High Complexity and Degree of Genetic Variation in *Brettanomyces bruxellensis* Population

Jean-Sébastien Gounot[1], Cécile Neuvéglise[2], Kelle C. Freel[1], Hugo Devillers[2], Jure Piškur[3], Anne Friedrich[1,*], and Joseph Schacherer[1,4,*]

[1]Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France

[2]Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France

[3]Department of Biology, Lund University, Sweden

[4]Institut Universitaire de France (IUF)

*Corresponding authors: E-mails: anne.friedrich@unistra.fr; schacherer@unistra.fr.

Accepted: April 13, 2020

Data deposition: The sequences related to the YJS5363, YJS5431, and YJS5447 were previously released in the European Nucleotide Archive under the accession number PRJEB21262 (Fournier et al. 2017). All sequencing data generated for this study have been deposited at the European Nucleotide Archive under the accession PRJEB28420. The annotations of the UMY321 isolate have been deposited at the European Nucleotide Archive under the accession PRJEB33245.

Abstract

Genome-wide characterization of genetic variants of a large population of individuals within the same species is essential to have a deeper insight into its evolutionary history as well as the genotype–phenotype relationship. Population genomic surveys have been performed in multiple yeast species, including the two model organisms, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. In this context, we sought to characterize at the population level the *Brettanomyces bruxellensis* yeast species, which is a major cause of wine spoilage and can contribute to the specific flavor profile of some Belgium beers. We have completely sequenced the genome of 53 *B. bruxellensis* strains isolated worldwide. The annotation of the reference genome allowed us to define the gene content of this species. As previously suggested, our genomic data clearly highlighted that genetic diversity variation is related to ploidy level, which is variable in the *B. bruxellensis* species. Genomes are punctuated by multiple loss-of-heterozygosity regions, whereas aneuploidies as well as segmental duplications are uncommon. Interestingly, triploid genomes are more prone to gene copy number variation than diploids. Finally, the pangenome of the species was reconstructed and was found to be small with few accessory genes compared with *S. cerevisiae*. The pangenome is composed of 5,409 ORFs (open reading frames) among which 5,106 core ORFs and 303 ORFs that are variable within the population. All these results highlight the different trajectories of species evolution and consequently the interest of establishing population genomic surveys in more populations.

Key words: intraspecific diversity, population genomics, genome evolution, yeast, Brettanomyces bruxellensis.

Introduction

The yeast species *Brettanomyces bruxellensis* (anamorph *Dekkera bruxellensis*) is a distant relative of *Saccharomyces cerevisiae* since lineages diverged >200 Ma. Interestingly, these species share characteristics that have been independently acquired during evolution, such as the metabolic ability to produce ethanol in the presence of oxygen and excess of glucose (Schifferdecker et al. 2014). Both species also share the capacity to efficiently catabolize the produced ethanol, their corresponding life style being described as "make–accumulate–consume (ethanol)" (Thomson et al. 2005). This strategy allows *B. bruxellensis*, which is associated with several human fermentation processes, to survive and develop in harsh and limiting environmental conditions. Until now, this yeast species has exclusively been isolated from anthropized niches.

The impact of *B. bruxellensis* on the fermentative processes is, however, diverse. As an example, it has a positive contribution to brewing of Belgium Lambic and Gueuze beers, but it is also well known as a major player leading to wine spoilage by producing odorant molecules (volatile phenol) described as having barnyard or horse sweat characteristics (Conterno et al. 2006). It has also been found associated with other food or industrial processes, such as kombucha or bioethanol, for which its contribution is still unclear (Teoh et al. 2004; Beckner et al. 2011). The opposite contributions of *B. bruxellensis* in fermentation processes raised a growing interest in this species, and high phenotypic variability regarding, for example, sugar metabolism or nitrogen source utilization has been highlighted among the isolates in various studies (Borneman et al. 2014; Crauwels et al. 2015, 2017).

The observed phenotypic variability is undoubtedly, at least in part, related to its genomic plasticity, which is another peculiarity of this species that defines it as a good model at the evolutionary level. Indeed, the first proof of *B. bruxellensis* high genomic variability was provided through the comparison of electrophoretic karyotypes from different isolates that showed extensive chromosomal rearrangements (Hellborg and Piskur 2009), indicating rapid evolution at the intraspecific level. Whole-genome sequencing of a small number of isolates revealed variation of the ploidy level among isolates with some triploid individuals derived from an allotriploidization event involving a moderately heterozygous diploid and a more distantly related haploid (Borneman et al. 2014). The level of ploidy is supposed to be linked to the substrate of isolation and geographical distribution (Albertin et al. 2014). Traditionally, genomic variability has essentially been restricted to the exploration of small regions of the genomes (Conterno et al. 2006; Curtin et al. 2007; Agnolucci et al. 2009; Vigentini et al. 2012; Crauwels et al. 2014; Avramova et al. 2018). More recently, the complete genomes of some isolates have been sequenced and assembled (Curtin et al. 2012; Piškur et al. 2012; Borneman et al. 2014; Crauwels et al. 2014; Olsen et al. 2015). Comparison of these genomes suggested that polyploidization and hybridization events might play a significant role in the evolution of this species (Borneman et al. 2014). The most recent population survey was performed on a large collection of almost 1,500 isolates based on 12 microsatellite regions and showed that the population is structured according to ploidy level, substrate of isolation, and geographical origin of the strain (Avramova et al. 2018).

With the currently available sequencing technologies, it is now possible to explore the intraspecific variability of a species at the genome-wide level. Such population genomic studies have been performed on multiple yeast species, including *S. cerevisiae* (Skelly et al. 2013; Bergström et al. 2014; Almeida et al. 2015; Strope et al. 2015; Gallone et al. 2016; Gonçalves et al. 2016; Zhu et al. 2016; Peter et al. 2018) and *Schizosaccharomyces pombe* (Fawcett et al. 2014; Jeffares et al. 2015) but also nonmodel yeast species (Almeida et al. 2014; Ford et al. 2015; Friedrich et al. 2015; Hirakawa et al. 2015; Leducq et al. 2016; Carreté et al. 2018; Ortiz-Merino et al. 2018; Ropars et al. 2018), granting better insights into their respective evolutionary histories as well as genotype–phenotype relationships.

Here, we conducted a population genomic survey of *B. bruxellensis* based on whole-genome sequencing data. In total, 53 worldwide collected isolates were completely sequenced using Illumina short-read technology. The data show a high complexity of genetic variation among isolates. This species displays a variable level of nucleotide diversity, heterozygosity, and copy number (CN) variants, depending on the ploidy of the isolates. This data set offers a first view of the genomic variants at a genome-wide scale within *B. bruxellensis* and overall, provides insights into the evolutionary history of the species as well as the identification of genetic contents linked to subpopulation adaptations.

Materials and Methods

Isolates and Sequencing

This study was performed using a collection of 53 *B. bruxellensis* isolates originating from diverse ecological and geographical origins (supplementary table S1, Supplementary Material online). Most of the samples were isolated from Europe, but some also originated from South Africa, Australia, and Chile. Although a subset of the samples had no ecological origins associated with them, approximately half of them were retrieved from fermentative and wine-related environments.

All *B. bruxellensis* isolates were subjected to Illumina paired-end sequencing. Yeast cell cultures were grown overnight at 30 °C in 20 ml of YPD medium to early stationary phase before cells were harvested by centrifugation. Total genomic DNA was then extracted using the QIAGEN Genomic-tip 100/G according to the manufacturer's instructions. Genomic Illumina sequencing libraries were prepared with a mean insert size of 280 bp and subjected to paired-end sequencing (2×100 bp) on Illumina HiSeq 2500 sequencers.

Reference Sequence Correction and Annotations

The annotation of the UMY321 genome sequence (Fournier et al. 2017) was performed using Amadea Annotation transfer tool (Isoft, France) with *Lachancea kluyveri* CBS3082^T and *Debaryomyces hansenii* CBS767^T genomes as references (revised versions available at http://gryc.inra.fr, last accessed April 25, 2020). This step was followed by a manual curation with RNA-Seq data from *B. bruxellensis* CBS2499 (Sequence Read Archive SRR427169[17]). Tophat2 aligner tool v. 2.1.0 was used to map reads against the assembled genome of UMY321 (Kim et al. 2013). Artemis v16.0.0 was used to visualize BAM files, to adjust exons and introns coordinates and to identify lncRNA (Carver et al. 2012). tRNA genes were detected using tRNAscan-SE v1.3.1 (Lowe and Chan 2016). Transposable elements were identified by BLAST with known yeast elements from different families such as *Ty1-copia*, *Ty3-gypsy*, and *hAT*, as queries.

A high number of out-of-frame genes were predicted, mostly related to long read sequencing errors. To correct these sequencing errors, an independent assembly was constructed by running SOAPdenovo from Illumina short reads (Luo et al. 2012). This latter was scaffolded with Redundans (Pryszcz and Gabaldón 2016) using the initial UMY321 assembly (Fournier et al. 2017) as template. The sequences corresponding to the frameshifted genes were detected in the new assembly by similarity searches and inferred in the original assembly if the correct frame was recovered. The subsequent annotations were modified accordingly in the UMY321 assembly.

Mapping, Variant Calling, and Annotation

For each sample, Illumina paired-end reads were mapped on the corrected sequence of UMY321 genome with BWA mem (v0.7.12, default parameters) (Li and Durbin 2009). Alignments were postprocessed using SAMtools fixmate (v1.1) (Li et al. 2009), GATK realignment (v3.3.0) (McKenna et al. 2010), and Picard MarkDuplicates (v1.1.140) (broadinstitute.github.io/picard, last accessed April 25, 2020).

Aneuploidies and segmental duplication events were visually identified based on read coverage profiles. To that end, coverage was computed for each position with SAMtools depth function and mean coverage values along the genome were plotted using 20-kb nonoverlapping sliding windows.

Single-nucleotide polymorphisms and small indels were called using GATK Haplotype Caller function with a ploidy set to 2. In order to determine the frequency of each variant for ploidy analysis, VCF file was annotated using GATK Variant Annotator function. Variants with a coverage <10× or a quality score <25 were removed using bcftools (v 1.5) filter (Li et al. 2009). Finally, variant files were merged into one single file using vcftools (v. 0.1.13) merge function (Danecek et al. 2011). To assess the impact of variants on protein sequences, annotation was processed with SnpEff (v 4.3i) (Cingolani et al. 2012) and the impact on protein function for nonsynonymous SNPs was predicted using SIFT (v 6.2.1) (Kumar et al. 2009).

Allele Frequency and Nucleotidic Diversity

The allele frequency of the heterozygous single-nucleotide variants was determined using the ABHet annotations generated by GATK Variant Annotator.

To get an estimate of the scaled mutation rate, the average pairwise nucleotide diversity (π), the proportion of segregating sites (θ_W), and Tajima's *D* value were computed with variscan (v. 2.064) (Tajima 1989; Vilella et al. 2005). It was run with mode 12, usemut 1, and 10-kb nonoverlapping windows on multiple alignments of the concatenated chromosomes with SNPs inferred from each sample. A similar process was used to determine these estimators for coding and noncoding regions.

Phylogenetic Relationships

The phylogenetic relationships between samples were estimated through the construction of a neighbor-joining tree based on the BioNJ algorithm provided by SplitsTree4 (Huson 2019). To that end, a multiple alignment of the concatenated scaffolds was produced in which SNPs were inferred using the IUPAC code, leading to a single sequence for each sample. The MatchState option was selected to properly manage the heterozygous positions while calculating the distances.

LOH Detection

The genome of each isolate was scanned (50-kb sliding windows with 25-kb overlap) to identify regions with ten heterozygous sites or less per 50 kb. These latter were considered to be under loss of heterozygozity (LOH).

For each strain, the level of heterozygosity (heterozygous SNPs per kb) was estimated without considering these LOH regions (i.e., the heterozygous SNPs found in LOH regions were not taken into account and these regions were subtracted from the total length of the genome).

CN Variants

Control-FREEC (v10.6) (Boeva et al. 2011) was used to determine the variation of the number of copy along the genome of each sample, through 1-kb window. The following parameters were used to run the software: breakPointThreshold = 0.6, window = 1,000, telocentromeric = 6,000, step = 200, minExpected GC = 0.3, and maxExpectedGC = 0.5. The ploidy parameter was set according to supplementary figure S1, Supplementary Material online. A gene was considered as duplicated (or lost) if >50% of its length was in a region detected as duplicated (or lost) by Control-FREEC.

Functional Insight into the Duplicated and Deleted Genes

To identify the CN of the 20 genes previously highlighted (Crauwels et al. 2014) within our population, we first identified each gene within our reference using BlastP similarity searches.

Supplemental Genes Identification

To determine the set of supplemental genes (genes not found in the reference genome) within our population, we used a pipeline similar to the one developed for the analysis of 1,011 *S. cerevisiae* strains (Peter et al. 2018). We first constructed an assembly for each sample using Abyss (v2.0.2) with 67 as kmer size (Jackman et al. 2017). Assemblies were compared with the reference sequence with BlastN allowing for the determination of nonreference segments. Gene prediction tools were applied on these segments through both SNAP (v2006.07.28) (Korf 2004) and Augustus (v. 3.2) (Stanke and Waack 2003) with training sets generated using the *B. bruxellensis* proteome. Genes showing an excess of low-complexity region as well as those whose product had a length <50 amino acids were removed. All genes were then compared with each other with BlastP and the application of a graph-based method leads to the identification of a nonredundant set of supplemental genes for the whole species.

Finally, these genes were annotated through similarity searches against the reference proteome and several protein databases such as Uniref90, SwissProt Fungi, UniProt Fungi as well as Saccharomycotina species found in the SSS (http://sss.genetics.wisc.edu/cgi-bin/s3.cgi, last accessed April 25, 2020) and GRYC (http://gryc.inra.fr/, last accessed April 25, 2020) databases.

Results and Discussion

In order to survey genome-wide variability within the *B. bruxellensis* species, we gathered a collection of 53 isolates from diverse origins (supplementary table S1, Supplementary Material online), with a large part of our collection being wine-related. This collection is representative of all five major clusters previously defined using microsatellites based on almost 1,500 isolates (Avramova et al. 2018). These strains were mostly isolated in Europe (e.g., Belgium, Italy, Spain) but also South Africa, Australia, and Chile. We sequenced the genomes of all strains with a short-read sequencing strategy at a 40-fold coverage at least, with a mean of 98-fold coverage. The genetic diversity was explored through the comparison of the isolate sequences with a recently published reference sequence of the species (Fournier et al. 2017), as well as through the construction of genome assemblies in order to define the gene repertoire of the species.

Gene Content of the *B. bruxellensis* Genome

The recent release of a highly contiguous reference assembly of *B. bruxellensis* (Fournier et al. 2017) opens the way to a genome-wide exploration of intraspecific variability. However, functional analyses of the genomic variability require genetic elements to be located on the genome sequence, which led us to carry out a complete annotation of this assembly. At first, a total of 5,206 protein-coding genes were predicted, among which 1,427 were interrupted by frameshifts or stop codons probably due to sequencing or assembly errors linked to the heterozygous diploid state of the sequenced isolate. This high proportion of out-of-frame genes led us to refine the reference sequence. This step was performed by retrieving the sequences of the concerned genes in an independent assembly constructed with Illumina reads and scaffolded with Redundans (Pryszcz and Gabaldón 2016). This procedure allowed us to infer 887 in-frame gene sequences in the initial assembly leading to a total of 4,666 in-frame protein-coding genes (89.6%) within the genome. The remaining 540 out-of-frame genes were considered to be pseudogenes.

Publicly available RNA-seq data (Piškur et al. 2012) allowed for the identification of 509 introns in 472 genes, that is, 9% of the genes, with 24 introns located in UTR. In total, 99 tRNA genes with 39 different anticodons were listed. The 26 transposable elements found in the UMY321 reference genome belong to the LTR retrotransposons. Most of them are closely related to *D. hansenii* Tdh5 (Neuveglise 2002). In addition to the intact and degenerate copies, at least 96 solo LTR were detected, sometimes grouped into large regions of up to 40 kb that remind the centromeres structures reported for *D. hansenii* (Lynch et al. 2010) as well as the ones described in the CBS2499 *B. bruxellensis* strain (Ishchuk et al. 2016).

The comparative analysis of the gene content between UMY321 and the recently annotated CBS11270 strain (Tiukova et al. 2019) through BlastN similarity searches showed that the gene content is only slightly variable. Indeed, 12 genes annotated on the UMY321 genome were absents from the CBS11270 assembly, whereas only three genes annotated on the CBS11270 chromosomes could not be found in the UMY321 assembly.

Genetic Diversity Variation Is Related to Ploidy Level within *B. bruxellensis*

The mapping of the Illumina paired-end reads on the UMY321 (YJS5431 in this study) reference sequence allowed for single-nucleotide variants detection. From 84,890 to 502,399 SNPs were detected per isolate, distributed among 811,159 polymorphic positions. A minimum of 84,500 heterozygous variants was called per sample, revealing that our data set was devoid of haploid or homozygous isolates. The allele frequency at heterozygous sites was used to infer the ploidy of each isolate: 39 isolates showed values centered ~0.5 and were therefore considered to be diploid. The remaining 14 isolates had an enrichment of values centered ~0.33 and 0.66 and were consequently defined as triploid (supplementary fig. S1, Supplementary Material online).

Interestingly, the number of homozygous variants was in the same range for all the strains in our collection with the exception of a group of ten diploid strains. This unique group included the UMY321 reference isolate, and <350 homozygous variants were called per strain (supplementary table S2, Supplementary Material online). By contrast, the number of heterozygous sites was much higher in triploid isolates compared with the diploid ones (supplementary fig. S2, Supplementary Material online), which is in accordance with previous studies, showing that two Australian triploid isolates were constituted of a moderately heterozygous diploid genome in combination with a divergent haploid genome

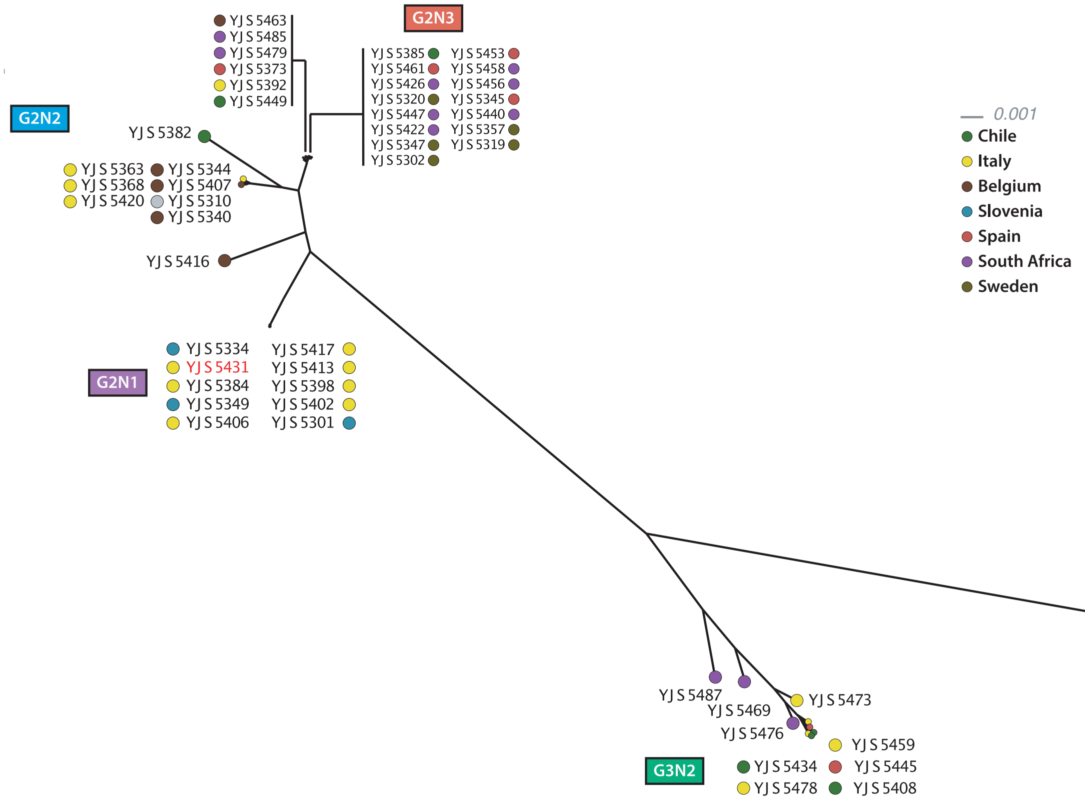

FIG. 1.—Neighbor-joining tree of the 53 isolates based on the 811,159 detected polymorphic positions. We identified five clusters composed of diploid (G2) or triploid (G3) strains. YJS5382 is the only triploid strain found among the diploid clusters (left side).

(Curtin et al. 2007; Borneman et al. 2014). The overall number of variants specific to a single subgenome (supplementary table S2, Supplementary Material online) allowed to estimate the genetic distance between the diploid and haploid subgenomes, which is ~2.4%.

At the population level, the genetic diversity estimated by the average pairwise difference between strains is relatively high ($\pi = 1.2\times10^{-2}$) compared with *S. cerevisiae* ($\pi = 3\times10^{-3}$), closer to what was observed for some other yeast species such as *L. kluyveri* ($\pi = 1.8\times10^{-2}$) or *Saccharomyces uvarum* ($\pi = 1.7\times10^{-2}$) (Almeida et al. 2014; Friedrich et al. 2015). As expected, the genetic diversity is lower in coding regions ($\pi = 1.0\times10^{-2}$) compared with intergenic regions ($\pi = 1.5\times10^{-2}$). In addition, SNPs found in CoDing Sequence (CDS) are mostly synonymous (64%), whereas nonsense mutations are rare (1%). These latest mutations also show a lower allele frequency within the population and a more contrasted but similar pattern for observed for missense variants (supplementary fig. S3, Supplementary Material online).

Phylogeny and Strain Relatedness in *B. bruxellensis*

The 811,159 polymorphic positions were used to infer phylogenetic relationships between the isolates. To handle heterozygosity within our samples, heterozygous SNPs were encoded using the IUPAC code and the average state method was used for the distance calculation. Five distinct clusters were highlighted in the neighbor-joining tree, three of them (G2N1-3) were composed of diploid isolates, whereas the remaining two (G3N1-2) exclusively contained triploid isolates (fig. 1).

These clusters are mostly in accordance with the five major clusters recently described using microsatellites on a very large population (Avramova et al. 2018). For example, the two triploid clusters, G3N1 and G3N2, correspond respectively to the so-called triploid beer AWRI1608-like cluster and the triploid wine AWRI1499-like cluster. Moreover, the three distinct diploid clusters can be observed, enhancing the delineation of this population. The G2N3 cluster corresponds to the wine CBS2499-like cluster, whereas the G2N1 cluster (including the wine reference YJS5341 strain isolated in Italy) is distinct, although both were in a single kombucha-like cluster in the former analysis.

The YJS5382 isolate could not be assigned to any cluster in our analysis, which is also in accordance with this previous study (Avramova et al. 2018), as it is the only representative of the wine L0308-like cluster in our data set. Interestingly, this isolate is the only triploid that does not group with the G3 clusters. By contrast, it is in fact closer to the diploid clusters, suggesting an independent triploidization event for this strain. Additionally, the YJS5416 strain is not closely related to any

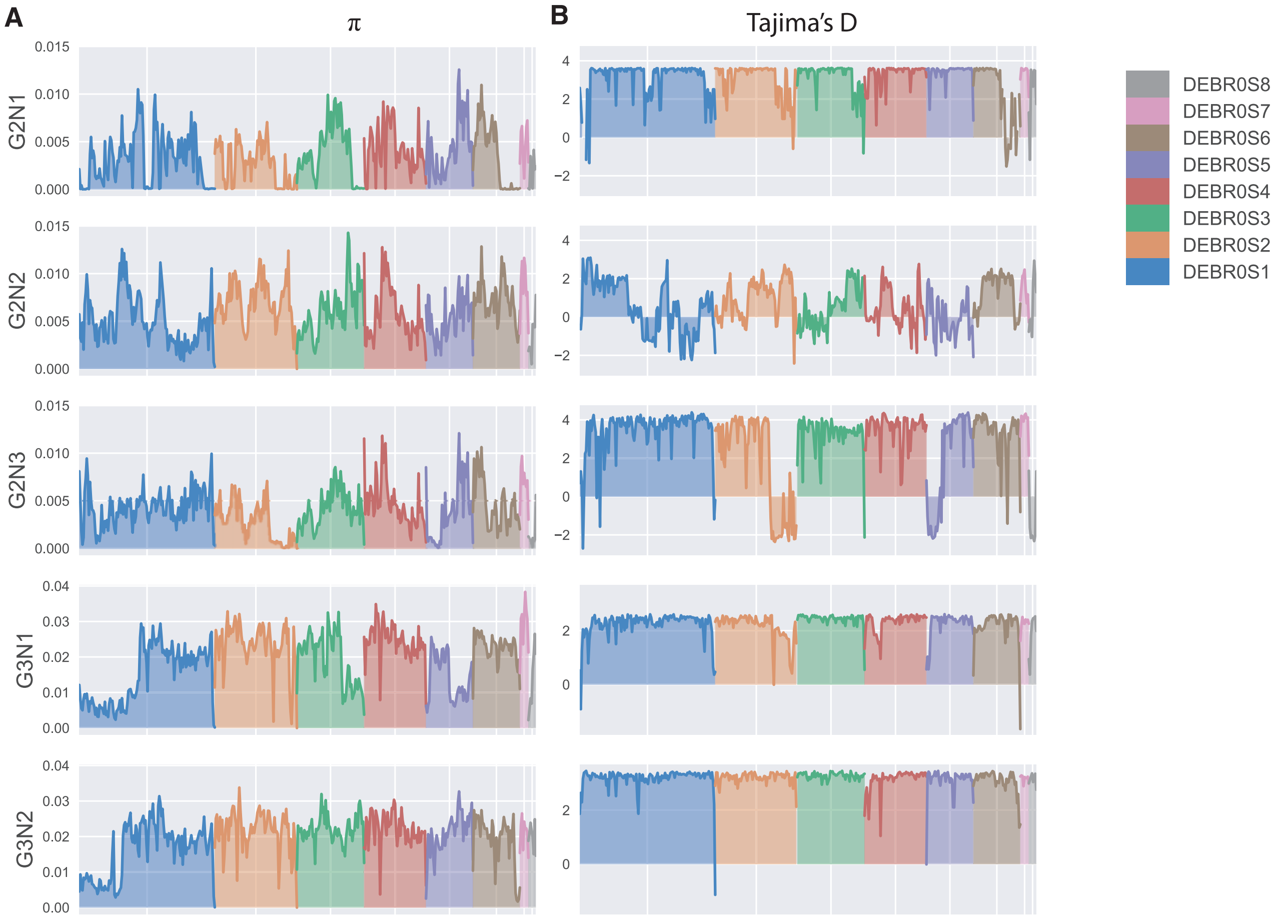

Fig. 2.—Variation of nucleotide diversity metrics for each cluster using nonoverlapping window (10 kb). (*A*) Pairwise nucleotide diversity (π) and (*B*) Tajima's *D* values.

cluster in our tree and is located between the G2N1 and G2N2 clusters and it might, therefore, be a representative of a new subpopulation.

Genomes Are Punctuated by a Few LOH Regions

To have a better insight into the nucleotide variation along the genome, we examined the genetic diversity across nonoverlapping windows of 10 kb for each cluster (fig. 2). Overall, the triploid subpopulations show a higher genetic diversity (π_{G3N1} and $\pi_{G3N2} = 2\times10^{-2}$) compared with diploid clusters ($\pi_{G2N1} = 3.5\times10^{-3}$, $\pi_{G2N2} = 6\times10^{-3}$, $\pi_{G2N3} = 4\times10^{-3}$). However, the nucleotide diversity is not homogeneous along the genome (fig. 2). In fact, the two triploid clusters display a lower genetic diversity ($\pi < 1\times10^{-2}$) within a large region of ~1.2 Mb on the left side of the scaffold 1. Interestingly, two similar regions can be observed in the triploid wine subpopulation (G3N1) on the right extremity of the scaffolds 3 and 5. In the diploid subpopulations, similar regions exhibiting a low genetic diversity can also be observed on almost all chromosomes. In the G2N2 and G2N3 clusters for which a notable decrease of the Tajima's *D* values is also observed, indicating a high proportion of rare alleles in these regions compared with the rest of the genome. Although the Tajima's *D* value never drops <0 with the exception of the right side of the scaffold 6 in the G2N1 cluster, the impacted region showed a lower Tajima's *D* value. This difference might be due to the close proximity between the strains belonging to this cluster and the reference sequence.

To better understand the origins of these regions with lower genetic diversity, we scanned the genomes for the presence of LOH events. This type of event leads to a decrease of the genetic variability allowing the expression of recessive alleles, and therefore can result in a beneficial adaptation in some environments. The prevalence of LOH events has recently been observed in several yeast species such as *S. cerevisiae* (Peter et al. 2018), *Candida albicans* (Ropars et al. 2018), and *Kluyveromyces marxianus* (Ortiz-Merino et al. 2018). Moreover, previous studies of *B. bruxellensis* isolates revealed the presence of several regions which underwent LOH. This was particularly clear in the case of the wine strains AWRI1613 and CBS2499 for which 17.9% and 16.3%

of the genome are impacted, respectively (Borneman et al. 2014). To detect LOH regions in our collection, the allele frequency of the polymorphic sites was plotted along their chromosomal location in the reference assembly and the number of heterozygous SNPs along the genome was further examined using 10-kb sliding windows (supplementary fig. S4, Supplementary Material online).

In all diploid isolates, several regions of LOH were observed representing a total of ∼0.7–2.7 Mb (mean value of 1.8 Mb, 13% of the genome) distributed among 6–18 regions (mean value of 11.7 regions) (supplementary table S3, Supplementary Material online). This level of LOH is low compared with *S. cerevisiae* for which LOH cover ∼50% of the genome on an average (Peter et al. 2018). The patterns of LOH are very well conserved within clusters and only very few regions are strain-specific. Some LOH events seem to have arisen before cluster expansion as they are shared between several clusters. However, GO-term analysis of detected genes in these regions did not reveal any functional enrichment. Moreover, most of the identified regions also showed a low nucleotide diversity, suggesting that LOH is the main reason for the nucleotide variability patterns mentioned earlier.

In triploid isolates, regions exhibiting a complete LOH are uncommon and only one to three LOH regions were detected covering 29–279 kb. These regions are well conserved between the strains. Interestingly, several regions in the triploid strains showed a reduced heterozygous rate and a lower nucleotide diversity. However, they did not meet our criteria to be considered as LOH. These regions could have resulted from an ancient LOH event that occurred before the separation of the two triploid clusters and the accumulation of new mutations, which is especially possible for the left part of the scaffold 1.

Masking LOH regions allowed us to precisely determine the level of heterozygosity, which ranged from 6.2 to 36.3 heterozygous sites per kb (supplementary table S4, Supplementary Material online). All diploid isolates share the same level of heterozygosity, with a mean value of 7.1. Much higher heterozygosity rates were associated with triploid strains: YJS5382 has 11.4 heterozygous SNPs/kb, and this level is much higher within the triploid clusters. Indeed, a mean of 34 heterozygous SNPs/kb was observed, this level was slightly higher in the G3N2 cluster compared with the other triploid subpopulations with 31 and 35 heterozygous SNPs/kb, respectively. This difference is most likely due to the regions that have undergone LOH in the G3N1 cluster and are not found in the G3N2 strains.

Aneuploidies and Segmental Copy Variants Are Uncommon in *B. bruxellensis*

To determine the frequency of aneuploidy as well as of segmental copy variants in our collection, we examined the coverage distribution along the reference genome using nonoverlapping windows of 20 kb. Coverage deviations were confirmed with the allele frequency variation in these regions. We observed such deviations at the whole scaffold level for only three isolates (5.6% of the 53 strains), and additionally only related to the smallest scaffolds (scaffolds 7 and 8). This result suggests that aneuploidies are rare in *B. bruxellensis* compared with other species such as *S. cerevisiae*, for which analysis of 1,011 strains revealed that roughly 20% of the population is affected by aneuploidies (Peter et al. 2018) or *C. albicans* (Hirakawa et al. 2015), but for which antifungal treatments are meant to distort the natural behavior of the species. Segmental variations were more prevalent and detected in 17 isolates (fig. 3), 9 of which harbored several regions that were affected. However, the relatively weak prevalence of these events and the size of the affected regions (mean size ∼350 kb) suggest that they could not explain the highly variable karyotypes observed within this species alone and that balanced rearrangements may also occur. Moreover, triploid clusters display a higher proportion of samples carrying segmental variants with more than half of the samples affected by such variants (supplementary fig. S5, Supplementary Material online). This result supports the idea that hybrids, and generally polyploid strains are affected by structural changes (Otto 2007).

We then focused on triploid isolates to determine whether these copy variants affected the haploid or diploid genomic version by taking advantage of the allele frequency within these regions. Indeed, although a duplication of the haploid version will result in an equal genome version ratio (2:2) and therefore in a frequency shift to 0.5, a supplemental copy of the diploid genome will lead to a frequency of 0.25 and 0.75 (3:1). On the other hand, a deletion of the haploid and diploid versions will result in an allele frequency of 1 (2:0) and 0.5 (1:1) in these regions, respectively. Eight triploid strains carrying segmental duplications were investigated (supplementary fig. S6 and table S5, Supplementary Material online). A total of seven duplicated regions showed a 3:1 ratio versus a 2:2 ratio suggesting that duplication of the diploid genome is more common. Moreover, six out of the seven deleted regions showed a ratio close to 0.5, corresponding to the loss of one copy of the diploid genome. This result suggests that the haploid version of the *B. bruxellensis* hybrids is mostly conserved compared with the diploid, in regard to deleterious structural variants.

Triploid Genomes Are More Subject to Gene CN Variation than Diploids

Our sequencing data provided the opportunity for a more precise view on the variation of gene CNs in the five groups of *B. bruxellensis*. The coverage along the genome of each strain was scanned with Control-FREEC (Boeva et al. 2011) to detect the genes impacted by a variation of the CN. Among the 5,206 annotated protein-coding genes in the reference

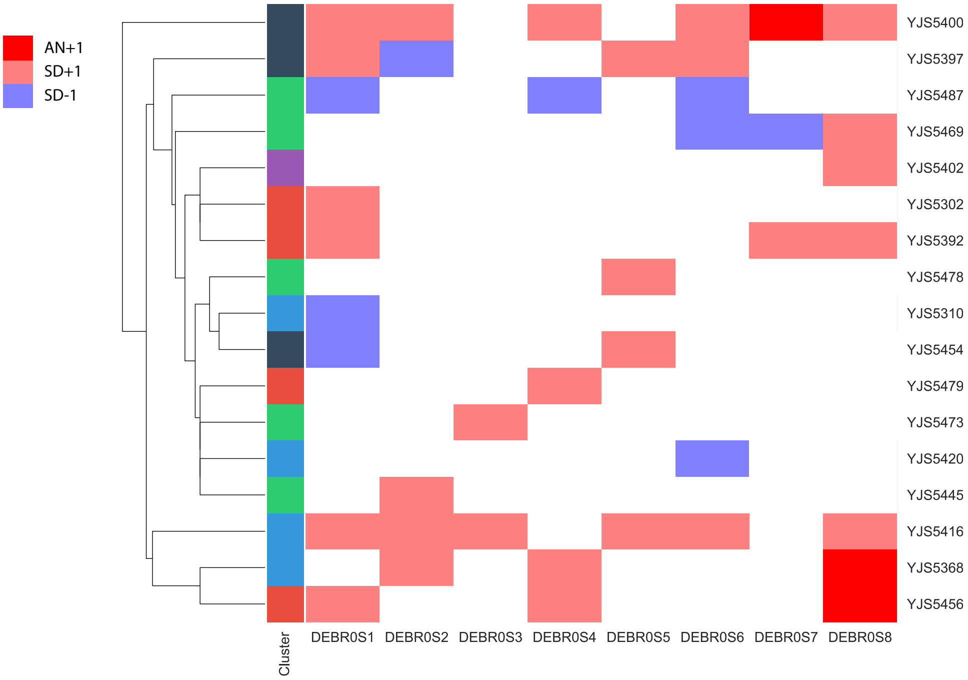

Fig. 3.—Distribution of chromosomal gain (CG), segmental duplications (SD), and segmental loss (SL) for each sample.

genome, 4,088 genes (78.5%) showed a variation in the number of copies in at least one strain. Among them, 3,587 genes were at least detected once as duplicated, whereas 1,734 genes were detected as lost. Among these genes, 100 were observed as homozygous deletion within the genome of one strain or more. The loss of a gene was observed as more shared across isolates compared with duplications (mean of 4.70 vs. 2.72 strains affected, respectively). In addition, CNs are enriched in subtelomeric regions as previously observed in other yeast species such as *S. cerevisiae* (Peter et al. 2018), and LTR, mobile elements, and tRNA are proportionally more impacted by these variants compared with CDS (supplementary fig. S7, Supplementary Material online).

Interestingly, CDSs in triploid isolates are more subjected to CN variation (fig. 4*A*, *P* value = 2.17×10^{-128}), which is consistent with the segmental duplication profiles. Although the number of CNVs in diploid strains is 5.6 times lower compared with triploid strains (mean of 169 vs. 881, respectively), this value is variable among strains and some diploid isolates, such as YJS5456, display a number of CNVs similar to that observed in triploid strains (fig. 4*B* and supplementary table S6, Supplementary Material online). Moreover, the ratio between deleted and duplicated CDS is variable among clusters (supplementary table S6, Supplementary Material online) and the G3N1 cluster shows a significantly higher number of duplicated CDSs compared with those that have been deleted (ratio = 2.6, supplementary table S6, Supplementary Material online). This could obviously be linked to the high number of segmental duplications found in the G3N1 cluster.

Functional Insight into the Duplicated and Deleted Genes

Gene copy variants are known to be a driving mechanism of genomic adaptation to changing environments in yeast (Kondrashov 2012) and are frequently found to be associated with domesticated processes. For example, in *S. cerevisiae*, duplications of the *CUP* genes have been repeatedly associated with resistance to copper (Strope et al. 2015; Peter et al. 2018) and the cluster of *MAL* duplicated genes has been highlighted as facilitating the utilization of maltose. It can be found in the ale beer isolates for which this sugar is the main carbon source during the fermentation process (Gallone et al. 2016). Within *B. bruxellensis*, the investigation of the genome assemblies related to different isolates already highlighted cases of gene content variation between strains. For example, an expansion of the alcohol dehydrogenase family, enabling greater control over ethanol formation and consumption, has been found in the wine isolate AWRI1499 (Curtin et al. 2012). Whole-genome comparison between two wine strains, AWRI1499 and CBS2499, and a beer strain, ST05.12/22 also highlighted 20 genes related to sugar metabolic processes and nitrogen consumption that are specifically present in the wine strains (Crauwels et al. 2014). To determine whether these variations are shared within clusters or are mostly independent events, we first investigated the gene CN of these 20 genes within the whole population (Crauwels et al. 2014) (fig. 5). We found that all these genes were missing in the YJS5392 strain (G2N3). This strain was isolated in Belgium and is closely related to the STO5.12/22 beer isolate. However, this strain is an exception within the whole

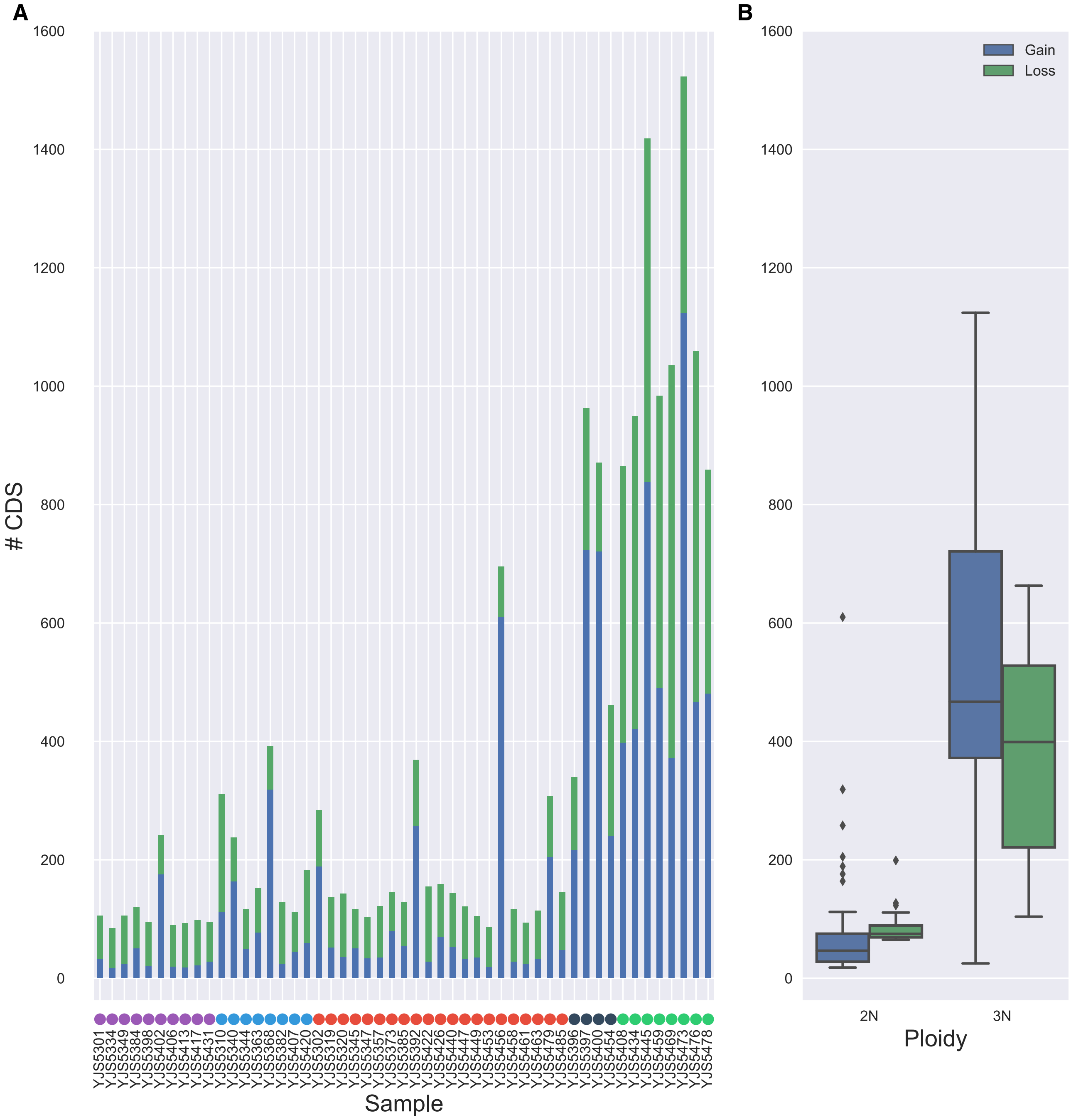

Fig. 4.—CNV distribution among the population. (*A*) Distribution of CNVs by ploidy. (*B*) Number of CNVs for each sample.

population and, as already observed through PCR amplification by Crauwels et al. (2015), most strains contains a majority of these genes. Our analysis also showed that the number of copies of these genes vary but is mostly stable within the different subpopulations. Interestingly, two genes coding for MFS drug transporters and one gene coding for a high-affinity glucose transporter have more than four copies in all triploid strains. Moreover, some strains from the triploid cluster G3N2 related to the wine strain AWRI1499 have multiple copies of genes coding for proteins responsible for galactose, glucose, and hexose transporter and metabolism, whereas diploid strains have two copies, and one copy is found in isolates from the G2N1 cluster. Finally, genes involved in nitrogen metabolism were completely absent from the seven diploid strains independently found in the three different clusters.

To determine if other genes are under selective pressure within subpopulations, we examined all duplicated or deleted genes within each cluster for which at least 70% of the strains

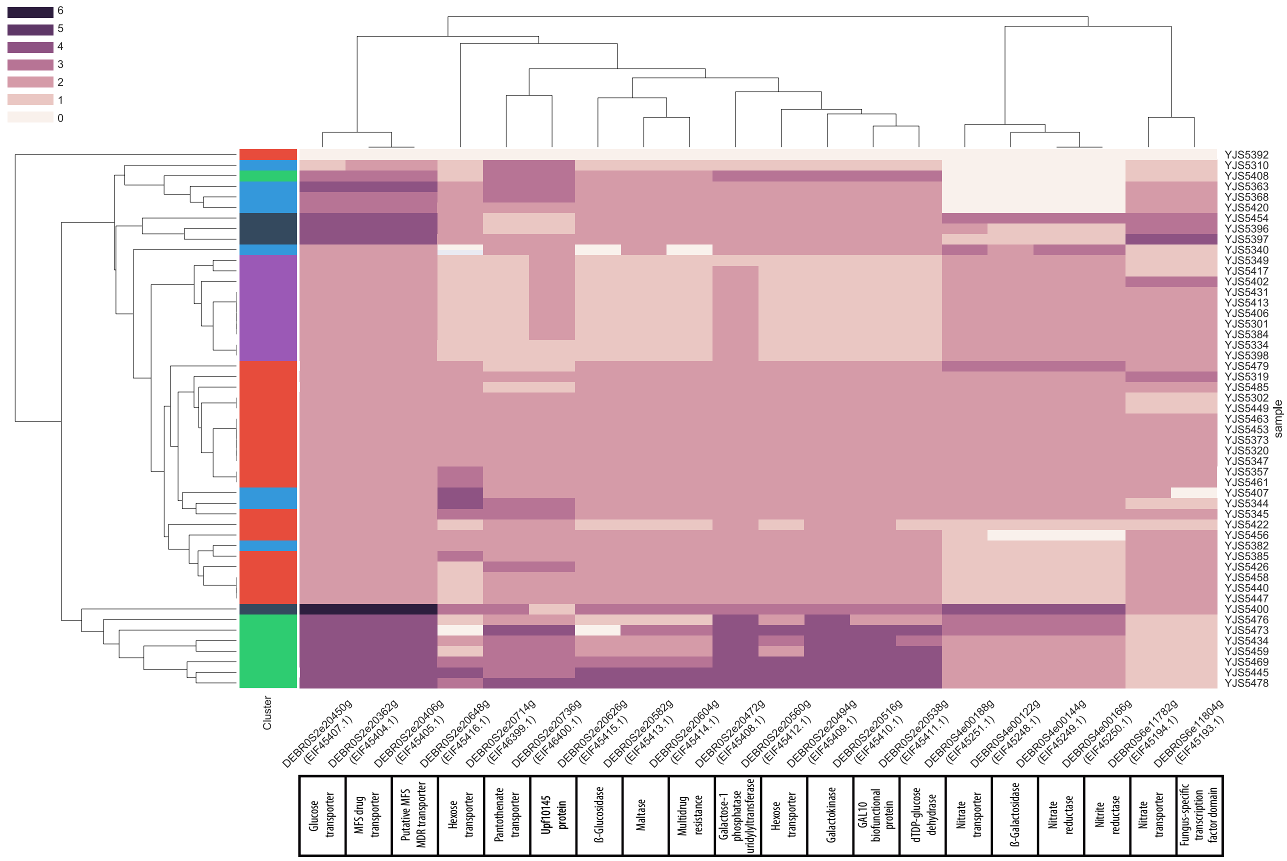

Fig. 5.—CNVs within the whole population for 20 genes previously found to be present in the two strains AWRI1499 and CBS2499 but missing in the ST05.12/22a isolate.

displayed the same structural changes (supplementary table S7, Supplementary Material online). Most of these genes were found in the triploid clusters, especially in the G3N2 cluster encompassing strains related to the AWRI1499 wine strain. Several genes involved in core pathways, such as histones H3 and H4 or ribosomal proteins can be found in the G2N3 and G3N2 clusters. Moreover, GO-term analysis of genes found in the G3N2 cluster revealed enrichment for several biological processes including galactose metabolism (*GAL7*, *GAL3*, *P* value = 0.003), nitrogen utilization (*UGA1*, *ATO2*, *P* value= 0.004), and transmembrane transport (9 genes, *P* value = 0.005).

The *B. bruxellensis* Pangenome Is Small with a Few Accessory Genes

To complete our analysis of the gene content within *B. bruxellensis*, we determined the species pangenome, that is, the global set of ORFs (open reading frame) present within the species. To that end, de novo assemblies for all the studied isolates were constructed and scanned to detect nonreference materials (see Materials and Methods). A total of 203 additional protein-coding genes were highlighted, leading to a pan-genome constituted of 5,409 protein-coding genes and a total of 303 accessory genes, as 100 genes were detected as fully deleted in at least one strain by our CNV analysis. This result shows that the pangenome is much smaller compared with that defined in *S. cerevisiae*, with a total of 1,712 accessory genes (Peter et al. 2018). In *B. bruxellensis*, 5,106 genes were found in all isolates and consequently were assigned to the core genome (94.4% of the pangenome). Supplemental ORFs were mostly found in triploid strains, with 154 and 49 of ORFs specific to the triploid and diploid strains, respectively. Moreover, these supplemental ORFs are poorly shared between the strains, as 67 (33%) are unique to specific isolates. However clustering analysis still revealed groups of genes associated with the different subpopulations (supplementary fig. S8, Supplementary Material online). To determine if these genes provide adaptive advantages, we searched for putative functions based on similarity searches from the protein sequences. Several transporters were found to be specifically shared within the triploid wine-related cluster (G3N2), such as two accessory genes with similarities to MFS drug transporters. However, no significant functional enrichment was found and multiple genes were linked to transposons or flocculation proteins, which are known to easily degenerate during evolution and could therefore be false positives in our analysis.

Although current data do not provide the opportunity to determine the precise origin of supplemental genes, their prevalence in the triploid isolates suggests that a significant number of these genes may result from hybridization events. Indeed, it is possible that the formation and selection of hybrids is linked to the contribution of the two genomes, sharing both new alleles and genes in a single individual, ultimately conferring beneficial advantages to specific environments. In this regard, further studies based on haplotype phasing in hybrid genomes need to be initiated, and will provide valuable insights into the genomic adaptations driven by hybridization events.

Conclusion

With the advent of affordable sequencing technologies, it is now possible to explore and analyze the genome-wide variability within nonmodel, but industrially relevant species. In this study, for the first time, we provide a comprehensive description of the genetic variability at the genome-scale of a population of *B. bruxellensis*, giving us a better view of its evolutionary history and the genetic variations underlying the phenotypic diversity within this species. Our results show the presence of at least two hybridization events, which is one of the main factors involved in the division classification of this species into subpopulations. It is likely that they are a driving mechanism of *B. bruxellensis* evolution as an adaptive response to the harsh environments found in the domestication processes. Interestingly, similar and common patterns of genetic variability are observed in both wine and beer triploid subpopulations. Nevertheless, significant differences between the genomes of the two subpopulations can also be found, suggesting the presence of potential genomic adaptations, which are specific to each of them. In addition, our results indicate that LOH is also present in *B. bruxellensis* evolution, impacting the genetic diversity within the species. Moreover, several aneuploidies, segmental duplications, and CNVs were also found in the whole population, especially in the triploid strains, indicating that these genome dynamics are important within the species and that triploid hybrids favor these types of structural variations. These observations are similar to what has been shown in the hybrid species *Saccharomyces pastorianus* resulting from the combination of the genomes of *S. cerevisiae* and *Saccharomyces eubayanus*, for which subpopulations show extensive chromosome loss and LOH events (Okuno et al. 2016). However, whether the hybridization events in *B. bruxellensis* derived from isolates of the same or two different species remains unknown and both a deeper analysis of triploid isolates genomes and the sequencing of closely related species genomes would be needed to investigate this aspect of their biology. At the whole population scale, our data provide the opportunity for a deeper view of the genetic variants involved in the phenotypic diversity of *B. bruxellensis*. Analysis of CNVs and accessory genes in the populations highlighted several genes involved in drug and sugar transports as previously found in other analyses. However, these genes were mostly found in the wine-related triploid clusters. Interestingly, the nitrogen pathway was independently lost in several diploid isolates within different subpopulations, suggesting that nitrate assimilation is not a common requirement for *B. bruxellensis* isolates. This result is in accordance with a phenotypic analysis in which up to a third of the isolates failed to grow on nitrate, which could result from the reduction of ethanol and the production of acetic acid during anaerobic fermentation nitrate assimilation (Galafassi et al. 2013).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We are grateful to Warren Albertin and Isabelle Masneuf-Pomarede for fruitful discussions and invaluable advice. We thank the B2I platform (IBMP-CNRS, Strasbourg, France) for their support. This work was supported by the Agence Nationale de la Recherche (ANR-18-CE20-0003-02 and ANR-18-CE12-0013-02) and the European Research Council (ERC Consolidator Grant 772505). J.-S.G. was supported by a grant from the French "Ministère de l'Enseignement Supérieur et de la Recherche." J.S. is a Fellow of the University of Strasbourg Institute for Advanced Study (USIAS) and a member of the Institut Universitaire de France.

Literature Cited

Agnolucci M, et al. 2009. Genetic diversity and physiological traits of *Brettanomyces bruxellensis* strains isolated from Tuscan Sangiovese wines. Int J Food Microbiol. 130(3):238–244.

Albertin W, et al. 2014. Development of microsatellite markers for the rapid and reliable genotyping of *Brettanomyces bruxellensis* at strain level. Food Microbiol. 42:188–195.

Almeida P, et al. 2014. A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. Nat Commun. 5:4044.

Almeida P, et al. 2015. A population genomics insight into the mediterranean origins of wine yeast domestication. Mol Ecol. 24(21):5412–5427.

Avramova M, et al. 2018. *Brettanomyces bruxellensis* population survey reveals a diploid-triploid complex structured according to substrate of isolation and geographical distribution. Sci Rep. 8(1):4136.

Beckner M, Ivey ML, Phister TG. 2011. Microbial contamination of fuel ethanol fermentations. Lett Appl Microbiol. 53(4):387–394.

Bergström A, et al. 2014. A high-definition view of functional genetic variation from natural yeast genomes. Mol Biol Evol. 31(4):872–888.

Boeva V, et al. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics 27(2):268–269.

Borneman AR, Zeppel R, Chambers PJ, Curtin CD. 2014. Insights into the *Dekkera bruxellensis* genomic landscape: comparative genomics reveals variations in ploidy and nutrient utilisation potential amongst wine isolates. PLoS Genet. 10(2):e1004161.

Carreté L, et al. 2018. Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. Curr Biol. 28(1):15–27.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28(4):464–469.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80–92.

Conterno L, Joseph CML, Arvik TJ, Henick-kling T, Bisson LF. 2006. Genetic and physiological characterization of *Brettanomyces bruxellensis* strains isolated from wines. Am J Enol Vitic. 57:139–147.

Crauwels S, et al. 2014. Assessing genetic diversity among *Brettanomyces* yeasts by DNA fingerprinting and whole-genome sequencing. Appl Environ Microbiol. 80(14):4398–4413.

Crauwels S, et al. 2015. Comparative phenomics and targeted use of genomics reveals variation in carbon and nitrogen assimilation among different *Brettanomyces bruxellensis* strains. Appl Microbiol Biotechnol. 99(21):9123–9134.

Crauwels S, et al. 2017. Fermentation assays reveal differences in sugar and (off-) flavor metabolism across different *Brettanomyces bruxellensis* strains. FEMS Yeast Res. 17:fow105.

Curtin CD, Bellon JR, Henschke PA, Godden PW, de Barros Lopes MA. 2007. Genetic diversity of *Dekkera bruxellensis* yeasts isolated from Australian wineries. FEMS Yeast Res. 7(3):471–481.

Curtin CD, Borneman AR, Chambers PJ, Pretorius IS. 2012. Pretorius IS 2012. *De novo* assembly and analysis of the heterozygous triploid genome of the wine spoilage yeast *Dekkera bruxellensis* AWRI1499. PLoS One 7(3):e33840.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

Fawcett JA, et al. 2014. Population genomics of the fission yeast *Schizosaccharomyces pombe*. PLoS One 9(8):e10424.

Ford CB, et al. 2015. The evolution of drug resistance in clinical isolates of *Candida albicans*. Elife 4:e00662.

Fournier T, et al. 2017. High-quality *de novo* genome assembly of the *Dekkera bruxellensis* yeast using nanopore minion sequencing. G3 (Bethesda) 7:3243–3250.

Friedrich A, Jung P, Reisser C, Fischer G, Schacherer J. 2015. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. Mol Biol Evol. 32(1):184–192.

Galafassi S, Capusoni C, Moktaduzzaman M, Compagno C. 2013. Utilization of nitrate abolishes the "Custers effect" in *Dekkera bruxellensis* and determines a different pattern of fermentation products. J Ind Microbiol Biotechnol. 40(3–4):297–303.

Gallone B, et al. 2016. Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. Cell 166(6):1397–1410.

Gonçalves M, et al. 2016. Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. Curr Biol. 26(20):2750–2761.

Hellborg L, Piskur J. 2009. Complex nature of the genome in a wine spoilage yeast, *Dekkera bruxellensis*. Eukaryot Cell. 8(11):1739–1749.

Hirakawa MP, et al. 2015. Genetic and phenotypic intra-species variation in *Candida albicans*. Genome Res. 25(3):413–425.

Huson DH 2019. Drawing rooted phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform. 6(1):103–109.

Ishchuk OP, et al. 2016. Novel centromeric loci of the wine and beer yeast *Dekkera bruxellensis* CEN1 and CEN2. PLoS One 11(8):e0161741.

Jackman SD, et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 27(5):768–777.

Jeffares DC, et al. 2015. The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. Nat Genet. 47(3):235–241.

Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14(4):R36.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc B. 279(1749):5048–5057.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5(1):59.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 4(7):1073–1081.

Leducq JB, et al. 2016. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. Nat Microbiol. 1:15003.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079.

Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 44(W1):W54–W57.

Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1(1):18.

Lynch DB, Logue ME, Butler G, Wolfe KH. 2010. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome Biol Evol. 2:572–583.

McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.

Neuveglise C. 2002. Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. Genome Res. 12(6):930–943.

Okuno M, et al. 2016. Next-generation sequencing analysis of lager brewing yeast strains reveals the evolutionary history of interspecies hybridization. DNA Res. 23(1):67–80.

Olsen RA, et al. 2015. *De novo* assembly of *Dekkera bruxellensis*: a multi technology approach using short and long-read sequencing and optical mapping. GigaScience 4(1):56.

Ortiz-Merino RA, et al. 2018. Ploidy variation in *Kluyveromyces marxianus* separates dairy and non-dairy isolates. Front Genet. 9:94.

Otto SP. 2007. The evolutionary consequences of polyploidy. Cell 131(3):452–462.

Peter J, et al. 2018. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature 556(7701):339–344.

Piškur J, et al. 2012. The genome of wine yeast *Dekkera bruxellensis* provides a tool to explore its food-related properties. Int J Food Microbiol. 157(2):202–209.

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44(12):e113.

Ropars J, et al. 2018. Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. Nat Commun. 9(1):2253.

Schifferdecker AJ, Dashko S, Ishchuk OP, Piškur J. 2014. The wine and beer yeast *Dekkera bruxellensis*. Yeast 31(9):323–332.

Skelly DA, et al. 2013. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res. 23(9):1496–1504.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2):ii215–ii225.

Strope PK, et al. 2015. The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res. 25(5):762–774.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595.

Teoh AL, Heard G, Cox J. 2004. Yeast ecology of kombucha fermentation. Int J Food Microbiol. 95(2):119–126.

Thomson JM, et al. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nat Genet. 37(6):630–635.

Tiukova IA, et al. 2019. Chromosomal genome assembly of the ethanol production strain CBS 11270 indicates a highly dynamic genome structure in the yeast species *Brettanomyces bruxellensis*. PLoS One 14(5):e0215077.

Vigentini I, et al. 2012. Intraspecific variations of *Dekkera/Brettanomyces bruxellensis* genome studied by capillary electrophoresis separation of the intron splice site profiles. Int J Food Microbiol. 157(1):6–15.

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21(11):2791–2793.

Zhu YO, Sherlock G, Petrov DA. 2016. Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. G3 (Bethesda) 6:2421–2434.

Associate editor: Kenneth Wolfe