcriptional repressor *YAP6*, which had a frameshift mutation the hyperadherent strains BG3995-96 relative to strain BG3993-94. Yap6 binds to subtelomeric genes in *S. cerevisiae* and regulates chromatin remodeling (Rodrigues-Pousada et al., 2019). Another lab member is following up the genetics to verify whether *YAP6* regulates cell adhesion.

Although the four isolates encode identical GPI-CWPs in general, the encoded GPI-CWPs in strain BG3993-96 differ both from the GPI-CWPs in strain CBS138 and those in strain BG2. We analyzed the recombination events between CBS138 and BG2 in chapter 5, and we will analyze the recombination events between BG3993-96 and BG2 as well as recombination events between BG3993-96 in the future.

Analysis of non-allelic mitotic recombination in *C. glabrata*

We developed a k-mer counting based approach to analyze the non-allelic mitotic recombination in *C. glabrata* (source code available at https://github.com/zhuweix/recombination_analysis). Our approach successfully identified recombination events in GPI-CWPs. We observed recombination events from simple gene conversion in non-repetitive region and recombination in C-terminal region, to complex events in which GPI-CWPs exchanged their tandem-repeats between strain BG2 and strain CBS138. We further quantitated genome-wide non-allelic mitotic recombination events between these two strains. We found that the subtelomeres of *C. glabrata,* and GPI-CWP genes encoded there undergo apparent high rates of recombination. In addition, non-subtelomeric GPI-CWPs have a higher apparent rate of recombination than other non-subtelomeric genes, although they have a lower rate of recombination than subtelomeric ones.

Recombination events are not limited to subtelomeres or GPI-CWPs. We also observed both intra and inter-chromosomal recombination in non-subtelomeric non-GPI-CWPs.

Future directions

C. glabrata is an asexual haploid yeast. The non-allelic mitotic recombination is one important mechanism for evolution. The subtelomeres are dynamic regions for recombination, and may be hotspots for adaptive evolution of *C. glabrata*. The subtelomeric GPI-CWPs undergo higher rate of recombination within the coding region. We will further characterize the individual recombination events between GPI-CWP genes to determine if there are rules for which regions (repetitive versus non-repetitive, for example) are preferred substrates for recombination.

In addition to recombination in repeat regions of ORFs, the terminal genes of the subtelomeres are also hotspots for larger translocations. Interestingly, non-reciprocal translocation of the terminal GPI-CWPs do not result in the gene next to the translocated terminal gene being terminal but rather replacement with another GPI-CWP gene that is homologous to other terminal GPI-CWPs. Changes in the GPI-CWPs complement and structure may have important adaptive function in *C. glabrata*. For instance, BG2 and BG3993-96 are vaginal isolates, and CBS138 was a urogenital tract isolate. It will be interesting to analyze GPI-CWP structure and complement in isolates from a wider range of host sites (blood, organs, GI tract) to understand if recombinational changes to the GPI-CWPs can help adapt to different host environments.

In addition to *C. glabrata*, other pathogenic fungi also benefit from mitotic recombination for adaptation. For example, mitotic recombination contributes to the gene diversity and stress adaptation in *Candida albicans* (Gusa and Jinks-Robertson, 2019). Our analysis method can be further applied to investigate recombination events in general pathogenic fungi.

Reference

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., et al. (2004). Genome evolution in yeasts. Nature *430*, 35–44.

Gusa, A., and Jinks-Robertson, S. (2019). Mitotic recombination and adaptive genomic changes in human pathogenic fungi. Genes (Basel) *10*.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Rodrigues-Pousada, C., Devaux, F., Caetano, S.M., Pimentel, C., da Silva, S., Cordeiro, A.C., and Amaral, C. (2019). Yeast AP-1 like transcription factors (Yap) and stress response: a current overview. Microb. Cell *6*, 267–285.

Appendices

Supplementary Tables

Chapter 2

Table S1. SNPs, indels and structural variations.

Table S2. PCR primers for validation of SNPs and indels.

Table S3. Oligos for cloning and sequencing subtelomeres.

Table S4. Genes in subtelomeric regions.

Table S5. Sequencing metrics.

Table S6. Chromosome length.

Table S7. Gene Comparisons.

Table S8. Putative GPI-anchored proteins.

Chapter 3

Table S1. SNPs in rRNA genes.

Table S2. Chromosome Rearrangements and structural variants.

Table S3. Oligos for cloning and sequencing.

Table S4. GPI-CWPs.

Table S5. LTRs.

Table S6. Sequencing Metrics.

Table S7. Chromosome Length.

Table S8. ORFs in subtelomeric regions.

Table S9. Gene Comparison.

Chapter 4

Table S1. SNPs and Structural Variants.

Table S2. Chromosome rearrangements between BG3993 and BG2.

Table S3. ORFs in subtelomeric regions.

Table S4. Gene Comparisons.

Table S5. GPI-CWPs.

Chapter 5

Table.S1 Minimal recombination events.

Table.S2 Recombined gene pairs.

<u>Supplementary Figures</u>

Chapter 2

Figure S1. Metrics of SMRT sequencing.

Figure S2. Chromosome sequence comparison between published CBS138 genome and our assembly.

Figure S3. Read depth of Illumina sequencing read in the published CBS138 genome.

Figure S4. Whole subtelomere alignment between subtelomeres in published CBS138 genome and subtelomeres in our assembly.

Figure S5. Subtelomere sequence comparison between *de novo* assembled subtelomeres and the Sanger sequenced cloned subtelomere.

Figure S6. Sequence comparison between structural variants between the published CBS138 genome and our assembly.

Figure S7. Sequence comparison between the mis-assembled or scrambled ORFs in the published CBS138 genome and the corrected ORFs in our assembly.

Figure S8. Structure of GPI-CWP genes in our assembly.

Chapter 3

Figure S1. Metrics of SMRT sequencing.

Figure S2. Subtelomere sequence comparison between *de novo* assembled subtelomeres and the Sanger sequenced cloned subtelomere.

Figure S3. Comparison of the subtelomere structure between strain BG2 and strain CBS138.

Figure S4. Change in subtelomere structure between strain BG2 and strain CBS138.

Figure S5. Sequence comparison between structural variants between strain BG2 and strain CBS138.

Figure S6. Sequence comparison between structural variants between strain BG2 and strain CBS138 with RNAseq data.

Figure S7. Phylogenetic tree of GPI-CWPs in strain BG2 and strain CBS138

Figure S8. Structure of GPI-CWP genes in strain BG2.

Chapter 4

Figure S1. Metrics of SMRT sequencing.

Figure S2. Chromosome sequence comparison between BG3994-96 and BG3993.

Curriculum Vitae

Zhuwei Xu

Personal information

Birthdate	Jan 12, 1989
Birthplace	Changchun, Jilin, China
Address	3501 Saint Paul Street, Apt 626 Baltimore, MD 21218
Cell phone	(410)302-4800
Work phone	(410)955-3651
Email	wszwei@gmail.com

Education

2012-Present	PhD student in Biochemistry, Cellular and Molecular Biology Graduate Program at Johns Hopkins University School of Medicine, Baltimore, MD
2008-2012	B.S. in Tsinghua University, School of Life Sciences Beijing, China
JUN-AUG 2011	2011 UM Chemistry-China Summer Exchange Program The University of Michigan, Ann Arbor, MI
JAN-FEB 2011	Immunology Summer Camp 2011 The University of Melbourne, Melbourne, Australia

Research Experience

2017-Present	Brendan Cormack Lab Johns Hopkins School of Medicine Thesis: *De novo* genome assembly and analysis of non-allelic recombination in pathogenic yeast *Candida glabrata*.
2012-2017	Jef Boeke Lab Johns Hopkins School of Medicine Research: Synthesis of a eukaryotic chromosome and a human mitochondrial genome.

2011-2012 Xiaohua Shen Lab

Tsinghua University

School of Life sciences

Thesis: Using in vitro single-molecule imaging technique to observe the recruitment of Polycomb Repressive Complex 2 to *XIST* locus

JAN-FEB 2011 Tom Kerppola Lab

The University of Michigan, Department of Biological Chemistry

Research: Amelioration of cross-linking and RNA immunoprecipitation method

Publications

Xu Z, Green G., Benoit N.,Schatz M., Wheelan S., Cormack B., De novo genome assembly of Candida glabrata reveals cell wall protein complement and structure of dispersed tandem repeat arrays. Molecular Microbiology (manuscript in revision)

Xu Z,Green G., Benoit N., Pan SJ., Schatz M., Wheelan S., Cormack B., De novo genome assembly of Candida glabrata strain BG2 (manuscript in preparation)

Xu Z, Schatz M., Wheelan S., Sobel J., Cormack B., De novo genome assembly of four serial clinical isolates of Candida glabrata (manuscript in preparation)

Xu Z, Schatz M., Wheelan S., Cormack B., Analysis of non-allelic mitotic recombination in Candida glabrata (manuscript in preparation)

Lin Q, Jia B, Mitchell LA, Luo J, Yang K, Zeller KI, Zhang W, **Xu Z**, Stracquadanio G, Bader JS, Boeke JD, Yuan YJ. RADOM, an efficient in vivo method for assembling designed DNA fragments up to 10 kb long in Saccharomyces cerevisiae. ACS Synth Biol. 2015 Mar 20;4(3):213-20

Oral Presentations

Xu Z, Schatz M., Wheelan S., Cormack B., Non-allelic recombination in genome evolution of the pathogenic yeast Candida glabrata, 2019 Bioinformatics and Genomics symposium, Oct 17, 2019

Posters

Xu Z, Schatz M., Wheelan S., Cormack B., Analysis of complex tandem repeats and intergenic recombination in the yeast Candida glabrata, 2019 AGBT precision health, Sept 6, 2019

Xu Z, Hwang-Wong E., Wheelan S., Cormack B. ChIP-seq analysis of Sir3 binding sites in C.glabrata, Baltimore Fungal Biology Center Symposium, Mar 20, 2019

Xu Z, Green B., Sobel J., Schatz M., Wheelan S., Cormack B., Insights from high quality sequence and assembly of the Candida glabrata genome. 14th ASM Conference on Candida and Candidiasis, April 16, 2018

Xu Z, Boeke JD, Synthesis of a human mitochondrial genome,Synthesis of a human mitochondrial genome, Second Annual Molecular Pharmacology Retreat, NYU School of Medicine, April 14, 2015

Xu Z, Deutsch S., Boeke JD, Synthesis of a eukaryotic chromosome,Synthesis of a eukaryotic chromosome,Synthesis of a eukaryotic chromosome,Synthesis of a eukaryotic chromosome, 2014 JGI User Meeting, Mar 18, 2014