Supplemental Information

Extensive sampling of *Saccharomyces cerevisiae* in Taiwan reveals ecology and evolution of pre-domesticated lineages

Tracy Jiaye Lee, Yu-Ching Liu, Wei-An Liu, Yu-Fei Lin, Hsin-Han Lee, Huei-Mien Ke, Jen-Pan Huang, Mei-Yeh Jade Lu, Chia-Lun Hsieh, Kuo-Fang Chung, Gianni Liti and Isheng Jason Tsai

Correspondence: Isheng Jason Tsai ijtsai@sinica.edu.tw

Supplemental Methods

Sampling strategies

In 2016, we began sampling by collecting barks from Fagaceae trees around Academia Sinica campus (Taipei City, Taiwan) using four different media that were previously reported to be successful in the selective enrichment of *Saccharomyces* yeast species(Sniegowski et al. 2002; Wang et al. 2012; Hyma and Fay 2013). It was not until July 2017, when we had our first successful isolation, that we started collecting samples from various substrates (including leaves, bark, litter and soil at the base of the tree) of amenity *Quercus glauca* trees near the National Taiwan University (NTU; Taipei City, Taiwan) campus. We noticed that different substrates yielded different levels of success, but the overall isolation success per tree host increased. Preliminary repeated sampling and isolation also showed that two out of four enrichment media (Sniegowski et al 2002) ,Hyma and Fay's high sugar (Hyma and Fay 2013) gave higher isolation success. We initially sequenced 23 isolates on the same trees and found that these isolates were primarily clones of each other, though genetically differentiated isolates could still be identified (**Supplemental Table S10**). To ensure effective sampling efforts and avoid redundancies in sequencing, we decided to i) sample various substrates in a given tree host, ii) used two enrichment media and iii) sequence only one isolate if multiple isolates were recovered from same substrate from a tree site. Following this decision, we began to expand the sampling sites to look for natural isolates.

We first made a trip around Taiwan in December 2017 each collecting samples from two to five trees from different regions of Taiwan. These trees were mostly but not entirely belonged to the *Fagaceae* family. When an isolate was successfully recovered from a site, we would revisit that region to sample adjacent trees at least tens of meters apart, if the environment permitted. The whole cycle of round trip to new sites, isolation and revisiting was repeated several times until October 2020, when we had recovered what we believed the be enough isolates covering enough sites to yield statistically significant results. Species identification was initially carried out by multiplex PCR to detect *Saccharomyces cerevisiae* species-specific amplicons as described by ref(Muir et al. 2011), but was subsequently modified to add step 6-8 in ref(Liti et al. 2017) detailed in Methods. For species identification, single colonies were picked out and lysed in QuickExtract DNA solution (epicentre). The ITS1F (5-

CTTGGTCATTTAGAGGAAGTAA -3') and ITS4 (5'-TCCTCCGCTTATTGATATGC-3') primer pair were used. The PCR cocktail consisted of 1 µl Colony lysate, 1 ul each of the primers (5 µmol), 12.5 µl KAPA Taq Ready Mix PCR Kit (KK1024, Kapa Biosystems, USA), and 9.5 µl double-distilled water. The following thermocycling conditions were used: an initial 3 min at 95°C, followed by 30 cycles at 95°C for 30s, 52°C for 30s,72°C for 1min, and a final cycle of 5 min at 72°C, cooling at 16°C.

Determining ploidy level of *S. cerevisiae* isolates

Frozen yeast stocks were first streaked out on YPD plates and kept in 30℃ until colonies became visible. For each isolate, a single colony was picked out and grown in liquid YPD overnight at 30℃ with agitation (200rpm). Cultures were diluted in YPD to obtain OD_{600} = 0.2, and incubated under the same condition for an additional 5-6 hours until OD_{600} reached around 1. Each cell culture was diluted again with YPD to achieve OD_{600} = 0.6, and 500µl of the liquid culture was transferred to a microcentrifuge tube. Cells were harvested with a benchtop centrifuge at 500 *x g* for 3 minutes at room temperature, then pellets were subsequently washed with 500µl sterilized ddH_2O. Cell pellets were collected again with centrifugation, resuspended in 1ml cold 70% ethanol and left to fix at 4℃ overnight. Cells were collected with centrifugation and resuspended in 50µl sodium citrate (50mM) with vortexing at max speed for 10 seconds, centrifuged again and resuspended in 200 µl sodium citrate (50mM) with a final concentration of 0.5mg/ml RNaseA (prepared in 40% glycerol). RNaseA treatment continued for 2 hours at 37℃. Cells were transferred into dark brown microcentrifuge tubes for subsequent staining. A final concentration of 25 µg/ml of propidium iodide was added to each sample, then samples were incubated at 37℃ overnight in the dark. Samples were vortexed at max speed for 10 seconds, then 25 µg/ml propidium iodide (in 50 mM sodium citrate) was added to create a 1:40 dilution of cells. Each diluted sample was then passed through a sterilized 30 µm pre-separation filter to remove any remaining large cell clumps. Cell cycles were recorded on a Beckman Coulter CytoFLEX S Cell Analyzer with a 610 nm filter. More than 20,000 events were recorded for each isolate, and standard samples with known DNA content were used as baseline to compare YL610-A(area) values at 2C and 4C. Ratios of mean YL610-A ranged between 1.78-2.24 for diploid isolates and 2.93 for one triploid isolate found in our collection.

Collection of additional geographical data

The last-level administrative divisions in Taiwan were retrieved from the GADM database (https://gadm.org/) and coordinates were grepped from Google Maps with R package RgoogleMaps(Loecher and Ropkins 2015) (v1.4.5.3).

DNA extraction from environmental samples

Bark 1.0 g of each bark sample was powdered using pestle and mortar with liquid nitrogen. 10 mL of lysis buffer (100mM Tris.Cl pH8, 1.4M NaCl, 2.0% w/v CTAB, 20mM EDTA and 1.0% w/v PVP)(Lee and Taylor 1990) was added to the powdered samples, which were then incubated for 1 hr at 65°C. Nucleic acids were extracted with 10mL of chloroform:isoamyl alcohol (24:1) mixed by inversion using a benchtop rotator at 30 rpm for 10min. Lysate mixture was centrifugated at 10,000 *x rcf* for 30 min at 25°C. The aqueous phase was transferred to a new tube, and nucleic acid precipitated with an equal volume of isopropanol. Precipitated nucleic acid was centrifugated at 10,000 *x rcf* for 10 min at 25°C. Supernatant was decanted, and the pellet air-dried for 5 min at room temperature. DNA was resuspended in 150 µL ddH_2O. Bark DNA was cleaned further using an equal volume of AMPure XP beads (Beckman Coulter, ID: A63881), per the manufacturer's protocol.

Leaf, twig and litter Due to variations in the field, collection sample size and surface were varied among the different tree families. Sample weights were not standardized to minimize handling and contamination. For each leaf, twig and litter sample, the quantity processed were weighed and transferred into a sterile 500 mL polypropylene centrifugation bottle (Beckman Coulter; Cat. 361691) using sterile tweezers. Samples were suspended in 250 mL of 1X PBS pH 7.4 with 0.1% Tween 20 and placed in a sonicating water bath (DELTA ULTRASONIC CO. LTD, Make: DELTA, Model: DC400) at a 40 kHz frequency for 20 min at 25°C as recommended by ref(Sare et al. 2020). Sonicated samples were transferred onto a horizontal shaker (Double Eagle Enterprise Co., Ltd, Make: TKS, Model: OSI 500) set to 120 rpm for 1 hour at 25°C. Large debris was removed by passing the suspension through a 0.25 mm sterile mesh. Flow-through was centrifugated at 10,000 x *rcf* for 1 hour at room temperature. The supernatant was filtrated using a 0.22µm PES membrane filtration cup (Jet Bio-Filtration Co., ID: FPE214250). Pellets were resuspended using 1 X PBS with 0.1%

Tween 20 and added to the filtration cup towards the end of filtration to minimise blockage. Filter membranes were excised using a scalpel, and total nucleic acid was extracted using a DNeasy PowerWater kit (QIAGEN; ID: 14900), per the manufacturer's instructions.

Soil. Soil samples were sieved through 2mm sterilized stainless steel mesh to remove visible rocks, insects, and plant materials. Sieved soils were homogenized by mixing using a sterile spatula. Total nucleic acid was extracted from 0.3g of processed soil samples using a DNeasy PowerSoil kit (QIANGEN, Cat. 12888) with minor modifications to the manufacturer's protocol. The incubation period after the addition of Solutions C2 and C3 was extended to one hour.

TreeMix analyses

We inferred the relationship between *S. cerevisiae* lineages using TreeMix(Pickrell and Pritchard 2012) (v.1.13). Lineages were defined by different ADMIXTURE genetic compositions at different K values (16 or 29). At ADMIXTURE K=16, the following criteria were used: i) isolates with >97.5% genetic ancestry from one single group was designated to that group, ii) CHN-VIII with genetic component from CHN-VI/VII.2 and Wine/European was designated as a group, iii) TW3 with CHN-VI/VII.2 and CHN_X/Malaysian component was designated as a group, iv) TW6 with genetic component from African beer, Wine/European and Qingkejiu/Sake was designated as a group, v) five isolates recovered from steamed buns (Mantou) containing genetic component from African beer and Qingkejiu/Sake, and vi) PD38A was designated as a group. At ADMIXTURE K=29, the following criteria were used: i) isolates with >97.5% genetic ancestry from one single group was designated to that group, ii) TW6 isolates were designated as a group, and iii) PD38A was designated as a group. To ensure the independence of the sites, PLINK(Chang et al. 2015) (v1.90b6.20) was first run with options --indep-pairwise 50 10 0.5 and the block size in TreeMix was set to 100. TreeMix was run from one to 10 migration events with TW1/CHN-IX was used as the outgroup of the phylogeny. Five independent replicates were run on each edge to assess the consistency of the inferences. When the number of migration events in the phylogeny was increased from one to 10, the variance explained in the model increased by up to 99.6% at K=16 (**Supplemental Fig S6a**). We inferred seven migration edges as the most likely model (**Figure 3a**), as it explained 99.3% of the

variance calculated using an *ad hoc* statistic based on change in the log likelihood between models of incrementing edges (**Supplemental Fig S6b**). Based on the K=29 designated grouping, we inferred eight migration edges to be the most likely model (**Supplemental Fig S8**) explaining 98.9% of the variance (**Supplemental Fig S9**).

Estimate divergence of *S. cerevisiae* lineages

To estimate divergence of *S. cerevisiae* lineages, we used two approaches: the first from a phylogeny calibrated using known molecular divergence and second with pairwise divergence between representative isolates.

To infer the *Saccharomyces cerevisiae* lineage phylogeny, amino acid, nucleotide sequences and annotation of proteomes from the following species in the *Saccharomyces sensu stricto* clade were downloaded: *S. eubayanus* FM1318 (GenBank accession GCA_001298625), *S. uvarum* CBS7001 (GenBank accession GCA_000166995.1), *S. kudriavzevii* NBRC 1802 and ZP 591 from (Macías et al. 2019), *S. jurei* from (Naseeb et al. 2018), *S. arboricolus* H6, *S. paradoxus* CBS432, N44, UFRJ50816, UWOPS91-917.1 and YPS138 from https://yjx1217.github.io/Yeast_PacBio_2016/data/. Orthology of proteomes from these outgroups and 45 *S. cerevisiaie* isolates was inferred considering synteny information using PoFF (v. 6.0.27) (Lechner et al. 2014). The protein alignment was constructed for each of the 1,594 single copy ortholog groups using MAFFT (Katoh et al. 2005), then back-translated into a codon sequence alignment using PAL2NAL (v.14) (Suyama et al. 2006). A maximum likelihood phylogeny was constructed with each of the 1,594 single copy ortholog protein alignments using RAxML-ng (v1.0.0, option --model LG+I+F+G4 --tree pars 10). The phylogeny and bootstrap support replicates were used together to infer a lineage phylogeny using ASTRAL-III (Zhang et al. 2018) (v5.6.3). A separate maximum likelihood phylogeny was built using RAxML-ng (v1.0.0) with the concatenated alignment of the single copy orthologs. The topology of both phylogenies was consistent and was used with the concatenated codon alignment of single copy orthologs as input for the MCMCtree method in the PAML (Yang 2007) package to estimate the divergence time among the *S. cerevisiae* lineages. The overall substitution rate was estimated from PAML (Yang 2007) based on the concatenated nucleotide alignment. The following molecular divergence estimates from reference (Shen et al. 2018) were used to calibrate the phylogeny: *S.*

cerevisiae-S. paradoxus 4–5.81 million years ago (Ma), *S. cerevisiae-S. mikatae* 6.97–9.47 Ma, *S. kudriavzevii-S. mikatae* 10.1–13 Ma, *S. arboricola-S. kudriavzevii* 11.7–14.8 Ma and *S. eubayanus-S. uvarum* 4.93–7.93 Ma.

We also calculated pairwise synonymous differences between strains which represented the average divergence time across the genome. A rate of divergence was translated to time using the *S. cerevisiae–S. paradoxus* 4–5.81 million years ago (Ma) from (Shen et al. 2018). An example of translation between *S. cerevisiae* and *S. paradoxus* lower bound divergence was 530,894 synonymous differences / 2,820,348 synonymous sites / 4.002 Ma = 0.047 changes per synonymous site per Ma. Using this number, lower bound of TW1–CHN-IX divergence was calculated as 3,620 synonymous differences / 2,820,348 synonymous sites / 0.047 changes per synonymous site per Ma = 0.027 Ma. The result of the calculations is shown in **Supplementary Table 9.**

Frequency of sex and selection tests

Diversity estimates for 16 nuclear chromosomes and corresponding coding/noncoding regions were examined by VariScan (Vilella et al. 2005) with RunMode 11 (n<4) and 12 (n>=4), and custom python scripts. The Watterson estimator θs from synonymous sites were also calculated. The recombination rate ρ (unit as Morgans per chromosome) for each chromosome of different designated lineages was estimated by rhomap (Auton and McVean 2007) as part of the LDhat program (v2.2) with 10,000 iteration and samples taken every 100 iterations. Inbreeding coefficients F was determined for each isolate by PLINK (Chang et al. 2015) (v.1.90b4; --ibc) on LD-trimmed SNP matrix. The per generation recombination rate (r with unit as centiMorgans per kb) was obtained from the *Saccharomyces* Genome Database (https://wiki.yeastgenome.org/index.php/Combined_Physical_and_Genetic_Maps_of_S._cerevisiae). All units were converted to per base pair prior to the calculations. The per chromosome diversities are shown in **Supplemental Table S12**.

The ratio of asexual to sexual divisions was estimated following the methods of (Tsai et al. 2008) to calculate effective population size N from mutational and recombinational diversity θ and ρ, respectively Using the relationship $N_\rho=\rho/4r(1-F)$ and $N_\theta=\theta(1+F)/4\mu$ (where θ was θs from synonymous sites and per generation

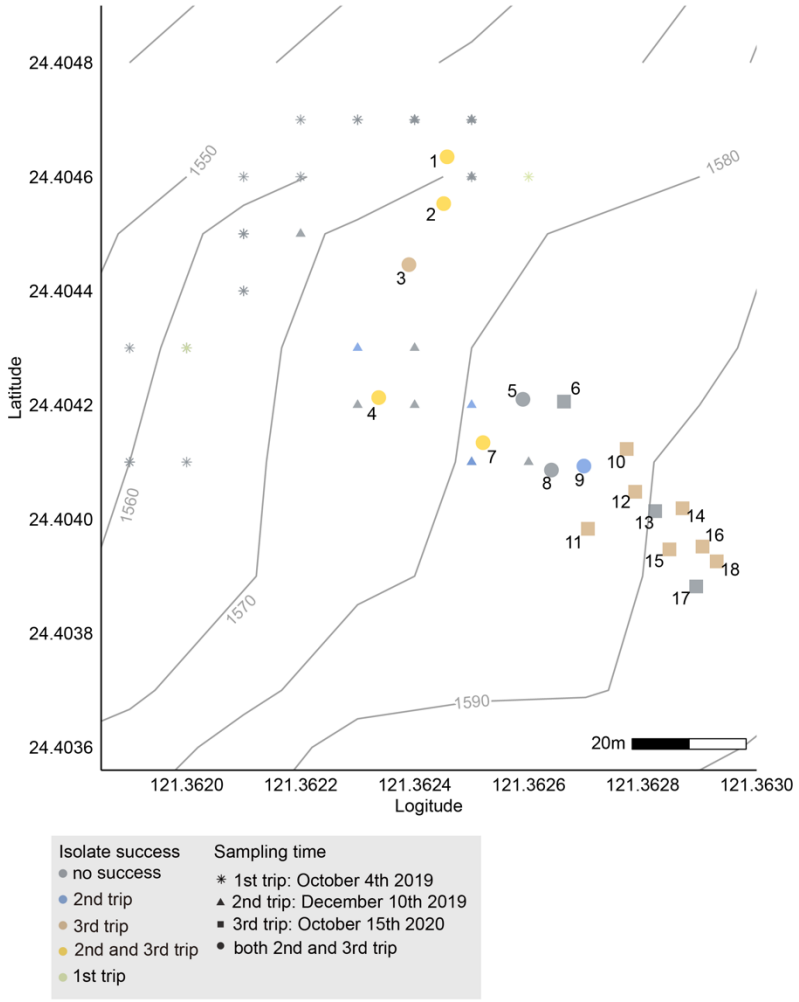
mutation rate $\mu=2.82\times10^{-10}$ per bp from (Tattini et al. 2019), frequencies of sexual reproduction can be estimated as N_ρ/ N_θ for each chromosome (**Supplemental Table S12).** The average of these ratios was reported as the frequency of in each lineage (**Supplemental Table S13**). Isolates within a population were selected having little admixture with other lineages, and were not clones of other isolates (**Supplemental Table S6**). Lineage TW6, CHN-III and Malaysian were excluded due to either insufficient non-clonal individuals within a lineage or the majority of variations being non-informative singletons.

Test of selection

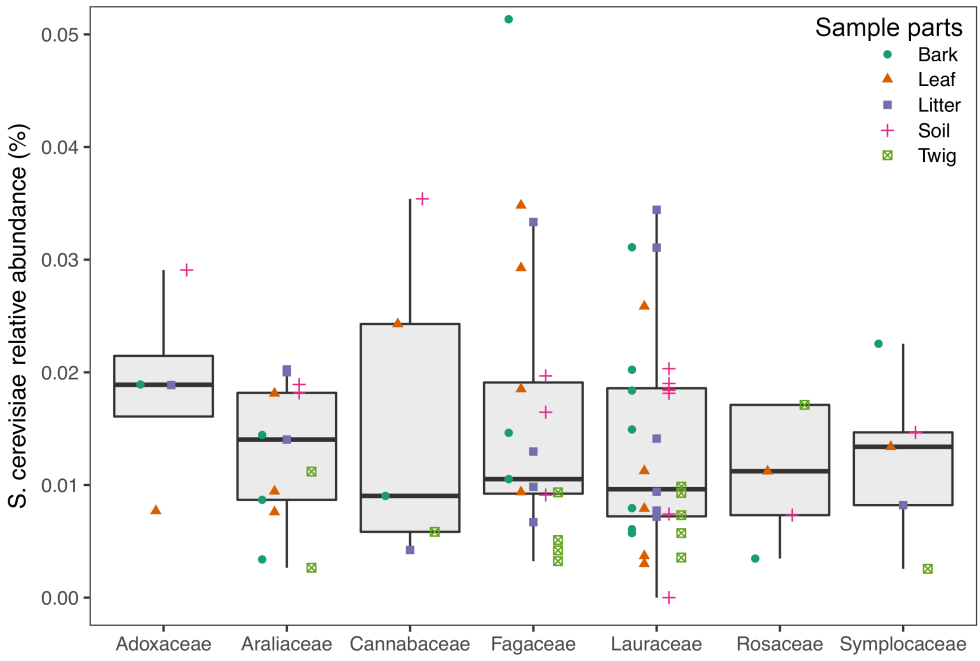
Sections of genome sequences were produced from S288C reference annotation with BEDTools (v2.27.1) (Quinlan and Hall 2010). Neutrality index and McDonald-Kreitman test was calculated from coding segregating sites within each lineage and *S. paradoxus* as a outgroup using the PopGenome package (Pfeifer et al. 2014). To estimate the ratio of nonsynonymous to synonymous substitution rates (d_N/d_S) for each gene, nucleotide sequence alignment and its translated protein sequence alignment were aligned using PAL2NAL (v.14) (Suyama et al. 2006) and estimated with the codeml program in PAML (Yang 2007).

Supplemental Figures

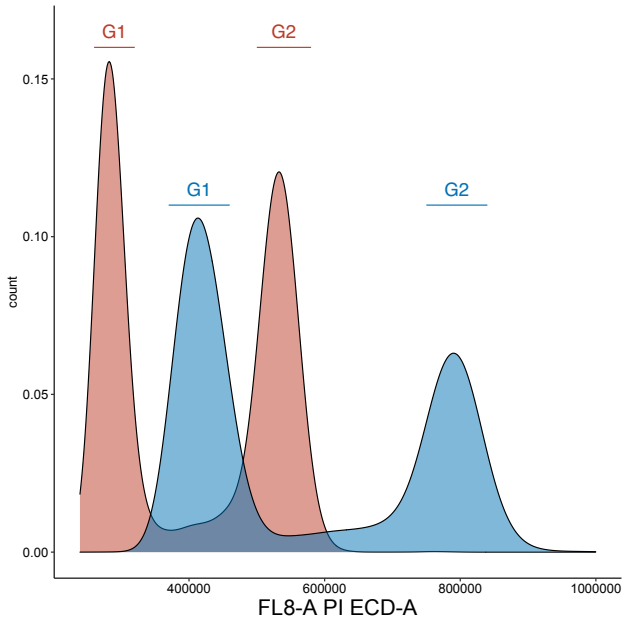
Supplemental Fig S1 – Sampling of 53 trees near Nanshan Village, Yilan County, Taiwan. Three sampling trips were made to this location. Point denotes each sampled tree, colors denote whether and when the isolate was recovered, and shape denotes the sampling time. A total of 18 trees were sampled in both the second and third trip; these were annotated with numbers.



Supplemental Fig S2 – *Saccharomyces cerevisiae* relative abundance amongst tree families. The percentage abundance of *S. cerevisiae* relative to the total fungal abundance per sample was calculated. The relative abundance of *S. cerevisiae* amongst tree families are shown in a boxplot. Samples collected from different parts of the trees were denoted by the different shapes and colours. No statistical significances was detected between tree families (Wilcoxon rank sum test, P=1.0).

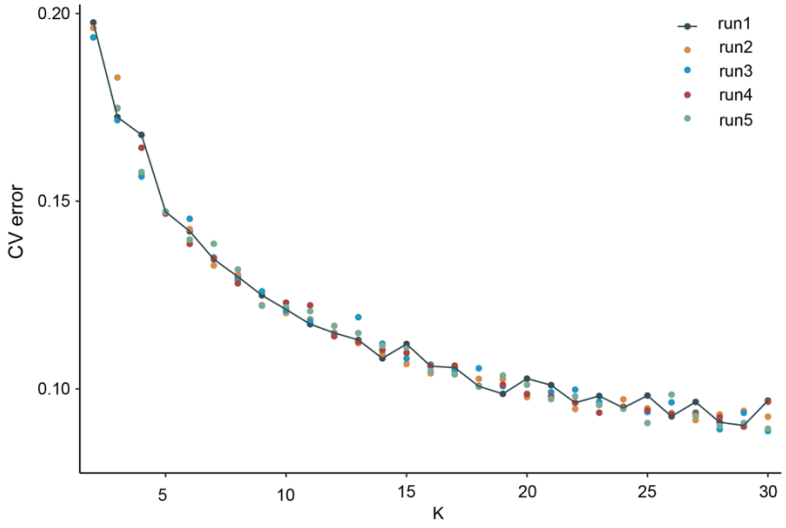


Supplemental Fig S3 – Fluorescence histogram of diploid and triploid isolates. Fluorescence histogram of yeast cells stained with propidium iodide of diploid (brick) and triploid (blue) isolates. Around 20000 events were sampled and filtered with successive gates: non-debris(P1), singlet cells(P2), and cells in G1/G2 phase.

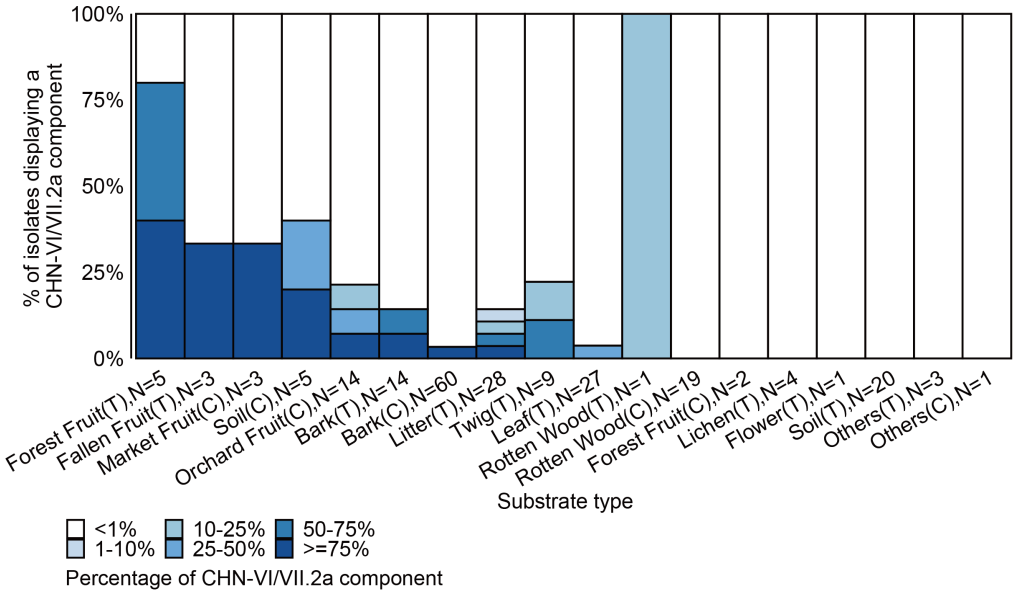


Sample	P1 Events	P2 Events	G1 Events	G2 Events	G1:Median PI-A	G2:Median PI-A
2N	20000	18333	9227	8506	276418.8	525596.9
3N	20012	18507	9943	7122	415769.8	792873.3

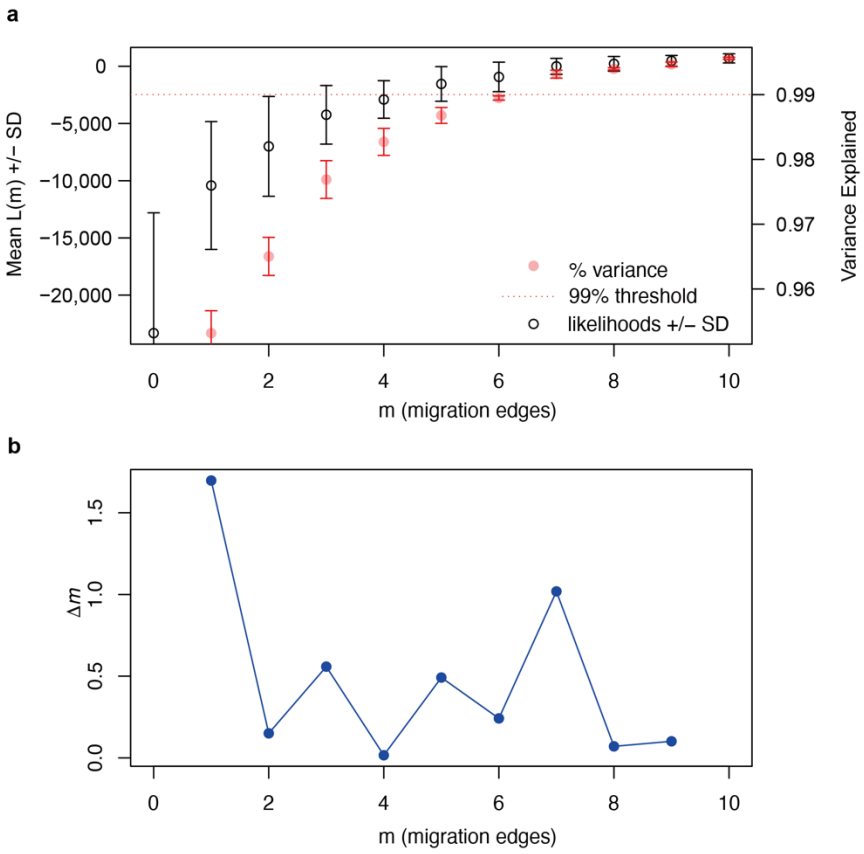
Supplemental Fig S4 – Cross validation (CV) error estimates from ADMIXTURE output for *S. cerevisiae* 340 isolates. Different point colors denote the five independent ADMIXTURE(Alexander et al. 2009) runs from K=2 to K=30 with a line connecting the CV error of the first run.



Supplemental Fig S5 – Percentage of CHN-VI/VII.2a genetic component of ADMIXTURE K=29 analysis in *S. cerevisiae* natural isolates. These isolates were categorized by substrate source and geographical origin (C and T denote Chinese and Taiwanese origin, respectively). N denote number of isolates.

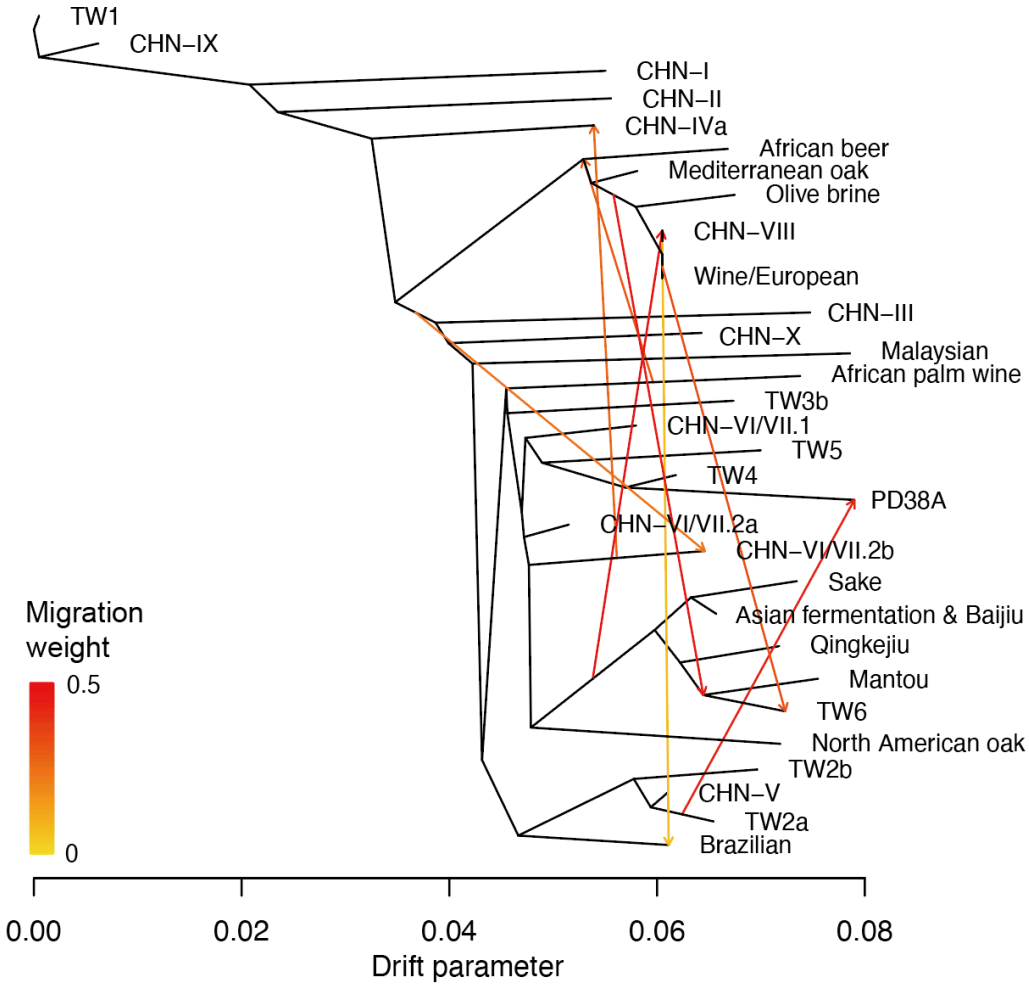


Supplemental Fig S6 – Selecting the most likely model of ADMIXTURE K=16 based on the OptM package (Fitak 2021). a. Distribution of log likelihood and variances explained by models with 0–10 edges. Standard deviations generated by independent TreeMix runs of varying k values (1, 5, 10, 50, 500, 1000). b. Distribution of deltaM–an ad hoc statistic based on the second order rate of change–in the log likelihood with standard deviation considered. We inferred seven edges (>99% variance explained plus the second highest deltaM) to be the most likely model based on ADMIXTURE K=16.

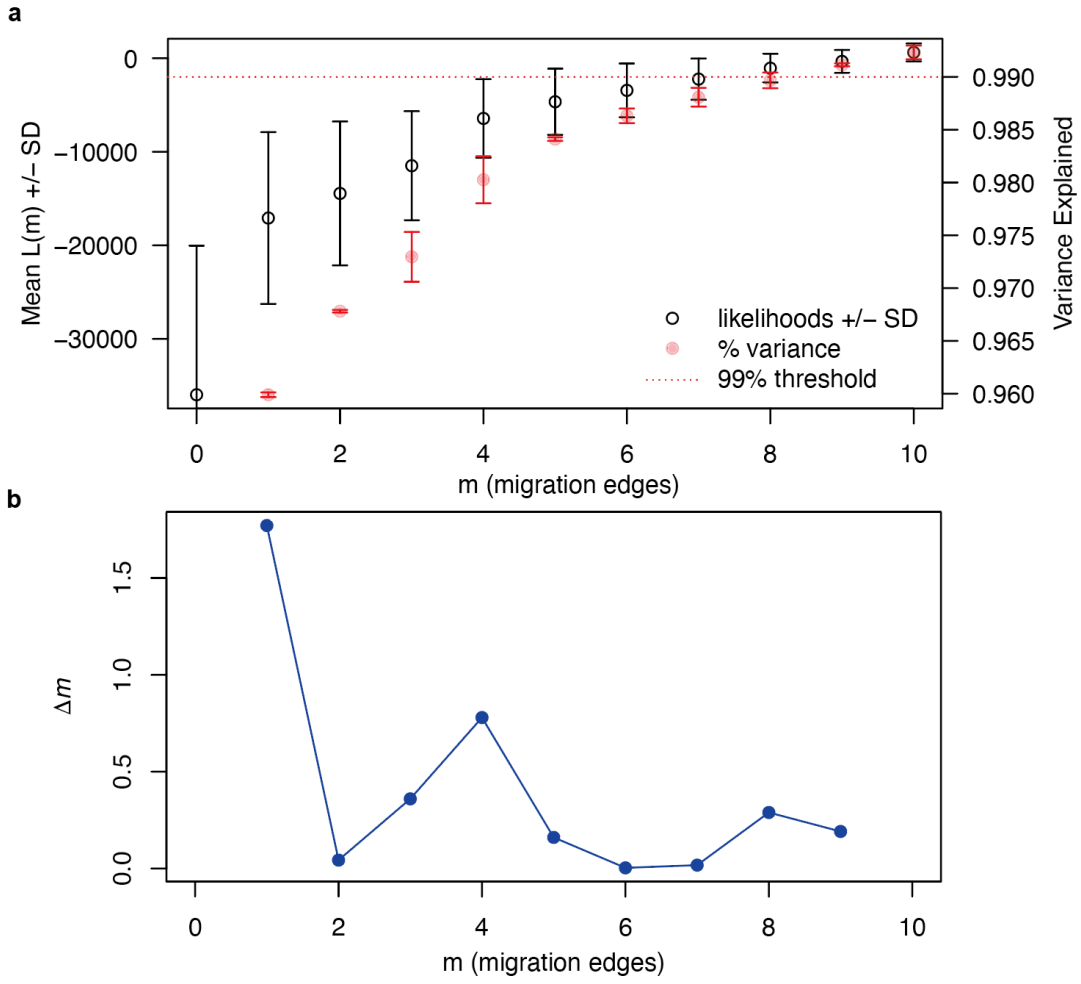


Supplemental Fig S7 – Pairwise divergence over 10kb non-overlapping windows across 16 nuclear chromosomes among isolate PD35A (TW2 lineage), PD36A (TW4 lineage) and PD38A (hybrid of TW2 and TW4 lineage). Divergence K was calculated using VariScan(Vilella et al. 2005) (v2.0.3., RunMode=21).

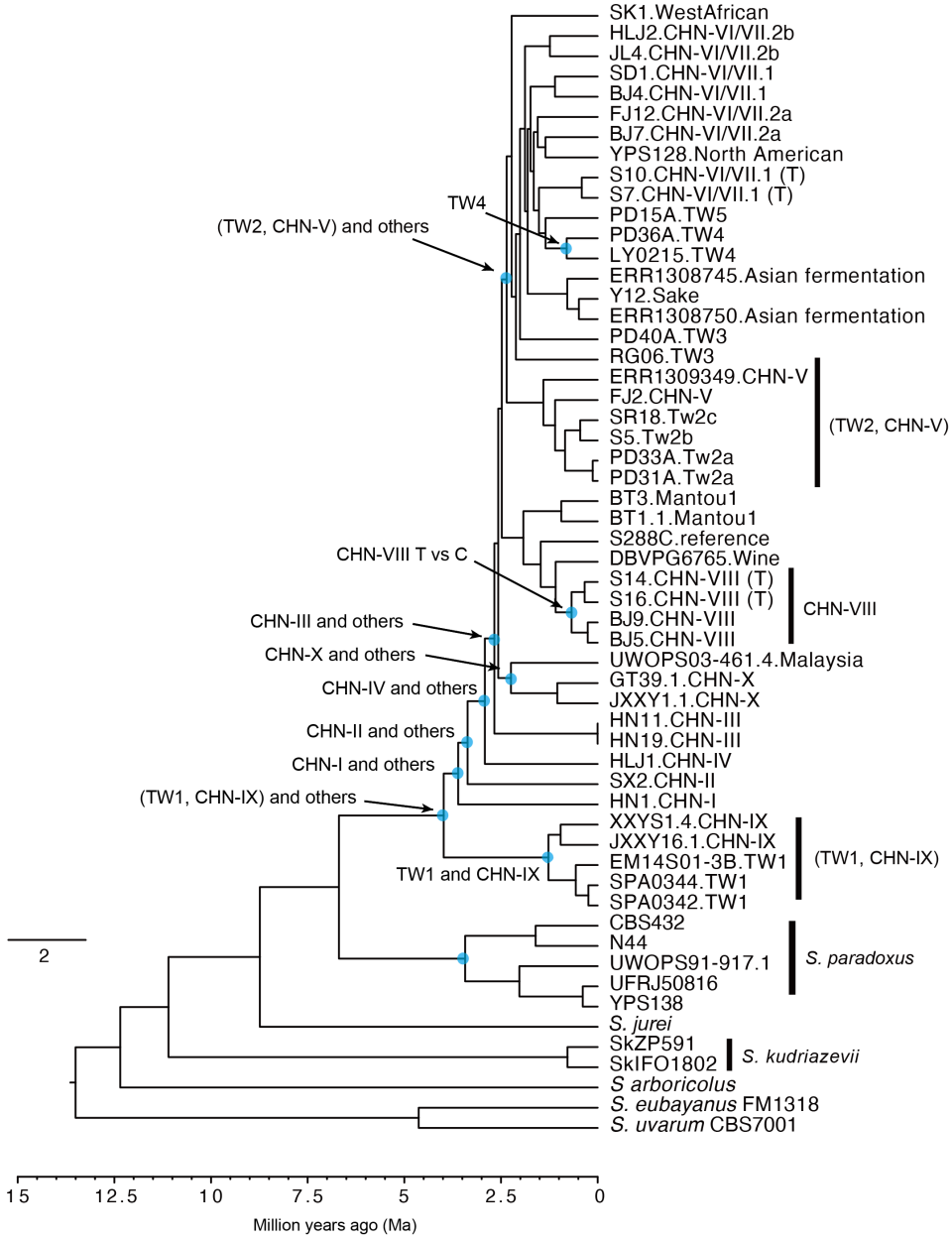
Supplemental Fig S8 – The estimated relationships among the *S. cerevisiae* lineages with eight migration edges based on ADMIXTURE K=29 results.



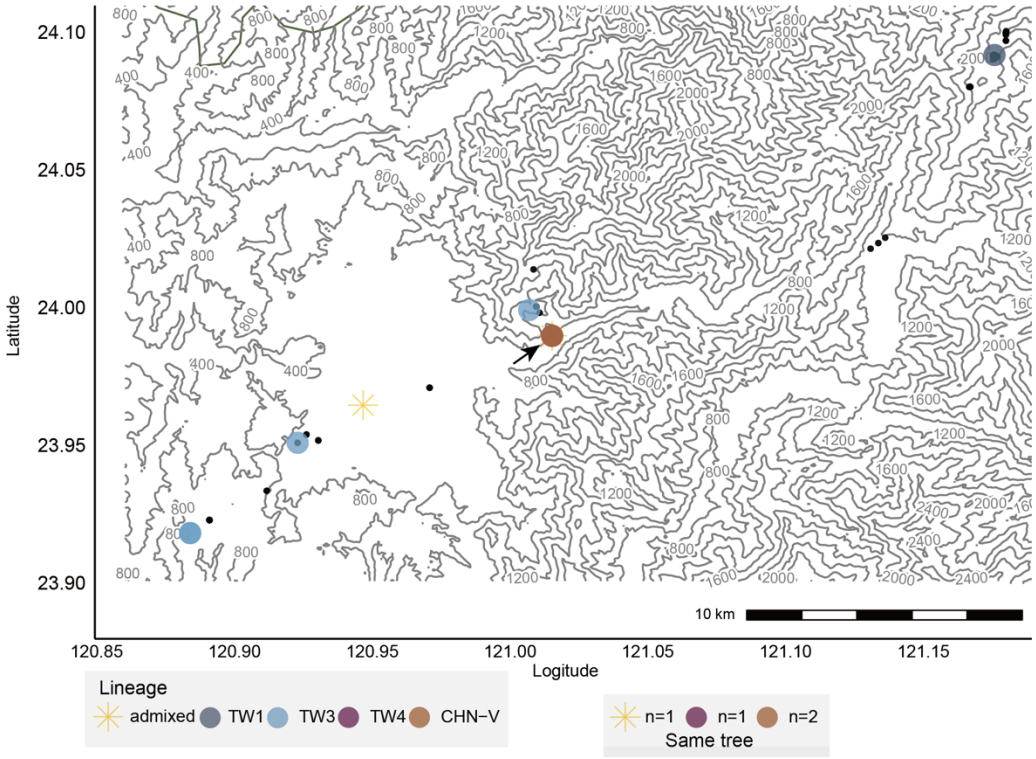
Supplemental Fig S9 – Selecting the most likely model of ADMIXTURE K=29 based on the OptM R package. a. Distribution of log likelihood and variances explained by models with 0–10 edges. Standard deviations generated by independent TreeMix runs of varying k values (1, 5, 10, 50, 500, 1000). b. Distribution of deltaM–an ad hoc statistic based on the second order rate of change–in the log likelihood with standard deviation considered. We inferred seven edges (~90% variance explained plus the third highest deltaM) to be the most likely model based on ADMIXTURE K=29.



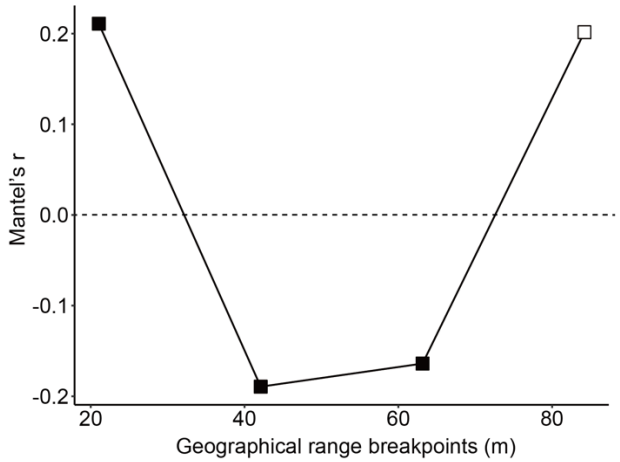
Supplemental Fig S10 – *S. cerevisiae* lineage phylogeny. The topology of the phylogeny was inferred using a coalescence of 1,594 single-copy orthogroup gene phylogenies from ASTRAL (Zhang et al. 2018) and divergence was estimated and calibrated using MCMCtree (Yang 2007). Blue points denote selected nodes shown on **Figure 3b**.



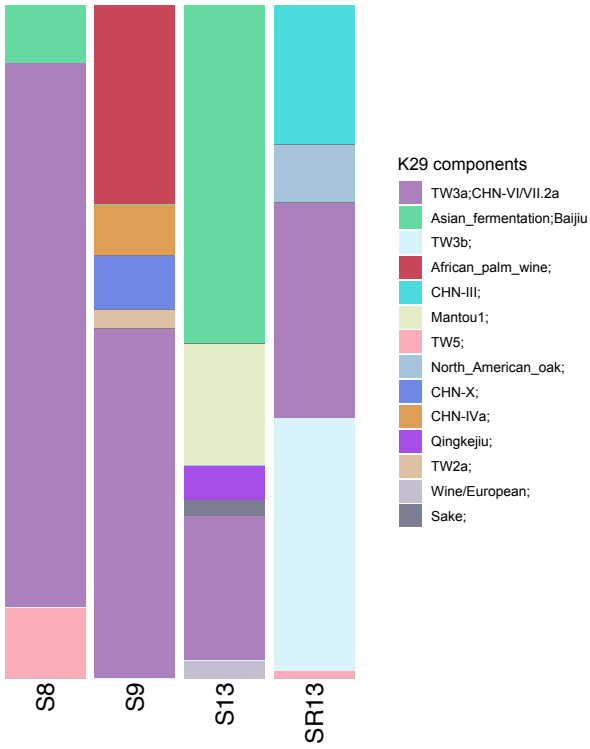
Supplemental Fig S11 – Sampling of 105 trees near Puli, Nantou County, Taiwan. Filled circles denote sampled trees without *S. cerevisiae* isolated. Some points overlap completely because of close proximity on the map. Arrow indicate the location of the tree that has five isolates from different lineages recovered.



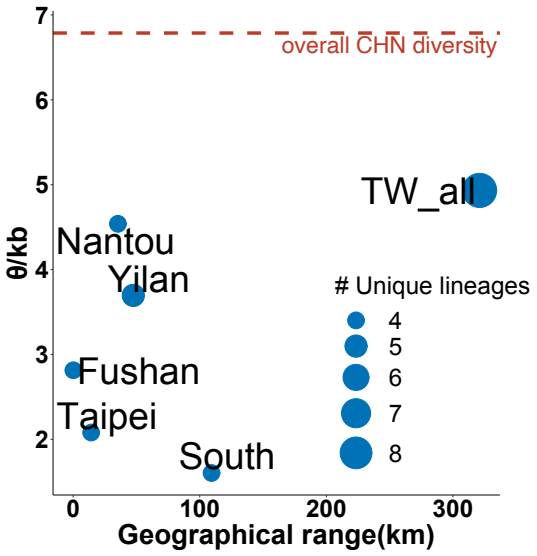
Supplemental Fig S12 – Mantel correlation r for each geographic distance class in Fushan Botanical Garden. Filled squares are statistically significant ($p<0.05$).



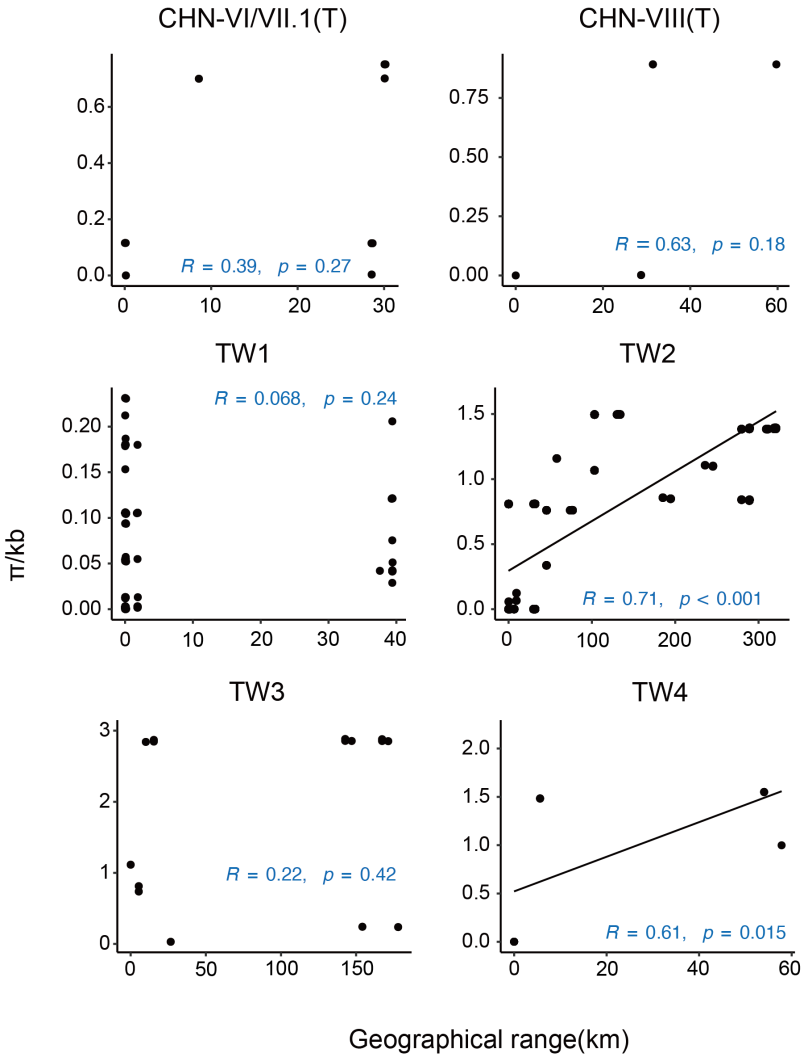
Supplemental Fig S13 – Admixture proportion of the four admixed isolates in Fushan. Genetic makeup of the four admixed isolates found in Fushan Botanical Garden estimated by ADMIXTURE at K=29.



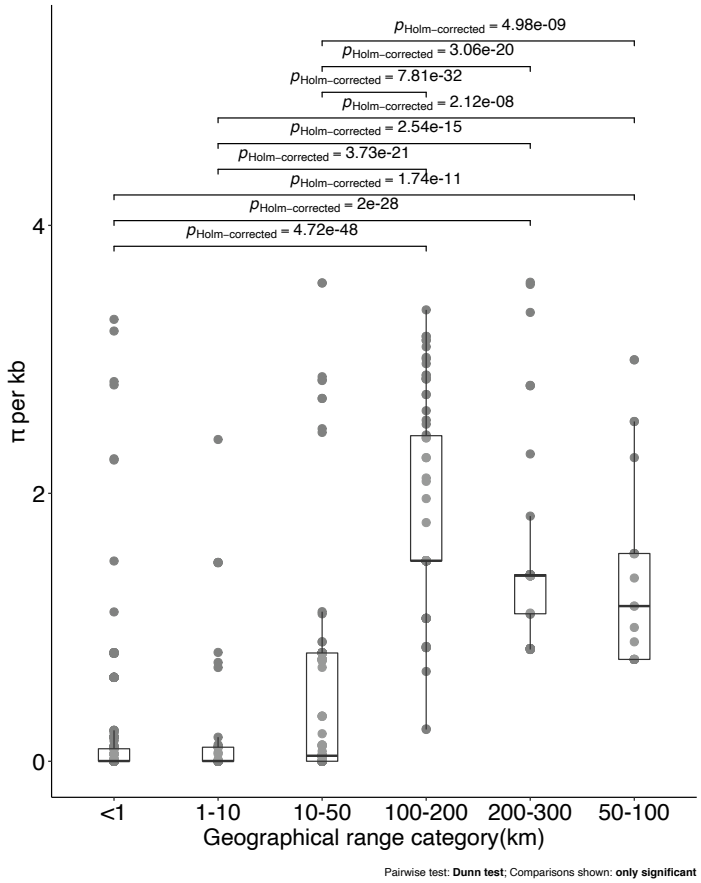
Supplemental Fig S14 – Genetic diversities θ_π in noncoding region among isolates within a sampling area, and overall diversity for all Taiwanese isolates. Dashed horizontal line indicates the overall diversity among natural Chinese isolates, which spanned over 3,500 km.



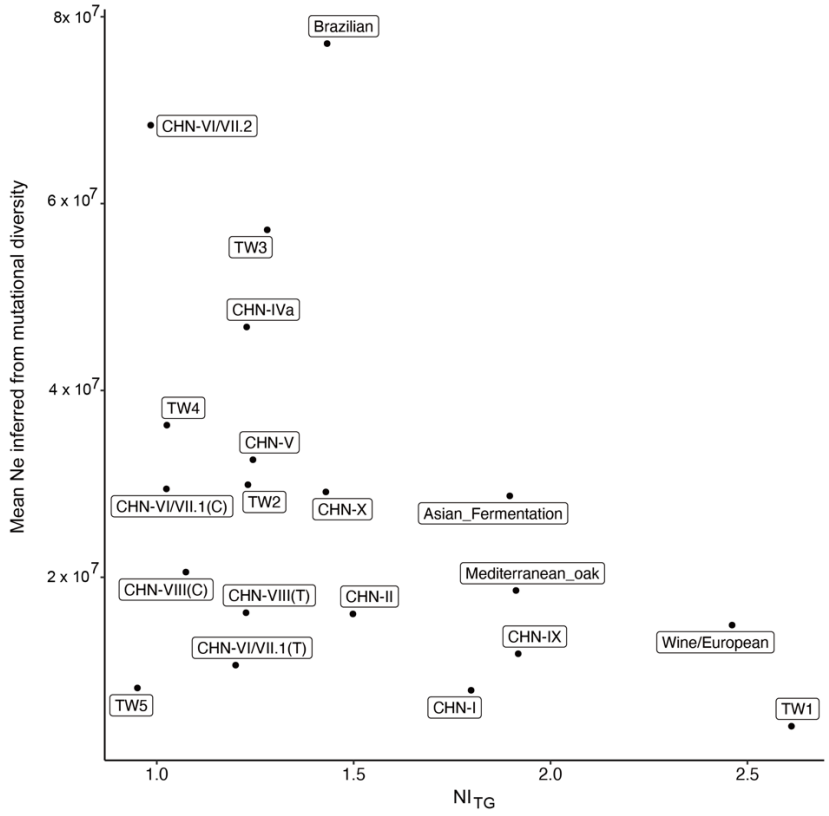
Supplemental Fig S15 – Pearson's r between genetics and geographical distance across natural Taiwanese lineages.



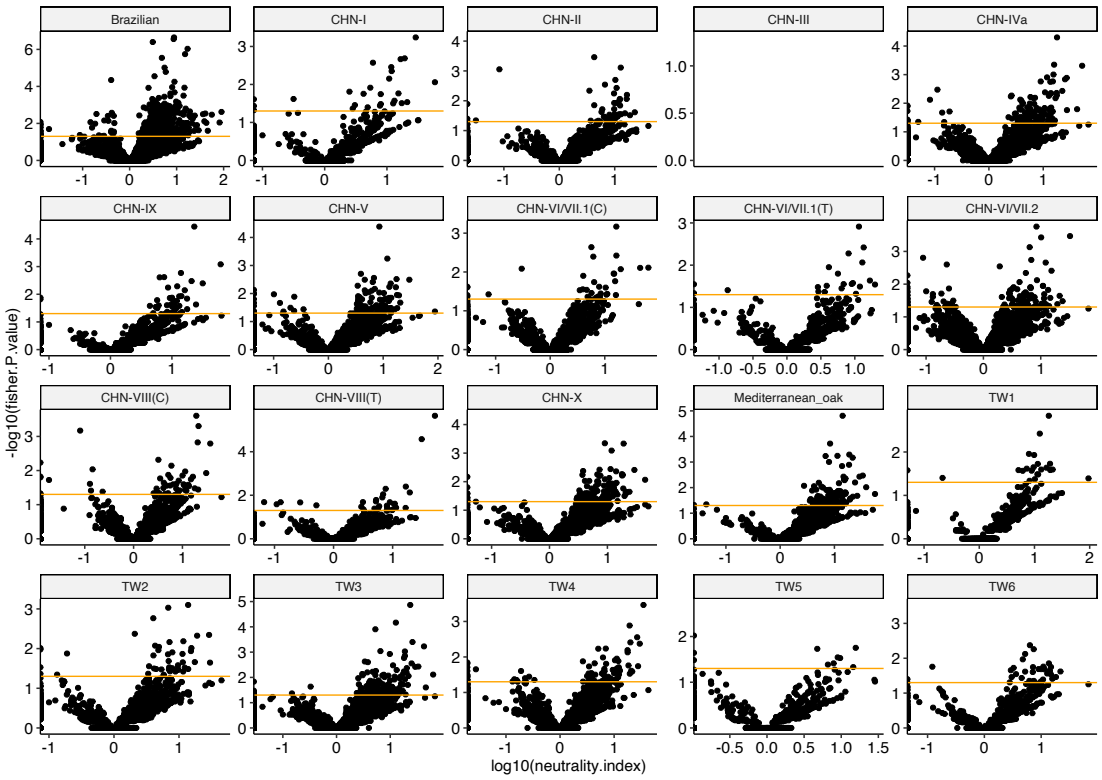
Supplemental Fig S16 – Pairwise sequence diversity in relation to increasing pairwise geographical distance Pairwise sequence diversity between natural Taiwanese isolates in incremental geographical distance categories.



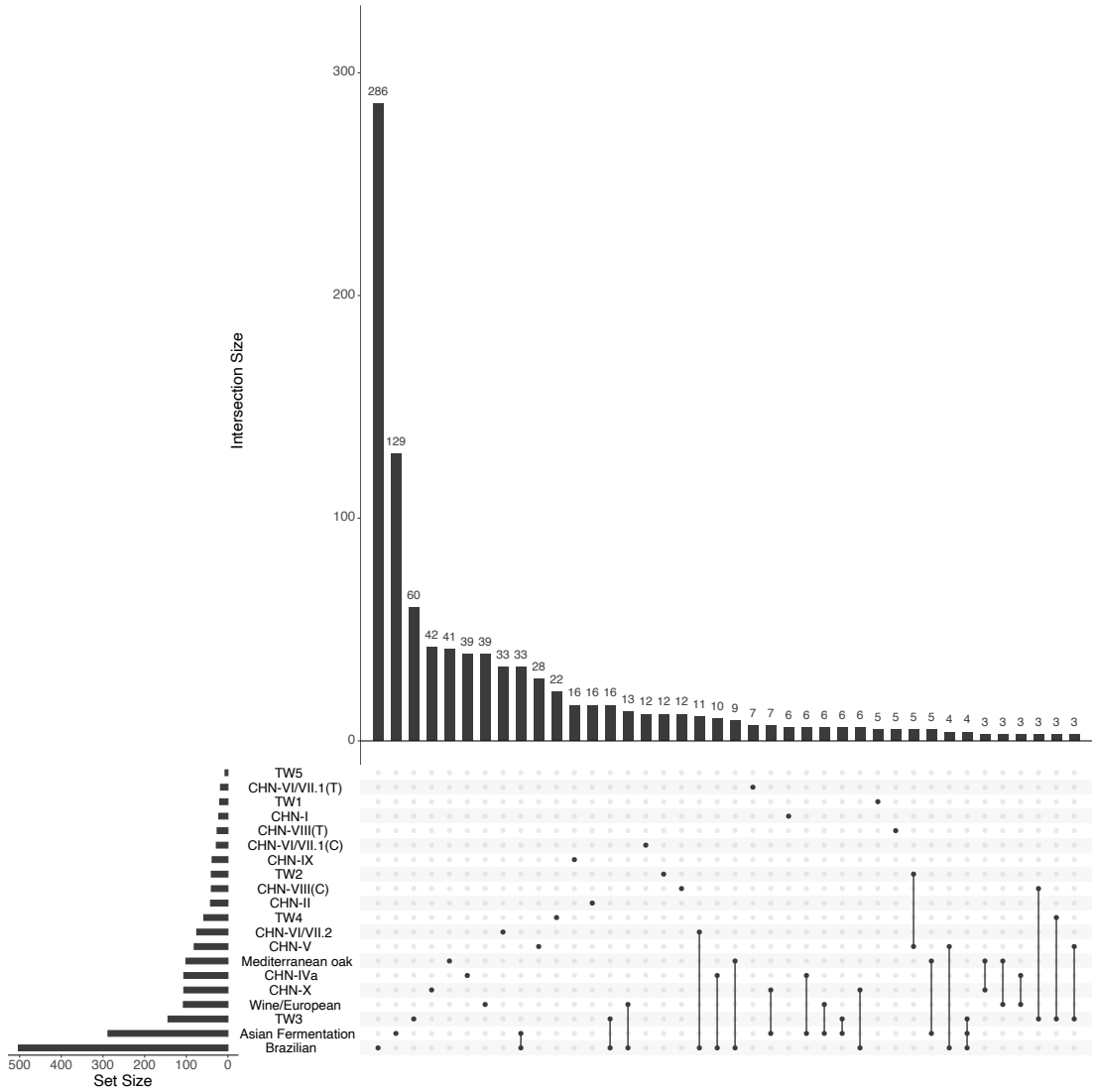
Supplemental Fig S17 – Lack of correlation between NI_{TG} and θw. Kendall's τ coefficient = -0.26, P=0.11.



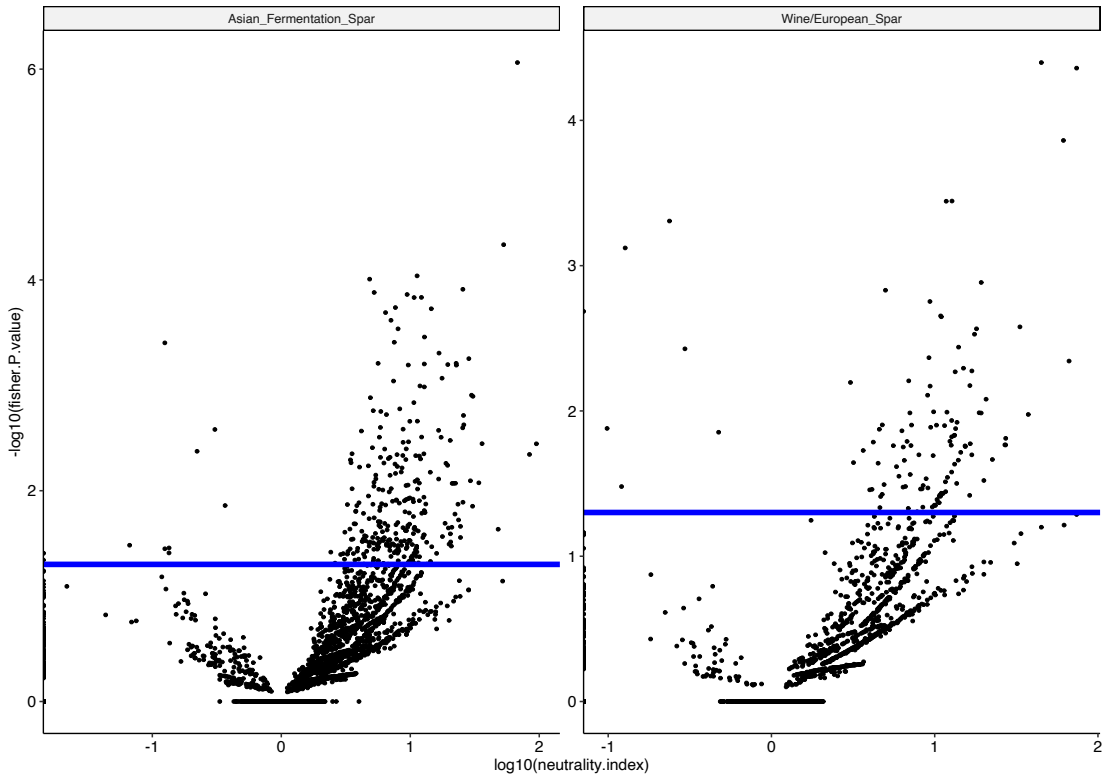
Supplemental Fig S18 – NI across lineage Neutrality index in natural lineages. Horizontal line indicates Fisher's exact test p-value at 0.05. Due to the lack of polymorphism in CHN-III, no genes were found to have valid NI values.



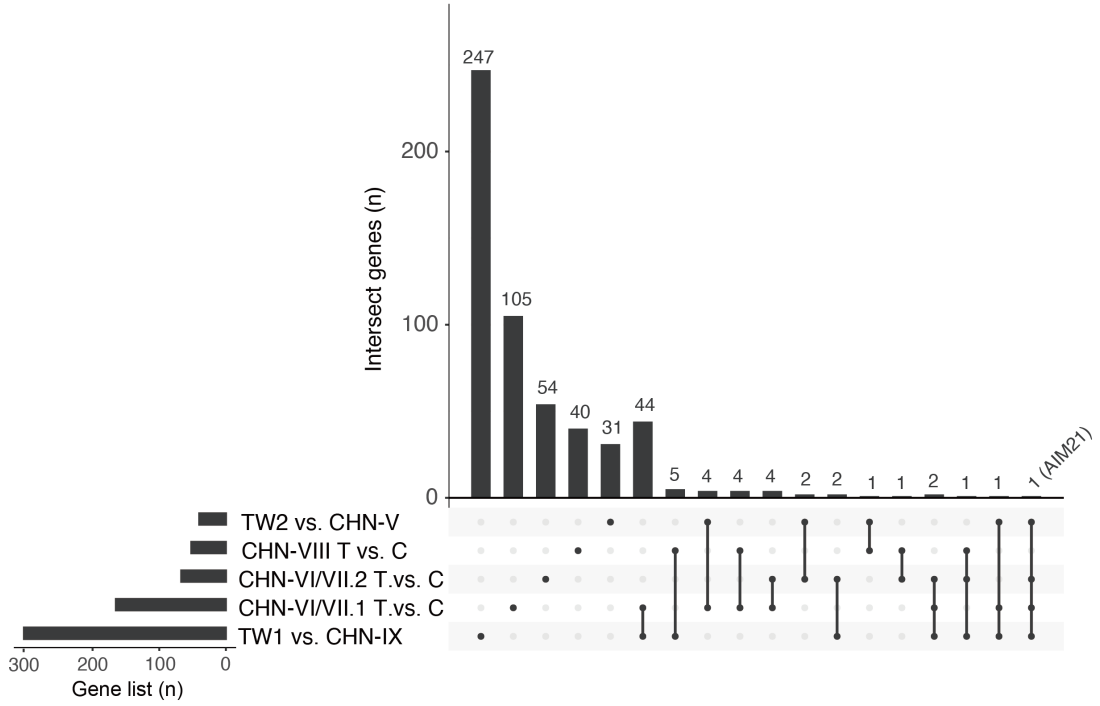
Supplemental Fig S19 – Number of genes with significant NI > 1 both unique in one lineage and common to multiple lineages.



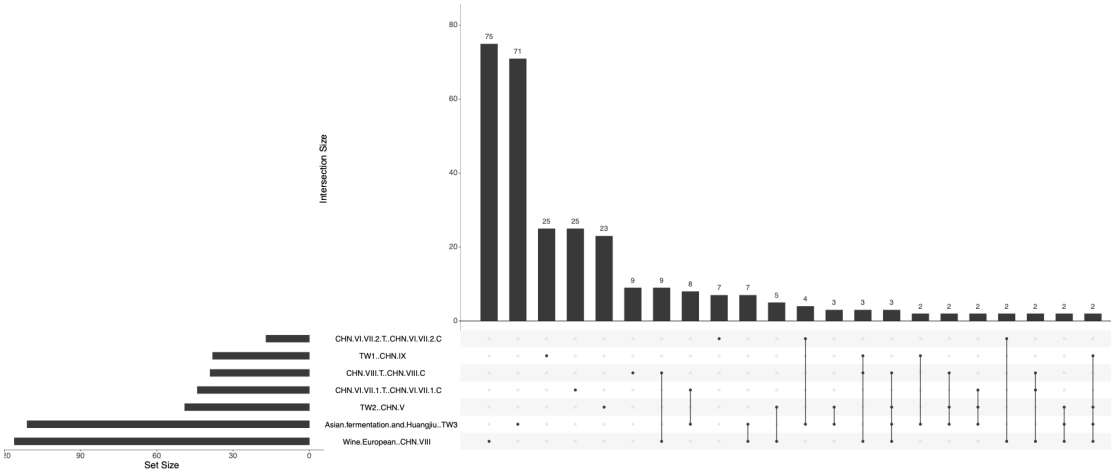
Supplemental Fig S20 – Neutrality index for each gene in two domesticated lineages, using *S. paradoxus* as outgroup species. Horizontal line indicates Fisher's exact test p-value at 0.05. One gene in Wine/European group (YOL081W, *IRA2*) with NI= 35.27419, p-value= 8.4E-11, was omitted on this plot for scale.



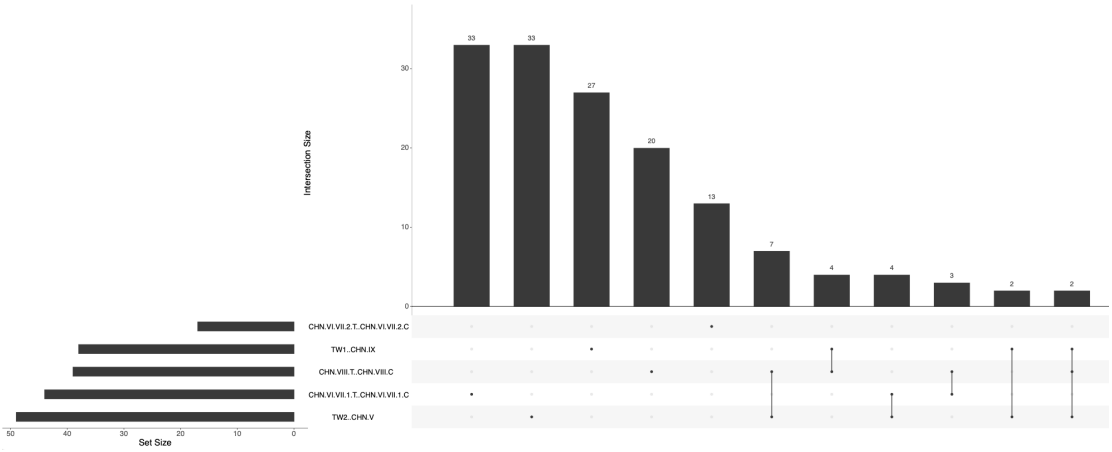
Supplemental Fig S21 – Number of genes with *dN/dS*>1 in each of the five groups. Number of genes with *dN/dS*>1 in isolate pairs from monophyletic TW-CHN groups. Single filled circles represent genes unique to only one lineage, connected circles show genes that are present in at least two lineages.



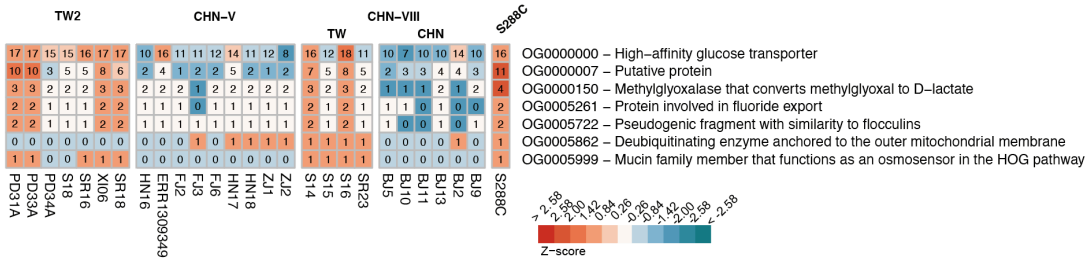
Supplemental Fig S22 – Number of orthogroup significant different in pairwise lineage comparisons. Single filled circles represent genes unique to only one lineage, connected circles show genes that are present in at least two lineages.



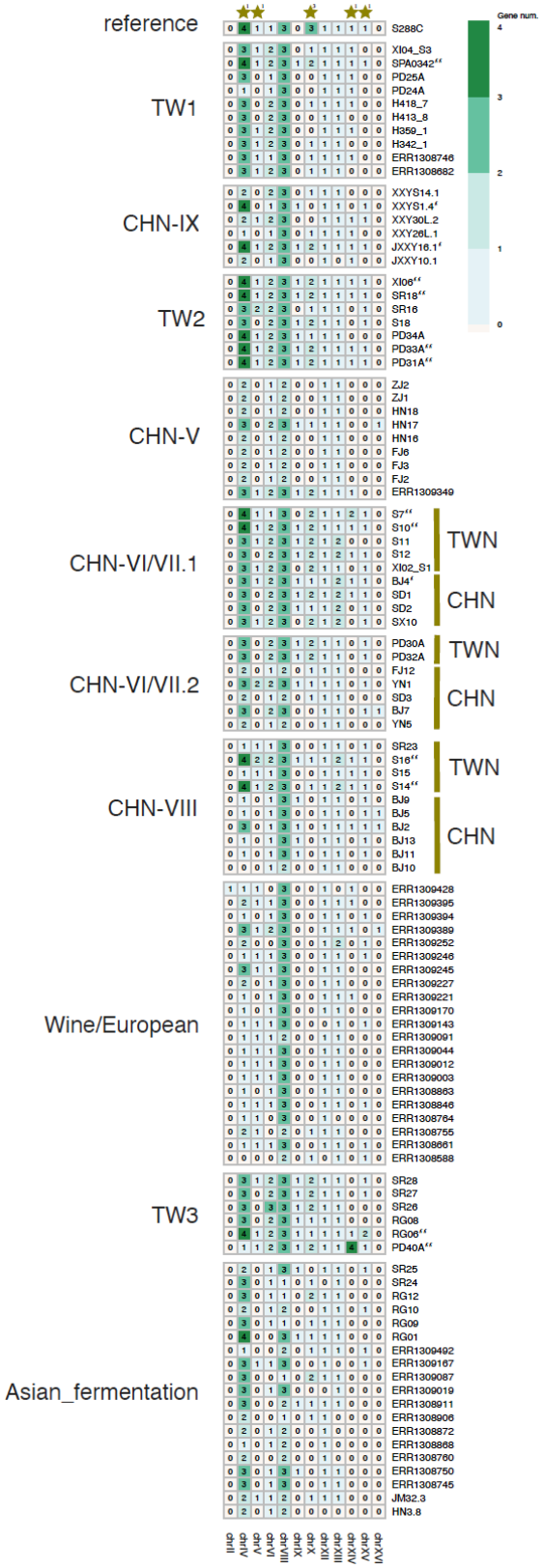
Supplemental Fig S23 – Number of orthogroup significant different in pairwise lineage comparisons. Only natural lineages were visualised. Single filled circles represent genes unique to only one lineage, connected circles show genes that are present in at least two lineages.



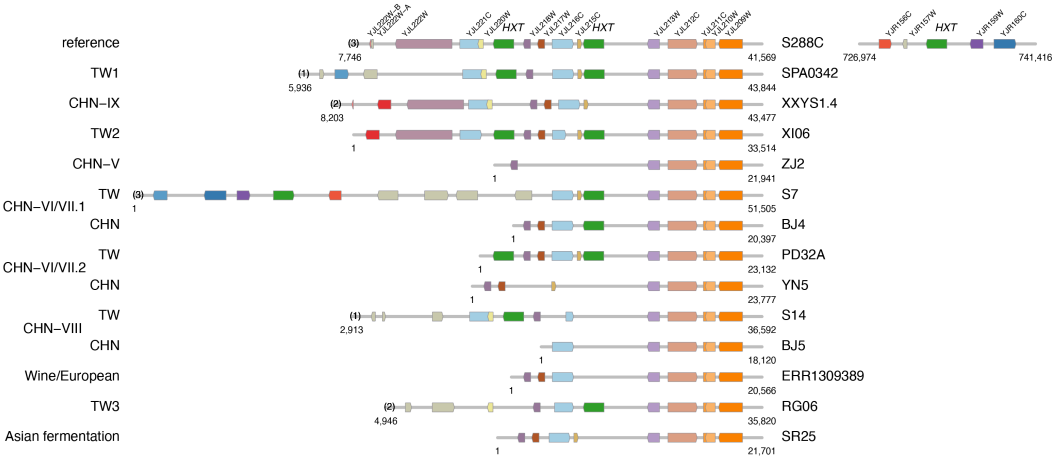
Supplemental Fig S24 – Copy number of orthogroup significant different in both pairwise CHN-VIII Taiwanese vs. Chinese isolates and TW2 vs. CHN-V lineage comparisons.



Supplemental Fig S25 – Distribution of members of HXT gene family on *S. cerevisiae* chromosomes Star with numbers indicate number of HXT locating on subtelomeres.



Supplemental Fig S26 – Distribution of members of HXT gene family on *S. cerevisiae* chromosome X



Supplemental Tables

All tables are saved in a merged Excel xlsx file

Supplemental Table S1 Descriptions of samples and yeast isolates collected in this study

Supplemental Table S2 Detailed information on sampled plant hosts and associated isolation success rates.

Supplemental Table S3 Isolation success rates in different types of samples, and in different months of the year from plant samples.

Supplemental Table S4 Isolation success rate for repeatedly sampled trees between two years

Supplemental Table S5 No significant difference in bioclimatic variables between regions with or without yeast isolates

Supplemental Table S6 Details on the 340 isolates used in this study including previously published genomes.

Supplemental Table S7 The 340 isolates, their phylogenetic groups and the lineages corresponding to their top three major components estimated by ADMIXTURE.

Supplemental Table S8 List of isolates with assemblies produced from nanopore reads

Supplemental Table S9 Range estimate for divergence time for selected nodes a) on the phylogeny in Supplemental Figure S10 and b) calculated using pairwise synonymous changes

Supplemental Table S10 Regional genetic diversity estimates by VariScan(Vilella et al. 2005) and maximum geographical range for natural lineages

Supplemental Table S11. Genetic diversity among isolates recovered within the same tree host

Supplemental Table S12 Diversity estimates of each chromosome across different lineages

Supplemental Table S13 Ranges of effective mutational and recombinational population size, rate of sexual reproduction and number of asexual generations per sexual generation in natural lineages

Supplemental Table S14 List of genes with neutrality index > 1 in each lineage

Supplemental Table S15 List of genes with neutrality index < 1 in each lineage

Supplemental Table S16 GO terms of biological processes found enriched in genes described in Supplemental Table S13. No significant GO enrichment was detected in genes with neutrality index < 1.

Supplemental Table S17 GO terms of biological processes found in genes with neutrality index >1 in both Wine/European and Asian fermentation lineages.

Supplemental Table S18 List of genes with dN/dS>1 in pairwise comparison of Taiwanese and Chinese isolate of shared lineages.

Supplemental Table S19 Percentage of single copy gain/loss in pairwise lineage orthogroup comparisons

Supplemental Table S20. Fisher exact test of whether significant differential orthogroup (OG) members was enriched on subtelomeres. Subtelomere region was defined from (Yue et al. 2017)

References

Alexander DH, Novembre J, Lange K. 2009. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome research* **19**: 1655-1664.

Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. *Genome research* **17**: 1219-1227.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**: 7.

Fitak RR. 2021. OptM: estimating the optimal number of migration edges on population trees using Treemix. *Biology Methods and Protocols* **6**.

Hyma KE, Fay JC. 2013. Mixing of vineyard and oak-tree ecotypes of Saccharomyces cerevisiae in North American vineyards. *Molecular Ecology* **22**: 2917-2930.

Katoh K, Kuma K-i, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research* **33**: 511-518.

Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thevenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF. 2014. Orthology detection combining clustering and synteny for very large datasets. *PLoS One* **9**: e105015.

Lee SB, Taylor JW. 1990. Isolation of DNA from Fungal Mycelia and Single Spores. In *PCR Protocols*, doi:10.1016/b978-0-12-372180-8.50038-x, pp. 282-287.

Liti G, Warringer J, Blomberg A. 2017. Isolation and Laboratory Domestication of Natural Yeast Strains. *Cold Spring Harbor Protocols* **2017**.

Loecher M, Ropkins K. 2015. RgoogleMaps and loa: Unleashing R Graphics Power on Map Tiles. *Journal of Statistical Software* **63**.

Macías LG, Morard M, Toft C, Barrio E. 2019. Comparative Genomics Between Saccharomyces kudriavzevii and S. cerevisiae Applied to Identify Mechanisms Involved in Adaptation. *Frontiers in Genetics* **10**.

Muir A, Harrison E, Wheals A. 2011. A multiplex set of species-specific primers for rapid identification of members of the genus Saccharomyces. *FEMS Yeast Res* **11**: 552-563.

Naseeb S, Alsammar H, Burgis T, Donaldson I, Knyazev N, Knight C, Delneri D. 2018. Whole Genome Sequencing, de Novo Assembly and Phenotypic Profiling for the New Budding Yeast Species Saccharomyces jurei. *G3 Genes|Genomes|Genetics* **8**: 2967-2977.

Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol* **31**: 1929-1936.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**: e1002967.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**: 841-842.

Sare AR, Stouvenakers G, Eck M, Lampens A, Goormachtig S, Jijakli MH, Massart S. 2020. Standardization of Plant Microbiome Studies: Which Proportion of the Microbiota is Really Harvested? *Microorganisms* **8**.

Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT et al. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**: 1533-1545.e1520.

Sniegowski PD, Dombrowski PG, Fingerman E. 2002. Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Research* **1**: 299-306.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609-612.

Tattini L, Tellini N, Mozzachiodi S, D'Angiolo M, Loeillet S, Nicolas A, Liti G. 2019. Accurate Tracking of the Mutational Landscape of Diploid Hybrid Genomes. *Molecular Biology and Evolution* **36**: 2861-2877.

Tsai IJ, Bensasson D, Burt A, Koufopanou V. 2008. Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 4957-4962.

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

Wang QM, Liu WQ, Liti G, Wang SA, Bai FY. 2012. Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Molecular Ecology* **21**: 5404-5417.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591.

Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC et al. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics* **49**: 913-924.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**.