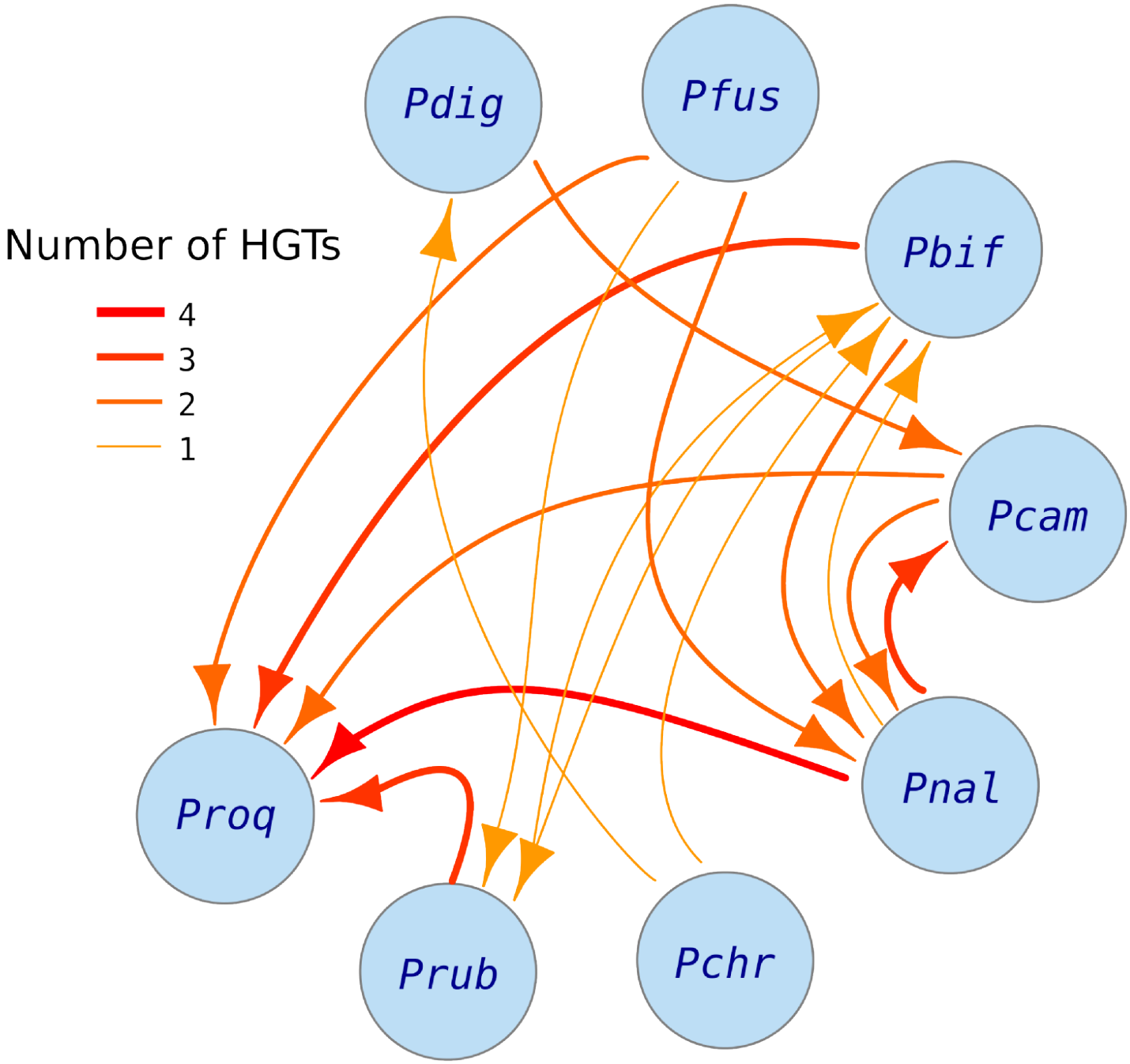
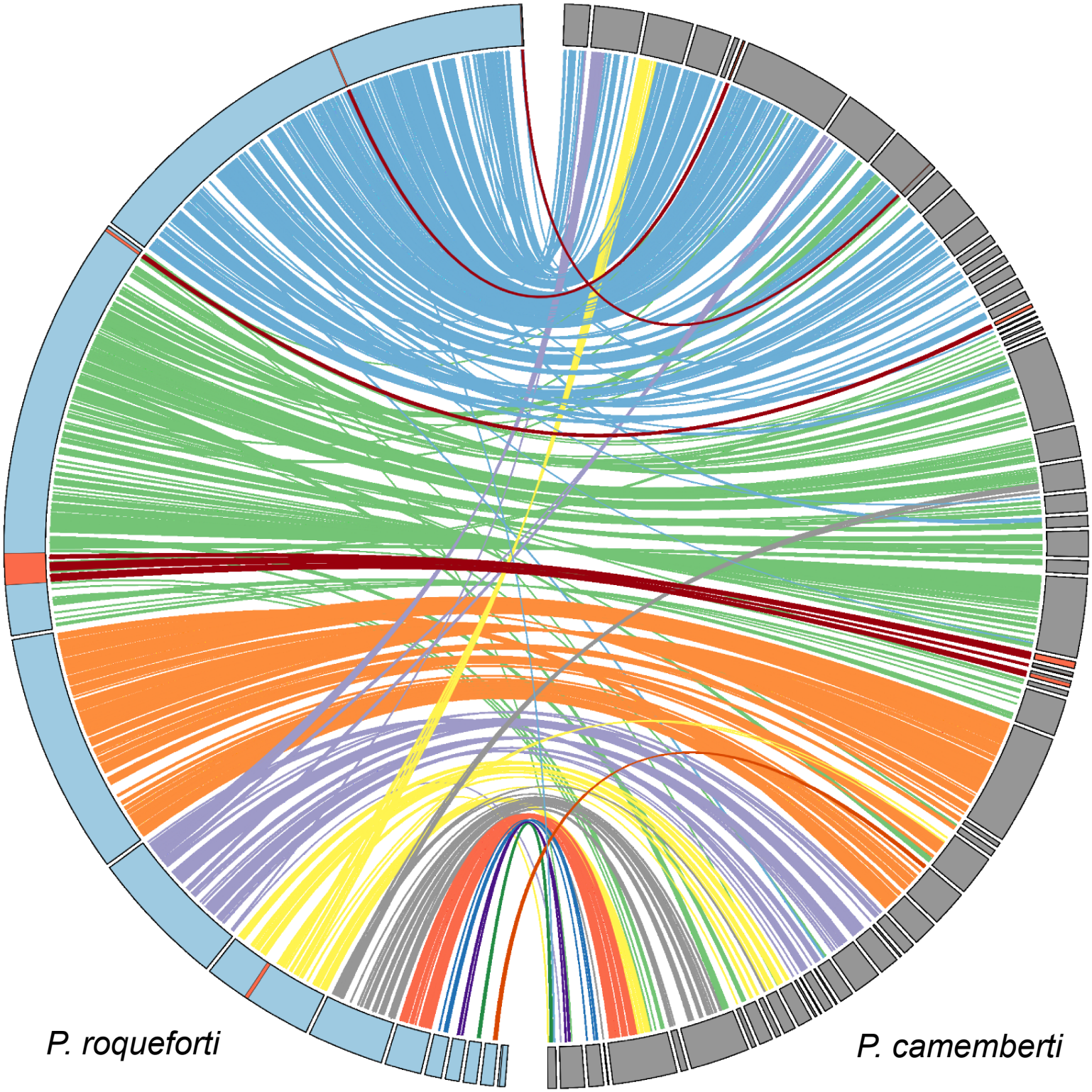
Current Biology

Supplemental Information

Adaptive Horizontal Gene Transfers between Multiple Cheese-Associated Fungi

Jeanne Ropars, Ricardo C. Rodríguez de la Vega, Manuela López-Villavicencio, Jérôme Gouzy, Erika Sallet, Émilie Dumas, Sandrine Lacoste, Robert Debuchy, Joëlle Dupont, Antoine Branca, and Tatiana Giraud


Number of HGTs
4
3
2
1
Pdig
Pfus
Pbif
Pcam
Proq
Pnal
Prub
Pchr


P. roqueforti
P. camemberti

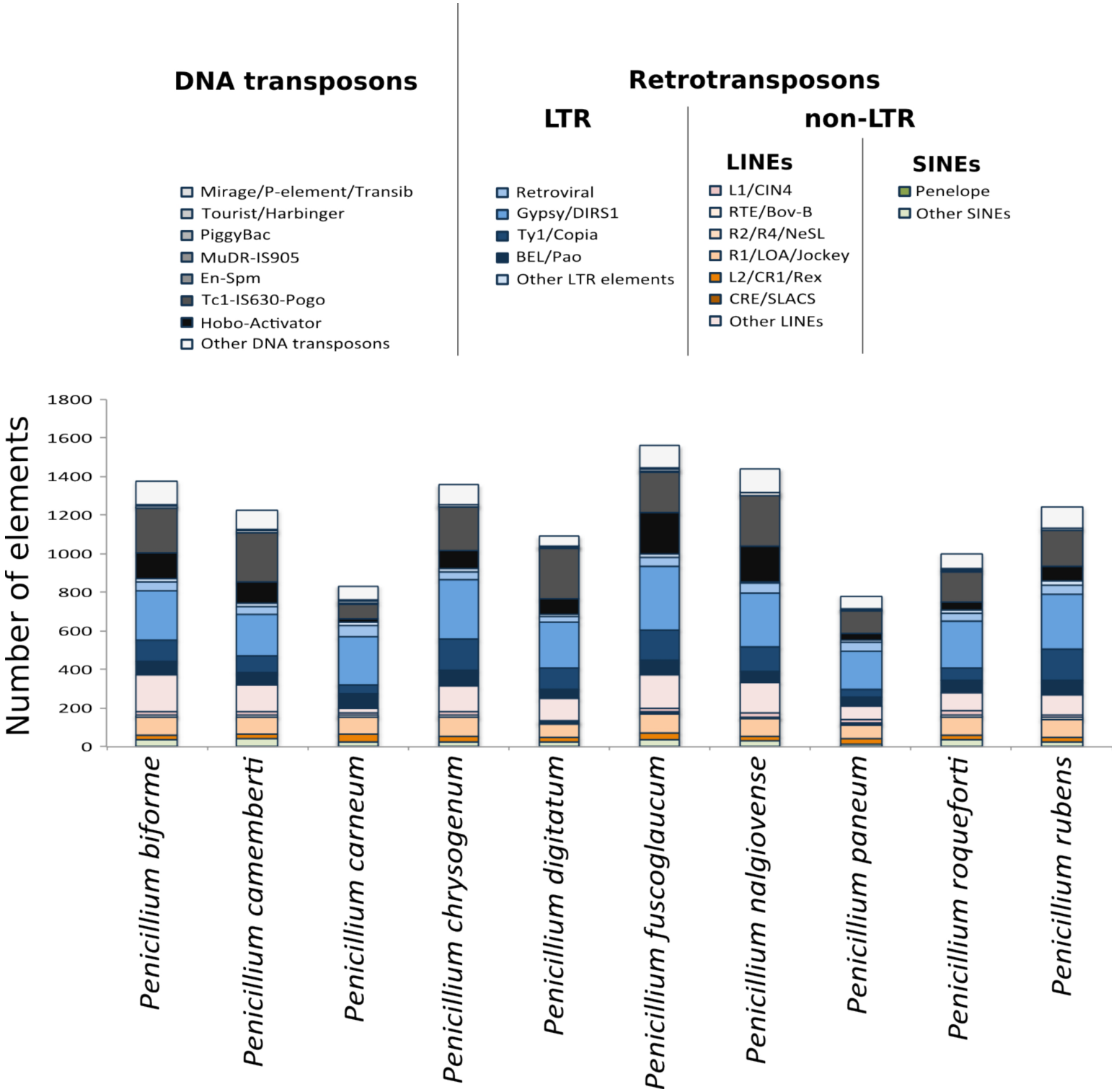

DNA transposons
Mirage/P-element/Transib
Tourist/Harbinger
PiggyBac
MuDR-IS905
En-Spm
Tc1-IS630-Pogo
Hobo-Activator
Other DNA transposons
Retrotransposons
LTR
Retroviral
Gypsy/DIRS1
Ty1/Copia
BEL/Pao
Other LTR elements
non-LTR
LINEs
L1/CIN4
RTE/Bov-B
R2/R4/NeSL
R1/LOA/Jockey
L2/CR1/Rex
CRE/SLACS
Other LINEs
SINEs
Penelope
Other SINEs
Number of elements
0
200
400
600
800
1000
1200
1400
1600
1800
Penicillium biforme
Penicillium camemberti
Penicillium carneum
Penicillium chrysogenum
Penicillium digitatum
Penicillium fuscoglaucum
Penicillium nalgiovense
Penicillium paneum
Penicillium roqueforti
Penicillium rubens

Figure S1 (related to figure 1): Oriented graph showing the direction and number of horizontal gene transfers between *Penicillium* species as inferred by the Notung software. The direction of each event is represented by the direction of the arrow, and the number of events by the width of the arrow and its color (the redder the more events). The figure was generated with the R package igraph [S1].

Figure S2 (related to figure 1): Synteny between the *Penicillium roqueforti* and *P. camemberti* genomes. Ortholog blocks larger than 10 kilobases are linked between the two genomes. Each color represents a specific *P. roqueforti* scaffold. Only the scaffolds 1 to 12 are represented for *P. roqueforti* as no matches of more than 10 kilobases have been found for the 36 other scaffolds. The red bands on the circle linked by dark red lines correspond to the horizontally transferred regions.

Figure S3 (related to figure 1): Classification and number of transposable elements in each *Penicillium* genome. In most of the genomes, *mariner* elements constitute the largest group of DNA transposons, while *gypsy* and *copia* elements represent most of the LTR retrotransposons, and *i* elements most of the LINE elements. Very few SINE elements were detected.

	P. biforme	*P. camemberti*	*P. fuscoglaucum*	*P.carneum*	*P. paneum*	*P. roqueforti*	*P. digitatum*	*P. rubens*	*P. chrysogenum*	*P. nalgiovense*
	FM169	FM013	FM041	LCP05634T	FM227	FM164	PHI26	Wisconsin54-1255		FM193
Substrate	Cheese environment	Cheese environment	Cheese environment	Mouldy rye bread	Cheese environment	Cheese environment	Citrus contaminant	Mouldy cantaloupe	Contaminant in the Postia Placenta MAD 698R	Cheese environment
Accession number	Genbank: HG813601-HG814182	Genbank: HG793134-HG793313	Genbank: HG814183-HG815135	Genbank: HG816029-HG818118	Genbank: HG813308-HG813531	Genbank: HG792015-HG792062	Genbank: JH993687-JH993786	Genbank: AM920416-AM920464	www.jgi.doe.gov	Genbank: HG815136-HG815288, HG815290-HG816004
Authors	This study	Cheeseman *et al.*, 2014*	This study	This study	This study	Cheeseman *et al.*, 2014*	Marcet-Houben *et al.*, 2012†	Van den Berg *et al.*, 2008‡	Department of Energy Joint Genome Institute	This study
Number of scaffolds	582	180	953	2,090	224	48	100	49	27	869
Total length of scaffolds	34,874,608	35,011,981	36,085,495	25,936,550	26,579,303	29,010,044	25,956,151	32,223,735	31,340,922	31,937,463
N50 (bp)	277,312	947,716	245,012	24,373	416,003	7,752,559	878,909	3,889,175	5,340,701	306,303
Maximum length (bp)	1,223,141	2,612,452	1,273,777	107,387	1,211,662	8,605,340	4,553,353	6,387,817	7,361,319	930.736
Minimum length (bp)	1,008	1,054	1,000	2,000	1,046	2,101	964	1,032	2,067	1,002
GC%	48.09	48.18	47.74	48.73	49.01	48.64	48.9	48.96	48.78	48.48
Nb of protein coding genes	15,246	14,525	15,615	13,412	12,201	12,319	9,133	12,591	11,396	14,503
Mean gene length (bp)	1264.8	1294.0	1253.2	1128.4	1243.9	1243.2	1390.1	1340.7	1407	1202.3
Transposable Elements (bp)	446,173	357,683	525,762	90,938	154,502	298,483	287,550	370,483	445,385	545,631

*Cheeseman, K. et al. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nat. Commun. 5, 2876 (2014).
†Marcet-Houben, M. et al. Genome sequence of the necrotrophic fungus Penicillium digitatum, the main postharvest pathogen of citrus. BMC Genomics 13, 646 (2012).
‡van den Berg, M. a et al. Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. Nat Biotechnol 26, 1161–1168 (2008).

Table S1 (related to figure 1): *Penicillium* genome sequencing, assembly, and annotation summary.

Table S2 (related to figure 2): Time-wise expression of *Penicillium camemberti* genes during cheese rind maturation; these data were generated in a previous study [42].

Table S3 (related to figures 1 and 2): Presence/absence of the horizontally-transferred regions *Wallaby* and *CheesyTer* in 416 different isolates from diverse origins belonging to 65 fungal species. Strains isolated from the cheese environment are shown in purple and positive PCR amplifications are represented in yellow. Type, neotype and epitype strains are indicated with respectively a "T", a "NeoT" and a "EpiT" after the strains number. All strains included in this study are available in the public fungal strain collection of the National Museum of Natural History in Paris, except the strains labeled "FM", corresponding to a private collection provided by French anonymous stakeholders. Only the six FM strains which genomes are available (*e.g.,* FM013, FM041, FM164, FM169, FM193 and FM227, indicated with an asterisk) have LCP numbers and are publicly available.

Table S4 (related to figure 3): Growth measures (in mm) obtained in the fitness experiments. **a)** Experiments assessing the ability of using the cheese substrate: growth measures of 50 *Penicillium roqueforti* strains on minimal medium (MM) and cheese medium. **b)** Experiments assessing competitive ability: growth measures of three challengers, i.e., *P. biforme*, *P. nalgiovense* and *Geotrichum candidum*, inoculated in the middle of Petri dishes on *P. roqueforti* lawns (N=23). **c)** Experiments assessing competitive ability: growth measures of 12 challengers belonging to *P. biforme*, *P. camemberti* or *P. rubens*, inoculated on *P. roqueforti* lawns (2 W+C+ and 2 W-C-). **d)** Experiments assessing competitive ability: growth measures of 19 *P. roqueforti* strains on cheese medium, inoculated pairwise in two opposite points of Petri dishes.

Supplemental Experimental Procedures

DNA extraction, sequencing and annotation

DNA was extracted as previously described [16]. The genomes of the strains *Penicillium biforme* FM 169, *P. fuscoglaucum* FM 041, *P. nalgiovense* FM 193, *P. paneum* FM 227 and *P. carneum* LCP05634 (CBS number: 112297) were sequenced at the BGI (Beijing Genomics Institute, Beijing, China), with Illumina HiSeq2000 Solexa sequencing technology. Paired-end genomic DNA libraries were prepared with an insert size of either 500 bp (*P. biforme* FM 169, *P. fuscoglaucum* FM 041, *P. nalgiovense* FM 193) or 700 bp (*P. paneum* FM 227, *P. carneum* LCP05634), and reads of 90 bp were obtained. Sequences were assembled at the BGI, with SOAPdenovo v1.05 [S2]. This software generated contigs, for kmers of 27, 35, 43, 47 and 51 for *P. carneum* LCP05634, *P. paneum* FM 227, *P. fuscoglaucum* FM 041, *P. biforme* FM 169, *P. nalgiovense* FM 193, respectively. Gap filling and single base correction were performed with SOAPaligner.

Gene models were predicted with EuGene [S3], a highly integrative eukaryotic protein-coding gene prediction platform previously trained for *Penicillium* [16]. Similarities to sequences present in a GenBank database of all Eurotiales proteins (release 2011/02/11), the proteome of *Saccharomyces cerevisiae* and the proteomes of three *Penicillium* genomes (*P. camemberti* FM013*, P. roqueforti* FM164 and *P. chrysogenum* Wisconsin 54-1255), were used as backbones to guide EuGene predictions. Initial functional annotations of protein domains on draft genomes were generated by applying InterProScan version v4.8 [S4] to the InterPro database (release 43.1). Functional and pathway annotations were obtained by directly mapping the InterPro

identifiers (InterPro release 48) onto GO. Information about the genomes is provided in Table S1.

Ortholog detection and multiple sequence alignments

Whole protein-coding gene sets from 10 *Penicillium* and four *Aspergillus* genomes (178,235 sequences in total) were blasted against each other with BLASTP and clustered with OrthoMCL V1.4 software and the default parameters [S5]. The *Aspergillus* ORF coding versions were retrieved directly from AspGD [18]: *A. fumigatus* Af 293 s03-m04-r21, *A. nidulans* FGSC_A4 s10-m03-r04, *A. niger* CBS_513_88 s01-m06-r09 and *A. oryzae* RIB40 s01-m08-r20. Almost 89% of the sequences clustered into one of the 18,536 orthologous groups (hereafter referred to as orthologous groups). There was a median of seven genes per group. The genes not assigned to clusters were considered to be species-specific. MUSCLE [S6] was then used to align the sequences of the proteins of each orthologous group, and the resulting amino-acid sequence alignments were used as templates for codon-based nucleotide alignments, generated with TranslatorX [S7].

Annotation and analyses of interspersed repeats

Transposable elements (TEs) were detected and classified with RepeatMasker version 4.0.3, using the rmblast algorithm [S8] and the repbase RepeatMasker libraries (01/31/2014 release). Total interspersed repeat content was low, ranging from 0.35% to 1.71% of the entire genome assembly for the 10 genomes. The three sister species, *P. carneum*, *P. paneum* and *P. roqueforti,* had genomes consisting of less than 1% TEs, this proportion being the lowest among the

Penicillium genomes. The most widespread classes of detected TEs were LTR retroelements and DNA transposons and, to a lesser extent, non-LTR retroelements (Figure S3). The most abundant elements were *mariner* DNA transposons of the *Tc-IS360-Pogo* type, which accounted for 16 to 34 % of the genome sequences attributed to TEs. The second most widespread elements were non-LTR *i* elements, corresponding to 10 to 40% of the total genome sequence attributed to TEs. The cheese fungi *P. camemberti* and *P. roqueforti* had the highest proportions of *i* elements relative to the total length of TE sequences, 39 and 36%, respectively, the mean value for *Penicillium* genomes being 25%. The third and fourth most widespread elements were *gypsy* and *copia* LTR retroelements, which accounted for 12 to 24% and 3 to 15 % of the genome sequences attributed to TEs, respectively. Together, the *mariner*, *i*, *gypsy* and *copia* elements accounted for 62 to 83% of the genome sequences attributed to TEs. We detected 37 TE families expanded only in the *P. digitatum* lineage. This represents a relatively large number of families, given that fewer than 10 TE families were specifically expanded in each of the other lineages. Nevertheless, *Penicillium* genomes generally had few interspersed repeats, slightly fewer, even, than the closely related *Aspergillus* [6].

Species and gene genealogy reconstruction

Species trees were reconstructed from the concatenation of alignments of single-copy genes present in at least 10 species, including either or both *P. roqueforti* and *P. camemberti* orthologs, and producing congruent phylogenetic trees. We first constructed maximum likelihood (ML) phylogenetic trees independently of all 6,275 single-copy ortholog alignments, using RAxML

[S9]. We constructed a meta-alignment of the concatenated 3,089 orthologous group alignments (176,4879 codons in total), to obtain a single maximum likelihood phylogenetic tree with RAxML [S9]. For each of the species missing from the 1,030 incomplete alignments, we completed the missing sequence with gap characters (-). Branch support was determined by a bootstrap analysis of 100 resampled data sets. All nodes in the tree based on the concatenated alignment had 100% bootstrap support. ML trees were also reconstructed for two subsets of orthologous groups containing more than one gene in either or both of *P. camemberti* and *P. roqueforti*.

Horizontal gene transfers detection through gene-species tree reconciliation

Gene genealogies were reconciled with the fully resolved species phylogeny using Notung in order to reliably infer HGTs [25-27]. First, to obtain the most conservative bifurcating gene genealogies, the root and the poorly supported nodes (Bootstrap values below 95%) in individual gene genealogies were resolved according to the species tree. Then, the Notung reconciliation mode was run for each gene genealogy with a duplication cost of 1.5, a loss cost of 1 and a transfer cost of 4.5 (thus assuming low HGT probability) and with the prune option enabled. Notung reconstructs each gene history by minimizing the total cost and also tests for the temporal feasibility of HGT events. Finally, the inferred transfer events that may have actually resulted from incomplete lineage sorting were filtered out, *i.e.,* we did not consider inferred transfers between lineages separated by a single node in the tree.

Macrosynteny

We aligned genomes pairwise with *P. roqueforti* genome, using YASS software [S10] and the following parameters: -O 100000 -C 5,-4,-3,-4 -E 10 -G -16,-4 -W 20,40000 -X 25 -c 1 -d 1 -e 2.8 -i 10 -m 25 -p "#@##-#---##---#--@####,####-#@#@#--##-##" -r 2 -s 70 -w 4. Only orthologous blocks of more than 10 kilobases were retained. The graphical representation of pairwise identity between orthologous blocks was generated with Circos [S11].

Penicillium camemberti **gene expression in cheese rind**

Transcriptome data for Camembert cheese rind were recovered from the NCBI short-read archive, for project SRP030470 samples SRR1002975 to SRR1002981 [42]. The first 25 positions of the 454 reads were trimmed, to remove adapter sequences. The reads were also trimmed on the basis of quality, by removing all 10 bp windows with a minimum aggregate quality of less than 20 from the reads. The reads were mapped onto the *P. camemberti* FM013 genome with Tophat2 [S12], and transcript abundances were estimated with cufflinks v2.2.0 [S13]. The following parameters were used for mapping: all positions, alignment of reads to all possible locations, maximum mismatch threshold of 15%, minimum alignment threshold of 0.5%, 8 processors, Smith-Waterman bandwidth of 235, alignment candidate threshold of 55 bp, hash position threshold of 200 and a homopolymer gap open penalty of 4. The levels of expression of the genes were estimated with cufflinks [S13] as Fragment per Kilobase per Millions of Reads (FPKM). Transcripts were clustered according to their expression level by hierarchical clustering

using the hclust() function of R.

An analysis of gene expression data from a previous study [42] showed that few genes were expressed constitutively during the 77 days of rind maturation. Over the full gene set of *P. camemberti*, we found 4,912 genes expressed on at least one day of the experiment (Table S2). We then focused on expressed genes encoded within the *Wallaby* and *CheesyTer* regions (Figure 2B).

Two genes located in *CheesyTer* were strongly expressed only at day 5, corresponding to the beginning of rind maturation: PcamFM013S059g000011, encoding a putative general substrate transporter, and PcamFM013S059g000012, encoding a putative β-galactosidase. During cheese maturation, lactose is present only during the first few days, during which it constitutes a primary source of carbon that is rapidly consumed by lactic acid bacteria, yeast and molds [36]. The possession of genes involved in lactose metabolism and transport, and the strong expression of these genes early in rind maturation, should be advantageous in the cheese environment. Two highly expressed genes from *Wallaby* were clustered next to each other, namely PcamFM013S057g00033 and PcamFM013S057g00034. None of these genes displayed significant similarity to sequences in public databases, and they were both found to be strongly expressed at later stages of rind maturation (days 21, 56 and 77). The patterns of expression of *Wallaby* and *CheesyTer* genes suggest that these regions could be involved in different stages of cheese maturation, reflecting changes in the substrate or in the organisms with which *Penicillium* fungi interact at different stages [14].

Finally, we found seven expressed genes in HTR2 and none in HTR1, the two other HTRs shared

between *P. camemberti* and *P. roqueforti*. Among the seven expressed genes of HTR2, PcamFM013S067g000021 was highly expressed at day 5. However no putative function could be assigned to this gene.

DNA extraction for PCR amplification, PCR amplification and sequencing of amplicons

Genomic DNA was extracted from fresh mycelium of the isolates listed in Table S3. Mycelium was obtained after 3–5 days on malt agar for *Aspergillus, Colletotrichum, Eupenicillium, Penicillium*, *Scopulariopsis* and *Fusarium* strains and on a confidential medium provided by starter producers for *Sporendonema casei*. The Qiagen DNeasy Plant Mini Kit (Qiagen Crawley, UK) was used for DNA extraction.

PCR was performed in a volume of 50 µL, containing 25 µL template DNA, 1.25 U AmpliTaq DNA polymerase (Roche Molecular Systems, Branchburg, NJ, USA), 5 µL 10 × *Taq* DNA polymerase buffer, 2 µL 5 mM dNTPs, 2 µL of each 10 µM primer. Primer sets designed to detect the presence of *Wallaby* (primers named Wallaby 1, 2 and 3) [16] and *CheesyTer* (primers named *CheesyTer* 1, 2 and 3) are shown in the section below. For *Wallaby,* results had been generated in our previous study [16]. For *CheesyTer,* amplifications were performed using a touchdown program with 20 cycles of 5 min at 94°C, a decreasing of 0.5°C every 30 sec from 60 to 50°C, and 30 sec at 72°C, followed by 35 additional cycles of 30 sec at 94°C, 30 s at 50°C, 30 sec at 72°C; the PCR program was followed by a final 30 s extension step at 72°C. PCR products were purified and sequenced by Eurofins (France). Further primer pairs were designed to confirm the *Wallaby* presence (as above) or absence (primer pair named *Wallaby* Ø) in 12 isolates of *P. roqueforti* (symbols in red in Table S3. PCR amplifications validating *Wallaby* absence were run with 30 cycles of 30 s at 95 ºC, 30 s at 60 ºC and 30 s at 72 ºC followed by a final extension of 1.5 min at 72 ºC.

Primer sets used to test the presence of *Wallaby* and *CheesyTer* in the fungal collection

Wallaby 1	Forward: 5'-GGCTCCTGCAGTAGACCAAG-3'
	Reverse: 5'-TTTGCCTTGGCTGAGTCTTT-3'
	Gene included: PROQFM164_S02g002809
Wallaby 2	Forward: 5'-CGCTATAACGGAAGCAGGAG-3'
	Reverse: 5'-CCCGTCCTTTTTAGACACCA-3'
	Gene included: PROQFM164_S02g002870
Wallaby 3	Forward: 5'-GCAACAGATCCAGCAGTTCA-3'
	Reverse: 5'-CGTTTCTCCTGAGTCCTTGC-3'
	Gene included: PROQFM164_S02g002951
Wallaby Ø	Forward: 5'-AATGATCCAGTGTACGGGCT-3'
	Reverse: 5'-AAGAGACAAGCAAAGGAGGC-3'
CheesyTer 1	Forward: 5'-ATCTCGACGACATAGGCAGG-3'
	Reverse: 5'-GGAGCGTGCAGATGAGAAAG-3'
	Gene included: PROQFM164_S02g000030
CheesyTer 2	Forward: 5'-CGTACGTCAAGCCTGGAATG-3'
	Reverse: 5'-AAGGCCATCATCATCGACCT-3'
	Gene included: PROQFM164_S02g000037
CheesyTer 3	Forward: 5'-CCCAAAGCCATGACCAACAA-3'

Reverse: 5'-ACGTCTTCCTGTCCGACAAT-3'

Gene included: PROQFM164_S02g000032

Fitness experiments: competition and use of the cheese substrate

Experiments were performed in duplicate, in 8.5 cm Petri dishes. Goat's cheese medium was prepared as described in a previous study [22]. We first compared the growth rates of 50 *P. roqueforti* strains (26 W+C+ and 24 W-C-, Table S4a). Experiments were performed on a minimal medium, in which sucrose was the only available carbon source, and the cheese medium. Plates were placed in an incubator, at 25°C, in the dark, for seven days. We then measured two perpendicular diameters for each colony, to assess colony size.

We then measured the impact of several *P. roqueforti* strains, carrying or lacking *Wallaby* and *CheesyTer* on the growth of three fungal strains: *P. biforme* LCP05529 (W-C-), *P. nalgiovense* FM193 (W-C-) and *Geotrichum candidum* FM074 (W-C-). We performed competition experiments on cheese, malt agar and minimal media. We first prepared a suspension of fungal spores from cultures of each of 23 *P. roqueforti* strains (11 W+C+, 12 W-C-). This last class included eight isolates from cheeses and five isolates from wild environments (see Table S4b for the list of strains used). Spores of *P. roqueforti* were collected from the surface of five-day cultures and vortexed in a solution of water and Tween. We then poured 50 µl of spore suspension for each *P. roqueforti* strain into Petri dishes containing cheese, malt agar and minimal media (two replicates for each treatment). The three competitors were inoculated at a

single point at the center of the plates (5 µL each), 24 h after the plating of *P. roqueforti.* We also inoculated the three different media with the three competitors in the absence of *P. roqueforti* lawns. Plates were incubated in a refrigerated and ventilated FOC 225I incubator at 15°C in the dark (mimicking cheese cellar conditions) for 11 days. Colony diameters were measured after 12 days (two perpendicular measurements per colony). In a second set of this experiment, we inoculated and measured, as described above, using four *P. roqueforti* strains as lawns (2 W+C+ and 2 W-C-), on cheese medium, different strains of species displaying a polymorphism in the *Wallaby* and/or *CheesyTer* presence. We used as challengers different strains of *P. camemberti* (W+C- or W+C+, N=4), of *P. biforme* (W-C- or W+C+, N=4) and of *P. rubens* (W-C- or W+C-, N=4); see Table S4c for the list of strains used.

In the third experiment, we compared the growth of two *P. roqueforti* strains in face-to-face confrontations, after inoculations in opposite points of Petri dishes. We used 11 W-C- strains and 8 W+C+ strains (Table S4d) and performed all pairwise face-to-face confrontations on cheese medium. Plates were incubated at 25°C in the dark and observed daily. Colony growths were measured after 8 days, when the mycelia of the two strains met in the Petri dishes. We measured the deviations from symmetrical growth that would lead to a boundary between strains exactly in the middle of the Petri dish. Therefore, deviation is defined as:

$$Deviation = \frac{D_{focal\ strain} - D_{other\ strain}}{2}$$

where $D_{focal\ strain}$ is the distance between the inoculation point and the edge of the growth of the focal strain and $D_{other\ strain}$ is the distance between the inoculation point and the edge of the growth of the competing strain.

For the W+C+ *versus* W+C+ and W-C- *versus* W-C- confrontations, deviations were measured as positive or negative values, the focal strain being chosen at random. For the W+C+ *versus* W-C- confrontations, relative deviations were measured as positive or negative values, taking the W+C+ strain as the focal strain. The distributions of the deviations did not significantly depart from normal distributions (Shapiro Wilk tests W-C- *versus* W-C-: W=0.99, *p*-value=0.96; W-C- *versus* W+C+: W=0.98, *p*-value=0.36; W+C+ *versus* W+C+: W=0.98, *p*-value=0.88).

Statistical analyses of competition and fitness experiments

All statistical analyses were performed with JMP, version 9 (SAS Institute). ANOVAs were carried out by first running full models including all factors and interactions. We then sequentially removed the non-significant interactions.

Supplemental References

S1. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research, http://igraph.org. InterJournal, Complex Systems 1695.

S2. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience *1*, 18.

S3. Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., Van de Peer, Y., Rouze, P., and Schiex, T. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Current Bioinformatics *3*, 87-97.

S4. Zdobnov, E., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.

S5. Li, L., Stoeckert, C.J., and Roos, D. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research *13*, 2178-2189.

S6. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. *32*, 1792-1797.

S7. Abascal, F., Zardoya, R., and Telford, M. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. *38*,

W7-13.
S8. Smit, A., Hubley, R., and Green, P. (2013-2015). RepeatMasker Open-4.0 http://www.repeatmasker.org
S9. Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics *21*, 456-463.
S10. Noé, L., and Kucherov, G. (2005). YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research *33*, W540-W543.
S11. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Research *19*, 1639-1645.
S12. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.
S13. Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics *27*, 2325-2329.