Extraordinary Genetic Diversity in a Wood Decay Mushroom

Maria A. Baranova,[1,2,3] Maria D. Logacheva,[2,3,4] Aleksey A. Penin,[2,4,5] Vladimir B. Seplyarskiy,[1,2,3] Yana Y. Safonova,[6] Sergey A. Naumenko,[1,2,3,4] Anna V. Klepikova,[2,4] Evgeny S. Gerasimov,[2,4,5] Georgii A. Bazykin,[1,2,3,4] Timothy Y. James,[7] and Alexey S. Kondrashov*,[1,4,7]

[1]School of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia

[2]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia

[3]Pirogov Russian National Research Medical University, Moscow, Russia

[4]A.N. Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia

[5]Department of Biology, Lomonosov Moscow State University, Moscow, Russia

[6]Algorithmic Biology Lab, St. Petersburg Academic University of the Russian Academy of Sciences, St Petersburg, Russia

[7]Department of Ecology and Evolutionary Biology, University of Michigan

***Corresponding author:** E-mail: kondrash@umich.edu.

Associate editor: Daniel Falush

Abstract

Populations of different species vary in the amounts of genetic diversity they possess. Nucleotide diversity π, the fraction of nucleotides that are different between two randomly chosen genotypes, has been known to range in eukaryotes between 0.0001 in *Lynx lynx* and 0.16 in *Caenorhabditis brenneri*. Here, we report the results of a comparative analysis of 24 haploid genotypes (12 from the United States and 12 from European Russia) of a split-gill fungus *Schizophyllum commune*. The diversity at synonymous sites is 0.20 in the American population of *S. commune* and 0.13 in the Russian population. This exceptionally high level of nucleotide diversity also leads to extreme amino acid diversity of protein-coding genes. Using whole-genome resequencing of 2 parental and 17 offspring haploid genotypes, we estimate that the mutation rate in *S. commune* is high, at 2.0×10^{-8} (95% CI: 1.1×10^{-8} to 4.1×10^{-8}) per nucleotide per generation. Therefore, the high diversity of *S. commune* is primarily determined by its elevated mutation rate, although high effective population size likely also plays a role. Small genome size, ease of cultivation and completion of the life cycle in the laboratory, free-living haploid life stages and exceptionally high variability of *S. commune* make it a promising model organism for population, quantitative, and evolutionary genetics.

Key words: **hyperdiversity, population genetics, genetic variation, de novo mutation rate.**

Introduction

Genetic diversity of a population is both a factor and an outcome of a wide range of evolutionary processes. In the vast majority of species, nucleotide diversity π, the probability that two alleles randomly chosen from the population at a nucleotide site differ from each other, is below 0.03; among the 167 eukaryotic species distributed in 14 phyla reviewed recently (Leffler et al. 2012), only two of the values are above 0.05. The highest value of π observed so far is 0.16 in the nematode *Caenorhabditis brenneri* (Dey et al. 2013). Here, we report a new record-holder, a fungus *Schizophyllum commune*, which exceeds this value, thus being the most polymorphic among all studied eukaryotic species.

Hypervariable species are of particular interest for population genetics. Indeed, population genetics deals with differences between genotypes, and the more the differences, the more we can learn about the factors that affect the population. Hyperdiversity could be caused by an unusually high effective population size (N_e), very high mutation rate (μ), or both.

The split-gill fungus *S. commune* has a cosmopolitan range, living on decaying wood in all continents except Antarctica. It can be easily cultivated in the laboratory, both in the haploid and in the diploid (dikaryon) phases of its life cycle (supplementary fig. S1, Supplementary Material online). The 38.5 Mb genome of *S. commune* (Ohm et al. 2010) contains approximately 13,500 protein-coding genes, and approximately 50% of the genome is protein-coding. An average gene contains approximately four introns with mean length 76 nt. We sequenced genotypes of haploid mycelia of *S. commune* from three locations in the United States and from three locations in Russia (fig. 1*a*).

Results and Discussion

Hyperdiversity of *S. commune*

Comparison of *S. commune* genotypes showed that at 4-fold degenerate (4FD) sites, nucleotide diversity π_{syn} is 0.13 and 0.09 within the US and the Russian population, respectively (fig. 1*b*, table 1). Estimates of neutral diversity at synonymous sites after Jukes–Cantor and codon bias corrections (supplementary fig. S2*a* and *b*, Supplementary Material online) result

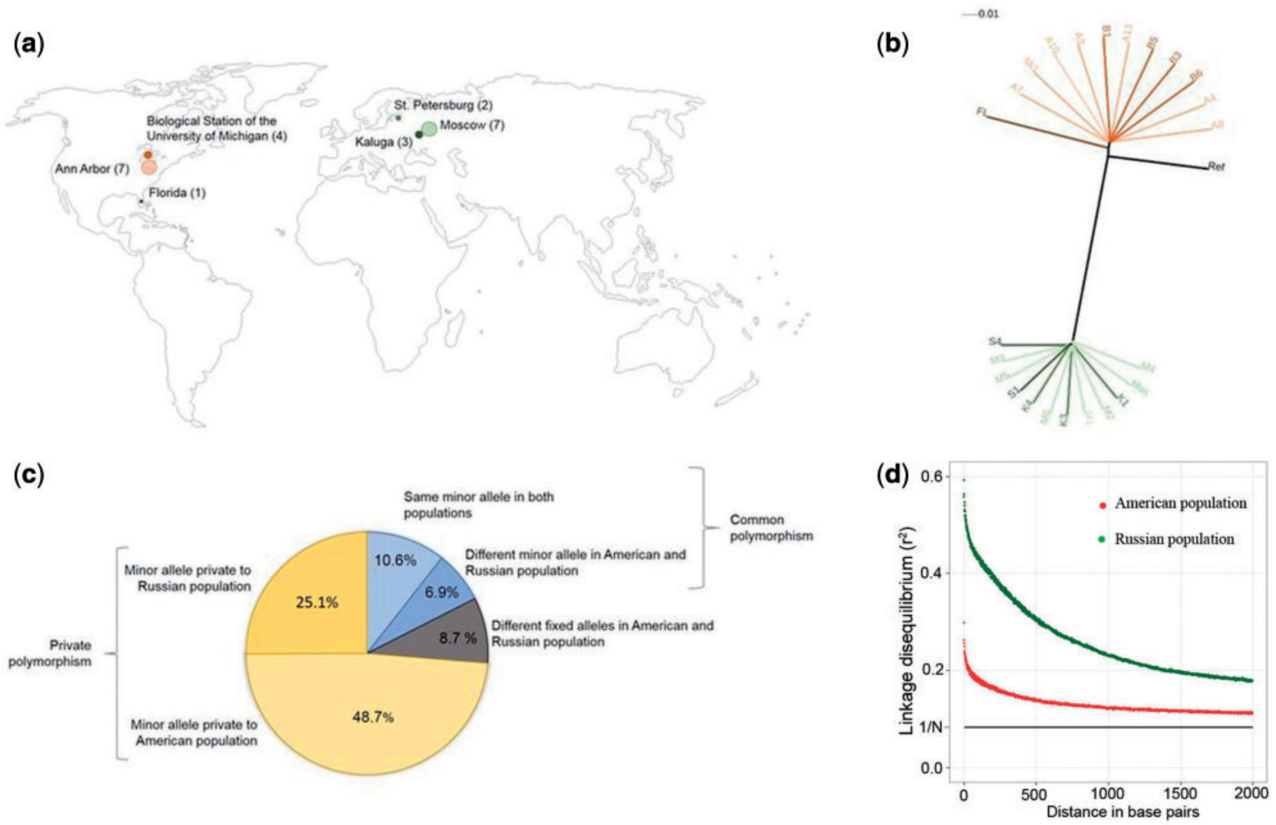

Fig. 1. Patterns of polymorphism in *Schizophyllum commune*. (*a*) Study populations, together with numbers of genomes sequenced from each location. (*b*) NJ tree based on π_{syn} at 4FD sites between all sequenced genotypes. Letters correspond to locations: Ann Arbor, MI (A) and its vicinity (Mi); Biological Station of the University of Michigan, MI (B); Florida (FL); Moscow, Russia (M); Kaluga Region, Russia (K); St. Petersburg, Russia (S). Ref, reference genome sequence based on an individual from MA. (*c*) The fractions of SNPs private to one of the populations, common to both populations, and distinguishing the two populations. (*d*) Patterns of LD between SNPs in the American (red) and in the Russian (green) population. The values of r^2 for different distances between SNPs are presented for all pairs of biallelic sites. The horizontal line is the expected asymptotic value (1/*N*) given the sample size (*N* = 12).

Table 1. Mean Proportion of Differences at 4FD Sites for All Possible Pairwise Comparisons of Individuals from Different Geographical Locations.

	A	B	FL	K	M	S
A	0.126					
B	0.125	0.126				
FL	0.143	0.143				
K	0.232	0.232	0.236	0.087		
M	0.232	0.233	0.236	0.086	0.085	
S	0.233	0.232	0.236	0.085	0.085	0.082

Note.—Letters correspond to locations: Ann Arbor, MI (A); Biological Station of the University of Michigan, MI (B); Florida (FL); Moscow, Russia (M); Kaluga Region, Russia (K); St. Petersburg, Russia (S). Only one genotype from Florida is available, so the FL–FL comparison is missing.

in $\pi_{syn\text{-}JC} = 0.20$ and $\pi_{syn\text{-}JC} = 0.13$ within the US and the Russian population, respectively, and the first of these figures is the highest reported. These extraordinarily high values cannot be explained by errors in assembly or alignment, as the quality of the assembly is high, and alignments of coding regions have few gaps (makarich.fbb.msu.ru/baranova/Population_genetics_Scommune, last accessed July 25, 2015; see Materials and Methods).

There is very little spatial genetic differentiation of *S. commune* within the United States and within Russia (table 1), but divergence between the two populations is substantial (0.48 mean pairwise difference between individuals for synonymous sites after Jukes–Cantor and codon bias corrections; $F_{st} = 0.65$; fig. 1*b* and supplementary fig. S2*c*, Supplementary Material online). Still, much of the polymorphism spans population boundaries. Among the sites that contain exactly two alleles in our sample of 24 genotypes from the United States and Russia, 73.8% are polymorphic in only one of the two populations, whereas in 17.5%, the same pair of alleles is observed in both populations, and only 8.7% are fixed for different alleles in the United States and in Russia (fig. 1*c* and supplementary fig. S3, Supplementary Material online). The best supported demographic history of the two populations based on joint site frequency spectra involves their separation approximately 2.7 million generations ago, and some subsequent admixture and population growth (supplementary text S1, table S1, and fig. S4, Supplementary Material online).

Hyperdiversity of *S. commune* also results in a large number of multiallelic single nucleotide polymorphisms (SNPs) (supplementary table S2, Supplementary Material online). In our sample of 24 genotypes, 5.32% of all nucleotide sites (17.7% of SNPs) are triallelic, and 0.76% of sites (2.5% of SNPs) carry all four possible alleles. In the alignment of 12 genotypes from the American population, tri- and tetraallelic sites comprise 2.80% and 0.25% of all sites, for a total of 600,674 and 52,806 sites, respectively. For the 12 genotypes from the Russian population, which is less polymorphic, the corresponding values are 1.04% and 0.06%. Such vast numbers of polyallelic SNPs even in a small sample allow studying detailed patterns of mutagenesis; for example, we were able to confirm the small-scale heterogeneity of the polymorphism density previously described in other species on the basis of many more individuals with substantially longer genomes (Hodgkinson and Eyre-Walker 2010; Seplyarskiy et al. 2012). Indeed, at 4FD sites in the American population, we observed a 1.4-fold excess of triallelic sites compared with the random expectation, suggesting that 4FD sites differ from each other in their mutation rates substantially (supplementary text S2 and table S3, Supplementary Material online). Additionally, we observed a higher excess of triallelic sites where both mutations are transversions (1.58), and a lower excess of triallelic sites comprised by a transition and a transversion (1.32), in line with the patterns described in Seplyarskiy et al. (2012) for *Drosophila* and Hominidae.

The Cause of the Extraordinary Diversity of *S. commune*

At neutral sites, $\pi = 4N_e\mu$, where N_e is the effective size of the diploid population, and μ is the mutation rate per nucleotide per generation. Thus, hyperdiversity could be caused by an unusually high N_e, μ, or both. Although the mutation rate has been measured previously for only a handful of species, all estimates so far fall between approximately 10^{-10} and 4×10^{-8} per nucleotide site per generation (Kondrashov FA and Kondrashov AS 2010; Gundry and Vijg 2012; Kong et al. 2012; Venn et al. 2014). We directly measured the mutation rate in *S. commune* by sequencing the genotypes of 2 parents (one individual from the United States and another from Russia) and 17 offspring (Seplyarskiy et al. 2014). We observed a total of 9 de novo single-nucleotide mutations (supplementary table S4, Supplementary Material online) per 17 offspring (0.6 per offspring genotype), corresponding to the mutation rate of 2.0×10^{-8} per generation (95% CI: 1.1×10^{-8} to 4.1×10^{-8}, assuming Poisson distribution; supplementary fig. S5, Supplementary Material online). This is about an order of magnitude higher than the mutation rate observed in *Drosophila melanogaster* (2.8×10^{-9}; Keightley et al. 2014). The mutation rate is generally positively correlated with the genome size (Lynch 2010); given the small genome of *S. commune*, the high observed μ is even more striking. Moreover, although the dikaryon and haploid mycelia grew only by several centimeters on the way from parental to offspring genotypes, it is likely that in nature, per-generation mycelial growth spans much larger distances and, thus, involves many more cell divisions; therefore, the mutation rate in natural populations of *S. commune* may well be substantially higher than in our measurement. Thus, the observed record-high π in *S. commune* is partially explainable by its high μ. The causes of the elevated μ are unclear. In particular, according to the reference annotation, *S. commune* has the key proteins for DNA repair processes, including mismatch repair (proteins MLH family, ATPases MutS family), nucleotide excision repair, and double strand break repair (RAD50, RAD51 proteins).

We measured directly the mutation rate and diversity for 4FD sites. Using the above equation for π and assuming that 4FD sites are neutral, we can calculate the effective population size for the American and for Russian populations of *S. commune*. The observed values of π and μ correspond to $N_e \sim 2.3 \times 10^6$ in the American population, and $N_e \sim 1.5 \times 10^6$ in the Russian population, which is only slightly higher than the effective size of the *D. melanogaster* population (1.4×10^6; Keightley et al. 2014). Moderate values of N_e are also consistent with the observed pattern of linkage disequilibrium (LD) between SNPs. In sexual populations, the rate at which the LD between two polymorphic loci decreases with distance between them depends on N_e, with high N_e associated with rapid decline (Hill and Robertson 1968). However, in the American population of *S. commune*, LD decays over the distance of hundreds of nucleotides (fig. 1*d*), similarly to what is observed in *D. melanogaster* (Mackay et al. 2012). Within the Russian population of *S. commune*, the rate of LD decay is even lower; this is consistent with its lower effective size suggested by its lower π (although in theory these differences between the two populations could also arise from differences in recombination).

Natural Selection in Populations of *S. commune*

Hypervariability of *S. commune* facilitates detailed studies of all forms of natural selection. Negative selection is the key force responsible for the differences in the levels of polymorphism between sites of different functional classes, with more functionally important classes being less polymorphic (fig. 2*a*). The genome-average ratios of amino acid-changing to synonymous polymorphism levels $\pi_{\text{rep-JC}}/\pi_{\text{syn-JC}}$ were 0.12 in the American population and 0.13 in the Russian population, which is similar to $\pi_{\text{rep-JC}}/\pi_{\text{syn-JC}} \approx 0.12$ in *D. melanogaster* calculated using the same method (fig. 2*b* and supplementary table S5, Supplementary Material online). This ratio has been reported to be 0.03 in *C. brenneri* (Dey et al. 2013); however, this low value was calculated only using the genes with gene-specific $\pi_{\text{rep}}/\pi_{\text{syn}} < 0.20$ and $\pi_{\text{rep}} < 0.02$ (Dey et al. 2013). This filtering biases the ratio downward; without it, the mean $\pi_{\text{rep-JC}}/\pi_{\text{syn-JC}}$ in *C. brenneri* is 0.10 (Cutter A, personal communication). Mean values are sensitive to outliers, and to the particularities of the filtering procedure; however, the median values of $\pi_{\text{rep-JC}}/\pi_{\text{syn-JC}}$ were also only approximately 20% lower in *S. commune* than in *D. melanogaster*, although they were more than twice as low in *C. brenneri*. $\pi_{\text{rep}}/\pi_{\text{syn}}$ ratio was similar between all genes and genes that had no paralogs in the genome, confirming that the high polymorphism level is not an artifact of erroneous alignments with paralogs. Because most amino

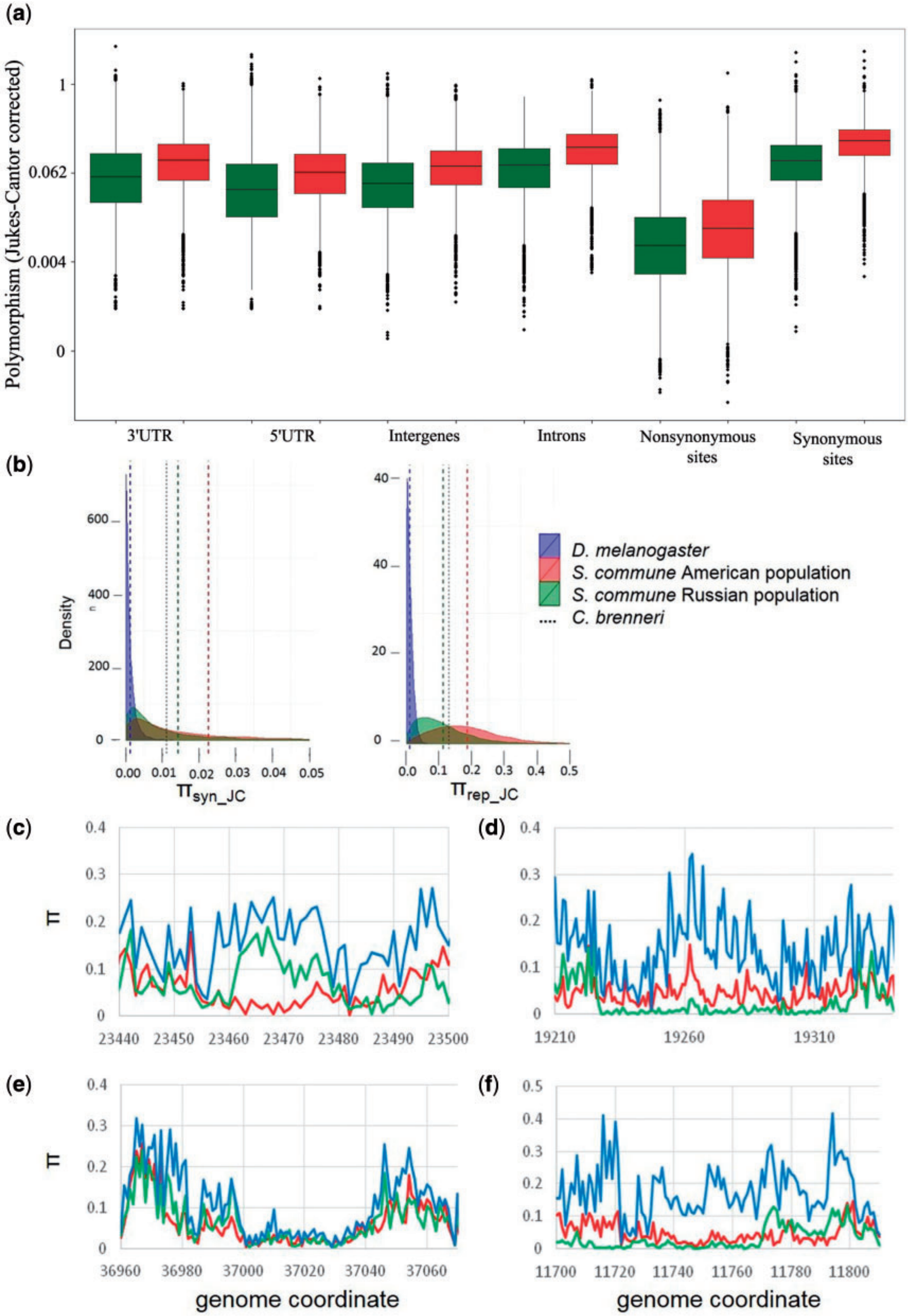

FIG. 2. Footprints of selection in *Schizophyllum commune* population. (*a*) Polymorphism at different classes of sites in *S. commune*. Red, American population; green, Russian population. Lines, boxes and whiskers correspond to medians, quartiles and inner fence. Lower inner fence = quartile 1 − 1.5 × IQR, and upper inner fence = quartile 3 + 1.5 × IQR, where IQR (interquartile range) = quartile 3 − quartile 1. Note the logarithmic scale (log2) of the vertical axis. (*b*) Density plots and mean values of $\pi_{\text{rep-JC}}$ and $\pi_{\text{syn-JC}}$ in 12 American individuals of *S. commune*, 12 Russian individuals of *S. commune*, and 12 American individuals of *Drosophila melanogaster*, together with the mean values of $\pi_{\text{rep-JC}}$ and $\pi_{\text{syn-JC}}$ in 33 individuals of *Caenorhabditis brenneri*. (*c*–*f*) Dips in polymorphism for all sites in populations of *S. commune*. Examples of dips specific to the American (*c*, scaffold 4) or to the Russian (*d*, scaffold 7) population, or affecting both populations (*e*, *f*, scaffolds 1 and 2), are shown. Red, π in the American population; green, π in the Russian population; blue, divergence between populations.

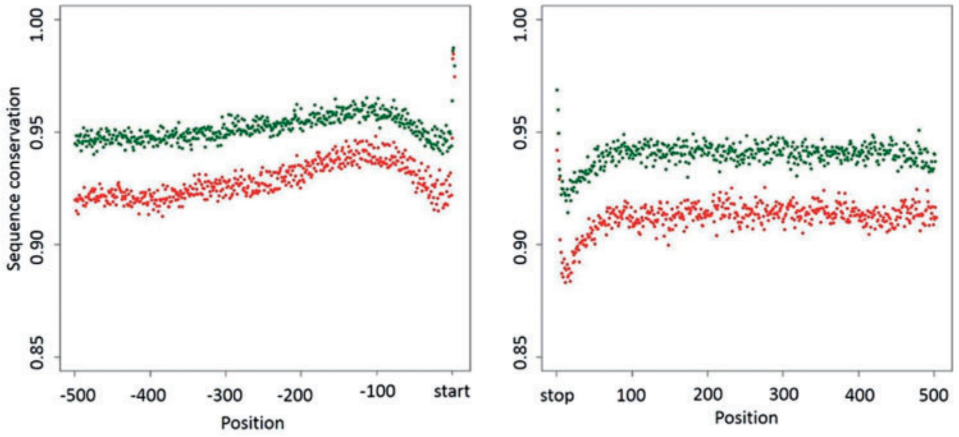

Fig. 3. Inferring selection from polymorphism data at single-nucleotide resolution. Sequence conservation ($1 - \pi$) observed along the 5′-UTR and the 3′-UTR at a range of distances from the start and the stop codon. Red, American population; green, Russian population. Each point presents mean value of conservation for corresponding position in 3′-UTR among 3,366 genes, and in 5′-UTR among 4,516 genes.

acid-changing mutations are slightly deleterious, these data suggest that hypervariability does not automatically lead to a radical increase in the strength of negative selection against amino acid-changing mutations. They also suggest that the higher π in *S. commune*, compared with *C. brenneri*, could not be explained only by higher N_e.

As a result of its high $\pi_{\text{rep-JC}}$, *S. commune* carries an exceptional amount of variation in amino acid sequences. Indeed, an average gene of a *S. commune* individual from the American population differs from the reference genome (from Massachusetts) at 14.8 amino acid sites. This is in stark contrast to *D. melanogaster* genes, which differ from the reference genome at an average of 1.4 amino acids, with the majority of the genes carrying no differences at all.

To estimate the effect that these amino acid differences may have on proteins, we compared the Grantham distances for amino acid-changing polymorphisms with those in *D. melanogaster*. Grantham distance is a measure of biophysical differences between two amino acids comprised from composition, polarity and molecular volume (Grantham 1974). The mean Grantham distance was substantially lower in the American population of *S. commune* (68 vs. 71 in *D. melanogaster*; Wilcoxon rank test, $P < 3.2 \times 10^{-63}$), suggesting that the amino acid changes in this species are less disruptive. If the difference in μ were the only factor different between *S. commune* and *D. melanogaster*, the spectrum of predicted phenotypic effects of polymorphic mutations would be identical in these two species. The fact that the amino acid-changing polymorphisms in *S. commune* appear to be more neutral is consistent with somewhat higher values of N_e in this species, leading to a lower proportion of segregating slightly deleterious mutations. Thus, the high π in *S. commune* is likely caused by a combination of high μ and high N_e. The Grantham distances are also lower in the American population than in the Russian population (68 vs. 69, Wilcoxon rank test, $P < 2.2 \times 10^{-26}$), consistent with a higher N_e of the former.

The exceptional level of polymorphism in *S. commune* allows studying negative selection with fine resolution. Using SNP data alone, we can infer genomic segments under varying degrees of negative selection (SNP shadowing; Cutter et al. 2013). For example, such an analysis applied to untranslated regions (UTRs) reveals a decrease in conservation immediately upstream of the start-codon and downstream of the stop-codon (fig. 3)—a pattern previously observed in divergence data of yeasts and mammals (Shabalina et al. 2004; Kovaleva et al. 2006). Moreover, with such a high level of polymorphism, lack of an SNP at a given nucleotide site is itself suggestive of function. To illustrate this point, we analyzed the extent to which nonsynonymous sites that are polymorphic or monomorphic in our *S. commune* sample are conserved in a moderately related basidiomycete fungus *Coprinopsis cinerea*. Conservation was substantially lower at sites carrying an SNP, compared with monomorphic sites (table 2), consistent with a higher fraction of selected nucleotides among the latter.

High polymorphism of *S. commune* also made it possible to detect short segments of reduced polymorphism, which are suggestive of recent selective sweeps (Smith and Haigh 1974) (fig. 2*c*–*f*). The width of a genomic region of reduced polymorphism caused by a sweep increases with the coefficient of positive selection that drove the corresponding allele replacement, and the reduction is the deepest for very recent sweeps. We identified regions of different widths (> 300, > 1,000, and > 3,000 nt) in which the polymorphism level is reduced to a certain extent (by 60%, 70%, 80%, and 90%, compared with the surrounding genomic segments). We found that most of such dips in polymorphism were specific to one of the populations, with the dips specific to Russia being approximately three times more prevalent than the dips specific to the United States, consistent with a higher LD and slower LD decay in the former population. Only a minority (~10%) of dips were observed in both populations (supplementary fig. S6, Supplementary Material online).

Because of a substantial divergence between the American and the Russian populations implied by the patterns in polymorphism (fig. 1*b* and *c*), we expected the dips to be population-specific, as sweeps predating the divergence

Table 2. Conservation of Polymorphic versus Monomorphic Nucleotide Sites.

	Conserved in *Coprinopsis cinerea*	Not Conserved in *Coprinopsis cinerea*	Fraction Conserved in *Coprinopsis cinerea*
With SNP in *Schizophyllum commune*	57,731	63,395	0.48
Without SNP in *Schizophyllum commune*	587,656	209,478	0.74

Note.—A nucleotide site was considered conserved in *C. cinerea* if the nucleotide observed in *C. cinerea* matched the reference nucleotide of *S. commune*.

would be masked by later mutation. However, dips common to both populations are wider (mean width 1,527 vs. 682 for population-specific dips, *t*-test $P < 5 \times 10^{-24}$) and deeper (mean π 0.002 vs. 0.004, *t*-test $P < 7.2 \times 10^{-48}$). This suggests substantial postdivergence admixture of the two populations that facilitated the spread of advantageous mutations, consistently with the inference of admixture from joint site frequency spectra (supplementary table S1, Supplementary Material online).

We cannot be sure that all the dips in polymorphism that we observe result from selective sweeps. In particular, many narrow and shallow dips may be spurious. Better methods for detection of positively selected regions tailored for hypervariable species are needed for a more formal analysis. Still, the wide dips of depths of 80% or 90% likely represent genuine recent selective sweeps. A higher number of such dips in the Russian population, as well as its lower polymorphism, are consistent with its origin from the American population followed by adaptation to a novel environment.

In summary, due to its extreme variation and ease of cultivation as well as asexual and sexual propagation in the laboratory, *S. commune* is a remarkable model organism for population and functional genomics studies. Hyperdiversity increases the precision of a range of genomic analyses, such as pinpointing the regions of weak and/or ancient positive selection, or localized negative selection (Boffelli et al. 2003; Cutter et al. 2013). Indeed, a sample of just two *S. commune* genotypes from the American population provides more SNPs per site than a sample of 162 *D. melanogaster* individuals (fig. 4).

Materials and Methods

Material Collection and Genome Sequencing

We collected fruiting bodies of *S. commune* at three locations in the United States and at three locations in Russia. A single meiospore was obtained from each fruiting body, and the genotypes of the haploid mycelia that grew from these spores (12 American and 12 Russian) were sequenced.

DNA was extracted from dried mycelia using CTAB method (Doyle JJ and Doyle JL 1987). Libraries were prepared from 1 mcg of total DNA using TruSeq DNA sample prep kit (Illumina) using manufacturer's instructions and then sequenced using HiSeq2000 (Illumina) from both ends with read length of 101. The coverage for all genomes varies between 30 and 40.

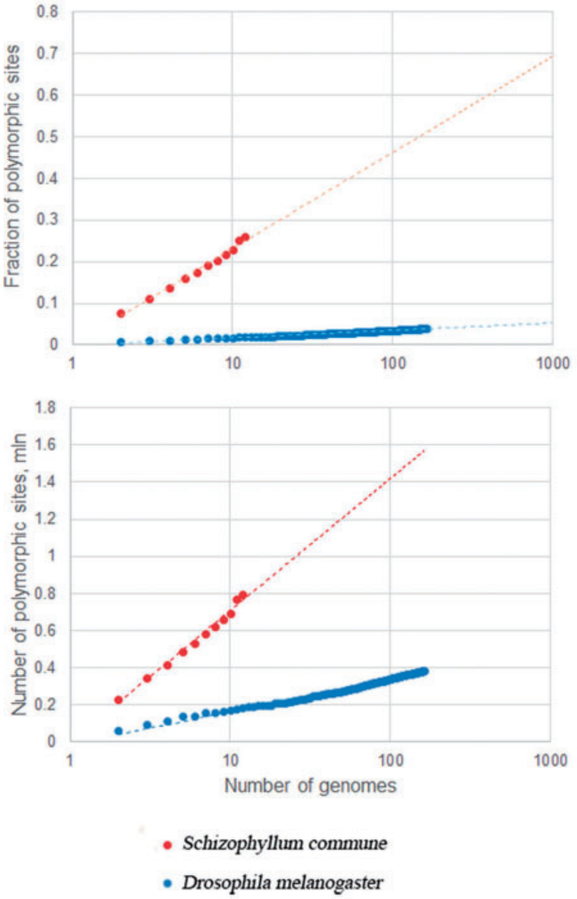

Fig. 4. Fractions and number of polymorphic sites in samples of different sizes of *Schizophyllum commune* and *Drosophila melanogaster*. Note the logarithmic scale of the horizontal axis.

Assembly, Alignment, and Annotation

Although a reference genome is available for *S. commune* (Ohm et al. 2010), for such a highly polymorphic species, mapping reads to the reference genome is difficult. Mapping with a small allowed number of mismatches results in few mapped reads, whereas mapping with a large number of allowed mismatches results in many reads mapped to two or more different genomic positions. Therefore, we decided to assemble each haploid genome de novo. The mean coverage depth was about 30–40 for each individual. Each genotype was assembled de novo using SPAdes (Bankevich et al. 2012) or Velvet (Zerbino and Birney 2008). SPAdes assembly was chosen for subsequent analyses, as it had much higher N50 values, compared with Velvet, varying between 48,928 and 104,000 between individuals, with the mean value across all 24 individuals of 68,688 (supplementary table S6, Supplementary Material online). The *S. commune* genome contains long diverged paralogous segments (supplementary fig. S7, Supplementary Material online) that usually result in chains of bulges in the de Bruijn graph. When paralogs differ substantially in read coverage, the assembler can regard the least covered paralog as erroneous and remove it in the course of the "Bulge Remover" procedure (Bankevich et al. 2012), leading to assembly errors. To minimize this problem and to keep genomic connections, we iteratively ran SPAdes with three different *k*-mer lengths ($k = 21$, 33, 55), and used soft parameters for de Bruijn graph cleaning. 21-mers extracted from contigs constructed at the iteration of $k = 21$ are included in the de Bruijn graph for $k = 33$. Similarly, the contigs constructed at the iteration of $k = 33$ are included in the de Bruijn graph for $k = 55$. This approach allows to fill some gaps in coverage. Paralogs correspond to chains of long highly covered bulges in the de Bruijn graph and, thus, can be detected based on de Bruijn graph structure. Multiple alignment with reference sequence of *S. commune* (Ohm et al. 2010) was created using multiz (Blanchette et al. 2004) and annotation 2.0 from JGI. Reliable segments of the alignment were selected for analyses, excluding regions not assembled in more than eight individuals and within 1,000 nt from such regions. In total, 7,987 genes were extracted according to the reference annotation from multiple alignment of reliable segments and realigned using macse (Ranwez et al. 2011). In an alternative approach, all genomes were annotated using Augustus (Stanke and Waack 2003) and 9,957 clusters of orthologous genes were obtained by orthomcl (Li et al. 2003) and realigned using macse (Ranwez et al. 2011); the results obtained using the two methods were nearly identical. In the results presented, the first method was used for all analyses except comparison with *C. cinerea*, where orthologs obtained from multiple alignments were used.

Polymorphism Estimation

Polymorphism level at 4FD sites was calculated as the average fraction of nucleotide differences over all pairwise genome comparisons of 4FD sites in each gene. Polymorphism level with Jukes–Cantor correction was measured for synonymous and replacement sites in exons, and for noncoding sites in introns, intergenic regions, and UTRs using Polymorphorama (Bachtrog and Andolfatto 2006). Jukes–Cantor substitution model assumes equal mutation rates between all possible pairs of nucleotides and equal nucleotide frequencies (Jukes and Cantor 1969), accounting for multiple mutations hitting the same site simultaneously segregating as polymorphism in our sample. π_{JC} was calculated as $-3/4 \times \ln(1 - 4/3\pi)$, where π is the fraction of differences between sequences, as implemented in Polymorphorama. Codon bias may lead to biased estimation of π; we corrected for this following (Dey et al. 2013). Specifically, we calculated the effective number of codons (ENC) using CodonW (Zhang et al. 2012) for one randomly chosen individual from the American population (A10) and from the Russian population (K1); plotted the linear regression between the Jukes–Cantor corrected values of π and ENC for each gene (supplementary fig. S2, Supplementary Material online); and estimated the value for codon bias-corrected $\pi_{\text{syn-JC}}$ from the regression line at ENC value of 61 which corresponds to uniform (neutral) codon usage.

Linkage Disequilibrium

We calculated LD as r^2 between all possible pairs of biallelic sites at given distance. For two biallelic sites *A*/*a* and *B*/*b*, r^2 is defined as

$$r^2 = \frac{([AB] \times [ab] - [Ab] \times [aB])^2}{[A][B][a][b]},$$

where values in square brackets are the frequencies of the corresponding haplotypes.

SNP Shadowing

The genome sequence and gene annotation of *C. cinerea* were downloaded from JGI (http://jgi.doe.gov/, last accessed July 25, 2015). Orthologous groups for *S. commune* and *C. cinerea* were determined using orthomcl (Li et al. 2003). In total, 3,096 orthologous groups that included the reference *S. commune* genome, all 24 individual *S. commune* genomes, and the *C. cinerea* genome were analyzed. Orthologous genes were aligned using macse.

Identification of Dips in Polymorphism

To identify dips in polymorphism, we calculated π separately for the American and Russian populations and for the joint sample in 100-bp nonoverlapping windows. We called a dip if π within a window was reduced by a given factor, compared with the previous window, and extended it as long as π in subsequent windows remained below this threshold.

Mutation Rate Estimation

For mutation rate estimation, we used the sequences of two parental individuals, one from the American population and one from Russian population, and their 17 offspring (Seplyarskiy et al. 2014). Offspring reads were mapped separately to both parental genotypes using SHRiMP (David et al. 2011). An offspring genotype consists of interleaved loci originating from different parents. Because of the high polymorphism level, a (haploid) genomic segment of moderate length in an offspring is typically identical to one of the parental genotypes (from which it originated), and differs from the other. Therefore, a single-nucleotide mutation would make the offspring genotype different from the more similar of the two parental genotypes at just one nucleotide site. We thus identified provisional de novo mutations as cases where the offspring genotype differs from one parental genotype at only one site, and from the other parental genotype, at one or more sites (supplementary fig. S5, Supplementary Material online). For this, we selected reads mapped to one parent with only one mismatch, and to the other parent, with one or more mismatches, and identified the positions of mismatches in the first parent. A high proportion of false SNP are known to be caused by mismapping. To reduce potential miscalls, we excluded the positions with very low (<10) or very high (>500) coverage and positions with polymorphism in more than 10% of reads using SamTools mpileup (Li et al. 2009). We also excluded positions in loci around crossover sites (Seplyarskiy et al. 2014), as reads at such loci could not be mapped to a parent unambiguously. All selected positions were curated manually using IGV (Thorvaldsdottir et al. 2013). All selected positions had good coverage, unambiguous alignment, and no reads supporting the alternative nucleotide in either parental or offspring genotypes. The only exception was a mutation in line 9, where 13 reads supported the alternative nucleotide variant. This position also had a suspiciously high coverage (251), and could not be supported by Sanger sequencing (see below; supplementary table S4, Supplementary Material online), leading us to exclude it from the count of de novo mutations. The mutation rate is estimated as the number of confirmed mutations divided by the number of callable sites. Callable sites exclude sites of low mapping quality or with high polymorphism level in mapped reads. To verify the mutations we used Sanger sequencing. For this, we designed primers that amplify the short (300–600 bp) region overlapping the provisional mutation site based on sequence alignment of offspring and parent. Polymerase chain reaction (PCR) was performed in two replicates using High-Fidelity 2X PCR Master Mix (New England Biolabs) and sequenced using ABI PRISM BigDye Terminator v. 3.1 on an Applied Biosystems 3730 DNA Analyzer (Life Technologies, USA). Nine of the ten provisional mutations were thus validated. The remaining mutation in line 9 met the formal requirements, but the final alignment with both parents had bad quality. The parental scaffold including this potential mutation was short (only 478 nucleotides), and could be an artifact of genotype assembly, consistent with abnormally high coverage and polymorphism in parental reads for this position, and small coverage by offspring reads (supplementary table S4, Supplementary Material online).

Data Accessibility

The data discussed in this publication have been deposited in the NCBI Sequence Read Archive (SRA). Alignments are available at makarich.fbb.msu.ru/baranova/Population_genetics_Scommune (last accessed July 25, 2015).

Acknowledgments

This work was partially supported by the Ministry of Education and Science of the Russian Federation (grant no. 11.G34.31.0008) and the Molecular and Cellular Biology Program of the Russian Academy of Sciences.

Supplementary Material

Supplementary text S1 and S2, figures S1–S9, and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

References

Bachtrog D, Andolfatto P. 2006. Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* 174:2045-2059.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19:455-477.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14:708-715.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299:1391-1394.

Cutter AD, Jovelin R, Dey A. 2013. Molecular hyperdiversity and evolution in very large populations. *Mol Ecol.* 22:2074-2095.

David M, Dzamba M, Lister D, Ilie L, Brudno M. 2011. SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics* 27:1011-1012.

Dey A, Chan CK, Thomas CG, Cutter AD. 2013. Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. *Proc Natl Acad Sci U S A.* 110:11056-11060.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.* 19:11-15.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862-864.

Gundry M, Vijg J. 2012. Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. *Mutat Res.* 729:1-15.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet.* 38:226-231.

Hodgkinson A, Eyre-Walker A. 2010. Human triallelic sites: evidence for a new mutational mechanism? *Genetics* 184:233-241.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. New York: Academic Press. p. 21–132.

Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics* 196:313-320.

Kondrashov FA, Kondrashov AS. 2010. Measurements of spontaneous rates of mutations in the recent past and the near future. *Philos Trans R Soc Lond B Biol Sci.* 365:1169-1176.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488:471-475.

Kovaleva GY, Bazykin GA, Brudno M, Gelfand MS. 2006. Comparative genomics of transcriptional regulation in yeasts and its application to identification of a candidate alpha-isopropylmalate transporter. *J Bioinform Comput Biol.* 4:981-998.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* 10:e1001388.

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

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178-2189.

Lynch M. 2010. Evolution of the mutation rate. *Trends Genet.* 26(8):345–352.

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482:173-178.

Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, de Vries RP, Record E, Levasseur A, Baker SE, et al. 2010. Genome sequence of the model mushroom *Schizophyllum commune*. *Nat Biotechnol.* 28:957-963.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* 6:e22594.

Seplyarskiy VB, Kharchenko P, Kondrashov AS, Bazykin GA. 2012. Heterogeneity of the transition/transversion ratio in *Drosophila* and *Hominidae* genomes. *Mol Biol Evol.* 29:1943-1955.

Seplyarskiy VB, Logacheva MD, Penin AA, Baranova MA, Leushkin EV, Demidenko NV, Klepikova AV, Kondrashov FA, Kondrashov AS, James TY. 2014. Crossing-over in a hypervariable species preferentially occurs in regions of high local similarity. *Mol Biol Evol.* 31(11):3016-3025.

Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ. 2004. Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res.* 32:1774-1782.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res.* 23:23-35.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl. 2):ii215–225.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14:178-192.

Venn O, Turner I, Mathieson I, de Groot N, Bontrop R, McVean G. 2014. Nonhuman genetics. Strong male bias drives germline mutation in chimpanzees. *Science* 344:1272-1275.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821-829.

Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, Yu J. 2012. Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics* 13:43.