nature genetics
OPEN

The wheat powdery mildew genome shows the unique evolution of an obligate biotroph

Thomas Wicker[1,9], Simone Oberhaensli[1,9], Francis Parlange[1], Jan P Buchmann[1,8], Margarita Shatalina[1], Stefan Roffler[1], Roi Ben-David[1,8], Jaroslav Doležel[2], Hana Šimková[2], Paul Schulze-Lefert[3], Pietro D Spanu[4], Rémy Bruggmann[5], Joelle Amselem[6], Hadi Quesneville[6], Emiel Ver Loren van Themaat[3], Timothy Paape[7], Kentaro K Shimizu[1,7] & Beat Keller[1]

Wheat powdery mildew, *Blumeria graminis forma specialis tritici*, is a devastating fungal pathogen with a poorly understood evolutionary history. Here we report the draft genome sequence of wheat powdery mildew, the resequencing of three additional isolates from different geographic regions and comparative analyses with the barley powdery mildew genome. Our comparative genomic analyses identified 602 candidate effector genes, with many showing evidence of positive selection. We characterize patterns of genetic diversity and suggest that mildew genomes are mosaics of ancient haplogroups that existed before wheat domestication. The patterns of diversity in modern isolates suggest that there was no pronounced loss of genetic diversity upon formation of the new host bread wheat 10,000 years ago. We conclude that the ready adaptation of *B. graminis* f.sp. *tritici* to the new host species was based on a diverse haplotype pool that provided great genetic potential for pathogen variation.

The onset of agriculture and the domestication of crops approximately 10,000 years ago resulted in drastic changes to plant pathogen environments. The genetically uniform agricultural ecosystems led either to rapid coevolution of the pathogen with its host during domestication (host tracking) or to the emergence of new pathogen species through host jump, host shift or hybridization[1–3]. For pathogens such as wheat leaf blotch *Mycosphaerella graminicola* and the potato blight *Phytophthora infestans*, the emergence of new pathogen species was accompanied by pronounced chromosomal changes and loss of genetic diversity[1,2] (**Supplementary Note**).

Powdery mildews are obligate biotrophic fungi that grow and reproduce only on living hosts. The accompanying disease occurs early in summer when haploid spores infect plants and asexually reproduce[4]. Sexual reproduction of isolates of opposite mating types in late summer results in the formation of overwintering chasmothecia (**Supplementary Note**). Cereal powdery mildew *B. graminis* has evolved into at least eight *formae speciales* that each specifically infect one host species[5]. It is assumed that the pathogen uses an arsenal of effector proteins to infect the host[6–8]. If such an effector is recognized by the plant, it renders the pathogen avirulent, and the effector gene becomes an avirulence gene[9,10]. This process allows selection for either effector changes or loss. It is postulated that the specific makeup of effectors determines the virulence spectrum of a particular mildew strain[7,11–13] (**Supplementary Note**). Here we wanted to study genetic diversity between and within powdery mildew *formae speciales* and explore the impact of the introduction of the new host (bread wheat) on *B. graminis* f.sp. *tritici* evolution (**Supplementary Note**).

The reference genome sequence of *B. graminis* f.sp. *tritici* isolate 96224 consists of a backbone of 250 BAC contigs to which Roche 454 sequence scaffolds were anchored (**Supplementary Table 1** and **Supplementary Note**). Having 250 large BAC contigs allowed the analysis of genome organization at the megabase level. In total, 82 Mb of the estimated 180-Mb genome could be assembled because many highly repetitive sequences were collapsed or removed from the assembly (**Supplementary Note**). We annotated 6,540 genes; however, over 90% of the genome was classified as transposable element (TE) sequences (**Supplementary Note**), making this the most repetitive fungal genome sequenced so far. Most of the gene space was covered, as 96% of the eukaryotic core genes were full length, and 98% were partially present (CEGMA evaluation; **Supplementary Note**). In comparison to non-obligate biotrophs, many gene families involved in primary and secondary metabolism were reduced or absent as in other obligate biotrophs[7,14–16] (**Supplementary Figs. 1–3** and **Supplementary Note**). Fewer than 50% of the genes had homologs in yeast. In the more closely related *Botrytis cinerea*, 72% of genes (4,731) had homologs. Almost 92% of the predicted

[1]Institute of Plant Biology, University of Zurich, Zurich, Switzerland. [2]Centre of the Region Hana for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc-Holice, Czech Republic. [3]Department of Plant Microbe Interactions, Max-Planck Institute for Plant Breeding Research, Cologne, Germany. [4]Department of Life Sciences, Imperial College London, London, UK. [5]Department of Biology, University of Bern, Bern, Switzerland. [6]Institut National de la Recherche Agronomique (INRA), Unité de Recherche Génomique Info (URGI), Versailles, France. [7]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland. [8]Present addresses: MTT/BI Plant Genomics Laboratory, University of Helsinki, Helsinki, Finland (J.P.B.) and Department of Agronomy and Natural Resources, Institute of Plant Sciences, Agronomy and Natural Resources (ARO), The Volcani Center, Bet Dagan, Israel (R.B.-D.). [9]These authors contributed equally to this work. Correspondence should be addressed to B.K. (bkeller@botinst.uzh.ch) or T.W. (wicker@botinst.uzh.ch).

Received 5 March; accepted 20 June; published online 14 July 2013; doi:10.1038/ng.2704

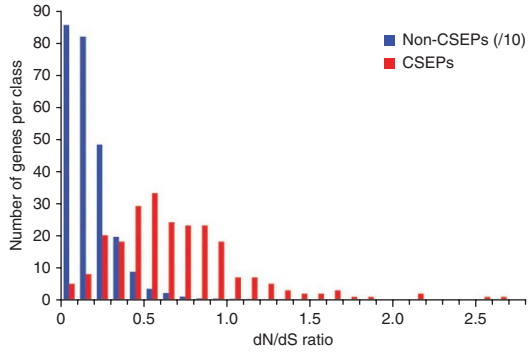

Figure 1 Comparison of 5,258 bidirectionally most closely related *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* homologs. The *x* axis indicates the ratio of nonsynonymous to synonymous substitutions (dN/dS) for all gene pairs, and the *y* axis indicates the number of gene pairs in each class. The red series represents the 237 gene pairs of bidirectionally most closely related *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* homologs encoding CSEPs, and the blue series represents all 5,021 other gene pairs. For better visibility, the numbers for non-CSEP genes were divided by 10.

B. graminis f.sp. *tritici* genes had homologs in *B. graminis forma specialis hordei*, indicating that these two *formae speciales* have very similar overall gene content and that there are a large number of genes that are specific to the *Blumeria* genus. Of these *Blumeria*-specific genes, 437 encoded candidate secreted effector proteins (CSEPs; **Supplementary Table 2** and **Supplementary Note**).

On the basis of substitutions in synonymous sites of the 5,258 bidirectionally most closely related *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* homologs, we estimated that *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* diverged 6.3 (± 1.1) million years ago (**Supplementary Note**). This finding narrows down previous estimates, which ranged from 4.7 to 10 million years ago[5,17], and indicates that the two *formae speciales* diverged several million years ago, after the divergence of their hosts 10–15 million years ago[18,19]. As in a previous study[17], we found gene order to be largely conserved between *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei*, whereas intergenic sequences were divergent owing to TE insertions and deletions (**Supplementary Fig. 4** and **Supplementary Note**).

Of the 5,258 *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* gene pairs, 96.6% had a ratio of nonsynonymous-to-synonymous substitutions (dN/dS) of less than 0.5 (average of 0.24). In contrast, CSEP genes showed much higher dN/dS ratios, with an average of 0.8, suggesting that they might be under diversifying selection (**Fig. 1**). Indeed, 55 of the 77 CSEP genes on which McDonald-Kreitman–like tests could be performed showed a positive direction of selection, which means that these genes are under selection pressure to evolve rapidly (**Supplementary Note**). This comparison of *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* genes allowed us to identify 165 novel genes that had no homologs in other fungi and lacked a sequence encoding a signal peptide but had a dN/dS ratio greater than 0.5. We propose that these genes may encode candidate effector proteins (CEPs; **Supplementary Note**) that are either non-secreted or secreted by non-conventional pathways[20]. Taking CSEP and CEP genes together, *B. graminis* f.sp. *tritici* has 602 putative effector genes, comprising 9.2% of its total gene complement (**Supplementary Note**). Post-infection transcriptome analysis showed expression of 99% of all CSEP and CEP genes, further supporting their potential involvement in the host-pathogen interaction.

In addition to the reference genome of isolate 96224 (collected in 1996 in Switzerland), we sequenced isolate JIW2 (collected in 1980 in England), isolate 70 (collected in 1990 in Israel) and isolate 94202 (collected in 1994 in Switzerland) (**Supplementary Note**). Sequencing of these isolates allowed us to sample genetic diversity of the wheat powdery mildew gene pool in different geographic regions (from the UK and Israel) as well as within the same country (Switzerland).

The gene content of the four *B. graminis* f.sp. *tritici* isolates was almost completely identical. Besides expected differences in the mating type locus (**Supplementary Fig. 5** and **Supplementary Note**), we identified 537 large deletions (>500 bp) in the 3 additional isolates. In 16 cases, these deletions led to presence-absence gene polymorphisms (**Table 1**). Notably, 13 of the 16 deleted genes were effector candidates. Considering that CEP and CSEP genes constitute only 9.2% of the gene content, they were highly overrepresented in these presence-absence polymorphisms. CSEP analogs were described in fungal pathogens of humans and animals[21,22], but specific loss of such genes has, to our knowledge, not been reported. It is possible that loss of CEP and/or CSEP genes reflects selective pressure resulting from breeding for pathogen resistance, which, unlike in animals and humans, is a normal process in crop plants (**Supplementary Note**). The 16 affected genes were lost in deletions ranging from 0.6 to 44 kb in size. Highly diagnostic sequence motifs, such as perfect or near-perfect direct repeats immediately flanking the deletion breakpoints, indicate that gene loss is the result of double-strand break repair, similar to what was described in grasses such as rice and *Brachypodium*[23,24] (**Fig. 2a** and **Supplementary Note**). One notable additional polymorphism was found in the *BgtE-5692* gene, where a highly variable sequence fragment was probably introduced in a gene conversion event (**Fig. 2b** and **Supplementary Fig. 6**). Sampling of 6 additional isolates

Table 1 Presence-absence polymorphisms of genes in the three *B. graminis* f.sp. *tritici* isolates JIW2, 94202 and 70 compared to reference isolate 96224

Gene	JIW2[a]	94202[a]	70[a]	Deletion	Gene product
Bgt-3306	–	–	+	>100 kb	Mating type (*Mat1-2-1*)
Bgt-2805	–	–	+	>100 kb	Mating type (*SLA-1*)
BgtE-5545	–	+	+	44 kb	CSEP
BgtE-5597[b]	–	–	–	25 kb	CSEP
BgtE-5802[b]	–	–	–	25 kb	CSEP
BgtE-5845	–	+	+	13 kb	CSEP
BgtE-5419	+	–	+	8 kb	CSEP
BgtE-3419	+	+	–	6 kb	CSEP
BgtAc-30466	+	–	+	5.3 bp	CSEP
BgtAc-31249	+	–	+	15 kb	CSEP
BgtAcSP-30824	+	+	–	4.5 kb	CSEP
BgtE-40100	–	+	+	1.3 kb	CSEP
BgtA-21525	–	+	–	0.6 kb	CEP
Bgt-4055	+	+	–	2.2 kb	CEP
BgtA-20784	+	+	–	9.4 kb	CEP
Bgt-369	–	+	+	13 kb	Peptidyl-prolyl isomerase
BgtAc-31336	+	–	–	0.8 kb	*ab initio*[c], no homolog
BgtA-20381	–	–	+	2.3 kb	*ab initio*[c], no homolog

[a]Presence or absence of a gene is indicated with plus and minus signs, respectively. [b]The two genes *BgtE-5802* and *BgtE-5597* are paralogs that were deleted in the same event. [c]The gene model originates from *ab initio* gene prediction.

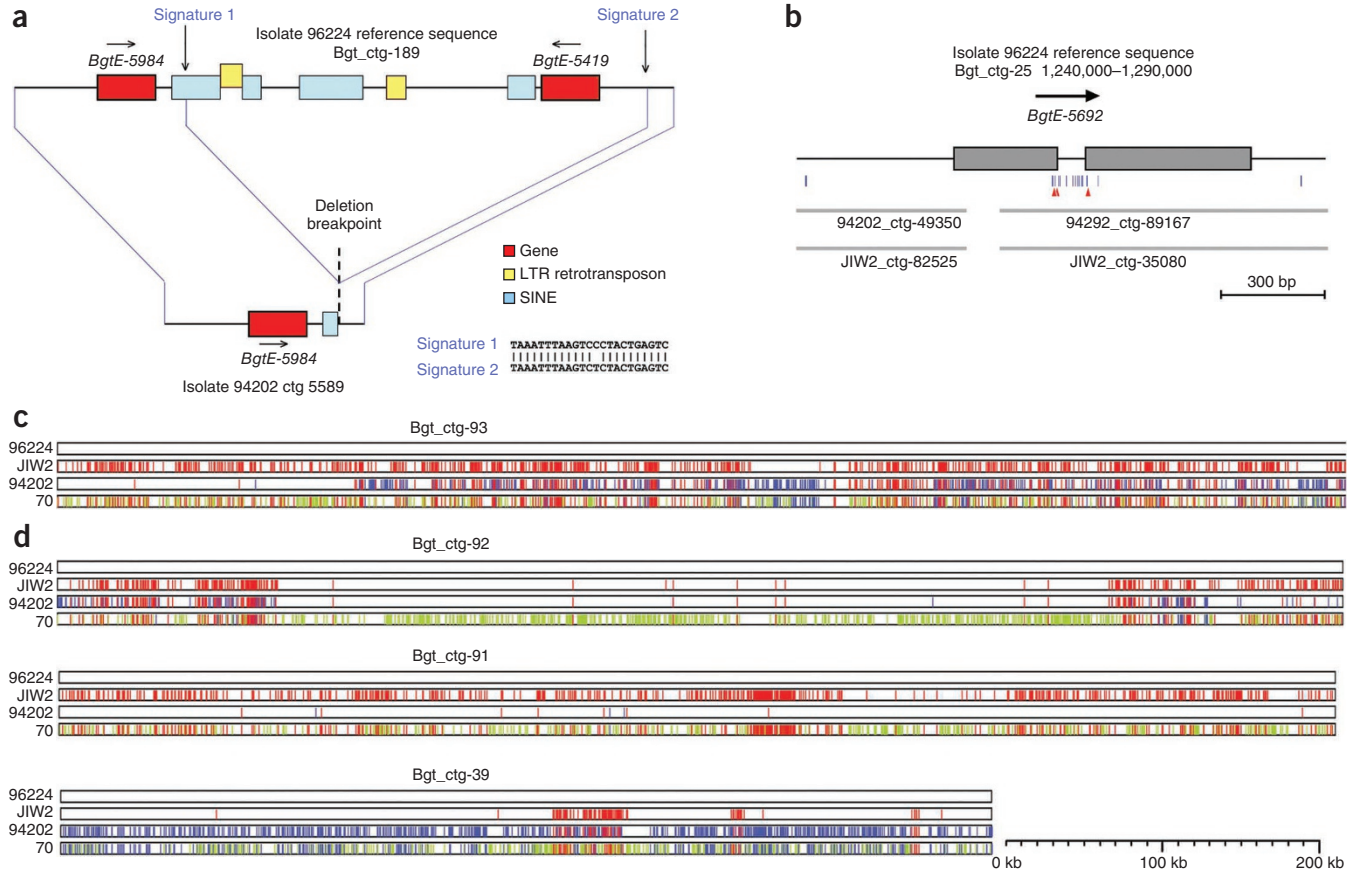

Figure 2 Presence-absence polymorphisms and genome sequence variation between *B. graminis* f.sp. *tritici* isolates. (**a**) A map of the reference genome sequence of isolate 96224 is shown at the top. Isolate 94202 differs from the reference genome in the absence of the *BgtE-5419* gene. The gene was lost in a deletion that removed over 8 kb. Homologous regions in the two isolates are connected with blue lines. The presence of a nearly identical 23-bp motif (signatures 1 and 2) precisely bordering the deleted fragment indicates that the deletion is the result of a double-strand break (**Supplementary Note**). SINE, short-interspersed nuclear element. (**b**) The candidate effector gene *BgtE-5692* contains a highly divergent segment covering parts of exons 1 and 2 (gray boxes) as well as the intron. The small size of the divergent fragment suggests that it was introduced through gene conversion. SNPs are represented by blue vertical lines. Three SNPs that result in amino acid changes are indicated with red arrowheads. The sequence assemblies of both isolates 94202 and JIW2 contain a 110-bp gap in the 5′ region of the gene, indicating a deletion. (**c**,**d**) The *B. graminis* f.sp. *tritici* genome is a mosaic of different haplogroups. The reference genome sequence of isolate 96224 is shown at the top, and the three resequenced isolates are shown below (in arbitrary order). Positions of SNPs are indicated with colored vertical lines. Priority was assigned top to bottom. For example, all nucleotide differences between JIW2 and the reference isolate 96224 are shown in red. If one of the other two isolates shares a SNP with JIW2, this SNP is also shown in red. Groups of SNPs of the same color indicate extensive chromosomal segments that originate from a different haplogroup. (**c**) Large parts of the *B. graminis* f.sp. *tritici* genome are a complex mosaic of haplogroup segments that are dozens of kilobases long. (**d**) Examples of extensive regions of shared haplogroups.

showed no correlation between the presence-absence of these genes and the geographic origin of the isolates (**Supplementary Table 3** and **Supplementary Note**).

The three resequenced isolates differed at 113,967 to 161,117 SNPs from the 96224 reference sequence, with the Israeli isolate 70 being the most divergent. Small insertions and deletions of 1 to 4 bp were almost 100 times less frequent than SNPs (**Supplementary Table 4** and **Supplementary Note**). Between 3.7 and 3.9% of the SNPs were found in the coding sequences of genes, and roughly 45% of these SNPs were nonsynonymous. For 57% of the genes, the predicted protein was identical. In 30% of all genes, we identified two protein variants, and 10% of genes had three different protein variants and 3% of genes four different protein variants (**Supplementary Table 5**). Candidate effector genes had more nonsynonymous substitutions than the average for all genes, indicating that they are under stronger diversifying selection, even within the same *forma specialis* (**Supplementary Fig. 7** and **Supplementary Note**).

We observed that the SNP frequency in all isolates varied strongly in different regions of the genome compared to the reference sequence. For example, in isolate JIW2, approximately 25% of the genome consisted of large segments that were nearly identical to the 96224 reference genome (0.11 SNPs/kb). These regions were distinct from regions with an approximately 10 times higher SNP frequency (**Fig. 2c**,**d**, **Supplementary Table 6** and **Supplementary Note**). This finding suggests that the isolates studied are mosaics of different haplogroups (chromosomal segments that are more closely related by descent than others). The average size of haplogroup segments ranged from 87.3 kb in isolate JIW2 to 150 kb in isolate 70. On the basis of the number of substitutions in the different haplogroup segments, we could distinguish two distinct groups representing more divergent haplogroups (H_{old}) and less divergent ones (H_{young}) (**Fig. 3a**). In approximately 40% of the genome, we could distinguish three different H_{old} haplogroups, whereas in about 25% of the genome four different H_{old} haplogroups were present. All four isolates shared the H_{young} haplogroup in only

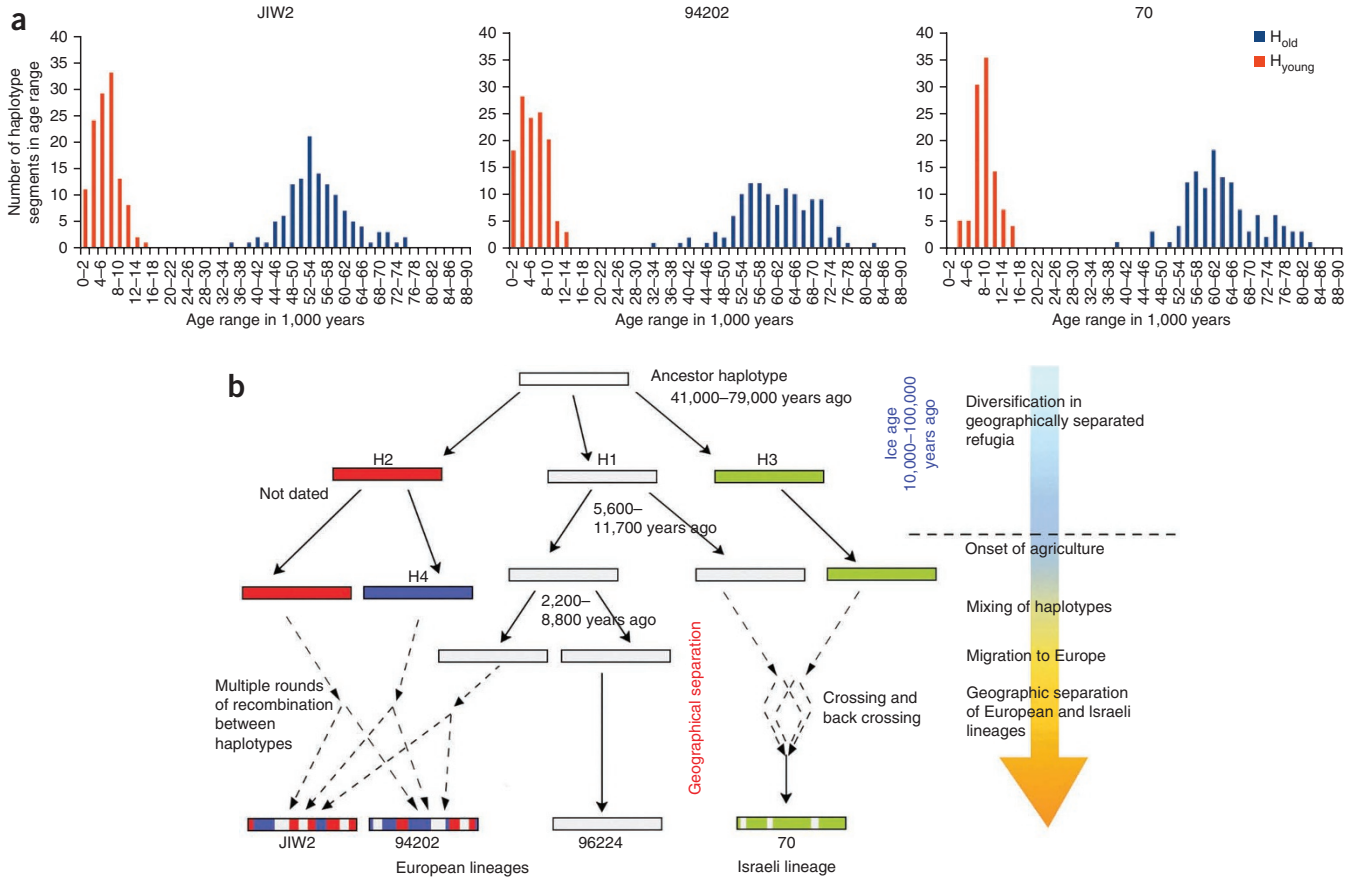

Figure 3 Divergence time estimates of genomic regions derived from different haplogroups. (**a**) Haplogroup segments in the genomes of the three resequenced *B. graminis* f.sp. *tritici* isolates were divided into regions derived from a young (H_{young}) and a more ancient (H_{old}) haplogroup relative to the 96224 reference isolate. Divergence time estimates were calculated individually for each of the 250 fingerprint (FP) contigs comprise the genome. The *x* axis shows ranges of divergence time estimates (with only every second age range labeled owing to space constraints), and the *y* axis shows how many FP contigs fall in the respective age categories. (**b**) Model for the evolution of powdery mildew isolates. The divergence and recombination of haplogroups is correlated to events such as the onset of agriculture and increasing temperatures (represented by the color of the arrow to the right).

2.2% of the genome. The H_{old} haplogroups diverged approximately 43,000 to 76,000 years ago from the 96224 reference. In contrast, H_{young} haplogroups diverged only approximately 2,100 to 8,600 years ago (for isolates JIW2 and 94202) and 5,600 to 11,700 years ago (for isolate 70) from the 96224 reference (**Fig. 3a** and **Supplementary Table 7**).

Notably, the divergence of the H_{old} haplogroups coincided with the last ice age (150,000–10,000 years ago), during which time it is assumed that wheat ancestors were restricted to the Fertile Crescent, which stretches from modern-day Israel to Iran[25]. We hypothesize that different *B. graminis* f.sp. *tritici* lineages (H1, H2 and H3 in the model in **Fig. 3b**) diverged by coevolving with different ancestral wheat populations in geographically separated areas and that the descendants of this diversification are represented in today's H_{old} haplogroup segments (**Fig. 3b**). In contrast, the H_{young} haplogroups diverged within the time period after agriculture was introduced. We speculate that northbound agricultural migration approximately 10,000 years ago could have restricted genetic exchange between European and Israeli *B. graminis* f.sp. *tritici* lineages. This hypothesis would explain why the youngest haplogroup segments shared by these isolates were fewer and diverged 8,700 (± 3,000) years ago, whereas the European isolates shared haplogroups that diverged more recently (**Fig. 3b** and **Supplementary Table 7**).

The large haplogroup segments indicate that the mildew isolates studied are descended from relatively few sexual recombination events and have since reproduced mainly clonally. *B. graminis* has very high sexual recombination rates (**Supplementary Note**). Thus, unrestricted mating of different *B. graminis* f.sp. *tritici* isolates would have completely homogenized SNP frequencies across the genomes and led to very low linkage disequilibrium (**Supplementary Note**). In contrast, our observations were consistent with clonal or near-clonal reproduction (for example, through inbreeding in small populations; **Supplementary Note**), which is key in pathogens, as it preserves successful combinations of genes and avoids the acquisition of undesirable avirulence genes[26–28]. We conclude that the distinct haplogroup patterns in the *B. graminis* f.sp. *tritici* isolates reflect strong selection for clonal propagation and/or inbreeding (**Supplementary Note**). A similar mechanism was suggested recently for barley powdery mildew[29].

The *Blumeria* genus shows unique evolutionary properties in that it has maintained high levels of adaptability and flexibility. The genomes of *B. graminis* f.sp. *tritici* isolates are composed of haplogroup segments that predate the formation of their hexaploid bread wheat host 10,000 years ago[30]. Thus, the shift from wild tetraploid to domesticated hexaploid wheat seemingly has not reduced genetic diversity in *B. graminis* f.sp. *tritici* (**Supplementary Note**), suggesting that the *B. graminis* f.sp. *tritici* gene pool provided all the necessary genetic diversity for adaption to a range of wheat species. This ability for adaptation is also demonstrated by its recent host range expansion

to the hybrid cereal Triticale (**Supplementary Note**). In contrast, in the *Phytophthora* (**Supplementary Note**) and *Mycosphaerella* genera, host changes were accompanied by the rapid formation of new species and loss of genetic diversity[1–3]. Indeed, the youngest two *Mycosphaerella* species probably date back merely 10,000 and 500 years[2,3] (**Supplementary Note**). Similarly, in *Magnaporthe oryzae*, possibly as few as three genes determine host specificity and incompatibility[31] (**Supplementary Note**). This scenario differs notably from that in powdery mildew: modern *B. graminis* f.sp. *tritici* isolates still maintain their ability to infect wild tetraploid wheat, even though their main host globally is hexaploid wheat. Additionally, *formae speciales* that diverged millions of years ago are still capable of mating[32]. Thus, the formation of reproductive barriers as a consequence of adaptation to new hosts might be detrimental to the lifestyle and evolutionary success of mildew.

URLs. The Broad Institute, http://www.broadinstitute.org/; Integrative Genomics Viewer, http://www.broadinstitute.org/igv/; NCBI, http://www.ncbi.nlm.nih.gov/; CEGMA, http://www.korflab.ucdavis.edu/Datasets/cegma/; PAML, http://abacus.gene.ucl.ac.uk/software/paml.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The genomes of the four powdery mildew isolates 96224, 94202, JIW2 and 70 have been deposited at the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL) and GenBank under accessions ANZE01000000, ASJK01000000, ASJL01000000 and ASJN01000000, respectively.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by an Advanced Investigator grant from the European Research Council (ERC-2009-AdG 249996, Durableresistance), Swiss National Science Foundation grant 310030B_144081/1 and the University Research Priority Programme (URPP) Systems Biology of the University of Zurich.

AUTHOR CONTRIBUTIONS

B.K., T.W. and K.K.S. designed the project. S.O., T.W., J.P.B., M.S., T.P. and S.R. designed software and analyzed the genome sequence. R.B. performed genome sequence assemblies. F.P. and R.B.-D. designed and performed crossing experiments. P.S.-L. and E.V.L.v.T. identified CSEPs. J.A. and H.Q. performed repeat analysis. J.D. and H.Š. constructed the BAC library. K.K.S. and P.D.S. discussed and commented on results and edited the manuscript. S.O., T.W. and B.K. wrote the manuscript and supplementary information and prepared the figures.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Raffaele, S. *et al.* Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* **330**, 1540–1543 (2010).
2. Stukenbrock, E.H. *et al.* The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome Res.* **21**, 2157–2166 (2011).
3. Stukenbrock, E.H., Christiansen, F.B., Hansen, T.T., Dutheil, J.Y. & Schierup, M.H. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proc. Natl. Acad. Sci. USA* **109**, 10954–10959 (2012).
4. Zhang, Z. *et al.* Of genes and genomes, needles and haystacks: *Blumeria graminis* and functionality. *Mol. Plant Pathol.* **6**, 561–575 (2005).
5. Inuma, T., Khodaparast, S.A. & Takamatsu, S. Multilocus phylogenetic analyses within *Blumeria graminis*, a powdery mildew fungus of cereals. *Mol. Phylogenet. Evol.* **44**, 741–751 (2007).
6. De Wit, P.J.G.M., Mehrabi, R., den Burg, H.A.V. & Stergiopoulos, I. Fungal effector proteins: past, present and future. *Mol. Plant Pathol.* **10**, 735–747 (2009).
7. Spanu, P.D. *et al.* Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* **330**, 1543–1546 (2010).
8. Hückelhoven, R. & Panstruga, R. Cell biology of the plant–powdery mildew interaction. *Curr. Opin. Plant Biol.* **14**, 738–746 (2011).
9. Michelmore, R.W. & Meyers, B.C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113–1130 (1998).
10. Jones, J.D.G. & Dangl, J.L. The plant immune system. *Nature* **444**, 323–329 (2006).
11. Godfrey, D. *et al.* Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. *BMC Genomics* **11**, 317 (2010).
12. Zhang, W.-J. *et al.* Interaction of barley powdery mildew effector candidate CSEP0055 with the defence protein pr17c. *Mol. Plant Pathol.* **13**, 1110–1119 (2012).
13. Pedersen, C. *et al.* Structure and evolution of barley powdery mildew effector candidates. *BMC Genomics* **13**, 694 (2012).
14. Raffaele, S. & Kamoun, S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* **10**, 417–430 (2012).
15. Duplessis, S. *et al.* Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. USA* **108**, 9166–9171 (2011).
16. Kemen, E. *et al.* Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biol.* **9**, e1001094 (2011).
17. Oberhaensli, S. *et al.* Comparative sequence analysis of wheat and barley powdery mildew fungi reveals gene colinearity, dates divergence and indicates host-pathogen co-evolution. *Fungal Genet. Biol.* **48**, 327–334 (2011).
18. Akhunov, E.D. *et al.* The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res.* **13**, 753–763 (2003).
19. Chalupska, D. *et al. Acc* homoeoloci and the evolution of wheat genomes. *Proc. Natl. Acad. Sci. USA* **105**, 9691–9696 (2008).
20. Nombela, C., Gil, C. & Chaffin, W.L. Non-conventional protein secretion in yeast. *Trends Microbiol.* **14**, 15–21 (2006).
21. Lee, S.A. *et al.* An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. *Yeast* **20**, 595–610 (2003).
22. Xiao, G. et al. Genomic perspectives on the evolution of fungal entomopathogenicity in *Beauveria bassiana*. *Sci. Rep.* **2**, 483 (2012).
23. Wicker, T., Buchmann, J.P. & Keller, B. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.* **20**, 1229–1237 (2010).
24. Buchmann, J.P., Matsumoto, T., Stein, N., Keller, B. & Wicker, T. Inter-species sequence comparison of *Brachypodium* reveals how transposon activity corrodes genome colinearity. *Plant J.* **71**, 550–563 (2012).
25. Pinhasi, R., Fort, J. & Ammerman, A.J. Tracing the origin and spread of agriculture in Europe. *PLoS Biol.* **3**, e410 (2005).
26. Tibayrenc, M. & Ayala, F.J. Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. *Proc. Natl. Acad. Sci. USA* **109**, E3305–E3313 (2012).
27. Heitman, J. Sexual reproduction and the evolution of microbial pathogens. *Curr. Biol.* **16**, R711–R725 (2006).
28. Bougnoux, M.-E. *et al.* Mating is rare within as well as between clades of the human pathogen *Candida albicans*. *Fungal Genet. Biol.* **45**, 221–231 (2008).
29. Hacquard, S. *et al.* Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. *Proc. Natl. Acad. Sci. USA* **110**, E2219–E2228 (2013).
30. Salamini, F., Özkan, H., Brandolini, A., Schäferr-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* **3**, 429–441 (2002).
31. Tosa, Y., Tamba, H., Tanaka, K. & Mayama, S. Genetic analysis of host species specificity of *Magnaporthe oryzae* isolates from rice and wheat. *Phytopathology* **96**, 480–484 (2006).
32. Hiura, U. Genetic basis of *formae speciales*, in *The Powdery Mildews* (ed. Spencer, D.M.) 101–128 (Academic Press, New York, 1978).

ONLINE METHODS

Roche 454 and Illumina genome sequencing. Genomic DNA from isolate 96224 was sequenced with Roche 454 Titanium technology at the Functional Genomics Center of the University of Zurich (Switzerland) to approximately 13× coverage using single-fragment (2.5 million reads, 900 Mb) and 3-kb insert (5 million reads, 1,653 Mb) paired-end libraries. Illumina sequencing was performed by GATC Biotech (Konstanz, Germany; isolates 96224 and JIW2) and The Genome Analysis Centre (Norwich, UK; isolates 94202 and 70). From each isolate, 5 µg of DNA was sequenced from paired-end libraries with insert sizes of 350–450 bp in length. Isolates 96224 and JIW2 were sequenced to approximately 24-fold coverage, and isolates 94202 and 70 were sequenced to approximately 50- to 70-fold coverage (**Supplementary Fig. 8**).

Production of the reference genome sequence of *B. graminis* f.sp. *tritici* isolate 96224. Quality-trimmed 454 reads were combined with 20,000 BAC end sequences[33] and assembled using Roche's Newbler assembler (version 2.5; default parameters, minimum overlap identity of 99%, minimum overlap length of 50 bp). Reference genome sequences were generated by integrating the scaffolds from the 454 assembly into a BAC library fingerprint assembly, which consisted of 266 contigs (called FP contigs) with a total size of 180 Mb (ref. 33). BAC end sequences of BACs present in the FP contigs were used as linker sequences between 454 scaffolds and FP contigs. Scaffolds were used as queries in BLAST searches against a database of all BAC end sequences. To avoid random anchoring of scaffolds to repetitive DNA in BAC end sequences, we used three different stringency levels (from very stringent to less stringent) for the BLAST searches. Sequence space between anchored scaffolds was filled with strings of N bases (representing any nucleotide) of a length estimated on the basis of the FP contigs. The BAC end sequences of 16 short FP contigs were all repetitive and could therefore not be used to anchor any 454 scaffolds.

Illumina sequences from isolate 96224 were used to correct the reference sequence for 454-specific sequencing errors. About 47.9 million reads (2 runs on the same 350-bp insert paired-end library, read size of 96 bp, 4.3 and 4.6 Gb of sequence data) were quality trimmed and aligned to the reference using CLC Assembly Cell version 3.2 (CLC bio) using the program clc_ref_assemble_long with parameters -s 0.98 -l 0.95. Nucleotide differences that were present in all the aligned Illumina reads and had minimal coverage of 2× were accepted as sequencing errors and were corrected in the reference sequence accordingly.

Gene annotation. Gene prediction in the *B. graminis* f.sp. *tritici* sequence was performed using two approaches. Genes conserved between *B. graminis* f.sp. *tritici and B. graminis* f.sp. *hordei* were identified by mapping the published *B. graminis* f.sp. *hordei* genes[7] onto the *B. graminis* f.sp. *tritici* sequence using GMAP[34]. The recently sequenced *B. graminis* f.sp. *hordei* genome contains 5,854 annotated genes[7]. Before mapping, the *B. graminis* f.sp. *hordei* gene set was carefully searched for sequences with homology to TEs or TE-related sequences (for example, EKA homologs[7]) by running BLAST searches of all *B. graminis* f.sp. *hordei* genes against an updated version of our *Blumeria* repeat database[33], which currently contains the predicted protein sequences of 74 TE families. On the basis of this analysis, 124 *B. graminis* f.sp. *hordei* genes were removed from the original *B. graminis* f.sp. *hordei* gene set. The remaining 5,730 *B. graminis* f.sp. *hordei* genes were mapped to the *B. graminis* f.sp. *tritici* genome using GMAP, which resulted in the annotation of 5,398 *B. graminis* f.sp. *tritici* gene models. Subsequently, the identified gene models and TEs were masked on the scaffolds of the 454 assembly. Augustus gene prediction software[35] was run on the masked sequences after it was trained on 3,143 coding sequences of identified *B. graminis* f.sp. *tritici* genes. *Ab initio* gene models that had homology to TEs in our repeat library were discarded, and the remaining models were mapped to the draft genome. In a final step, the structure and location of all genes, including the *ab initio* models, were visualized on the draft genome using IGV (Integrative Genomics Viewer; see URLs) for manual curation.

To assign functions to gene models, we performed gene ontology (GO) analysis with Blast2Go software[36] with the entire gene set using default settings. In addition, we performed a BLAST search of the protein sequences against the PFAM database and *Botrytis cinerea* genes (Broad Institute; see URLs). BLAST searches were performed with the BLASTALL program from NCBI (see URLs) on local Linux servers with local databases. For all analyses, BLAST hits with *E* values smaller than 1×10^{-10} were considered significant. We combined all the information available to provide detailed annotation in the definition line of each gene in the fasta file. These BLAST hit cutoffs were also used when gene families were determined (**Supplementary Tables 8** and **9** and **Supplementary Note**).

CEGMA (Core Eukaryotic Genes Mapping Approach; see URLs)[37] evaluation was run on the 454 scaffolds using CEGMA version v2.4.010312. CEGMA uses a reference set of conserved protein families that occur in a wide range of eukaryotes. The degree to which the gene set of a genome covers the CEGMA reference set is a measure of how completely the gene space of the genome is covered. Gene annotation was performed in the same way on the *de novo* assemblies of the three resequenced powdery mildew isolates (**Supplementary Table 10**).

Transcriptome sequencing. RNA was extracted from wheat leaves infected with *B. graminis* f.sp. *tritici* at 4, 8, 12, 24 and 48 h after infection. Equal amounts of RNA from each time point were mixed and sequenced with Illumina sequencing technology. About 1,109 million reads (50-bp read length) that represented fungal and wheat RNA from all 5 time points were pooled and mapped to the genome of isolate 96224. Mapping was performed with CLC Genomics Workbench version 6.0.1, thereby allowing only one mismatch per read and counting only reads that mapped to exons. A total of 7,442,144 reads (0.6%) could be mapped to the powdery mildew genome. This proportion was expected because most of the extracted RNA comes from wheat. For each gene, the number of reads per gene and the average coverage (total reads in base pairs divided by exon length in basepairs) was calculated to obtain a rough estimate of the overall expression level.

Evaluation of gene prediction using transcriptome data. The quality of annotation was evaluated with the exon discovery function of CLC used under default settings. We used a set of 4,000 genes to which at least 50 mRNA reads per kilobase of coding sequence were mapped. For each gene, we counted the number of reads that covered predicted exon-intron boundaries (thus contradicting the predicted exon-intron structure) and compared this number with the number of reads that mapped in exons or connected predicted exons. A low background number of reads that cover predicted exon-intron boundaries is expected owing to incorrectly or unprocessed mRNA. However, a high number of such reads indicates incorrect annotation (or alternative splicing). For more than 75% of the genes tested, fewer than 2% of transcriptome reads contradicted the predicted intron-exon structure, and, for 90% of genes, fewer than 6% of reads fell into this category. Furthermore, the 4,000 genes tested comprised 10,782 exons. Transcriptome data indicated the presence of only 45 additional exons. These data indicate high accuracy in gene prediction.

dN/dS analysis and tests for direction of selection. The aligned coding sequences of the bidirectionally most closely related homologs of *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* (**Supplementary Note**) were processed with the yn00 program of the PAML package[38] (see URLs). yn00 implements the method of Yang and Nielsen[39] for estimating synonymous and nonsynonymous substitution rates. For each of the gene pairs, the dS rate (synonymous substitutions per synonymous site) and dN rate (nonsynonymous substitutions per nonsynonymous site) was calculated. dN/dS ratios were assessed separately for the 5,021 non-CSEP and 237 CSEP gene pairs to test whether the group of CSEP genes showed characteristics of positive selection.

The dS values of all non-CSEP and CSEP genes were compared to test whether some of the bidirectionally most closely related homologs might represent deep paralogs. The distribution of dS values in the 5,021 non-CSEP alignments was used as a reference with which the dS values of CSEP genes were compared (**Supplementary Fig. 9**).

We used the dN/dS ratio as a new criterion to identify previously unknown classes of candidate effector genes. We chose the cutoff of 0.5 for the dN/dS ratio for the following reasons. First, 96.6% of the non-CSEP genes had dN/dS values smaller than 0.5. Second, the dN/dS ratio distribution of CSEP genes had its peak at 0.5 (**Fig. 1**). Therefore, this value can be viewed as the expected dN/dS ratio of a given candidate effector. Third, the average dN/dS value of CSEP genes was 0.8, whereas the average dN/dS value of non-CSEP genes was 0.24; the average between the two was 0.52.

McDonald-Kreitman–like tests[40] were employed to estimate the proportion of adaptive substitutions and the direction of selection in CSEP genes. We selected those bidirectional *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei* CSEP homologs for which we had complete sequences for all four *B. graminis* f.sp. *tritici* isolates (**Supplementary Tables 11** and **12**).

Divergence time estimate of *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei*. To estimate the divergence time of *B. graminis* f.sp. *tritici* and *B. graminis* f.sp. *hordei*, we used synonymous sites in the coding sequences of the bidirectional most closely related homologs. We used only alignment positions corresponding to the third base of codons for Ala, Gly, Leu, Pro, Arg, Ser, Thr and Val. For Leu, Arg and Ser (which each have six possible codons), we used only the codons starting with CT, TC and CG, respectively. These are the codons in which the third base can be exchanged without causing an amino acid change. We concatenated all the synonymous sites into one alignment and applied the 1.3×10^{-8} substitutions per site per year rate to obtain a single estimate for the whole genome. The error estimate of 2.29×10^{-9} for the substitution rate was then applied to the calculated divergence time. The Kimura two-parameter criterion was applied to weight the transition-to-transversion ratio as previously described[41].

Automated identification of *B. graminis* f.sp. *tritici* haplogroup segments. Positions of all SNPs in the three resequenced isolates were mapped to the 96224 reference genome sequence and visualized with an in-house Perl script (**Fig. 2**). The genomes of the three resequenced isolates are mosaics of segments that are nearly identical to the 96224 reference sequence (referred to as H_{young}) and regions that have roughly five- to tenfold higher SNP density (referred to as H_{old}).

The identification of the different haplogroup segments was automated as follows. SNP distribution was surveyed in 20-kb sliding windows across the genome. Because the genome sequence contains large gaps caused by sequence scaffolds that were anchored on opposite ends of a BAC, sequence gaps larger than 2,000 bp were excised from the genome sequence and replaced by stretches of 200 N bases for the analysis. Using a 20-kb sliding window, gaps of 2,000 bp or less influenced SNP density by 10% at most. Because SNP densities of the different haplogroups differed roughly by a factor of ten, the different haplogroups could still be clearly distinguished. However, one has to be aware that large sequence gaps could sometimes contain additional haplogroup breakpoints that would be missed in this analysis.

This analysis was performed on the 128 largest FP contigs that contained at least 200 kb of sequence without gaps (10 times the size of the sliding window). The resulting SNP density distribution was an overlay of the densities of the SNP-rich and SNP-poor regions. For all three resequenced isolates, density in SNP-rich regions peaked at approximately 22 SNPs per 20 kb (1.1 SNPs/kb; example in **Supplementary Fig. 10**). For simplicity, we divided the genome into segments with average SNP densities of 22 SNPs per 20 kb or higher (H_{old}) and segments with a lower SNP density. To determine a suitable cutoff between the two groups, we simulated SNP densities, assuming a random distribution of SNPs at an average density of 22 SNPs per 20 kb. This simulation showed that practically no segments with nine or fewer SNPs per 20 kb (approximately one SNP every 2,300 bp) could be expected by chance. Thus, segments with lower SNP density were defined as the H_{young} haplogroup (**Supplementary Figs. 11** and **12**).

Employing this cutoff value, we used distances between neighboring SNPs to identify breakpoints between the H_{old} and H_{young} haplogroups. Regions containing SNPs that were spaced at distances of at least 2,300 bp were assigned to haplogroup H_{young}. Single incidents of too closely spaced SNPs (for H_{young}) or too widely spaced SNPs (for H_{old}) were ignored. A single large spacing in a SNP-rich H_{young} region could, for example, be caused by a gap in a sequence scaffold (454 scaffolds may contain gaps from a few hundred base pairs to 2,000 bp in length owing to linking of paired-end reads). Likewise, two SNPs could be closely spaced by chance in an otherwise SNP-poor region.

The mapping of haplogroup segments resulted in a table with start and end positions of H_{old} and H_{young} haplogroup segments for each of the four isolates. These coordinates were then used for pairwise comparisons to determine the genomic regions where isolates shared the same haplogroup and the genomic regions in which they differed. From these data, we also calculated the average size of haplotype fragments. In this analysis, the full lengths of the contigs (including the large gaps that were removed earlier) were used (**Supplementary Fig. 13** and **Supplementary Note**). Tables with haplogroup positions for all isolates can be obtained via FTP upon request.

Molecular dating of *B graminis* f.sp. *tritici* haplogroups. The genomic segments assigned to haplogroups H_{old} and H_{young} were used for molecular dating. For dating, genes and the 1-kb regions up- and downstream of them were removed to avoid sequences that are under selection pressure. For the calculation of divergence times, we used the same synonymous substitution rate described above ($1.3 \times 10^{-8} \pm 2.29 \times 10^{-9}$ substitutions per site per year[42]). To obtain an estimate for variance and standard deviation, haplogroup data were processed individually for each of the 250 FP contigs. For example, FP contig Bgt_ctg-2 had a size of 898 kb of non-N bases. In isolate JIW2, this contig contained six segments that corresponded to haplogroup H_{old}. These 6 segments added up to 527 kb (59% of the FP contig), and they contained a total of 729 substitutions. From these numbers, two estimates for the divergence time of the H_{old} haplogroup from the 96224 isolate were derived: one with a substitution rate of 1.071×10^{-9} and one with a substitution rate of 1.529×10^{-9}. This was done to factor in the error in the substitution rate. In this case, two estimates are calculated of 43,800 and 62,600 years, respectively. The distribution of the individual divergence estimates for all FP contigs was used to calculate the overall standard deviation of the age estimate of the respective haplogroup. The variance was calculated as the square of the sum of all the differences from the average ($\Sigma(X_i - X_{average})^2$). The standard deviation was the square root of the variance.

All Perl programs used in this study are available upon request.

33. Parlange, F. *et al.* A major invasion of transposable elements accounts for the large size of the *Blumeria graminis* f.sp. *tritici* genome. *Funct. Integr. Genomics* **11**, 671–677 (2011).
34. Wu, T.D. & Watanabe, C.K. Gmap: a genomic mapping and alignment program for MMA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
35. Stanke, M. & Waack, S. Gene prediction with a Hidden Markov Model and a new intron submodel. *Bioinformatics* **19** (suppl. 2), ii215–ii225 (2003).
36. Conesa, A. *et al.* Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
37. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
38. Yang, Z. Paml 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
39. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
40. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the *adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
41. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
42. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **101**, 12404–12410 (2004).

doi:10.1038/ng.2704