European derived *Saccharomyces cerevisiae* colonization of New Zealand vineyards aided by humans

Velimir Gayevskiy[1], Soon Lee[1] and Matthew R. Goddard[1,2]

[1]School of Biological Sciences, The University of Auckland, Auckland, New Zealand.

[2]The School of Life Sciences, The University of Lincoln, Lincoln, United Kingdom

Running head: Phylogenomics of *S. cerevisiae*

Address correspondence to Matthew Goddard: mgoddard@lincoln.ac.uk

Keywords: Phylogenomics, yeast, *Saccharomyces cerevisiae*, population genetics, genomics

Abstract

Humans have acted as vectors for species and expanded their ranges since at least the dawn of agriculture. While relatively well characterized for macrofauna and macroflora, the extent and dynamics of human-aided microbial dispersal is poorly described. We studied the role which humans have played in manipulating the distribution of *Saccharomyces cerevisiae*, one of the world's most important microbes, using whole genome sequencing. We include 52 strains representative of the diversity in New Zealand to the global set of genomes for this species. Phylogenomic approaches show an exclusively European origin of the New Zealand population, with a minimum of ten founder events mostly taking place over the last 1,000 years. Our results show that humans have expanded the range of *S. cerevisiae* and transported it to New Zealand where it was not previously present, where it has now become established in vineyards, but radiation to native forests appears limited.

One sentence summary: Genome sequencing shows that humans have unwittingly transported wine yeast to the other side of the planet, where this species has become established in vineyards.

Introduction

Humans have transported other species beyond their natural ranges for thousands of years, both intentionally for agricultural purposes (Diamond, 2002) and unintentionally as a consequence of human migration (Wichmann *et al.*, 2009). Other than disease agents, whose effects are apparent once transposed (Mazzaglia *et al.*, 2012), the extent to which humans have manipulated the species ranges of microbes is poorly characterized (Litchman, 2010). Previous studies suggested microbes have virtually limitless dispersal abilities (de Wit & Bouvier, 2006). However, while some microbes, such as marine bacteria, appear globally distributed (Pedrós-Alió, 2006), others, such as hot spring communities, are certainly not (Martiny *et al.*, 2006; Valverde *et al.*, 2012; Almeida *et al.*, 2014; Talbot *et al.*, 2014; Taylor *et al.*, 2014; Tripathi *et al.*, 2014), and the forces which give rise to these microbial patterns are not clear (Hanson *et al.*, 2012; Morrison-Whittle & Goddard, 2015). Microbes are key components of both natural and agricultural ecosystems, but we are generally ignorant of the means by which microbes might be dispersed, let alone the degree to which humans influence microbial species ranges (Talbot *et al.*, 2014; Tripathi *et al.*, 2014).

Phylogeography is the primary method used to study the distributions of organisms in relation to their genetic diversity (Avise *et al.*, 1987), and allows inference of movements and speciation events. Phylogenomics follows this approach but utilizes large portions of genomes, as opposed to a few markers (Delsuc *et al.*, 2005). To date phylogenomic studies have mainly been applied to plant and animal species (del Campo *et al.*, 2014). While a vast array of robust biogeography studies have examined the variance in microbial species distributions (reviewed in Hanson *et al.*, 2012), there are relatively few that have employed a phylogeographic approach, and those

that exist have largely used mtDNA, microsatellite or single-locus genetic markers which can be biased or lack adequate resolution (Beheregaray, 2008). However, recent studies have examined the population genomics of *Saccharomyces* yeast species to infer their origin and signals for domestication (Almeida *et al.*, 2014; 2015; Barbosa *et al.* 2016; Ludlow *et al.*, 2016). The *Saccharomyces* genus is composed of seven species and originated 10-20 million years ago (Hittinger, 2013). All species have complete genomes available and have been used for numerous functional (Skelly *et al.*, 2013; Bergstrom *et al.*, 2014), phylogenetic (Drummond *et al.*, 2006; Scannell *et al.*, 2011), biochemical (Piskur *et al.*, 2006) and evolutionary (Novo *et al.*, 2009) studies. *S. cerevisiae* was the first eukaryote sequenced in its entirety due to its small 12Mb genome. Since then, it has become the best annotated eukaryotic genome (Cherry *et al.*, 2011) and remains a cornerstone of the genomics community with over well 200 genomes available that are being added to consistently (Cherry *et al.*, 2011; Skelly *et al.*, 2013; Bergstrom *et al.*, 2014; Almeida *et al.*, 2015; Strope *et al.*, 2015, reviewed in Peter & Schacherer 2016), and a further 37 available for its sister species *S. paradoxus* (Liti *et al.*, 2009; Bergstrom *et al.*, 2014).

The global distribution of *S. cerevisiae* is becoming increasingly well characterized, as demonstrated by the recent revelation of major basal clades in China (Wang *et al.*, 2012), discovery of an ancient European population (Almeida *et al.*, 2015), the discovery of hybrid populations associated with coffee and coca (Ludlow *et al.*, 2016) and novel linages in Brazil (Barbosa *et al.* 2016). One pattern consistently found in all studies to date is the close relatedness and short divergence time of a "Wine/European" group (Liti *et al.*, 2009; Schacherer *et al.*, 2009; Wang *et al.*, 2012; Cromie *et al.*, 2013). This group includes commercial winemaking strains, strains sampled in vineyards and wineries worldwide, as well as strains

from European forests, and the proposed ancestral European group inhabiting Mediterranean oak (Almeida *et al.*, 2015). Using micro-satellite profiles, it has been suggested that dispersal from Europe by humans in association with the global spread of viticulture and winemaking explains this pattern (Legras *et al.*, 2007). Together this suggests that *S. cerevisiae* is a species with some clades that are closely associated with and dispersed by humans, but there are other clades present in natural environments and probably dispersed only locally by other means, such as insects (Stefanini *et al.*, 2012; Buser *et al.*, 2014), or not at all. A very recent study described genetically distinct populations of *S. cerevisiae* associated with coffee and cocoa in Africa and South America (Ludlow *et al.*, 2016); intriguingly these do not contain novel alleles, but are inferred to have been created by the mixing of existing populations associated with European vineyards, American oak trees and the ancestral seat of this species in Far East Asia. Ludlow *et al.* (2016) reasonably suggests that the movement of these strains, and thus creation of these populations, was facilitated by humans. Similarly, the recently discovery of a novel lineage in Brazil shows it was formed in part by hybridization of migrants from the European/wine group with endemic *S. paradoxus,* which presumably then facilitated the colonization of native Brazilian trees (Barbosa *et al.* 2016); it also seems reasonable to infer that humans facilitated this radiation event.

New Zealand (NZ) is the last major landmass colonized by humans ~1,000 years ago (Hurles *et al.*, 2003) and represents a unique environment to investigate questions concerning species range expansion. Māori were the first humans to settle in New Zealand, and Europeans did not arrive until Captain Cook's voyage of 1769 (though Abel Tasman sighted NZ in 1642). Viticulture was introduced into NZ around 1800. Many of NZ's endemic macroscopic flora and fauna have been studied (Wallis & Trewick, 2009); however, extremely limited work has been

conducted on the biogeography of microbial species in NZ. Previous analyses, based in microsatellite profiles and RAD-seq, suggest that NZ harbours a diverse and globally genetically distinct metapopulation of *S. cerevisiae*, with some geographically distinct localized populations that are also connected by various levels of gene flow (Goddard *et al.*, 2010; Gayevskiy & Goddard, 2011; Cromie *et al.*, 2013; Knight & Goddard, 2015). Strains have been isolated from vineyard and winemaking associated niches (Goddard *et al.*, 2010; Knight & Goddard, 2015), oak trees planted by European migrants (Zhang *et al.*, 2010), and from native NZ forests and fruiting trees (Knight & Goddard, 2015; Gayevskiy & Goddard, 2016). While small in terms of global production, the New Zealand wine industry commands a strong position in the premium market and this sector is significant to the NZ economy. *Saccharomyces* yeast play a role in the production of wine, including potentially being part of the process that geographically differentiates wines (Knight *et al.* 2015). Along-side the academic interest in *Saccharomyces* ecology and population biology (Goddard & Grieg, 2015) their role in winemaking adds economic interest to understand the origin of *Saccharomyces cerevisiae* populations. One recent study suggests the presence of an ancient population of the yeast *Saccharomyces uvarum* in Australasia, but this species is certainly not endemic to NZ, nor is there evidence for an NZ-specific population (Almeida *et al.*, 2014). Another study only recently reported the presence of *Saccharomyces eubayanus* and *Saccharomyces arboricola* in New Zealand, but the age and origin of these species is uncertain (Gayevskiy & Goddard, 2016).

Here we ask why *S. cerevisiae* is present in New Zealand, and use phylogenomic methods to evaluate its history and range expansion. Two extremes present themselves, either: 1) there was an ancient *S. cerevisiae* population present in NZ prior to humans arriving under 1,000 years ago; or 2) that this species was transported to NZ by humans with winemaking who

unwittingly expand this species' range along with exotic fruit bearing plants and trees. Of course some mix of the two is also possible.

Materials and Methods

Strain Selection and Sequencing

The K-means clustering algorithm used to identify maximally divergent genotypes was implemented in R (R Development Core Team, 2011). Sulfite tolerance was assayed by plating onto YPD with either 10, 15 mM or 20 mM sodium metabisulfite in triplicate and scoring the growth of colonies as full, partial or none after 2 days at 28 °C. Each strain was propagated in YPD and high molecular weight genomic DNA was extracted using the Qiagen™ Blood & Cell Culture DNA Kit. Libraries were constructed using the Illumina TruSeq Nano DNA Sample Prep Kit with 550 bp insert size. Sequencing was carried out at the Beijing Genomics Institute (China) on a single 150 bp paired-end lane of an Illumina HiSeq 2000.

Genome Mapping and Quality Control

Each sequenced genome was treated identically using a custom bioinformatics pipeline written in Perl. This pipeline is outlined below.

Quality Control and Trimming FASTQC (v0.10.1; Andrews, 2012) was used for quality control of each library and to determine optimal trimming parameters. Trimming was conducted with Trimmomatic (v0.25; Lohse *et al.*, 2012) using the following parameters: "LEADING:3

TRAILING:3 SLIDINGWINDOW:3:20 MINLEN:30". Following trimming, FASTQC was executed on the trimmed reads for comparison with the initial reports.

Mapping and Variant Calling All trimmed reads were mapped against the *S. cerevisiae* reference strain S288C using Bowtie2 (v0.12.7; Langmead & Salzberg, 2012). Following mapping, samtools (v0.1.18; Li *et al.*, 2009) was used for alignment conversion, sorting and indexing. A variant call file was produced using the mpileup command within samtools with the "-Bu" parameters. The variant call file was used to create a consensus genome without the reference to allow for INDELs using the vcf2fq Perl script within samtools. Putative heterozygous positions were conservatively called as 'N' as the phylogenetics and population genetics methods utilized do not support ambiguous calls. Heterozygous positions were quantified with a custom Perl script which filtered out positions with a sequencing depth below 10 or above 100 and a genotype quality below 20.

Data Availability We have made our raw sequence data and consensus genomes aligned to S288C (Goffeau et al., 1996; EBI:GCA_000146045.2) publicly available at SRA: SRP042301 and BioProject: PRJNA247448.

Sequence Extraction from Sequenced and International Genomes

In addition to the 52 genomes sequenced here, a further 72 *S. cerevisiae* and 37 *S. paradoxus* genomes were obtained from the *Saccharomyces* Genome Database (http://yeastgenome.org), NCBI, the *Saccharomyces* Genome Resequencing Project (https://sanger.ac.uk/research/projects/genomeinformatics/sgrp.html) and from Huang *et al.*,

(2014) in the form of consensus genomes and/or raw data. To obtain an accurate estimation of the relatedness of the genomes, we extracted the well-known set of 106 orthologous loci spread through the genome of *S. cerevisiae* and present in all *Saccharomyces* species (Rokas *et al.*, 2003). The sequences of these 106 loci were extracted by searching the S288C sequence for each locus against each consensus genome using the BLAST algorithm. Only genomes with complete sets of 106 loci were retained for phylogenetic analysis.

All sets of 106 loci were subjected to a multiple sequence alignment using clustalw (v.2.1; Larkin *et al.*, 2007) within Geneious (v6; Biomatters Ltd., 2012). Alignments were manually curated within Geneious due to the frequent homopolymer indels present in some of the genomes due to older sequencing technology. We created a second dataset comprising 13 loci sequenced from 99 *S. cerevisiae* strains isolated in China (Wang *et al.*, 2012). Only these loci were sequenced from these Chinese strains with no overlap with our main dataset. The consensus sequence for each of these loci was used to search against all available genomes outlined above. Genomes with complete sets of all 13 loci were retained for phylogenetic analysis. We used five *S. paradoxus* genomes as an outgroup, although four of the loci include intergenic regions and the *S. paradoxus* genomes did not yield these loci. The remaining nine loci were sufficient for phylogenetic analysis. Multiple sequence alignments were carried out in the same way as for the 106 loci dataset.

Phylogenetics

Phylogenetic analyses were conducted using BEAST (v1.7.5; Drummond & Rambaut, 2007) on the finalized sequence alignments for both data sets. A number of scenarios were run to explore

the relationships between genomes and to determine the stability of inferred relationships by locus and dataset.

Substitution and clock models were unlinked for all loci in all analyses to facilitate their independent estimation. Trees were linked to obtain a consensus tree using all loci. All substitution model and rate options were left on default due to the large increase in processing time observed when any were changed. A lognormal relaxed clock (uncorrelated) was used with an exponential distribution of mean 0.3. All runs were conducted with 1 billion iterations due to the size of the data sets. We verified MCMC convergence by examining the effective sample sizes of all parameters in each analysis and with visual inspection of the traces. 10 to 40% of each run was discarded as burn-in depending on the convergence of the MCMC trace. Separate phylogenetic analyses were conducted for the two clades found housing NZ strains in the 106 loci dataset. These used S288C as the outgroup to determine high-resolution structure within these clades. Parsimony analyses including permutations of NZ and Europe terminal taxa status, and calculations of the minimum change of this state over these phylogenies, were conducted in Mesquite (Maddison & Maddison, 2014).

We used the published divergence date estimates between *S. cerevisiae* and its sister species *S. paradoxus* as a calibration point for the divergences of clades within our phylogenies (Liti *et al.*, 2006). Divergences between clades within phylogenies are typically estimated using molecular clocks and/or by calibration time points of established species divergences using fossils. Molecular clocks for *S. cerevisiae* are not in wide use due to the difficulty of estimating clock-like rates of evolution in a species with unknown generation times in its natural environment and high rates of inbreeding. The time of the common ancestor of *S. cerevisiae*

and *S. paradoxus* has been estimated at 0.4 to 3.4 mya (Liti *et al.*, 2006). The molecular substitution rate observed between the split of the *S. cerevisiae* and *S. paradoxus* genomes was assumed to correspond to this time period. To estimate the divergence date of a particular clade, the proportional substitution rate for the clade was calculated against the calibration point to give a date estimate.

Population Metrics and Structure

We utilized ANGSD (v0.588; Nielsen *et al.*, 2012) to generate population genomic metrics. ANGSD operates on short read alignment bam files which affords statistical robustness in calculating the site frequency spectrum in comparison with traditional tools operating on a set genotype. Given this requirement, we could only use genomes with raw data available from Illumina sequencing technology. This was aided by the recent resequencing of diverse worldwide strains (Bergstrom *et al.*, 2014). We thus created three superset subsets: the NZ strains (52), the previous and the Wine/European strains (66) and the previous with all remaining strains (75). The number of sites and the number of segregating sites for each population was determined from the Mean Allele Frequency calculations in ANGSD with the minInd parameter set to the number of strains per population and the minMaf alternatively set to 0 and 0.01. Only high quality data were used (minQ=20 and minMapQ=30). Watterson's Estimator (θ) (Watterson, 1975), Tajima's Pi (π) (Tajima, 1989) and Tajima's D (Tajima, 1989) were calculated by first calculating the site allele frequency likelihood, then the maximum likelihood estimate of the SFS, then the thetas per site and finally summarized with the thetaStat utility in ANGSD.

Tests for admixture within the *S. cerevisiae* genomes were conducted with Structure (Pritchard *et al*., 2000) due to the haploid nature of some of the genomes in our dataset. We chose to include all strains of *S. cerevisiae* where a consensus genome consisting of entire chromosomes was available to capture entire genomic diversity. The chromosomes of the 93 strains that met these criteria were aligned using Mauve (v2.3.1; Darling *et al*., 2004) within Geneious (v6; Biomatters Ltd., 2012) and any nucleotide positions where either all strains showed no variation or at least one gap was present were removed. The remaining positions were run through Structure using the admixture model with 10,000 iterations of burn in followed by 20,000 iterations of analysis. K values between 2 and 20 were used with 3 replicate chains for each value of K to check for convergence, and the optimal number of sub-populations inferred using the Evanno method (Evanno *et al*., 2005). Population classifications were not used for the prior. Resulting ancestry profiles were objectively analyzed using ObStruct (Gayevskiy *et al*., 2014) to determine the extent that geographic origin, niche of isolation and our phylogenomic analysis explains inferred population structure.

Results

Strain Selection

We collated data from six recent studies that have surveyed for *S. cerevisiae* across New Zealand (Table 1; Serjeant *et al*. 2008; Goddard *et al.* 2010; Zhang *et al*. 2010; Gayevskiy & Goddard 2012; Knight & Goddard 2015; Gayevskiy & Goddard 2016). *S. cerevisiae* has been isolated from over 99% of spontaneous ferment samples, 10% of vineyard samples and only 1% of forest/tree samples. The order of magnitude difference in recovery of this species is not due

to differential sampling effort as most effort was spent sampling native forests, then vineyards and least for spontaneous ferments (Table 1). Just six genotypes (characterized at 9 microsatellite loci**;** Richards *et al*., 2009) have been recovered from trees/forests – one of these was from an exotic oak tree (Zhang *et al*. 2010), but microsatellite profiling showed this to be very closely related to DBVPG1106 – a strain isolated from Australian vineyards which clusters with the wine/European group for which whole genome sequence is already available, and is included in this study (Zhang *et al*. 2010). Population genetic analyses of the remaining five genotypes isolated from native NZ forests show these to be homogeneous with their regional vineyard and spontaneous ferment populations: there is no evidence for genetic differentiation between strains isolated from native forests and their vineyard counterparts (Knight & Goddard, 2015). Thus, it is clear that *S. cerevisiae* is very common in NZ spontaneous ferments, and at reasonable abundance in vineyard habitats, but rare in NZ native forests. This observation, and the fact that *S. cerevisiae* populations in these three habitats are connected within each region of NZ (Knight & Goddard, 2015), means the most likely explanation is that there is just one *S. cerevisiae* meta-population in NZ closely associated with vineyards and ferments, but members of this population are transposed to native habitats at some low rate. There is no evidence that NZ harbours another genetically distinct *S. cerevisiae* population that is not primarily associated with ferments and vineyards. The question we ask here concerns the origin of this group. To address this question we need data from a set of genomes that best reflects the genetic diversity in this population: this is the most pertinent parameter relevant to elucidating the origin of this species in NZ. We thus identified a set of 52 maximally divergent *S. cerevisiae* genotypes from the of 724 in our database using k-means clustering of microsatellite profiles; these are detailed in Table S1.

Sequencing and Mapping

Sequencing of genomes derived from clonally expanded diploid populations yielded an average of 5.1 million 150bp paired-end reads per strain for a total of 39.8Gbp of data. An average mapping rate of 97.16% was obtained for all genomes using the S288C genome as a reference, with an average coverage of 61X, and mapping quality of 38.78 (Phred score). An average of 52,421 (SE=532) SNPs and 4,915 (SE=36) INDELs were obtained for each genome. This number of SNPs is entirely consistent with other *S. cerevisiae* strains sequenced on the same platform from a diversity of international locations and niches (e.g. Bergstrom *et al.*, 2014). The average number of heterozygous SNPs per strain was 7,274 (SE=805) which is consistent with that found for other vineyard isolates (Magwene *et al.*, 2011), but heterozygosity levels ranged from ~3,000 to ~31,000 (6% to 43% of all SNPs) across the 52 NZ genomes. Cursory analyses of large-scale copy number variations indicate the genome of 6-Sol7-2 contains three copies of chromosome 4 and 27-WI_S_JASA_13 contains three copies of the first half of chromosome 7. Further, 37 of the 52 NZ genomes show large copy numbers (1.5-50X) of locus YGR201C (unknown function) on chromosome 7. We stress that the main hypothesis under test is the phylogeography of *S. cerevisiae*, and thus we do not concentrate on the details of fine-scale differences between genomes any further here.

Population Genomic Statistics

First we compared the NZ derived genomes to one another and then to 14 previously published genomes that either derived from Europe or are associated with winemaking, which form a tight clade, and finally to a further nine genomes derived from a diversity of locations and niches (Table S2). Only genomes with high quality data (Illumina sequencing with ~30X or greater

coverage) were included in this analysis. We only included sites that had high quality read data for all genomes, and thus the number of comparable homologous sites reduced as more genomes were added due to the increase in missing data for some of the existing genomes. The number of segregating sites is proportionately similar between the Wine/European group and the NZ population, but nearly doubles when the other genomes are added, indicating their relative divergence (Table 2). Pi is a measure of nucleotide diversity (Tajima, 1989) and the NZ and Wine/Europe populations appear identical in terms of nucleotide diversity, but again the inclusion of the non-wine/Europe genomes leads to a 30% increase in this statistic. This suggests the NZ derived genomes are more similar to the genomes deriving from Europe or associated with winemaking than to genomes derived from elsewhere. It is tempting to further investigate comparative population genetics, but since the international samples are not random representatives of a population this defies many of the assumptions underlying population genetic calculations, and so we have not pursued this here.

Phylogenomic Approaches

First we employed phylogenomic methods to evaluate long-term within-species population structure. Initially we chose a comprehensive set of 106 orthologous loci compiled by Rokas *et al.* (2003) for analyses due to their distribution across the genome, presence in all *Saccharomyces* species, and their proven capability to provide a robust phylogenetic signal. We did not use the entirety of the genomic data due to potential problems with identifying orthologs and paralogs. Of the existing 72 *S. cerevisiae* and 37 *S. paradoxus* genomes, 60 and 36 respectively contained these 106 loci; the remaining genomes had insufficient or low quality

sequencing coverage for at least some of the loci (Table S2). All 52 NZ genomes contained complete sets of these 106 loci.

Rampant recombination among these genomes, which would be indicated by more of network-like than tree-like relationships, would significantly decrease the validity of using a phylogenetic approach for the analyses of these genomes due to its assumptions regarding bifurcating relationships. Neighbour-net (Huson & Bryant, 2006) analysis (Figure S1) reveals a topology that to a first approximation shows a more tree-like than network-like structure, suggesting little recombination between major groups, and thus lends greater confidence for the use of phylogenetic approaches to evaluate some of the relationships between these genomes. To place the NZ population in a global context, we reconstructed a phylogeny using Bayesian approaches and included the 36 *S. paradoxus* genomes for calibration and rooting purposes (Figure 1). The inclusion of NZ genomes reproduces an overall global topology that is comparable to earlier analyses (Liti *et al.*, 2009; Schacherer *et al.*, 2009). Strikingly, 85% of the NZ strains, including the strain isolated from a native forest, are interspersed within the Wine/European clade (Figure 1). The resolution within this clade is extremely poor, suggesting this comprises a contemporaneous population experiencing gene flow and recombination, and the relatively short branch lengths show little time since divergence. This represents the first piece of evidence that the *S. cerevisiae* in NZ have a significant portion of ancestry, and thus derive from, and are in fact part of, the European population. Not all NZ strains fall within this Wine/European group however. The remaining 15% of NZ strains form a sister clade to the Wine/European group, with the inclusion of I14 and Y55, which are two soil isolates from Europe. Resistance to sulphite is a key defining phenotype of the *S. cerevisiae* lineage associated with viticulture and winemaking as sulphur is and has been used as an anti-microbial

in both vineyards and wineries (Pretorius, 2000; Aa *et al.*, 2006). To evaluate whether this smaller group might represent a population not associated with wine, we tested the sensitivity of these to sulphite. There is no significant difference in resistance to 10, 15 and 20 mM sulphite between the NZ and European groups as determined by plate assays ($F_{[1,105]}$ 1.59, 0.53, 0.00 respectively and all P<0.21), but these two groups are significantly more tolerant to 10, 15 and 20 mM sulphite than the rest of the non-wine associated strains ($F_{[2,129]}$ 54.3, 20.9, 8.9 respectively and all P>0.0002); Figure 2.

Previously identified clades (Liti *et al.*, 2009; Schacherer *et al.*, 2009) are reconstructed and expanded with our analyses due to the inclusion of further recently sequenced genomes. The North American clade includes additional strains sampled from Missouri (T7) and Bahamas (UWOPS83_787_3), the West African clade contains the PW5 strain sampled from Nigerian palm wine and the Sake clade contains an additional three strains (UC5, Kyokai7 and ZTW1). Several new clades are present for laboratory and bioethanol strains. Apart from the Sake clade, the ordering of the clades in relation to the *S. paradoxus* outgroup places strains isolated from non-agricultural niches as basal while agricultural and biotechnological associated strains are relatively derived indicating their more recent formation. Strains not residing in these clades are interspersed through the tree and tend to be positioned at the ends of longer branches and could indicate the presence of further under-sampled populations or represent chimeric strains with ancestry in multiple clades.

Recently, a large novel diversity within *S. cerevisiae* was revealed by the sequencing of 13 loci from 99 strains isolated in China, leading to suggestions this species originated on the Asian continent (Wang *et al.*, 2012). We extracted these loci from all available whole genomes,

resulting in 214 *S. cerevisiae* comprising: 99 Chinese strains, 52 NZ strains, 60 strains used in the first analysis, and a further three international strains containing these six loci (Table S2). We reconstructed a phylogeny with these 13 loci (Figure S2). Posterior probabilities for all labelled clades shown in Figure S2 were >0.92 indicating adequate resolution, but the posterior probabilities for relationships of individuals within these clades, particularly within the Wine/European/NZ clade, was very poor, likely due to gene tree incongruence. Broadly, our analyses agree with earlier findings, and all eight Chinese lineages previously identified (Wang *et al.*, 2012) were reconstructed. The tree features a large split between strains that have been sampled from non-agricultural environments regardless of sampling location, and those that are closely associated with human activity. The exception to this is the Sake clade which tends to cluster with non-agricultural strains due to a hypothesized secondary domestication event (Fay & Benavides, 2005). The genetic diversity (branch lengths) within the human-associated clades are significantly lower than for the other clades, which, taken with low posterior probabilities, implies incomplete lineage sorting and/or high rates of admixture for human-associated strains.

Population Genetic Approaches

S. cerevisiae is a sexual eukaryote, and along with the reasonable rates of heterozygosity revealed here, previous analyses show that while it tends to inbreed, there is clearly a reasonable amount of outcrossed recombination and gene-flow between sub-populations occurring in the NZ population (Goddard *et al.*, 2010; Gayevskiy & Goddard, 2011; Knight & Goddard, 2015), and the inference of recombination and hybridisation in global studies suggest this may well be the case at larger scales (Liti *et al.*, 2009; Cromie *et al.* 2013; Barbosa *et al.*

2016; Ludlow *et al.* 2016). The degree to which phylogenomic methods are able recover any signal when there is diffuse population structure is not clear: i.e. when some population differentiation is present but with reasonable gene flow and recombination between sub-populations. To enable us to analyse a spectrum of possible population structures, from completely homogenized through to highly structured due to ancient divergences, we use complementary Bayesian-based population genetic methods capable of inferring finer degrees of population structure that account for recombination implemented in STRUCTURE (Pritchard *et al.*, 2000), and the subsequent analyses of ancestry profiles by OBSTRUCT (Gayevskiy *et al.*, 2014). From the 11,059,143 nucleotide positions in the 93 aligned concatenated genomes, any which were uninformative or had missing data were conservatively removed leaving a total 66,316 robustly informative positions for population structure analysis. We employed Bayesian population structure approaches that account for and incorporate recombination (admixture) between strains. These population genetic approaches infer the presence of four populations using the Evanno method (Evanno *et al.*, 2005) as implemented in Structure Harvester (Earl & vonHoldt, 2011). Figure 3 shows the resulting ancestry profiles: each vertical column represents a strain and the colours show the proportion of ancestry of each strain to each of the four inferred populations. Strains that have different degrees of ancestry in different sub-populations are a result of mating and recombination between strains (or their ancestors) originating from different sub-populations. There is a progressive and gradual increase in ancestry to the orange inferred population as one moves from the assumed 'natural' strains on the left to the increasingly 'human-associated' strains on the right. Again, the NZ strains fall together and with the European strains, but with varying degrees of ancestry. It is clear that these various populations are not discrete: there are signals for some gene flow and genetic mixing among the species as a whole.

We went on to analyse the inferred ancestry profiles (Gayevskiy *et al.*, 2014) to determine whether geographic origin or niche of isolation might correlate most strongly with population structure. This analysis shows that variance in genetic structure in the NZ population correlated with niche of isolation (R^2=0.51, P<0.0001) only marginally more than geographic origin (R^2=0.45, P<0.0001). Unsurprisingly, the amount of genetic variance explained (R^2) is greater (R^2=0.74, P<0.0001) when ancestries are compared to those groups revealed from the independent phylogenetic analyses shown in Figure 1, rather than simply geographic origin or niche of isolation. This shows that neither geographic location nor niche/'use' alone is sufficient to describe the observed population structure, and is exactly in line with the recent conclusion of Almeida *et al.* (2015) who examined the global population but included European oak population. Thus, the most accurate picture is one of a global metapopulation of connected sub-populations inhabiting various places and niches, and recapitulates the picture seen at national levels (Knight & Goddard, 2015).

No evidence for hybridization with other *Saccharomyces* species

Barbosa *et al* (2016) recently reported novel *S. cerevisiae* lineages in Brazil that are related to Japanese and North American lineages, but the Brazilian populations also contain signals for mating and introgression with the wine/European group, as well as hybridization and introgression with American *S. paradoxus.* This hybridization and subsequent introgression conceivably provided novel genetic combinations better adapted to inhabit Brazilian native biomes. Ludlow *et al* (2016) reported lineages associated with coffee and cocoa were created by the hybridization of genomes from the European/wine with north American and Chinese

populations. Might the NZ *S. cerevisiae* have undergone a similar process – where hybridization with an endemic *Saccharomyces*, or some other *Saccharomyces cerevisiae* population, provided an opportunity for more effective adaptive radiation in New Zealand? For the *Saccharomyces* species know to be present in NZ, *S. paradoxus* has been inferred to have recently migrated from Europe with oak trees (Zhang et al, 2010). The single representative of *S. eubayanus* is also inferred to have recently arrived from South America (Gayevskiy & Goddard 2016). However, there is evidence to suggest that *S. uvarum* and *S. arboricola* may have more ancient populations in NZ (Almeida *et al.*, 2015; Gayevskiy & Goddard 2016). Following Gayevskiy and Goddard (2016), alignment for all 52 NZ *S. cerevisiae* genomes to reference genomes for the other *Saccharomyces* species show that all align best to *S. cerevisiae*, with an average of 97.2%, and a minimum of 93.2% (Table S3). The two *Saccharomyces* species to which the NZ *S. cerevisiae* align the poorest are the two candidates for potential endemic NZ species, with a maximum of just 43%. Further, there is no evidence for large blocks of NZ *S. cerevisiae* genomes to be more greatly related to any species other that *S. cerevisiae* (Table S3). Together, this provides no evidence for recent hybridization or introgression event in the NZ *S. cerevisiae* group from other species. Thus, given the data available, it appears the NZ *S. cerevisiae* population derive exclusively from the European/wine group. Further, the earlier ancestry profile analysis shows most of the NZ strains have the majority of their ancestry in the wine/European group – i.e. most NZ strains are 'clean' wine/European strains.

Number and Timing of NZ Incursion Events

It is clear that the NZ *S. cerevisiae* population derived from Europe. But how many times might strains have been transferred from one side of the planet to the other, and when did this occur? We define incursion events as transfers to NZ that have become established enough for us to detect strains, or related lineages deriving from such strains, which are thus founder events. The theoretical number of incursion events ranges from just one to approximately 2,000 as this represents the best estimate for the number of different *S. cerevisiae* genotypes currently present in NZ (Knight & Goddard, 2015). First we evaluate whether there is any evidence to support a single founder event, versus multiple incursions given these data. A single founder event would mean that all current NZ *S. cerevisiae* would coalesce to a single ancestor. One signal for this would be the presence of shared fixed alleles in the NZ but not European population. Of the 66,316 SNPs none are fixed in the NZ genomes, providing no strong evidence for a single founder event.

We acknowledge the tentative nature of this analyses given the relatively few strains for which we have sequences, but wished to estimate the likely minimum number of incursion events given our data to provide a lower bound to this rate, and appropriately used a maximum parsimony approach to analyse this. Under this minimal change parsimony framework, the best explanation for clades entirely comprised of NZ derived genomes is that the ancestor of this clade was transported to NZ from Europe. Thus, we modelled the minimum possible incursion events into NZ by minimizing the change from 'Wine/European' to 'NZ' strain status over the phylogeny (Figure 4). The minimum number of transfer events inferred by this analysis is ten (one and nine in each of the two clades where NZ isolates are present). By comparisons to null distributions, this observed number of incursion events is significantly less than we would expect to see by chance (P=0.0116 and P< 0.0001 for each clade given 10,000 permutations of

terminal taxa status) given these phylogenies and proportions of NZ and European derived genomes. As an alternate approach, strains that survived transport to and establishment in NZ might tend to sire independent lineages, and this signal might be revealed by the presence of separate sub-populations in NZ. To test this we analysed the population structure in these genomes using STRUCTURE (Pritchard *et al.*, 2000), and the optimal number of inferred sub-populations is 11 across the two clades that harbour NZ derived genomes. The inferred number of sub-populations is in line with the number of incursion events suggested by parsimony analyses.

It appears the movement of *S. cerevisiae* from Europe to NZ is thus not only detectable but also constrained. The question under scrutiny here is the extent to which humans have expanded microbial species ranges. Just because we infer at least ten incursion events from Europe, this does not necessarily prove that humans were the agents of transfer; *S. cerevisiae* might have been moved by other means and been present before humans arrived. Humans only arrived in NZ about 1,000 years ago, and winemaking only in the last ~200 years (Hurles *et al.*, 2003), so next we attempted to estimate the ages of the NZ clades. Again, under a parsimony framework, these potentially represent the ages of lineages and populations that have expanded since their ancestors arrived in NZ. Given our data, the substitution rate between *S. paradoxus* and *S. cerevisiae* is 0.3366. Dating microbial phylogenies is difficult, and the time to the common ancestor of *S. cerevisiae* and *S. paradoxus* has been estimated between 0.4 and 3.4 mya (Liti *et al.*, 2006). If the molecular substitution rate is assumed to be constant across this time period, then the potential ages of these ten clades may be simply estimated by calculating the proportional distance of the relevant nodes compared to the node defining the *S. paradoxus/S. cerevisiae* split. With this approach the lower bound timing estimates of *S. cerevisiae* incursions

into NZ from Europe spans from approximately 60 to 5,000 years ago (Figure 4). However, we note that large confidence limits around the timing of the *S. paradoxus* and *S. cerevisiae* split (Liti *et al.*, 2006) clearly translate into large limits around the estimates for incursions into NZ. Without wanting to overly extrapolate these tentative timings, it is interesting to note that most inferred incursion events are just above or well below the 1,000 year cut-off, and just one is substantially older. This one 'older' event is the inferred incursion event from the smaller sister clade to the Wine/European group, where eight NZ derived genomes cluster with soil isolates from Europe (Figure 4). Given the uncertain nature of the dating of this phylogeny combined with the assumptions of constant substitution rates, apart from one possible exception, there is no compelling evidence to suggest that *S. cerevisiae* has been in NZ significantly longer than humans have. Thus, human introduction appears the most likely explanation for *S. cerevisiae*'s presence in NZ.

Lastly, we estimated the node containing the entire Wine/European/NZ group to be between 4,635-39,394 years old. This estimate overlaps with the earliest evidence for humans producing fermented drinks some 9,000 years ago in China (McGovern *et al.*, 2004), and this places all the *S. cerevisiae* found in NZ in the group that expanded along with the human passion for viticulture and winemaking.

Discussion

These data reveal that the *S. cerevisiae* inhabiting NZ originated from Europe. We estimate a minimum of ten incursion events founded the NZ population. It appears that this species has been transported to NZ and has become successfully established in vineyards, but that radiation to native forest habitats is rare, but detectable. This may be due to low rates of movement or that this *S. cerevisiae* population is poorly adapted to NZ native forest niches, or both. *S. arboricola* inhabits NZ native forests, but unlike *S. cerevisiae* populations in habiting Brazilian native forests, which hybridized with endemic *S. paradoxus*, there is no evidence that the NZ S. cerevisiae have hybridized with endemic *S. arboricola*. Due to the very recent arrival of *S. cerevisiae* to NZ, perhaps this is occurring currently, or will do soon.

Permutation analyses suggest the rate of incursion into NZ is not rampant. However, analyses estimating the number of incursions might produce an erroneous result not necessarily due to analytical reasons, but mostly due to the unequal number of samples deriving from NZ and European populations. Recall the NZ strains were deliberately chosen to represent the genetic diversity in NZ based on surveys from both vineyard and native forest habitats, but it might well be that with increased sampling in the European wine group one finds strains that interdigitate among NZ strains in various clades. This would elevate the number of estimated transitions from Europe. However, additional data may not show this pattern; either way, here we estimate the minimum number of incursion events given the data available. Under the assumption that the largest NZ clade was founded by a single strain, which has then radiated in NZ, one can compare metrics that might provide insights into the evolution of *S. cerevisiae* since arriving in NZ. The largest NZ clade (defined by node 2 in Figure 4) has values of $\pi = 1909$, $\theta w = 2114$,

and Tajima's D = -0.41. This compares to π = 1155, θw = 1305, and Tajima's D = -0.43 for a set of 10 Wine/European Genomes, and implies possibly greater genetic diversity, but no more compelling signal for selection, in this NZ clade compared to the Wine/European group generally.

The estimates concerning timings of these incursion events are less certain. This is due to the problems associated with dating microbial phylogenies in general, owing to the lack of fossils, and then extrapolating the uncertain estimates we have to relatively recent divergence events. Whilst the mutation rate of *S. cerevisiae* has been estimated (Lang & Murray, 2008), this has been deduced using a few strains under laboratory conditions. A further complication is that we have very little idea of absolute mitotic and meiotic generation times in nature, making calibrations of absolute timings using mutation rates a fruitless way forward (Goddard & Grieg, 2015). Here we make assumptions about the constancy of the rates of molecular evolution. Given these caveats, we estimate the likely timings of these incursions, and their lower bounds are not greatly above, and indeed are mostly below 1,000 years ago (Figure 4). Another possible reason for an inaccurate inference in terms of transfer timings also stems from a lack of sampling: strains might have migrated only very recently to NZ even though their last common ancestor with a European strain occurred thousands of years ago.

Previous analyses, using repeat regions, showed the NZ *S. cerevisiae* population as internationally genetically distinct (Goddard *et al.*, 2010). The analysis of whole genomes here does not agree with this. This discrepancy might be explained by the fact that repeat loci evolve rapidly and thus are capable of resolving finer levels of population differentiation than average signals from whole genomes (or many loci) can. Significant signals for differentiation revealed

by analyses using repeat regions would occur if rates of gene-flow (incursion events) between Europe and NZ are relatively low, and less than the rates of evolution at these repeat regions: analysis here suggests the number of incursion events into NZ have not been that great, and thus correlates with this idea. The sequencing of diploid genomes here allows rates of heterozygosity to be calculated, and these are on average similar to those previously estimated (Magwene *et al.*, 2011); however, the variance in rates of heterozygosity in the NZ genomes is substantial (from 6% to 43% of SNPs). Such rates of heterozygosity may be explained, at least in the NZ group, by the inference that ~20% of mating events are outbred combined with reasonable levels of gene-flow between sub-populations (Knight & Goddard, 2015), and that the European/ wine group more generally have elevated outcrossing rates (Peter & Schacherer, 2016).

Together, these estimates of origins and timings strongly suggest that humans introduced *S. cerevisiae* into NZ recently, and thus expanded the range of this species. This pattern correlates with the previous observation of *S. cerevisiae* presence in new oak barrels from Europe once arrived in NZ (Goddard *et al.*, 2010). The signals provided by these phylogenomic analyses are also in line with work showing trends in *S. cerevisiae* population division that correlates with the expansion of viticulture globally (Legras *et al.*, 2007). This extent of human mediated movement is also consistent with the analyses of cocoa and coffee populations in Africa and Europe, where approximately three significant movements from Europe, North America and Asia to Africa and South America were inferred, and migration of wine strains to Brazil which hybridized with *S. paradoxus*. However, there were no estimations regarding the timing of either the Brazil or cocoa and coffee strain movements (Barbosa *et al.*, 2016; Ludlow *et al.*, 2016). In addition, analysis of a handful of *S. paradoxus* isolates from NZ also infer transfer from Europe to NZ

associated with the movement of another plant species by humans: *Quercus* (oak trees) (Zhang *et al.*, 2010). It is interesting to note that and *S. uvarum* is inferred to have been present in Australasia, and possibly *S. arboricola* in NZ, well before humans might have been, and so it seems that the ranges, modes and ages of dispersal of these sister taxa differ.

The earliest evidence for the human use of *S. cerevisiae* for fermentation has been dated to approximately 9,000 years ago and comes from pottery jars in China (McGovern *et al.*, 2004). The earliest evidence for wine production comes from Iran approximately 7,400 years ago and seeds of domesticated grapes have been found in Georgia and Turkey and dated to approximately 8,000 years ago (This *et al.*, 2006). Winemaking then spread to adjacent areas and around the Mediterranean approximately 5,000-5,500 years ago (This *et al.*, 2006). Our results show strains associated with winemaking are closely related to one another regardless of geographic location. The archaeological dates allow another way to calibrate the dating on this phylogeny and indicate that the split between *S. cerevisiae* and *S. paradoxus* is closer to the lower bound of 0.4 mya than the upper of 3.4 mya.

Overall the analyses conducted here add further support to the concept that humans have facilitated the global transfer of this microbial species through our agricultural activities, and thus have significantly expanded this species' range. In doing so, it appears humans have provided an opportunity for one lineage of *S. cerevisiae* to radiate to and become established in areas well beyond the ancestral range for this species. Not only has the transfer of this species provided an opportunity for it to become established in New Zealand's agricultural ecosystems, but it is now also found in natural forest ecosystems. Whether *S. cerevisiae* has or may become established in NZ native forest ecosystems is debatable as the low rate of recovery may simply

reflect rare transposition events, by humans or insects perhaps (Buser *et al.* 2014), that will perhaps ultimately fail to seed successful lineages in that inhabit. Indeed, we do not have a good understanding of the niches to which *S. cerevisiae* is adapted, if any at all (Goddard & Grieg 2015). Alternatively, it is possible that *S. cerevisiae* may become established in native habitats. New Zealand has a list of human introduced invasive species that has decimated portions of endemic ecosystems – stoats and rats destroying native NZ birds as a prime example (Norton 2009). The interesting question is whether *S. cerevisiae* is classed as an invasive species in NZ: while this species has been introduced by humans, at the moment its invasion is primarily restricted to agricultural ecosystems, where it arguably adds value.

Funding

This work was supported by University of Auckland FRDF (grant 3700513), Plant and Food Research Ltd., the Ministry of Business, Innovation and Employment, and NZ Winegrowers grants to MRG.

Acknowledgements

The completion of this research would not have been possible without the support and assistance of the many collaborating companies who allowed access to their land and donated juice: Amisfield, Ata Rangi, Coal Pit, Constellation, Delegats, Domain Road, Frey Vineyard, Misha's Vineyard, Mt Difficulty, Mt Riley, Neudorf, Palliser, Pernod Ricard, Rippon, Seifried, Te Kairanga, Tohu, Trinity Hill, Villa Maria and Vita Brevis.

Bibliography

Aa E, Townsend JP, Adams RI, Nielsen KM & Taylor JW (2006) Population structure and gene evolution in *Saccharomyces cerevisiae*. *Fems Yeast Res* **6**: 702–715.

Almeida P, Barbosa R, Zalar P, *et al.* (2015) A population genomics insight into the Mediterranean origins of wine yeast domestication. *Molecular Ecology* **24**: 5412–5427.

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

Andrews S (2012) FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Avise J, Arnold J, Ball R, Bermingham E, Lamb T, Neigel J, Reeb C & Saunders N (1987) Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review in Ecology and Systematics* **18**: 489–522.

Barbosa R, Almeida P, Safar SVB, *et al.* (2016) Evidence of Natural Hybridization in Brazilian Wild Lineages of *Saccharomyces cerevisiae*. Genome Biol. Evol. 2016; 8:317–329.

Beheregaray LB (2008) Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology* **17**: 3754–3774.

Bergstrom A, Simpson JT, Salinas F, *et al.* (2014) A high-definition view of functional genetic variation from natural yeast genomes. *Molecular Biology and Evolution* **31**: 872–888.

Biomatters Ltd. (2012) Geneious.

Buser CC, Newcomb RD, Gaskett AC & Goddard MR (2014) Niche construction initiates the evolution of mutualistic interactions. *Ecol Letters* **17**: 1257–1264.

Cherry JM, Hong EL, Amundsen C, *et al.* (2011) *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research* **40**: D700–D705.

Cromie GA, Hyma KE, Ludlow CL, Garmendia-Torres C, Gilbert TL, May P, Huang AA, Dudley AM & Fay JC (2013) Genomic sequence diversity and population structure of *Saccharomyces cerevisiae* assessed by RAD-seq. *G3: Genes, Genomes, Genetics* **3**: 2163–2171.

Darling ACE, Mau B, Blattner FR & Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14**: 1394–1403.

de Wit R & Bouvier T (2006) “Everything is everywhere, but, the environment selects;” what did Baas Becking and Beijerinck really say? *Environmental Microbiology* **8**: 755–758.

del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R & Ruiz-Trillo I (2014) The others: our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution* **29**: 252–259.

Delsuc F, Brinkmann H & Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**: 361–375.

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* **418**: 700–707.

Drummond AJ & Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214–218.

Drummond AJ, Ho SYW, Phillips MJ & Rambaut A (2006) Relaxed Phylogenetics and Dating with Confidence Penny D, ed. *PLoS Biol* **4**: e88.

Earl DA & vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet Resour* **4**: 359–361.

Evanno G, Regnaut S & GOUDET J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**: 2611–2620.

Fay JC & Benavides JA (2005) Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PLoS Genet* **1**: 66–71.

Gayevskiy V & Goddard MR (2011) Geographic delineations of yeast communities and populations associated with vines and wines in New Zealand. *ISME J* **6**: 1281–1290.

Gayevskiy V, Klaere S, Knight S & Goddard MR (2014) ObStruct: A Method to Objectively Analyse Factors Driving Population Structure Using Bayesian Ancestry Profiles Pajewski NM, ed. *PLoS ONE* **9**: e85196.

Gayevskiy V & Goddard MR (2016) *Saccharomyces eubayanus* and *Saccharomyces arboricola* reside in North Island native New Zealand forests. *Environmental Microbiology* **18**: 1137–1147

Goddard MR, Anfang N, Tang R, Gardner RC & Jun C (2010) A distinct population of *Saccharomyces cerevisiae* in New Zealand: evidence for local dispersal by insects and human-aided global dispersal in oak barrels. *Environmental Microbiology* **12**: 63–73.

Goddard MR & Greig D. *Saccharomyces cerevisiae*: a nomadic yeast with no niche? FEMS Yeast Research,2015; 15: fov009

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., et al. (1996) Life with 6000 Genes. *Science* **274**: 546–567.

Guindon S & Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**: 696–704.

Hanson CA, Fuhrman JA, Horner-Devine MC & Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology* **10**: 497–506.

Hittinger CT (2013) Saccharomyces diversity and evolution: a budding model genus. **29**: 309–317.

Huang C, Roncoroni M & Gardner RC (2014) MET2 affects production of hydrogen sulfide during wine

fermentation. *Applied Microbiology and Biotechnology* **98**: 7125–7135.

Hurles ME, Matisoo-Smith E & Gray RD (2003) Untangling Oceanic settlement: the edge of the knowable. *Trends in Ecology & Evolution* **18**: 531–540.

Huson DH & Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**: 254–267.

Knight S & Goddard MR (2015) Quantifying separation and similarity in a Saccharomyces cerevisiae metapopulation. *ISME J* **9**: 361–370.

Knight S, Klaere S, Fedrizzi B & Goddard MR (2015) Regional microbial signatures positively correlate with differential wine phenotypes: evidence for a microbial aspect to terroir. *Scientific Reports* **5**:14233

Lang GI & Murray AW (2008) Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics* **178**: 67–82.

Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**: 357–359.

Larkin MA, Blackshields G, Brown NP, *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Legras J-L, Merdinoglu D, Cornuet J-M & Karst F (2007) Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Molecular Ecology* **16**: 2091–2102.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R & Subgroup 1GPDP (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Litchman E (2010) Invisible invaders: non-pathogenic invasive microbes in aquatic and terrestrial ecosystems. *Ecol Letters* **13**: 1560–1572.

Liti G, Barton DBH & Louis EJ (2006) Sequence Diversity, Reproductive Isolation and Species Concepts in *Saccharomyces*. *Genetics* **174**: 839–850.

Liti G, Carter DM, Moses AM, *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M & Usadel B (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research* **40**: W622–W627.

Ludlow CL, Cromie GA, Garmendia-Torres C, Sirr A, Hays M, Field C, Jeffery EW, Fay JC & Dudley AM (2016) Independent Origins of Yeast Associated with Coffee and Cacao Fermentation. *Curr Biol* **26**: 965–971.

Maddison WP & Maddison DR (2014) Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org.

Magwene PM, Kayikci O, Granek JA, Reininga JM, Scholl Z & Murray D (2011) Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*.

PNAS **108**: 1987–1992.

Martiny JBH, Bohannan BJM, Brown JH, *et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology* **4**: 102–112.

Mazzaglia A, Studholme DJ, Taratufolo MC, Cai R, Almeida NF, Goodman T, Guttman DS, Vinatzer BA & Balestra GM (2012) *Pseudomonas syringae* pv. actinidiae (PSA) isolates from recent bacterial canker of kiwifruit outbreaks belong to the same genetic lineage. *PLoS ONE* **7**: e36518.

McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, Nuñez A, Butrym ED, Richards MP & Wang C (2004) Fermented beverages of pre-and proto-historic China. *PNAS* **101**: 17593–17598.

Morrison-Whittle P & Goddard MR (2015) Quantifying the relative roles of selective and neutral processes in defining eukaryotic microbial communities. *ISME J* **9**: 1–9.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y & Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE* **7**: e37558.

Norton, D (2009) Species Invasions and the Limits to Restoration: Learning from the New Zealand Experience. *Science* **325**:569-571

Novo M, Bigey F, Beyne E, *et al.* (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proceedings of the National Academy of Sciences* **106**: 16333–16338.

Pedrós-Alió C (2006) Marine microbial diversity: can it be determined? *Trends in Microbiology* **14**: 257–263.

Peter J & Schacherer J (2016) Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast* **33**: 73–81.

Piskur J, Rozpedowska E, Polakova S, Merico A & Compagno C (2006) How did Saccharomyces evolve to become a good brewer? *Trends Genet* **22**: 183–186.

Pretorius IS (2000) Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. *Yeast* **16**: 675–729.

Pritchard JK, Stephens M & Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/.

Richards KD, Goddard MR & Gardner RC (2009) A database of microsatellite genotypes for Saccharomyces cerevisiae. *Antonie van Leeuwenhoek* **96**: 355–359.

Rokas A, Williams BL, King N & Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.

Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M & Hittinger CT (2011) The awesome power of yeast evolutionary genetics: new genome sequences and strain

resources for the *Saccharomyces* sensu stricto genus. *G3: Genes, Genomes, Genetics* **1**: 11–25.

Schacherer J, Shapiro JA, Ruderfer DM & Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* **458**: 342–345.

Serjeant K, Tang R, Anfang N, Beggs J R, Goddard M (2008) Yeasts associated with the New Zealand *Nothofagus* honeydew system. *New Zealand Journal of Ecology* **32**: 209-213.

Skelly DA, Merrihew GE, Riffle M, *et al.* (2013) Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Research* **23**: 1496–1504.

Stefanini I, Dapporto L, Legras J-L, *et al.* (2012) Role of social wasps in Saccharomyces cerevisiae ecology and evolution. *Proceedings of the National Academy of Sciences* **109**: 13398–13403.

Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, Dietrich FS & McCusker JH (2015) The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Research* **25**: 762–774.

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

Talbot JM, Bruns TD, Taylor JW, *et al.* (2014) Endemism and functional convergence across the North American soil mycobiome. *PNAS* **111**: 6341–6346.

Taylor MW, Tsai P, Anfang N, Ross HA & Goddard MR (2014) Pyrosequencing reveals regional differences in fruit-associated fungal communities. *Environmental Microbiology* **16**: 2848–2858.

This P, Lacombe T & Thomas MR (2006) Historical origins and genetic diversity of wine grapes. *Trends Genet* **22**: 511–519.

Tripathi BM, Lee-Cruz L, Kim M, Singh D, Go R, Shukor NAA, Husni MHA, Chun J & Adams JM (2014) Spatial scaling effects on soil bacterial communities in Malaysian tropical forests. *Microb Ecol* **68**: 247–258.

Valverde A, Tuffin M & Cowan DA (2012) Biogeography of bacterial communities in hot springs: a focus on the actinobacteria. *Extremophiles* **16**: 669–679.

Wallis GP & Trewick SA (2009) New Zealand phylogeography: evolution on a small continent. *Molecular Ecology* **18**: 3548–3580.

Wang Q-M, Liu W-Q, Liti G, Wang S-A & Bai F-Y (2012) Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Molecular Ecology* **21**: 5404–5417.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 256–276.

Wichmann MC, Alexander MJ, Soons MB, Galsworthy S, Dunne L, Gould R, Fairfax C, Niggemann M, Hails RS & Bullock JM (2009) Human-mediated dispersal of seeds over long distances. *Proceedings of the Royal Society B: Biological Sciences* **276**: 523–532.

Zhang H, Skelton A, Gardner RC & Goddard MR (2010) *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* reside on oak trees in New Zealand: evidence for migration from Europe and interspecies hybrids. *Fems Yeast Res* **10**: 941–947.

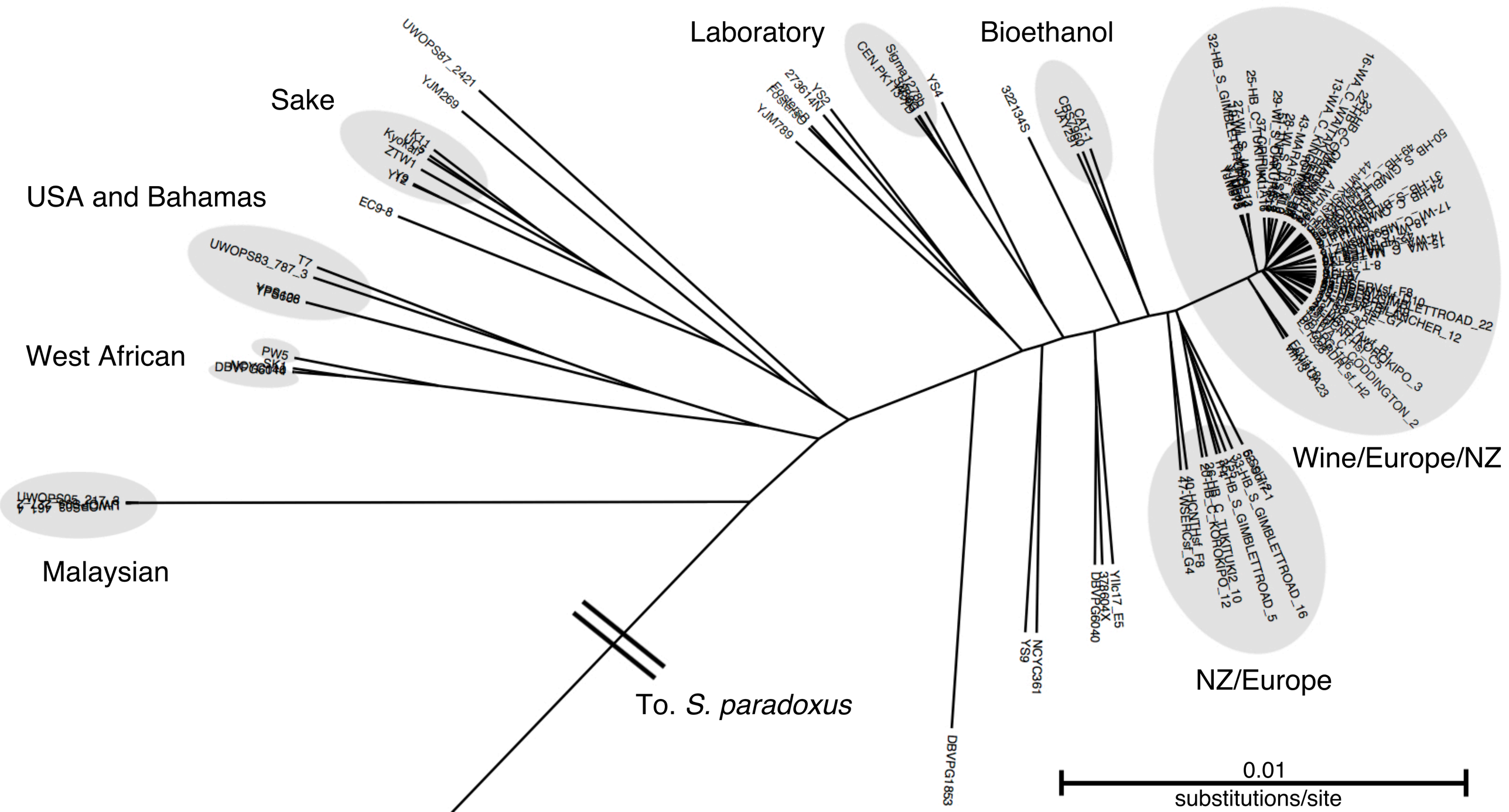
840

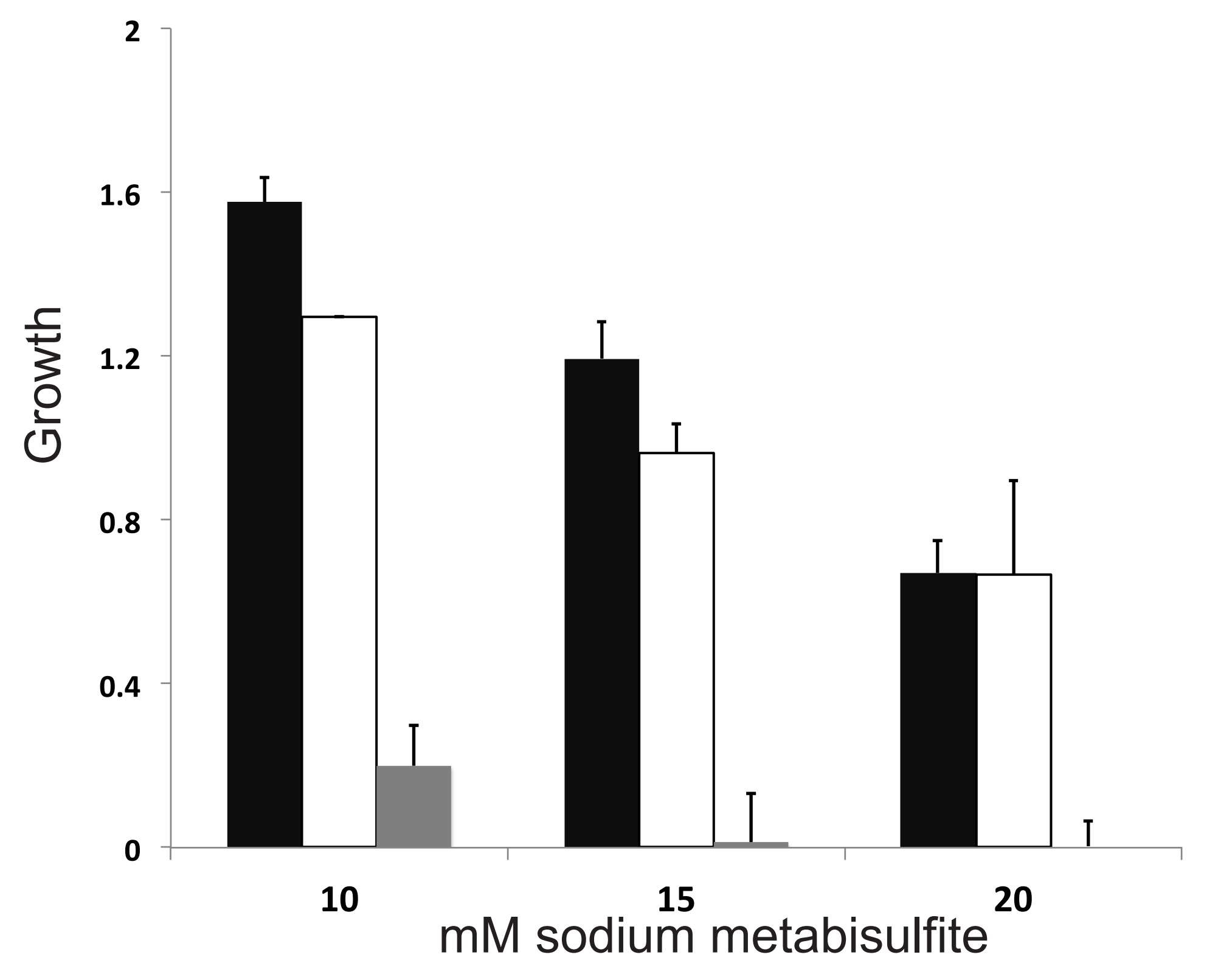
Habitat	Samples	Colonies analysed	Samples yielding Sc	Number of Genotypes	Recovery rate (%)
Exotic oak	190	1140	2	1	1.05
Native forest	523	7522	5	5	0.96
Vineyard	360	10833	39	62	10.83
Ferment	160	11590	159	656	99.38
Total	**1233**	**31085**	**205**	**724**	

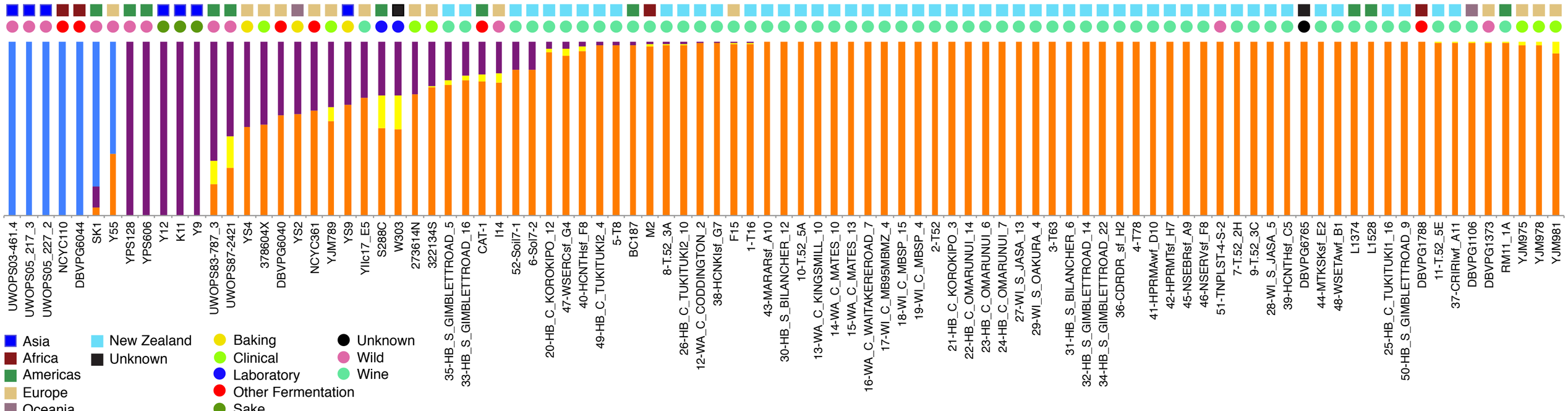
Table 1. The collated sampling effort and recovery of *Saccharomyces cerevisiae* (Sc) from six recent studies in New Zealand (Serjeant *et al*., 2008; Goddard *et al.,* 2010; Zhang *et al*., 2010; Gayevskiy & Goddard 2012; Knight & Goddard 2015; Gayevskiy & Goddard 2016). Native forest samples derived from soil, honeydew and fruits of native trees (Serjeant *et al*., 2008; Knight & Goddard 2015; Gayevskiy & Goddard 2016); vineyard samples derived from soil, fruit and flowers (Goddard *et al.,* 2010; Gayevskiy & Goddard 2012; Knight & Goddard 2015); ferment samples derived from spontaneous ferments of Sauvignon blanc, Chardonnay and Syrah (Goddard *et al.,* 2010; Gayevskiy & Goddard 2012; Knight & Goddard 2015).

Population	Sites[a]	Segregating Sites[b]	θ (x 1,000)	π (x 1,000)	Tajima's D
NZ (52)	11,084,457	124,566 (1.1%)	3.3	2.3	-1.27
Wine/Europe/NZ (66)	6,505,396	91,543 (1.4%)	3.8	2.3	-1.65
All Available (75)	6,397,673	134,386 (2.1%)	5.2	3	-1.77

Table 2: Nucleotide diversity in three superset populations of *S. cerevisiae*. [a] Sites where all genomes have at least one high quality read; [b] Sites where all genomes have at least one high quality read and at least one genome differs from the rest, percentage in brackets is from the total number of sites.


Laboratory
Bioethanol
Sake
USA and Bahamas
West African
Malaysian
Wine/Europe/NZ
NZ/Europe
To. S. paradoxus
UWOPS87_2421
YJM269
EC9-8
T7
PW5
DBVPG6044
DBVPG1853
NCYC361
YS9
0.01
substitutions/site


2
1.6
1.2
0.8
0.4
0
Growth
10
15
20
mM sodium metabisulfite


UWOPS03-461.4
UWOPS05_217_3
UWOPS05_227_2
NCYC110
DBVPG6044
SK1
Y55
YPS128
YPS606
Y12
K11
Y9
UWOPS83-787_3
UWOPS87-2421
YS4
378604X
DBVPG6040
YS2
NCYC361
YJM789
YS9
YIIc17_E5
S288C
W303
273614N
322134S
35-HB_S_GIMBLETTROAD_5
33-HB_S_GIMBLETTROAD_16
CAT-1
I14
52-Soil7-1
6-Sol7-2
20-HB_C_KOROKIPO_12
47-WSERCsf_G4
40-HCNTHsf_F8
49-HB_C_TUKITUKI2_4
5-T8
BC187
M2
8-T.52_3A
26-HB_C_TUKITUKI2_10
12-WA_C_CODDINGTON_2
38-HCNKIsf_G7
F15
1-T16
43-MARARsf_A10
30-HB_S_BILANCHER_12
10-T.52_5A
13-WA_C_KINGSMILL_10
14-WA_C_MATES_10
15-WA_C_MATES_13
16-WA_C_WAITAKERERoAD_7
17-WI_C_MB95MBMZ_4
18-WI_C_MBSP_15
19-WI_C_MBSP_4
2-T52
21-HB_C_KOROKIPO_3
22-HB_C_OMARUNUI_14
23-HB_C_OMARUNUI_6
24-HB_C_OMARUNUI_7
27-WI_S_JASA_13
29-WI_S_OAKURA_4
3-T63
31-HB_S_BILANCHER_6
32-HB_S_GIMBLETTROAD_14
34-HB_S_GIMBLETTROAD_22
36-CDRDR_sf_H2
4-T78
41-HPRMAwf_D10
42-HPRMTsf_H7
45-NSEBRsf_A9
46-NSERVsf_F8
51-TNPLST-4-S-2
7-T.52_2H
9-T.52_3C
28-WI_S_JASA_5
39-HCNTHsf_C5
DBVPG6765
44-MTKSKsf_E2
48-WSETAwf_B1
L1374
L1528
25-HB_C_TUKITUKI1_16
50-HB_S_GIMBLETTROAD_9
DBVPG1788
11-T.52_5E
37-CRIRIwf_A11
DBVPG1106
DBVPG1373
RM11_1A
YJM975
YJM978
YJM981
Asia
Africa
Americas
Europe
Oceania
New Zealand
Unknown
Baking
Clinical
Laboratory
Other Fermentation
Sake
Unknown
Wild
Wine

(a)

S288C
Y55
I14
35-HB_S_GIMBLETTROAD_5
33-HB_S_GIMBLETTROAD_16
1
40-HCNTHsf_F8
47-WSERCsf_G4
20-HB_C_KOROKIPO_12
26-HB_C_TUKITUKI2_10
52-Soil7-1
6-Soil7-2
9.0E-4

(b)

S288C
BC187
RM11_1A
18-WI_C_MBSP_15
17-WI_C_MB95MBMZ_4
28-WI_S_JASA_5
37-CRIRIwf_A11
25-HB_C_TUKITUKI1_16
22-HB_C_OMARUNUI_14
23-HB_C_OMARUNUI_6
2
2-T52
31-HB_S_BILANCHER_6
46-NSERVsf_F8
10-T.52_5A
9-T.52_3C
3-T63
13-WA_C_KINGSMILL_10
7-T.52_2H
16-WA_C_WAITAKEREROAD_7
21-HB_C_KOROKIPO_3
48-WSETAwf_B1
39-HCNTHsf_C5
1-T16
11-T.52_5E
38-HCNKIsf_G7
41-HPRMAwf_D10
34-HB_S_GIMBLETTROAD_22
24-HB_C_OMARUNUI_7
45-NSEBRsf_A9
30-HB_S_BILANCHER_12
Vin13
EC1118
LalvinQA23
YJM981
YJM978
YJM975
DBVPG1106
L_1528
L_1374
6
27-WI_S_JASA_13
32-HB_S_GIMBLETTROAD_14
8
19-WI_C_MBSP_4
36-CDRDR_sf_H2
VL3
3
43-MARARsf_A10
9
8-T.52_3A
F15
51-TNPLST-4-S-2
29-WI_S_OAKURA_4
10
M2
AWRI796
7
5-T8
12-WA_C_CODDINGTON_2
4
44-MTKSKsf_E2
42-HPRMTsf_H7
4-T78
14-WA_C_MATES_10
15-WA_C_MATES_13
DBVPG1373
DBVPG1788
DBVPG6765
50-HB_S_GIMBLETTROAD_9
49-HB_C_TUKITUKI2_4
5
8.0E-4

(c)