Year: 2011

A major invasion of transposable elements accounts for the large size of the Blumeria graminis f.sp. tritici genome

Parlange, F ; Oberhaensli, S ; Breen, J ; Platzer, M ; Taudien, S ; Šimková, H ; Wicker, T ; Doležel, J ; Keller, B

DOI: https://doi.org/10.1007/s10142-011-0240-5

Originally published at:
Parlange, F; Oberhaensli, S; Breen, J; Platzer, M; Taudien, S; Šimková, H; Wicker, T; Doležel, J; Keller, B (2011). A major invasion of transposable elements accounts for the large size of the Blumeria graminis f.sp. tritici genome. Functional Integrative Genomics, 11(4):671-677.
DOI: https://doi.org/10.1007/s10142-011-0240-5

A major invasion of transposable elements accounts for the large size of the *Blumeria graminis* f.sp. *tritici* genome

Francis Parlange[1], Simone Oberhaensli[1], James Breen, Matthias Platzer, Stefan Taudien, Hana Šimková, Thomas Wicker, Jaroslav Doležel, and Beat Keller

Francis Parlange[1], Simone Oberhaensli[1], James Breen, Thomas Wicker, Beat Keller (corresponding author)

Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland

Phone: +41 44 634 8230

Fax: +41 44 634 8204

E-mail: bkeller@botinst.uzh.ch

Matthias Platzer, Stefan Taudien

Leibniz Institute for Age Research – Fritz Lipman Institute, Beutenbergstrasse 11, D-007745 Jena, Germany

Hana Šimková, Jaroslav Doležel

Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Sokolovská 6, CZ-77200 Olomouc, Czech Republic

[1] Both authors contributed equally to this work

Abstract

Powdery mildew of wheat (*Triticum aestivum* L.) is caused by the ascomycete fungus *Blumeria graminis* f. sp. *tritici*. Genomic approaches open new ways to study the biology of this obligate biotrophic pathogen. We started the analysis of the *Bg tritici* genome with the low-pass sequencing of its genome using the 454 technology and the construction of the first genomic BAC library for this fungus. High-coverage contigs were assembled with the 454 reads. They allowed the characterization of 56 transposable elements and the establishment of the *Blumeria* Repeat Database. The BAC library contains 12,288 clones with an average insert size of 115 kb, which represents a maximum of 7.5-fold genome coverage. Sequencing of the BAC-ends generated 12.6 Mb of random sequence representative of the genome. Analysis of BAC-end sequences revealed a massive invasion of transposable elements accounting for at least 85 % of the genome. This explains the unusually large size of this genome which we estimate to be at least 174 Mb, based on a large-scale physical map constructed through the fingerprinting of the BAC library. Our study represents a crucial step in the perspective of the determination and study of the whole *Bg tritici* genome sequence.

Keywords

Blumeria graminis; BAC library; BAC-end sequences; transposable elements

Abbreviations

Bg: *Blumeria graminis*; BES: BAC-End Sequences; TE: transposable element

Introduction

Powdery mildew fungi are some of the most damaging plant pathogens. They affect a wide range of dicotyledonous and monocotyledonous host species, and cause significant economic losses in crop plants worldwide (Glawe 2008). Powdery mildews belong to the family *Erysiphaceae* in the order *Erysiphales* (*Ascomycota*) (Inuma et al. 2007). Their interactions with the host are characterized by the establishment of structures called haustoria inside epidermal plant cells, allowing the pathogen to maintain a parasitic relationship and to take up nutrients from the host. This results in a complete dependence of powdery mildew growth on living plant cells (Glawe 2008).

The fungal pathogen *Blumeria graminis* is an ascomycete species sub-divided in seven *formae speciales* (ff. spp.), each highly specialized for different host species (Inuma et al. 2007). *Blumeria graminis* f. sp. *tritici* (hereafter called *Bg tritici*) is the causal agent of powdery mildew on wheat (*Triticum aestivum* L.). Little is known about the biology of this fungus and, therefore, methods and resources are needed to identify genes promoting virulence and determining *Bg tritici*-wheat interaction, and to understand the mechanisms underlying host specialization of *Blumeria graminis*.

Recently, the genome sequencing of *B. graminis* f. sp. *hordei* (*Bg hordei*), closely related to *Bg tritici* and causal agent of the powdery mildew of barley (*Hordeum vulgare*), has been completed (Spanu et al. 2010). This work, together with the reports of other obligate biotrophs genome sequences (Baxter et al., 2010; Duplessis et al., 2011) revealed genomic hallmarks possibly driven by adaptations to the obligate biotrophic lifestyle. Those include a massive proliferation of transposable elements correlated with expansion of the genome size, and the loss of genes which are not essential for the biotrophic lifestyle, such as genes encoding enzymes devoted to plant cell wall degradation, or nitrate and sulfur assimilation pathways (Spanu et al., 2010; Baxter et al., 2010; Duplessis et al., 2011).

In order to determine its genomic features, we initiated the exploration of the *Bg tritici* genome with the construction and characterization of the first Bacterial Artificial Chromosome (BAC) library from this fungus. We fingerprinted the library and produced a physical map of the genome which allowed a first estimation of the genome size. Based on low-pass 454 sequencing of the genome and 20,001 BES representing approximately 7 % of the nuclear genome, we were able to build a *Blumeria* Repeat Database and to obtain a first insight into the *Bg tritici* genome.

Materials and methods

Plant and fungal material

The construction of the BAC library and 454 sequencing were performed using DNA from *Bg tritici* isolate 96224 (Brunner et al. 2010). Cultures of 96224 were propagated by infecting fresh leaf segments of the susceptible bread wheat cultivar Kanzler, kept on agar supplemented with benzimidazole at a concentration of 30 mg/L.

BAC library

Construction and characterization of the BAC library are described in the Supplementary Text.

Assembly of a physical contig map of *Bg tritici*

Fingerprinting was performed at the Instituto di Genomica Applicata (http://www.appliedgenomics.org). High Information Content Fingerprints (HICF) were produced and processed through FPB software (Scalabrin et al. 2009) for fingerprint background removal, and GenoProfiler software (You et al. 2007) for removal of contaminants and batch processing of fingerprints into size files that can be input into FPC (Soderlund et al. 1997). Fingerprinted clones were initially assembled using FPC at a Sulston cutoff score of 1e^{-60} (initial incremental contig build) and Q-clones were split using 3 DQ steps at slightly lower Sulston scores. Singleton clones were then added to contigs and ends were merged (when applicable) by increasing the cutoff score by 1e^{-5} in a stepwise manner to 1e^{-20} (final cutoff). The approach to control experimentally the accuracy of the FPC assembly is described in Supplementary Text.

BAC-end sequencing

BAC-end sequencing was made at the Arizona Genomics Institute, University of Arizona (www.genome.arizona.edu). Sequencing was performed at both ends. Sequence trimming was conducted by processing traces files using the Phred program for base-calling and a quality score of 20 (Ewing et al. 1998). Vector sequences were masked using CROSS_MATCH (www.genome.washington.edu) and removed from the analysis. Only reads with a length of at least 100 bp were retained, providing 20,001 high-quality BAC-end sequences.

Construction of the *Blumeria* Repeat Database

The low-pass genome sequencing of the *Bg tritici* isolate 96224 was performed using the GS FLX platform (Roche) (Supplementary Text). Reads were assembled using MIRA software with defaults settings for assembly of 454 sequences. Contigs with a 10-25x coverage and a minimal length of 7 kb were used for the manual characterization of full-length transposable element (TE) sequences.

The strategy for the identification of TEs was the following: BLASTN and BLASTX searches (Altschul et al. 1997) against specialized databases such as RepBase (www.girinst.org) and TREP (wheat.pw.usda.gov/ITMI/Repeats/) were performed in order to reveal typical features characterizing the different superfamilies of TEs. LINE (Long Interspersed Nuclear Elements) elements were identified by their generally well conserved ORF2 sequence. Presence of associated ORF1 and poly-A sequences allowed further identification of complete elements. SINE (Short Interspersed Nuclear Elements) elements were identified by the presence of internal A and B promoter boxes necessary for RNA polymerase III binding as well as a poly-A tail at the 3' end. For LTR (Long Terminal Repeat) retrotransposons, typical patterns of the terminal repeats were revealed using DOTTER (WICKERsoft software). Target site duplications (TSD) and LTR-borders were determined manually. The classification into *copia* or *gypsy* superfamilies was done according to similarity of the ORF-encoded proteins with the PTREP database, and their internal organization within the element (Wicker et al. 2007). Additionally, we used contigs of the *Bg hordei* draft genome (version June 2007) which were made available for us by the BluGen consortium (www.blugen.org) for homology search to identify the *Bg hordei* homologs of *Bg tritici* repeats.

In order to reduce redundancy within the different families, we set a threshold of 80 % similarity at the nucleotide level for the definition of a family. Finally, elements were named according to the nomenclature of Wicker et al. (2007).

BES analysis

The 20,001 BES were first analyzed for their repeat content through BLASTN and BLASTX searches (Altschul et al. 1997) against the *Blumeria* Repeat Database. Only hits with a minimal alignment of 100 bp, 80 % of nucleotide identity (for BLASTN) and an E-value < 10 e^{-10} (for BLASTX) were considered. For the identification of additional high-copy sequences, sequences matching the Repeat Database were removed and the remaining ones were searched against themselves using the same BLASTN parameters.

Access to sequence data

All BAC-end sequences can be accessed through accession numbers FR776010 to FR796010 in the EMBL Nucleotide Sequence Database. An FTP server (address available on request; for review purposes: 130.60.201.71) provides access to the complete set of sequences of the 56 identified *Bg* repeats (files Bg_repeats_fasta and Bg_repeats_hypothetical_proteins_fasta).

Results

Fingerprinting of the *Bg tritici* BAC library provides a physical map of the genome and an estimate of the minimal genome size

A large-insert BAC library was constructed with *Bg tritici* reference isolate 96224 (Supplementary Text). Fingerprinting of the complete library (12,288 clones) generated 6,831 High Information Content Fingerprints (HICF) which were assembled to produce 266 BAC contigs (Table 1). Only 146 (2.1 %) BAC clones remained as singletons. The largest contig is 5.8 Mb, and 50 % of the assembly is contained in contigs larger than 1 Mb. By comparison with experimentally tested overlaps of BAC clones at two genomic regions (Supplementary Fig. 3 and Supplementary Table 1), we could confirm the accuracy of the fingerprint assembly and its relevance for establishing contigs spanning large genomic regions. The total length of the assembly is 174 Mb, giving a first estimate of the *Bg tritici* minimal genome size.

Construction of a *Blumeria* Repeat Database

In order to study the fraction of repetitive DNA in the *Bg tritici* genome, we established a *Blumeria* Repeat Database, exploiting two datasets of sequence information. First, whole genome sequencing of the *Bg tritici* genome was carried out by one full 454 GS FLX run. This resulted in 491,163 reads with an average size of 226 bp. Assembly of these reads produced 39,363 contigs and contigs with a very high coverage were selected, as this indicates sequences corresponding to high-copy repeats. Additionally, we also exploited few contigs belonging to the first *Bg hordei* draft genome sequence (version June 2007) which were made available to us by the BluGen consortium (www.blugen.org).

Composition of the *Blumeria* Repeat Database is presented in Table 2. We identified 20 families of LINEs (Long Interspersed Nuclear Element) and two *Bg tritici* SINEs (Short Interspersed Nuclear Element), Bgt_RSX_Yhi and Bgt_RSX_Lie, homologs of the previously characterized *Bg hordei* SINE elements EGH-24-1 (Rasmussen et al. 1993) and EG-R1 (Wei et al. 1996), respectively. A total of 27 LTR (Long Terminal Repeat) retrotransposons were found (Table 2), of which thirteen families could be classified as members of the *gypsy* superfamily and nine as members of the *copia* superfamily. Five sequences showed characteristics of solo-LTRs, but the complete retrotransposon they originated from could not be characterized. Finally, seven sequences exhibited characteristics of transposable elements (TE) and a high copy number, but could not be classified into any order of repeat ("unclassified" in Table 2). Among them were two *Bg tritici* sequences for which we could identify two homologous sequences in *Bg hordei* (both *Bg tritici* and *Bg hordei* homologs are in the database).

In conclusion, our *Blumeria* Repeat Database is composed of 56 TE families, including some elements which are conserved in *Bg tritici* and *Bg hordei* (Table 2).

BAC-end sequencing and TE content analysis

All the 12,288 BAC clones of the library were sequenced from both ends. After trimming the individual sequencing reads for length (threshold of 100 bp) and low quality bases, vector and bacterial contaminant sequences were eliminated. In the end, the *Bg tritici* BAC-end database consisted of 20,001 sequences with an average read length of 633 bp (Supplementary Fig. 4). The total BES length is 12,662,922 bp with an average GC content of 44.3 %. This large dataset of representative, random sequence was subsequently used to analyze the composition of the *Bg tritici* genome.

Sequences corresponding to TEs were first identified in the 20,001 BES by BLASTN search against our *Blumeria* Repeat Database. The cumulative length of sequences with homology to the 56 repeat families represented 24.1 % of the BES database (Supplementary Fig. 5), suggesting that the characterized repeat families could contribute approximately one fourth of the genome. The 10 most abundant elements represented half of the TE-fraction (49.8 %), and accounted for around 12 % of the genome (Supplementary Fig. 5). Five LINE elements represented all together 6.2 % of the genome. The most abundant element of all was the SINE Bgt_RSX_Yhi (2 %).

We then masked the sequences matching the *Blumeria* Repeat Database at the nucleotide level, and performed with the remaining sequences a second search against the *Blumeria* Repeat Database at the protein level, in order to evaluate the representation of TE superfamilies. A cumulative length representing 23.7 % of the BES set gave hits. Taken together with the previous analysis, the fraction of the BES set matching TEs of the *Blumeria* Repeat Database is 47.8 %, i.e. 6.04 Mb. The analysis of these sequences revealed the predominance of non-LTR retrotransposons over LTR-retrotransposons, mainly due to LINE elements (Table 2).

In order to identify additional unknown repeats, we masked all the sequences which previously matched our Repeat Database at the nucleotide and protein level, and kept only the BES if the remaining unmasked sequence was longer than 50 bp. This resulted in 13,270 remaining BES which were searched against themselves by BLASTN. Repeats or high-copy sequences were defined as sequences with at least 2 copies in the 13,270 BES set. Considering that the complete BES database represents 7.2 % of the *Bg tritici* minimal genome size, a high-copy sequence according to our definition would then be expected to occur in more than 28 copies in the genome. This search revealed 8,880 high-copy BES with a total length of 4.74 Mb. Together with the 6.04 Mb matching the *Blumeria* Repeat Database, we estimate the total repeat content in the BES database, and by extension in the *Bg tritici* genome, to be 85 %.

Discussion

In this paper, we report on the construction and characterization of the first *Blumeria graminis* f. sp. *tritici* large-insert BAC library. The majority of BAC libraries constructed from fungal or oomycete pathogens have a relatively small average insert size between 40 and 80 kb, and those constructed from the barley powdery mildew *Bg hordei* were reported to have average insert sizes of 30 and 41 kb (Ridout and Brown 1999; Pedersen et al. 2002). The *Bg tritici* BAC library consists of 12,288 clones of 115 kb on average with 87 % of the inserts larger than 100 kb. This result is remarkable for DNA obtained from a true obligate biotrophic fungus which cannot be cultivated *in vitro*, and is comparable with the largest libraries reported for ascomycete or oomycete species (Zhu et al., 1997; Zhang et al. 2006; Chang et al. 2007). With a 7.5x-coverage of the genome, our BAC library thus represents a powerful tool for the exploration of the *Bg tritici* genome.

Taking advantage of this library, we could show that *Bg tritici* possesses an expanded genome of at least 174 Mb, much larger than what is commonly observed for fungal genomes (Gregory et al., 2007). This observation is in accordance with the recently reported genome size of the closely related barley powdery mildew pathogen *Bg hordei*, which is estimated to be 120 Mb (Spanu et al., 2010), and demonstrates that the *formae speciales* of the *Blumeria graminis* species have an atypically large genome size. The high percentage of repeats in *Bg tritici* (85 %) seems to be the explanation for the unusually large size of its genome, which is possibly also true for the genome of *Bg hordei* as hypothesized by Spanu et al. (2010). We observed that non-LTR retrotransposons in the form of LINEs are predominant over LTR-retrotransposons in the *Bg tritici* genome. SINEs are also surprisingly abundant in *Bg tritici* and could represent at least 3 % of the genome, although they are relatively small in size (Wicker et al. 2007). Similarly, Spanu et al. (2010) observed that LINEs and SINEs are largely predominant over LTR-retrotransposons. This picture is different than what was recently reported in other repeat-rich oomycete and fungal genomes such as *Hyaloperonospora arabidopsis* (Baxter et al., 2010), *Melampsora larici-populina* and *Puccinia graminis* f.sp. *tritici* (Duplessis et al., 2011). In *Bg hordei* as well as in *H. arabidopsis*, only a small fraction of class II transposable elements was detected (Spanu et al., 2010; Baxter et al., 2010), which is not the case for *M. larici-populina* and *P. graminis* f.sp. *tritici* where the proportion of class I and class II elements is more equal (Duplessis et al., 2011).

The very stringent parameters we used to assess the fraction of repeat DNA (80 % identity) indicates that repeat copies are very similar, which could suggest that proliferation of repetitive DNA in *Bg tritici* is the consequence of a high rate of recent transposon activity. Recently, Oberhaensli et al. (2011) sequenced and annotated three *Bg tritici* BAC clones. They found a large difference of TE content in a comparative analysis with *Bg hordei*, indicating that indeed most of the TE activity in the two genomes occurred after divergence of the two *formae speciales*, around 10 million years ago. In the same study, it was found that TEs accounted for 48.8 and 51.4 % of the contigs length, respectively. However, those clones were specifically screened to encompass gene-containing regions. On a third locus, TEs were shown to occupy up to 69 % of the sequence (F. Parlange, unpublished results), which is closer to the estimation presented in the current study. This suggests that repeated elements may not be equally distributed along the genome, and proves the importance of generating large and randomly dispersed sets of sequences to draw an accurate picture of the composition of large and highly repetitive genomes.

The reports on genome sequences from three powdery mildew species, including *Bg hordei*, *Erysiphe pisi* and *Golovinomyces orontii* (Spanu et al., 2010), and the “downy mildew” *H. arabidopsis* (Baxter et al., 2010) highlighted striking signatures of convergent evolution to an obligate biotrophic lifestyle, in particular marked by an unusual expanded genome size correlated with a proliferation of transposable elements. Recently, the same observation was reported in two other obligate biotrophic parasites, the rust fungi *M. larici-populina* and *P. graminis* f.sp. *tritici* (Duplessis et al., 2011). Those observations in different evolutionary lineages support the hypothesis of Spanu et al. (2010) that large genome size and high repetitive DNA content are common hallmarks associated with obligate biotrophy. Transposable elements affect the genome by their ability to move and replicate. They can generate high levels of genetic variation independently of sexual recombination, and could contribute to genome flexibility responsible for rapid adaptation of populations to selection imposed by resistance genes in the case of phytopathogenic fungi, or to environmental constraints for symbionts. The genomes of the basidiomycete fungus *Laccaria bicolor* and the ascomycete *Tuber melanosporum*, which form ectomycorrhizal symbiosis with their host plant, were also reported to be 65 and 125 Mb respectively, with a high proportion of repeats (21 and 58 % respectively; Martin et al. 2008; Martin et al. 2010).

A convergent biotrophic adaptation was also observed at the genetic level, with a common reduction of genes which are not essential for the biotrophic lifestyle, such as genes encoding enzymes involved in the primary and secondary metabolism (Spanu et al. 2010), enzymes devoted to plant cell wall degradation (Spanu et al. 2010; Baxter et al., 2010; Duplessis et al., 2011) and transporters (Spanu et al. 2010). Absence of genes involved in the inorganic nitrate and sulphur assimilation pathways also seems to be a feature of obligate biotrophic genomes (Spanu et al. 2010; Baxter et al., 2010; Duplessis et al., 2011). However, little is still known about the molecular mechanisms involved in the establishment of the interaction between obligate biotrophic fungi and their hosts. Investigations on those aspects represent the major challenge in the study of this class of pathogens.

The future sequencing and annotation of the complete *Bg tritici* genome are the next steps in the exploration of this genome. Sequencing can now be considered through next generation sequencing technologies (Nowrousian et al. 2010) and the physical map and BES generated in this study should greatly facilitate assembly of the genome. The updated *Blumeria* Repeat Database will also help to overcome difficulties related to the massive presence of TEs and simplify the identification of gene coding sequences. This should provide the opportunity for comparative studies with the other recently sequenced powdery mildew genomes or, at a broader scale, with obligate biotrophic genomes, and contribute to the understanding of the molecular features determining the pathogenesis of those parasites.

Acknowledgements

We would like to thank Gabriele Büsing for excellent technical assistance. We thank the Blugen consortium (www.blugen.org) and especially Dr. P. Spanu for access to the barley powdery mildew genome sequences. This work was supported by the Swiss National Science Foundation Grant 3100A-127061/1 (BK), an Advanced Grant of the European Research Council (Durable resistance 249996, BK) and by European Union grant No. ED0007/01/01 Centre of the Region Haná for Biotechnological and Agricultural Research (HŠ, JD).

Ethical standards

The experiments presented in this study comply with the current laws of the country in which they were performed.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402

Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, Kemen E, Thines M, Ah-Fong A, Anderson R, Badejoko W, Bittner-Eddy P, Boore JL, Chibucos MC, Coates M, Dehal P, Delehaunty K, Dong S, Downton P, Dumas B, Fabro G, Fronick C, Fuerstenberg SI, Fulton L, Gaulin E, Govers F, Hughes L, Humphray S, Jiang RH, Judelson H, Kamoun S, Kyung K, Meijer H, Minx P, Morris P, Nelson J, Phuntumart V, Qutob D, Rehmany A, Rougon-Cardoso A, Ryden P, Torto-Alalibo T, Studholme D, Wang Y, Win J, Wood J, Clifton SW, Rogers J, Van den Ackerveken G, Jones JD, McDowell JM, Beynon J, Tyler BM (2010) Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. Science 330:1549-1551

Brunner S, Hurni S, Streckeisen P, Mayr G, Albrecht M, Yahiaoui N, Keller B (2010) Intragenic allele pyramiding combines different specificities of wheat *Pm3* resistance alleles. Plant J 64:433-45

Chang YL, Cho S, Kistler HC, Hsieh CS, Muehlbauer GJ (2007) Bacterial artificial chromosome-based physical map of *Gibberella zeae* (*Fusarium graminearum*). Genome 50:954-962

Duplessis S, Cuomo CA, Lin YC, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL, Chiu R, Coutinho PM, Feau N, Field M, Frey P, Gelhaye E, Goldberg J, Grabherr MG, Kodira CD, Kohler A, Kües U, Lindquist EA, Lucas SM, Mago R, Mauceli E, Morin E, Murat C, Pangilinan JL, Park R, Pearson M, Quesneville H, Rouhier N, Sakthikumar S, Salamov AA, Schmutz J, Selles B, Shapiro H, Tanguay P, Tuskan GA, Henrissat B, Van de Peer Y, Rouzé P, Ellis JG, Dodds PN, Schein JE, Zhong S, Hamelin RC, Grigoriev IV, Szabo LJ, Martin F (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci USA 108:9166-9171

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred I Accuracy assessment. Genome Res 8:175-185

Glawe DA (2008) The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens. Annu Rev Phytopathol 46:27-51

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD (2007) Eukaryotic genome size databases. Nucleic Acids Res 35:D332-D338

Inuma T, Khodaparast SA, Takamatsu S (2007) Multilocus phylogenetic analyses within *Blumeria graminis*, a powdery mildew fungus of cereals. Mol Phylogenet Evol 44:741-751

Martin F, Aerts A, Ahren D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buee M, Brokstein P, Canback B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbe J, Lin YC, Legue V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Oudot-Le

Secq MP, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kues U, Lucas S, Van de Peer Y, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouze P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature 452:88-92

Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R, Porcel B, Rubini A, Amicucci A, Amselem J, Anthouard V, Arcioni S, Artiguenave F, Aury JM, Ballario P, Bolchi A, Brenna A, Brun A, Buee M, Cantarel B, Chevalier G, Couloux A, Da Silva C, Denoeud F, Duplessis S, Ghignone S, Hilselberger B, Iotti M, Marcais B, Mello A, Miranda M, Pacioni G, Quesneville H, Riccioni C, Ruotolo R, Splivallo R, Stocchi V, Tisserant E, Viscomi AR, Zambonelli A, Zampieri E, Henrissat B, Lebrun MH, Paolocci F, Bonfante P, Ottonello S, Wincker P (2010) Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. Nature 464:1033-1038

Nowrousian M, Stajich JE, Chu ML, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC, Osiewacz HD, Poggeler S, Read ND, Seiler S, Smith KM, Zickler D, Kuck U, Freitag M (2010) De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genet 6:e1000891

Oberhaensli S, Parlange F, Buchmann JP, Jenny FH, Abbott JC, Burgis TA, Spanu PD, Keller B, Wicker T (2011) Comparative sequence analysis of wheat and barley powdery mildew fungi reveals gene colinearity, dates divergence and indicates host-pathogen co-evolution. Fungal Genet Biol:doi:101016/jfgb201010003

Pedersen C, Wu B, Giese H (2002) A *Blumeria gramini*s fsp *hordei* BAC library--contig building and microsynteny studies. Curr Genet 42:103-13

Rasmussen M, Rossen L, Giese H (1993) Sine-like properties of a highly repetitive element in the genome of the obligate parasitic fungus *Erysiphe-Graminis* f sp *Hordei*. Mol Gen Genet 239:298-303

Ridout CJ, Brown JKM (1999) Physical mapping of avirulence genes in the barley powdery mildew pathogen *Erysiphe graminis* fsp *hordei* (abstract). The First International Powdery Mildew Conference, Palais des Papes, Avignon, France, 29 August–2 September

Scalabrin S, Morgante M, Policriti A (2009) Automated FingerPrint Background removal: FPB. BMC Bioinformatics 10:127

Soderlund C, Longden I, Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci 13:523-535

Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Ver Loren van Themaat E, Brown JKM , Butcher SA, Gurr SJ, Lebrun M-H, Ridout CJ, Schulze-Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski P, Bindschedler LV, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson MM, Collemare J, Cramer R, Lopez-Ruiz F, Frenkel O, Godfrey D, Harriman J, Hoede C, King BC, Klages S, Kleemann J, Knoll D, Koti PS, Kreplak J, Lu X, Maekawa T, Mahanil S, Milgroom MG, Montana G, Noir S, O'Connell RJ, Oberhaensli S, Parlange F, Pedersen C, Quesneville H, Reinhardt R, Rott M, Sacristán S, Schmidt SM, Schön M, Skamnioti P, Sommer H, Stephens A, Takahara H, Thordal-Christensen H, Vigouroux M, Weßling R, Wicker T, Panstruga R (2010) Genome expansion and gene loss in powdery mildew fungi reveal functional tradeoffs in extreme parasitism. Science 330:1543-1546

Stojiljkovic I, Bozja J, Salajsmic E (1994) Molecular-cloning of bacterial-DNA in-vivo using a transposable R6k ori and a P1vir phage. J Bacteriol 176:1188-1191

Wei YD, Collinge DB, SmedegaardPetersen V, ThordalChristensen H (1996) Characterization of the transcript of a new class of retroposon-type repetitive element cloned from the powdery mildew fungus, *Erysiphe graminis*. Mol Gen Genet 250:477-482

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-82

You FM, Luo MC, Gu YQ, Lazo GR, Deal K, Dvorak J, Anderson OD (2007) GenoProfiler: batch processing of high-throughput capillary fingerprinting data. Bioinformatics 23:240-242

Zhang X, Scheuring C, Tripathy S, Xu Z, Wu C, Ko A, Tian SK, Arredondo F, Lee MK, Santos FA, Jiang RHY, Zhang HB, Tyler BM (2006) An integrated BAC and genome sequence physical map of *Phytophthora sojae*. Mol Plant Microbe Interact 19:1302-1310

Zhu H, Choi SD, Johnston AK, Wing RA, Dean RA (1997) A large-insert (130 kbp) bacterial artificial chromosome library of the rice blast fungus *Magnaporthe grisea*: Genome analysis, contig assembly, and gene cloning. Fungal Genet Biol 21:337-347

Tables

Table 1 Characteristics of the *Bg tritici* contig assembly

Total clones	12,288
Useful fingerprints	6,831
Assembled contigs	266
Clones in contigs	6,685
Singletons	146
Maximum # of clones per contig	325
Largest contig	5,825 kb
N50 (# contigs)	51
Length of N50 contig	1,002 kb
Total length of assembly	174 Mb

Table 2 Transposable element families of the *Blumeria* Repeat Database and representation of the superfamilies in the BES dataset

Order	Superfamily	Families in the database	Percentage of the BES database in length
LINE		20	21.6
SINE		2	3.0
LTR retrotransposons	gypsy	13	8.3
	copia	9	8.3
	solo-LTRs	5	0.6
unclassified		7	6.0
Total		56	47.8

Supplementary Text

Material and Methods

DNA isolation

For 454 sequencing, conidiospores were ground with glass beads (1.7-2.0 mm) in a Mixer Mill MM300 (Retsch GmbH), then mixed with 2 ml of pre-warmed (65°C) 2x CTAB buffer (2% CTAB, 200 mM Tris/HCl pH 8.0, 20 mM EDTA, 1.4 M NaCl, 1% PVP, 0.28 M β-Mercaptoethanol) and incubated for 1h at 65°C. The volume was adjusted to 6 ml with 2x CTAB. The homogenate was extracted with an equal volume of dichloromethane : isoamylalcohol (24:1) and centrifuged for 15 min at 2,800 rpm. This step was repeated twice. RNA was digested by RNase A (10 mg/µl). DNA was precipitated with 0.7 volume of cold isopropanol and centrifuged for 10 min at 3,200 rpm. The pellet was washed for 15 min with Solution I (76% ethanol, 200 mM sodium acetate, 100 mM Tris/HCl pH 7.4), then 2 min with Solution II (76% ethanol, 10 mM NH_4 acetate) and centrifuged for 2 min at 2,800 rpm. DNA was air-dried and resuspended in 50 µl TE (10 mM Tris, 1 mM EDTA) buffer.

For the BAC library construction, High Molecular Weight (HMW) DNA was prepared according to the protocol used by Pedersen et al. (2002) with some modifications. One gram of conidiospores was lyophilized (220 mg dried material), washed twice in 50 mM EDTA (pH 8.0), 0.5 % Tween 20 followed by centrifugation for 10 min at 3,500 rpm. A third wash was performed without Tween 20. The pellet was resuspended in 100 µl of 50 mM EDTA (pH 8.0) containing a cocktail of lysing enzymes (Sigma L1412) at 48 mg/ml. The suspension was incubated at 40°C during 20 min, mixed with an equal volume of pre-warmed 1.8% Incert Agarose (Lonza, Rockland, USA) prepared in 50 mM EDTA (pH 8.0) and transferred to plastic moulds at 4°C. After solidification, agarose plugs were incubated at 37°C for 20h in LET solution [0.5 M EDTA (pH 8.0), 10 mM Tris-HCl (pH 8.0), 5 mM DTT] containing 48 mg/ml of lysing enzymes. Plugs were then incubated 2 x 24h at 50 °C in NDA solution [0.5 M

EDTA, 10 mM Tris-HCl (pH 9.5), 1% sodium N-lauroyl sarcosinate] with 1mg/ml proteinase K. Plugs were washed 3 x 1 h in 100 mM EDTA (pH 8.0). For long time storage, plugs were equilibrated in 70% ethanol during 8h at room temperature and stored at -20°C.

BAC library construction and characterization

Agarose plugs were equilibrated in ice-cold TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) buffer (6 ml/plug) for 20h to remove ethanol. Before digestion, plugs were washed 3 x 1h in ice-cold TE supplemented with 100 mM PMSF, then 3 times in TE without PMSF, and finally stored overnight in TE at 4°C. The library construction was performed in two rounds, using 5 and 6 plugs for the first and the second round, respectively.

The library was constructed as described in Peterson et al. (2000) with some modifications (Šimková et al. 2011). To evaluate digestibility of HMW DNA, preliminary tests were performed on 3 plugs cut into 9 pieces each: the first plug was a control, the second was partially digested by 10 U/ml *Hin*dIII during 20 min, and the third plug was completely digested by 100 U/ml *Hin*dIII during 6h. For library construction, plugs were cut into pieces, distributed three by three into tubes and partial digestion of the HMW DNA was performed with 4 to 10 U/ml *Hin*dIII (6.5 U/ml on average). For the first 5 plugs, the partially digested DNA was size-separated by PFGE (Pulsed-Field Gel Electrophoresis) in a 1% SeaKem Gold Agarose gel (Lonza, Rockland, USA) in 0.25x TBE under the following conditions: 12.5 °C, 6V/cm, switch time 1-50 s, 17h. The size fraction of 100-150 kb was excised from the gel and subjected to a second step of size selection in a 0.9% SeaKem Gold Agarose gel in 0.25x TBE (12.5 °C, 6V/cm, switch time 3 s, 17h). The size fraction of 90-150 kb was excised from the gel and split into fractions of 100-120 kb (B) and 120-150 kb (M1), respectively. The DNA of particular fractions was electroeluted from the gel and amount of the released DNA was estimated in standard 1% agarose gel by comparing with dilution series of phage λ. Each of the fractions was used to ligate with *Hin*dIII-digested cloning-ready pIndigoBAC-5 vector (Epicentre, Madison, USA) in1:3.6 molar ratio (DNA:vector). For the second batch of 6 plugs, only the M fraction (M2) was used for ligation. The recombinant vector was used to transform *E. coli*

ElectroMAX DH10B competent cells (Invitrogen, Carlsbad, USA). Bacterial colonies were picked using Qbot (Genetix, New Milton, UK) and ordered in 32 x 384-well plates filled with 75 µl of freezing medium (2YT supplemented with 6.6% glycerol and 12.5 mg/l chloramphenicol). The BAC library has been stored at -80°C, and is permanently maintained at the Institute of Experimental Botany in Olomouc.

Three hundred BAC clones (60 from the B, 160 from the M1 and 80 from the M2 fraction) were used to estimate the average insert size. The DNA was isolated using standard alkaline lysis method and digested with *Not*I (0.02 U/µl). DNA fragments were separated in 1% agarose gel in 0.25x TBE buffer by PFGE at 12.5°C, 6V/cm, switch time ramp 1-40 s, 15h. Insert sizes were estimated by comparing with Lambda Ladder PFG Marker and MidRange Marker I (New England Biolabs, Beverly, USA).

For the screening of the library, three dimensional (3-D) pools have been prepared. The 32 plates were subdivided into 4 stacks (8 plates each). Clones of each stack were combined to create a superpool of clones. Further, 48 3-D pools (8 plate, 16 row, 24 column) were prepared for each stack. Thus, the entire library is represented by 192 3D-pools. The pools were processed as described in Šimková et al. (2011).

For fingerprinting and BAC-ends sequencing, two replica of the BAC library were prepared by inoculating new 384-well plates filled with freezing medium with clones of the master copy. After 20h growth at 37°C, the replica were frozen and sent for fingerprinting and BAC-ends sequencing.

Assessment of FPC assembly accuracy

The two loci used to control the accuracy of the FPC assembly correspond to overlapping BAC clones identified by PCR-screening of the 3-D pools (plates 1 to 16, 3.75x genome coverage) using distinct molecular markers. The first region is called locus 2 according to Oberhaensli et al. (2011) who previously described the screening approach at this locus. For the second region, we exploited a genetic map of *Bg tritici* (Parlange and Keller, unpublished data) and chose arbitrarily the AFLP marker GTCA_E4 (Forward primer: CAAAGGTAATTTCATCCACTGGT; Reverse primer:

CATGACATGAGCAATATCAATACA) for the screening of the 3-D pools. Accordingly, the locus was named GTCA_E4. Supplementary Fig. 3a and b were produced by parsing the FPC files using the WICKERsoft software (available on request).

Results

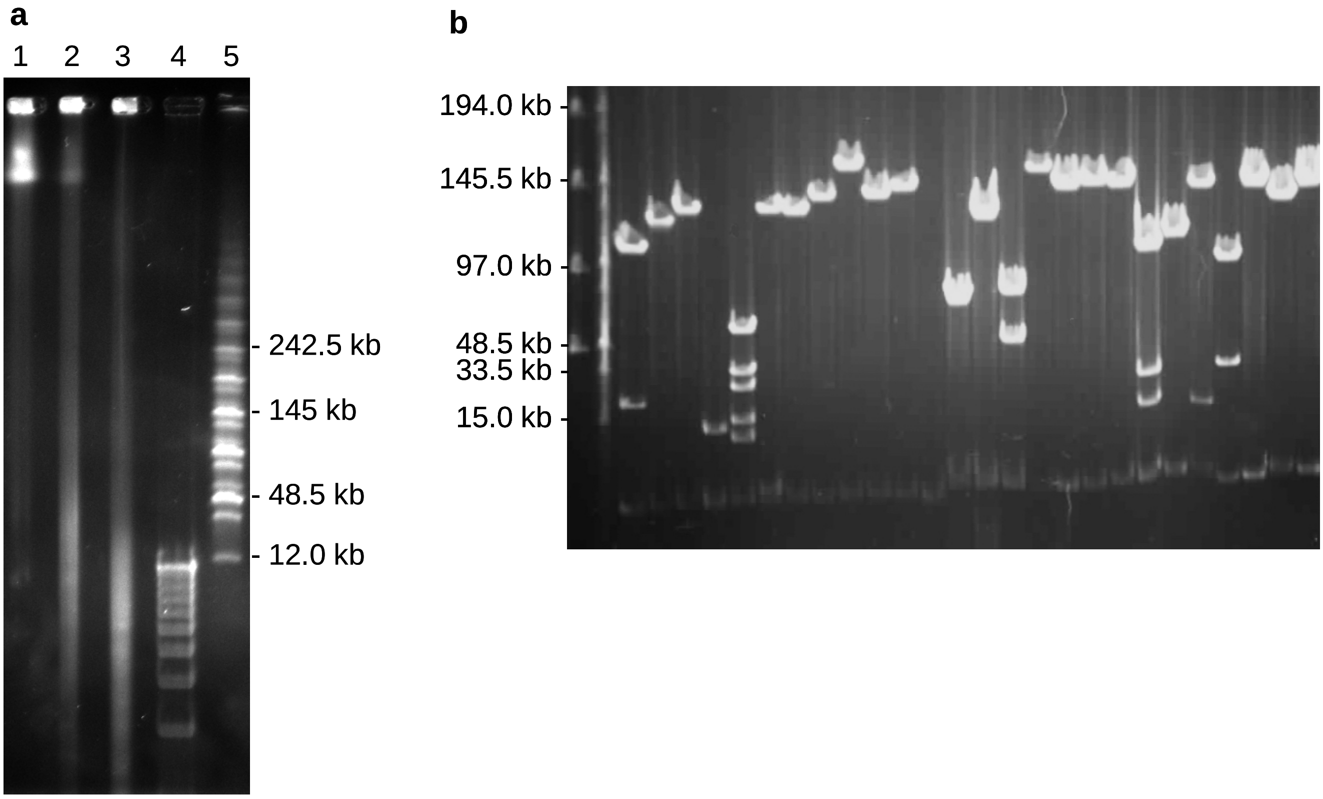
Construction and characterization of the *Bg tritici* BAC library

High Molecular Weight (HMW) DNA was isolated from conidiospores of isolate 96224 and partially digested using *Hin*dIII. Although relatively large amounts of HMW DNA were present in our preparations, only a fraction was accessible to the restriction enzyme (Supplementary Fig. 1a). This could be a consequence of an incomplete lysis of the conidiospores by the cocktail of lysing enzymes employed. Two size selection steps were applied, and the library was prepared from three independent ligations in the pIndigoBAC-5 vector: one for the DNA size fraction of 100-120 kb (B) and two for the fraction of 120-150 kb (M1 and M2).

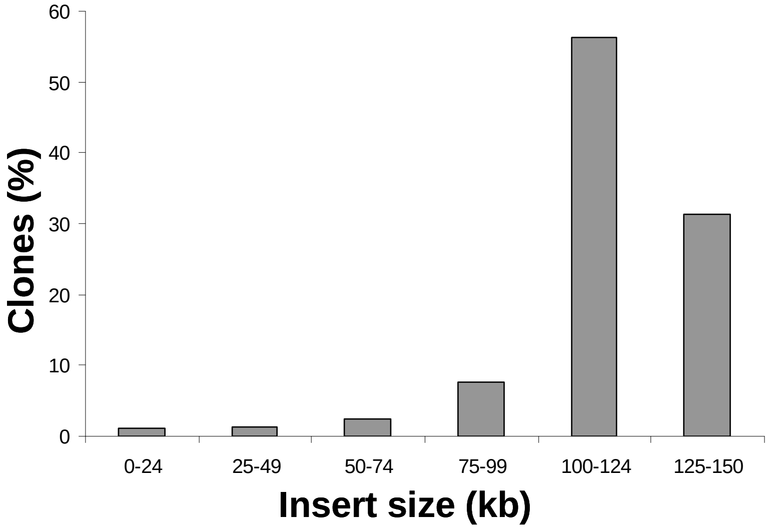
The complete library comprises 12,288 clones ordered in 32 x 384-well plates. To evaluate the quality of the BAC library and its suitability for physical mapping, insert sizes were estimated by restriction analysis of a set of 300 BAC clones randomly-selected from all fractions of the library (B, M1, M2). Insert sizes were determined by adding up the sizes of all *Not*I fragments and subtracting the size of the vector (Supplementary Fig. 1b). A significant proportion of empty clones was observed (7 %). The B-fraction sub-library comprises 2,688 clones (7 plates) with an average insert size of 105 kb, the M1 fraction provided 4,992 clones (13 plates) with an insert size of 113 kb on average, and the M2 fraction represented 4,608 clones (12 plates) with an insert size of 123 kb. Overall, the calculated average insert size was 115 kb. The distribution of insert sizes, calculated by combining data from the three fractions and considering their proportion in the library, revealed 87 % of inserts larger than 100 kb (Supplementary Fig. 2).

Assessment of FPC assembly accuracy

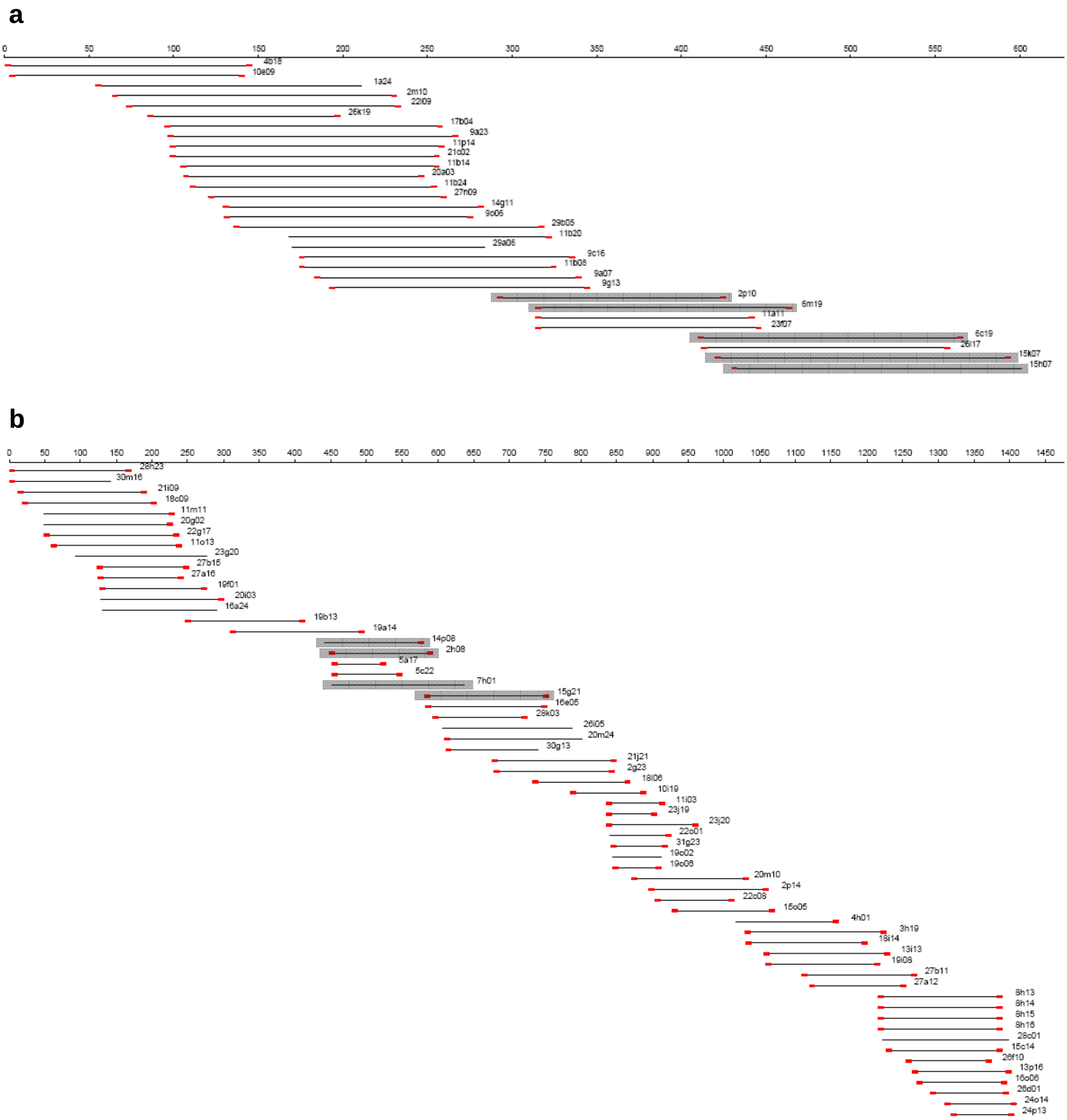
In order to control the accuracy of this FPC assembly, we determined experimentally overlapping BAC clones at two genomic regions. We performed PCR-screening of the 3D-pools for plates 1 to 16 of the library (3.75x genome coverage) using several molecular markers. The first genomic region corresponds to the locus 2 described previously by Oberhaensli et al. (2011). The screening identified six clones listed in Supplementary Table 1. Five out of the six clones screened by PCR were overlapping in contig ctg5 (Supplementary Fig. 3a). The sixth BAC (9N7) was missing from the 6,831 BACs used for physical map assembly (Supplementary Fig. 3a). No overlap was detected between clones 2P10 and 15H07. However, because the FPC information is based on common bands and not on sequence, this graphical representation cannot be considered as an accurate reflection of overlap between the clones. BAC clone 2P10 has been sequenced by Oberhaensli et al. (2011). We performed a BLASTN search of 2P10 sequences against the BES database, and found hits (99 to 100 % identity) matching BES 6C19_R, 9N07_F and 15K07_F (corresponding clones were also identified in the PCR-screening) but also 26I17_R, and 11A11_F, confirming further the accuracy of the ctg5 assembly. For the second genomic region, called locus GTCA_E4, we took advantage of a genetic map of *Bg tritici* (Parlange and Keller, unpublished data). One randomly-chosen molecular marker, named GTCA_E4, was exploited in a screening of the library and revealed five overlapping BACs (Supplementary Table 1). At this locus, four of the BACs screened by PCR were located in the contig ctg25 (Supplementary Fig. 3b). The fifth BAC clone (8O11) was absent from the assembly. According to the assembly, the region covered by the BACs identified by PCR screening was only 300 kb, and the entire FPC contig was 1.4 Mb (Supplementary Fig. 3b). The results observed on both loci confirmed the high quality of the fingerprint assembly and its importance for the construction of contigs spanning large genomic regions and possibly large genetic intervals.



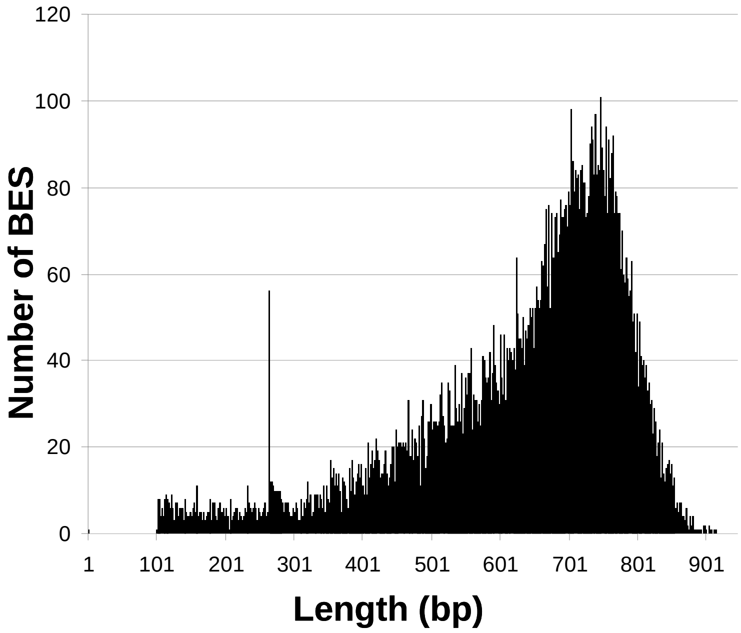
Supplementary Fig. 1 Construction of a BAC library from *Bg tritici* DNA. a. Digestibility tests of HMW DNA. DNA was tested for digestibility using restriction enzyme *Hin*dIII. Lane 1: undigested control; lane 2: partial digestion with 10U/ml for 20 minutes; lane 3: complete digestion with 100U/ml for 6 hours; lane 4: 1kb ladder; lane 5: MidRange PFG Marker I. b. Insert size analysis of 27 randomly selected BAC clones. BAC DNA was digested with *Not*I and separated by PFGE. Markers on the left are Lambda Ladder PFG Marker and MidRange Marker I



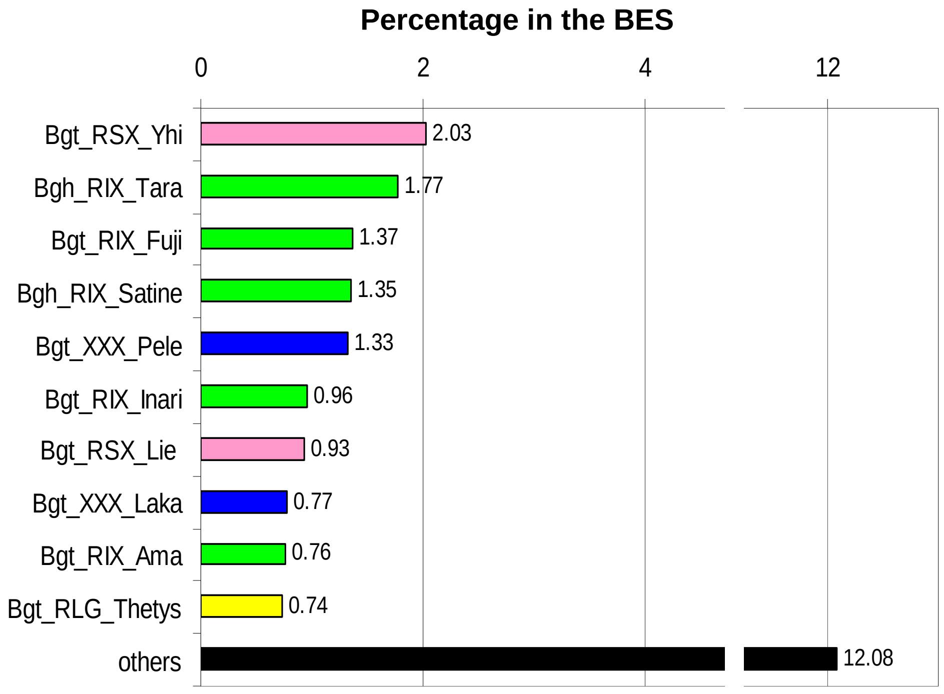
Supplementary Fig. 2 Insert size distribution in the BAC library. Insert sizes were analysed in 300 BAC clones randomly selected from the three fractions of the library (B, M1, M2). The overall distribution of insert sizes was calculated by combining data from the three fractions considering their proportion in the library



Supplementary Fig. 3 BAC contigs comprising locus 2 and locus GTCA_E4. a. Contig ctg5, harbouring the locus 2 (Oberhaensli et al. 2011). b. Contig ctg25, harbouring the locus GTCA_E4. Scale is in kilobases. BAC clones identified by PCR-screening of the library are highlighted in gray. Black boxes at the BAC-ends indicate the availability of the respective BES (note that the orientation is unknown). The graphical representation was produced based on the FPC files and using WICKERsoft software



Supplementary Fig. 4 Size distribution of BAC-end sequence (BES) length. A total of 20,001 BES were generated by the sequencing from both ends of the entire BAC library. The average read length is 633 bp with 82 % of the reads being above 500 bp. The peak observed at 263 bp is caused by 54 identical sequences which were shown by BLAST analyses to correspond to the sequence of the origin of replication "transposable R6K ori" (Stojiljkovic et al. 1994)



Supplementary Fig. 5 Distribution of the ten most abundant TE families of the *Blumeria* Repeat Database in the BES. Bgt and Bgh indicate origin of TE (*Bg tritici* and *Bg hordei*, respectively). Names are according to the nomenclature of Wicker et al. (2007): RSX, SINE (pink); RIX, LINE (green); RLG, Gypsy (yellow); XXX, unclassified (blue).

Supplementary Table 1 Overlapping BAC clones for two genomic regions. PCR screening of the library was done on plates 1 to 16. Molecular markers used to screen for locus 2 have been described in Oberhaensli et al. (2011). Screening for locus GTCA_E4 was done with molecular marker GTCA_E4 (Parlange and Keller, unpublished data)

	Clones screened by PCR	Presence in the FPC assembly
Locus 2	2P10	yes
	6M19	yes
	9N07	no [a]
	15H07	yes
	15K07	yes
	6C19	yes
Locus GTCA_E4	2H08	yes
	7H01	yes
	8O11	no [a]
	14P08	yes
	15G21	yes

[a] Those BACs were missing from the 6,831 BACs used for physical map assembly