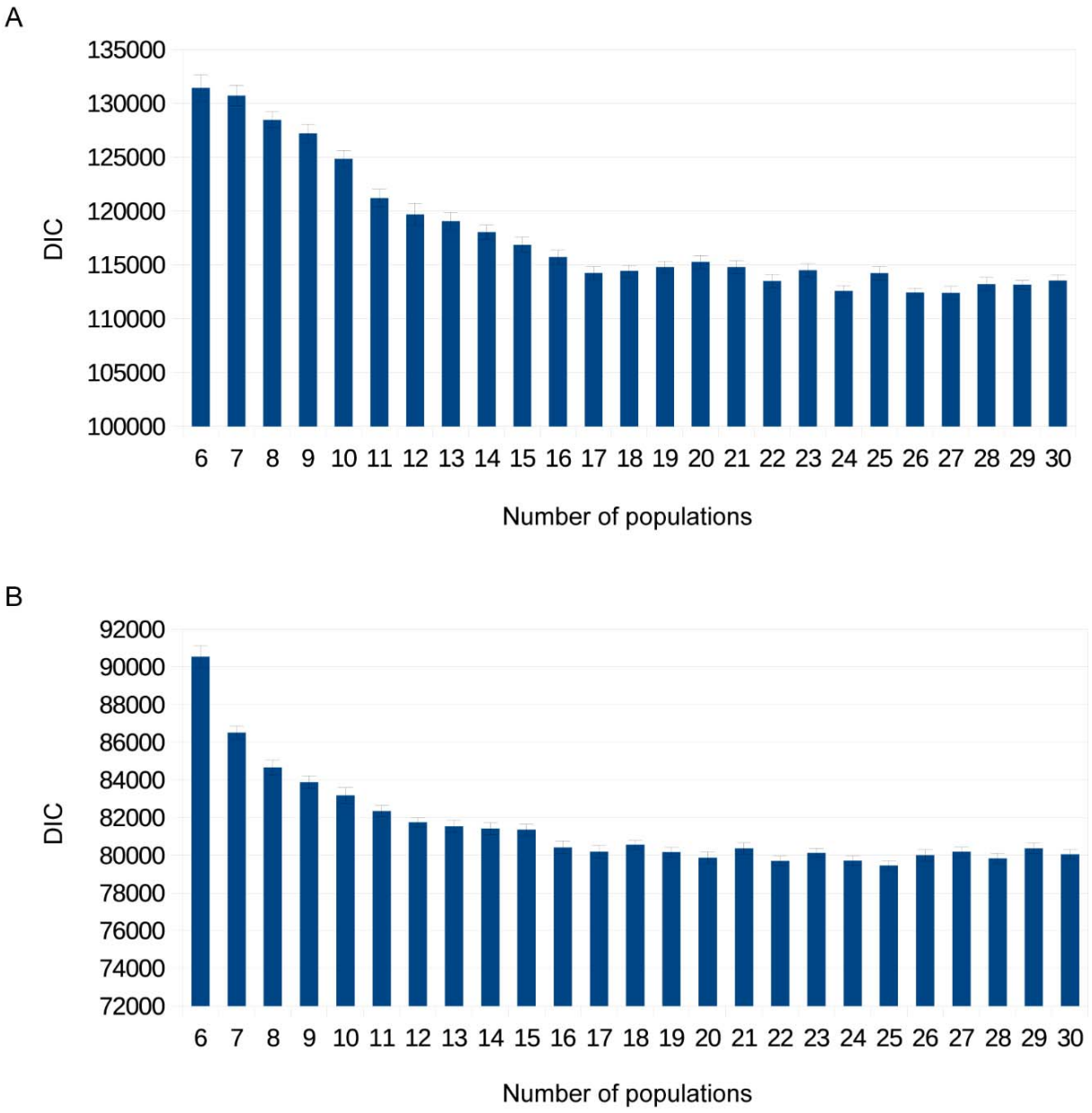

Figure S1, related to Figure 2. Fit of the yeast population genetic data to models with different numbers of populations. Fit is measured by the deviance information criterion (DIC) with lower values indicating a better fit and the standard error (gray error bars) of DIC is shown for 20 independent runs of InStruct. **A.** DIC for 843 variable sites in 438 strains. **B.** DIC for 680 variable sites and 322 strains not identified as aneuploid or polyploid.

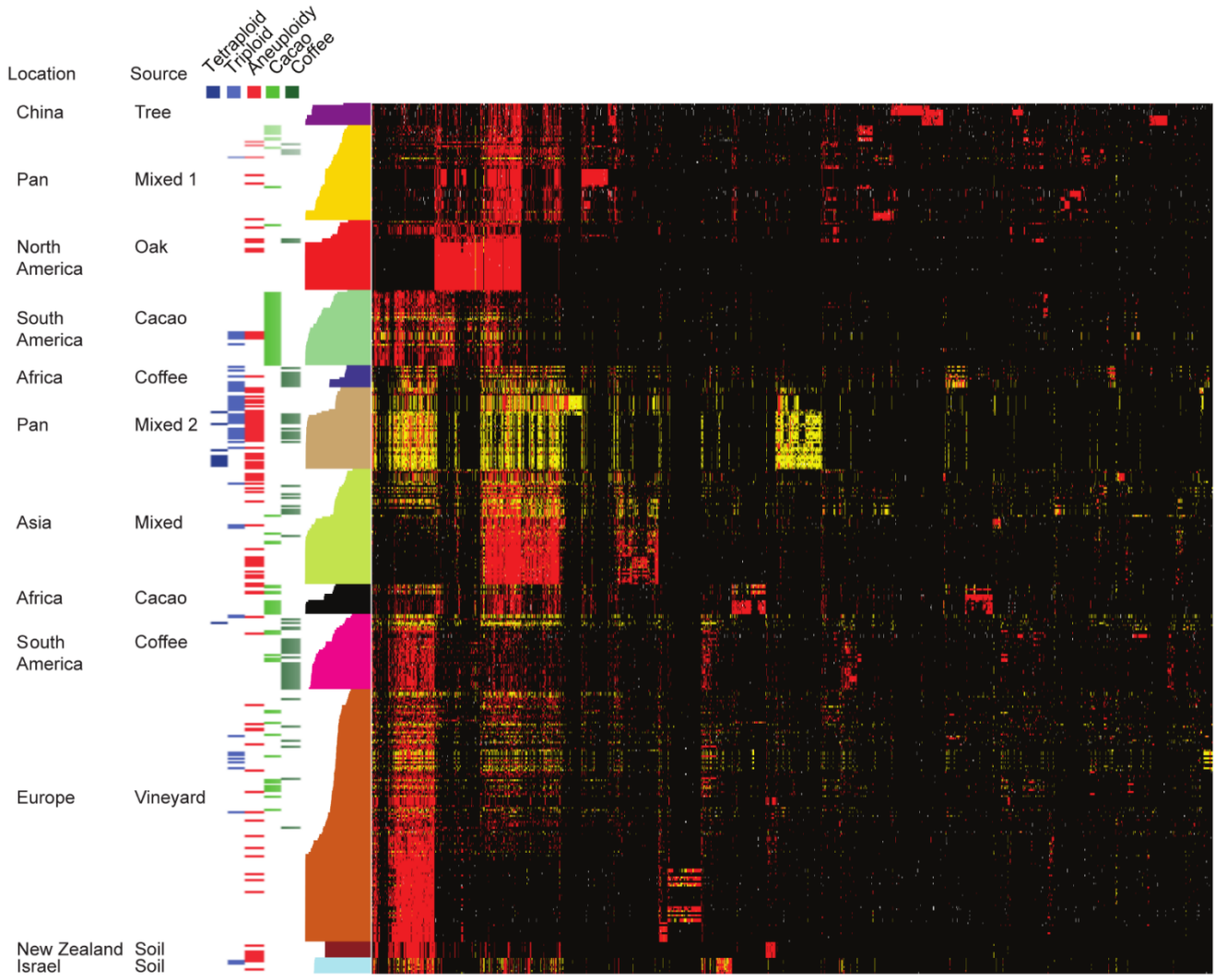

Figure S2, related to Figure 2. Clustered genotypes illustrating shared and population specific alleles. Each row shows one of 438 strains and each column shows one of 2,615 variable sites: homozygous/hemizygous rare allele (red), heterozygous (yellow), homozygous/hemizygous common allele (black) and missing (white). Strains are ordered based on ancestry to each of 12 populations inferred by InStruct and sites are ordered by hierarchical clustering. Population labels (location, source), tetraploidy, triploidy, chromosome aneuploidy, and strains isolated from beans (cacao, coffee) are shown on the left by colored bars. Colored bars to the immediate left of the graph indicate the highest ancestry of each strain, with the width proportional to the percent ancestry.

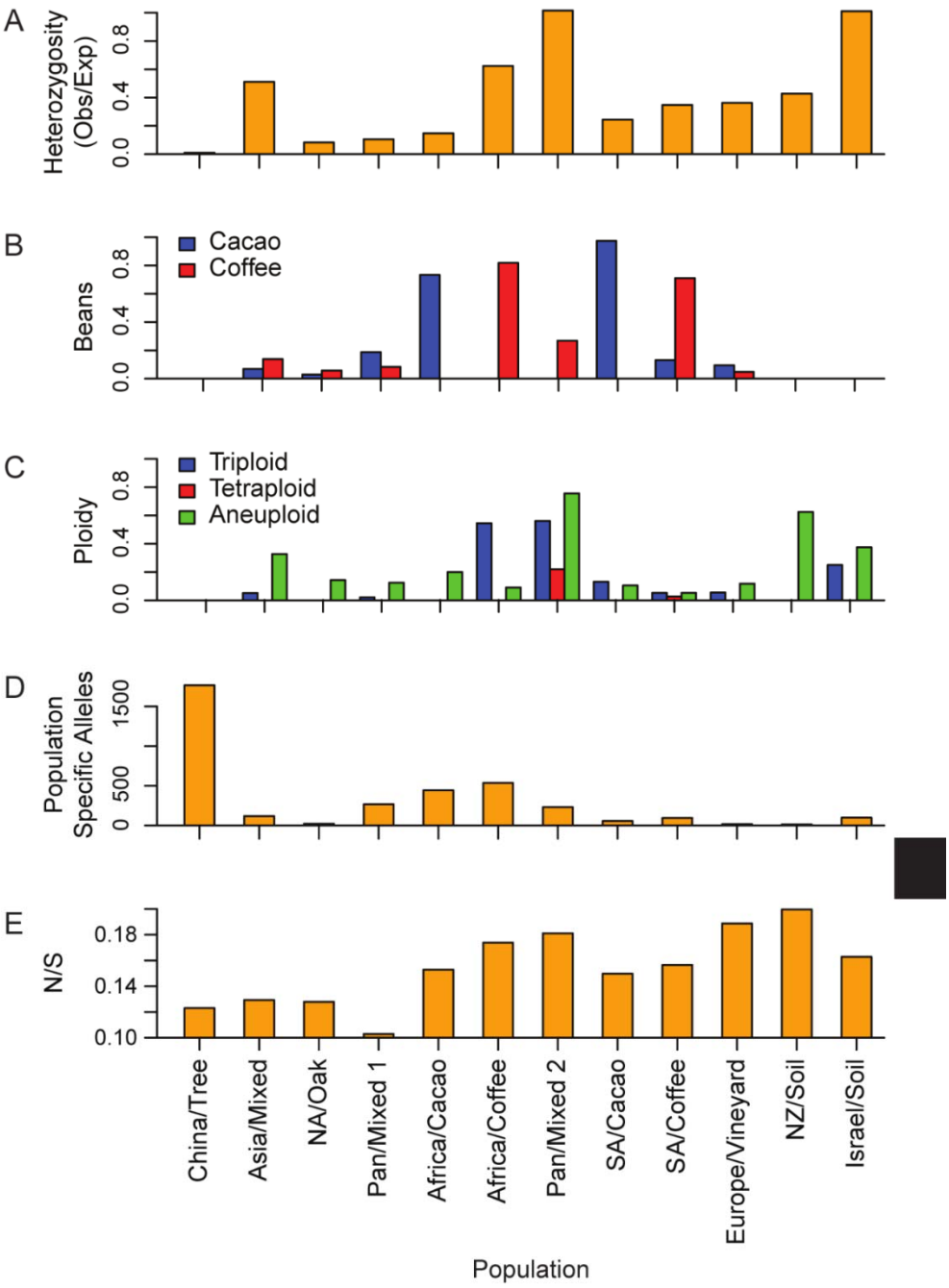

Figure S3, related to Figure 2. Characteristics of the 12 populations. **A.** Ratio of observed to expected heterozygosity within each population, with expected heterozygosity calculated using allele frequency. **B.** Fraction of strains isolated from coffee or cacao beans. **C.** Fraction of triploid, tetraploid and chromosome aneuploid strains. **D.** Number of population-specific minor alleles, defined by an allele frequency of 10% or more in a subpopulation, but less than 1% in all other populations combined. **E.** Ratio of non-synonymous (N) to synonymous (S) SNPs at 20% frequency or more in each population divided by the ratio of non-synonymous to synonymous sites (2.57).

Table S1, related to Table 1. Detailed information about the strains used in our analysis, including their population membership and whether the RAD-seq data was generated in this study. See Excel spreadsheet Table S1.

Table S2, related to Figure 2. Relationships between the populations described in this paper and those in previous publications that used a subset of the same strains.

Populations from this study	Corresponding Cromie, *et al.* [S1]	Corresponding Liti, *et al.* [S2, S3]
1. China Tree	none	none
2. Pan Mixed 1	North America oak (2) some Africa/S.E. Asia (6)	North America
3. North America Oak	North America oak (1)	none
4. South America Cacao	none	none
5. Africa Coffee	none	none
6. Pan Mixed 2	Human associated (4)	none
7. Asia Mixed	Asia (3)	Sake
8. Africa Cacao	some Africa/S.E. Asia (6)	West African
9. South America Coffee	none	none
10. Europe Vineyard	Europe wine (8), Human associated (5), some New Zealand soil (9)	European wine
11. New Zealand Soil	New Zealand soil (9)	none
12. Israel Soil	Israel soil (7)	none

Table S3, related to Figure 2. Percentage membership of coffee and cacao strains in each of the InStruct derived populations.

	Cacao	Coffee
China Tree	0%	0%
Pan Mixed 1	19%	8%
North America Oak	3%	6%
South America Cacao	97%	0%
Africa Coffee	0%	82%
Pan Mixed 2	0%	27%
Asia Mixed	6%	14%
Africa Cacao	73%	0%
South America	13%	71%
Europe Vineyard	9%	5%
New Zealand Soil	0%	0%
Israel Soil	0%	0%

Table S4, related to Figure 1. The frequency of aneuploidy and polyploidy in vineyard, cacao, and coffee strains

	Vineyard	Cacao	Coffee
Aneuploid	30%	10%	19%
Triploid	0%	6%	22%
Tetraploid	0%	0%	1%
Total strains	27	79	67

Table S5, related to Figure 2. Genetic polymorphisms in all strains. Variant sites across all strains (after filtering, as described in the Materials and Methods) are encoded as 0 (common allele), 2 (rare allele), 1 (heterozygous) or NA (missing). This data was used to construct the heat map in Figure S2 and calculate the pairwise genetic distances. See tab delimited text file Table S5.

Supplemental Experimental Procedures

Identification of aneuploid and polyploid strains

Variation in relative sequencing depth between chromosomes was used to identify aneuploid strains [S4]. To determine ploidy for each strain, raw allele (read) counts (A vs. B allele) were plotted for all 15,426 variable sites, and the lowest ploidy consistent with the observed distribution of ratios was assigned to the strain. For example, a strain displaying frequency peaks at A/(A + B) count ratios of 0, 0.33,0.66 and 1 was assumed to be triploid. Where available, chromosome allele ratios were used to confirm aneuploidy.

Virtual predictions of coffee and cacao provenance

Virtual predictions of coffee and cacao provenance were carried out using the Euclidian identity-by-state distance matrix between all of the coffee strains and all of the cacao strains, separately. For each strain, other strains isolated from the same crop were scanned to identify its closest neighbor and this strain was used to "infer" the origin (continent and country) of the original strain. This was then compared to the declared origin of that strain to score "inference" success or failure. For the purpose of this analysis Haiti and the Dominican Republic were treated as a single country, as were Central American countries other than Mexico.

The influence of coffee fermentation method on yeast isolation

Coffee growers use one of three types of fermentation to digest the pulp surrounding the beans. In the "wet" process, the skin is removed from the coffee cherries and they are immersed in water for a spontaneous 24-48 hour fermentation that allows the pulp to be washed away [S5, S6]. In the "dry" method, whole cherries are spread on a platform, heaped at night and spread again each day for 10-25 days until the beans can be hulled [S6]. The less common semi-dry method combines elements of both processes. Our ability to recover live *S. cerevisiae* from unroasted coffee beans correlated with the fermentation method. Of 55 coffee samples for which the processing method was known, *S. cerevisiae* was only isolated from 7% of dry processed (1 of 14), 50% of semi-dry processed (3 of 6), and 54% of wet processed (19 of 35). These results are consistent with previous observations that the inclusion of washing steps in the fermentation might favor yeast growth [S5-S7]. Information about the time of harvesting and processing was unavailable for most beans used in our experiments. However, in some cases live yeast could be cultured from cacao beans that had been processed two or three years earlier.

A human-associated population with extensive aneuploidy, polyploidy and heterozygosity

Because the Pan Mixed 2 population harbored a substantial number (27%) of coffee strains, we examined it more closely. While some strains in the Pan Mixed 2 group were isolated by our laboratory or were gifts from

other researchers, most came from the U.S. Department of Agriculture Agricultural Research Service's (USDA/ARS) yeast strain collection. Although some of these were obtained from natural sources such as soil, plant material or insects, many were human-associated isolates from food fermentations, industrial processes or clinical specimens (Table S1). Records available in the USDA/ARS strain catalog indicate that members of this almost clonal group of strains have been isolated from nearly every continent over a time span of approximately seven decades.

This population bears a mixture of Asian and European alleles, as well as a number of alleles specific to the population. However, all of the strains in this population also exhibited heterozygosity at a much higher frequency than other populations (Figures S2 and S3). The potential for heterozygosity rises with increased chromosome copy number and both aneuploidy and polyploidy have been documented among strains isolated from both natural and industrial sources of *S. cerevisiae* [S3, S8-S10]. To determine the frequency of altered ploidy in our strains, we used relative sequencing depth across different chromosomes to identify aneuploidy [S4] and examined allele ratios, via sequencing read counts, at heterozygous sites to assess polyploidy and aneuploidy. Using these methods, we observed relatively high frequencies of aneuploidy and polyploidy in the Pan Mixed 2 population (78% polyploid, 63% aneuploidy, 21% both) as well as among coffee strains generally (Table S4). For comparison, we performed the same analysis on all strains in our collection and found aneuploidy and polyploidy to be present at much lower frequencies (Figure S3). The significant levels of altered ploidy and heterozygosity found among these natural and industrial yeast isolates reveal yet another potentially rich source of phenotypic and genetic diversity.

References

S1. Cromie, G.A., Hyma, K.E., Ludlow, C.L., Garmendia-Torres, C., Gilbert, T.L., May, P., Huang, A.A., Dudley, A.M., and Fay, J.C. (2013). Genomic sequence diversity and population structure of Saccharomyces cerevisiae assessed by RAD-seq. G3 *3*, 2163-2171.

S2. Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337-341.

S3. Strope, P.K., Skelly, D.A., Kozmin, S.G., Mahadevan, G., Stone, E.A., Magwene, P.M., Dietrich, F.S., and McCusker, J.H. (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.

S4. Tan, Z., Hays, M., Cromie, G.A., Jeffery, E.W., Scott, A.C., Ahyong, V., Sirr, A., Skupin, A., and Dudley, A.M. (2013). Aneuploidy underlies a multicellular phenotypic switch. Proceedings of the National Academy of Sciences of the United States of America *110*, 12367-12372.

S5. Agate, A.D., and Bhat, J.V. (1966). Role of pectinolytic yeasts in the degradation of mucilage layer of Coffea robusta cherries. Applied microbiology *14*, 256-260.

S6. Silva, C.F., Schwan, R.F., Sousa Dias, E.S., and Wheals, A.E. (2000). Microbial diversity during maturation and natural processing of coffee cherries of Coffea arabica in Brazil. Int J Food Microbiol *60*, 251-260.

S7. Masoud, W., Cesar, L.B., Jespersen, L., and Jakobsen, M. (2004). Yeast involved in fermentation of Coffea arabica in East Africa determined by genotyping and by direct denaturating gradient gel electrophoresis. Yeast *21*, 549-556.

S8. Codon, A.C., Benitez, T., and Korhola, M. (1998). Chromosomal polymorphism and adaptation to specific industrial environments of Saccharomyces strains. Applied microbiology and biotechnology *49*, 154-163.

S9. Kvitek, D.J., Will, J.L., and Gasch, A.P. (2008). Variations in stress sensitivity and genomic expression in diverse S. cerevisiae isolates. PLoS genetics *4*, e1000223.

S10. Muller, L.A., and McCusker, J.H. (2009). Microsatellite analysis of genetic diversity among clinical and nonclinical Saccharomyces cerevisiae isolates suggests heterozygote advantage in clinical environments. Mol Ecol *18*, 2779-2786.