Supplementary methods

Resequencing, assembly and SNP calling of 92 *Neurospora tetrasperma* genomes

Illumina 500 bp paired-end libraries were prepared from each sample at BGI in Hong Kong and each sample was sequenced on Illumina HiSeq 2000 (Illumina, San Diego, CA), producing paired-end reads of 90bp in length. Filtering of the resulting fastq files included trimming of the first 5bp of low quality bases from each read, removal of duplicate reads, reads with a proportion of N's greater than 10% and a number of low quality bases (Q≤20) greater than 40. This filtering of the fastq files resulted in between 1,102 – 1,905 Mbp of sequence data produced per *N. tetrasperma* strain (Table S2).

The filtered reads for each strain were mapped to the *N. tetrasperma* 2509 (mating type a) reference genome (http://genome.jgi-psf.org/Neute_mat_a1/Neute_mat_a1.download.ftp.html) using BWA (v0.6.1) with -I flag for Illumina quality scores, -n set to 5 and the remainder of parameters set to their default values (Li and Durbin 2009). This procedure resulted in reference-assemblies with a genome-wide average depth ranging from 26X to 48X (Table S1). Following read mapping, indel realignment was carried out using the GATK RealignerTargetCreator and IndelRealigner tools. The realigned BAM files produced were then used for variant calling and genotyping with the GATK Unified Genotyper. The Unified Genotyper was run with the following parameters --ploidy1 and--output_mode EMIT_ALL_SITES (DePristo et al. 2011). Variant sites with a QUAL < 60, QD < 2.0 and FS > 60 were filtered out and a genotype calls for a strain was included if the individual read depth at that sites was greater than or equal to 8 (DP ≥8) and the genotype quality was greater than or equal to 40 (GQ≥40). Sites within 5bp of a called indel and sites falling within repetitive regions of the genome were also filtered out.

In addition, we performed *de novo* assemblies of all strains by using SOAPdenovo (version 1.05)(Li et al. 2010). Based on the comparison of number of scaffolds and N50 values obtained when using different kmers, we chose 37 as kmer size in the assemblies, and kept other parameters as default.

Analyses of divergence times, intron gain and loss rate and positional biases

Divergence time (MY) between *N. crassa* and *N. tetrasperma* was calculated as (genomic synonymous divergence between the species, dS)/(2*synonymous nucleotide substitution rate,

μ). Genomic dS was calculated based on the 5,723 autosomal ortholog alignments among *N. crassa*, *N discreta* and *N. tetrasperma a* by the program codeml implemented in the PAML package, with assumption of a constant molecular clock (Yang 2007). The synonymous nucleotide substitution rate (μ) was retrieved from the study by Kasuga et al. (2002) in their use of the Langley Fitch algorithm and the calibration time points of 400 and 670 Mya between Eurotiomycetes and Sordariomycetes. Rates estimated by Kasuga et al., (2002) using the 310 Mya calibration time was discarded, as it is too recent judging from fossil record evidence (Taylor et al. 1999). Furthermore, in order to be conservative, loci and species pairs with substitution rates higher than $2.0x10^{8}$ was considered outside of the regular range (Kasuga et al. 2002), and thus not included in the calculation. Intron gain and loss rates were calculated for the *N. crassa* and *N. tetrasperma* lineages as follows: gain rate = No. gained introns/((total coding nucleotide sites – ancestral intron No.)*divergence time) and loss rate = 1 - (No. of retained ancestral intron/No. of ancestral introns)$^{(1/\text{divergence time})}$ (Roy and Gilbert 2005; Lynch 2007).

We compared the differences in intron gain and loss numbers between the two species, and between the SR and R regions in *N. tetrasperma*, by Pearson chi-square tests. The expected number of intron gains/losses for a given lineage was calculated from the product of the total number of gains/losses multiplied by the number of nucleotide sites in a given lineage, divided by the sum over numbers of nucleotide sites for all lineages.

References:

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**(5): 491-498.

Kasuga T, White TJ, Taylor JW. 2002. Estimation of nucleotide substitution rates in Eurotiomycete fungi. *Mol Biol Evol* **19**(12): 2318-2324.

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

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**(2): 265-272.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Associates, Sunderland, Mass.

Roy SW, Gilbert W. 2005. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci U S A* **102**(16): 5773-5778.

Taylor TN, Hass H, Kerp H. 1999. The oldest fossil ascomycetes. *Nature* **399**(6737): 648.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8): 1586-1591.