# *Trichoderma reesei* complete genome sequence, repeat-induced point mutation and partitioning of CAZyme gene clusters

**Wan-Chen Li[1,2,3,#], Chien-Hao Huang[3,4,#], Chia-Ling Chen[3], Yu-Chien Chuang[3] , Shu-Yun Tung[3] and Ting-Fang Wang[1,3,*]**

1. Taiwan International Graduate Program in Molecular and Cellular Biology, Academia Sinica. Taipei 115, Taiwan
2. Institute of Life Sciences, National Defense Medical Center, Taipei 115, Taiwan
3. Institute of Molecular Biology, Academia Sinica. Taipei 115, Taiwan
4. Institute of Genome Sciences, National Yang-Ming University, Taipei 112, Taiwan

[#] These two authors contributed equally to this work

[*]Corresponding author: Ting-Fang Wang, email: tfwang@gate.sinica.edu

Wan-Chen Li: wanwan9121@hotmail.com

Chien-Hao Huang: ases87287@gmail.com

Chia-Ling Chen: chialing.chen1118@gmail.com

Yu-Chien Chuang: kifoiser@gmail.com

Shu-Yun Tung: mbsytung@imb.sinica.edu.tw

## Abstract

*Trichoderma reesei* (Ascomycota, Pezizomycotina) QM6a is a model fungus for a broad spectrum of physiological phenomena, including plant cell wall degradation, industrial production of enzymes, light responses, conidiation, sexual development, polyketide biosynthesis and plant-fungal interactions. The genomes of QM6a and its high-enzyme producing mutants have been sequenced by second-generation-sequencing methods and are publicly available from the Joint Genome Institute (JGI). While these genome sequences have offered useful information for genomic and transcriptomic studies, their limitations and especially their short read lengths make them poorly suited for some particular biological problems, including assembly, genome-wide determination of chromosome architecture and genetic modification or engineering. We integrated Pacific Biosciences and Illumina sequencing platforms for the highest-quality genome assembly yet achieved, revealing seven telomere-to-telomere chromosomes (34,922,528 bp; 10877 genes) with 1630 newly-predicted genes and >1.5 Mb of new sequences. Most new sequences are located on AT-rich blocks, including 7 centromeres, 14 subtelomeres and 2329 interspersed AT-rich blocks. The seven QM6a centromeres separately consist of 24 conserved repeats and 37 putative centromere-encoded genes. These findings open up a new perspective for future centromere and chromosome architecture studies. Next, we demonstrate that sexual crossing readily induced cytosine-to-thymine point mutations on both tandem and unlinked duplicated sequences. We also show by bioinformatic analysis that *Trichoderma reesei* has evolved a robust repeat-induced point mutation (RIP) system to accumulate AT-rich sequences, with longer AT-rich blocks having more RIP mutations. The widespread distribution of AT-rich blocks correlates genome-wide partitions with gene clusters, explaining why clustering of genes has been reported to not influence gene expression in *Trichoderma reesei*. Compartmentation of ancestral gene clusters by AT-rich blocks might promote flexibilities that are evolutionarily advantageous in this fungus' soil habitats and other natural environments. Our analyses, together with the complete genome sequence, provide a better blueprint for biotechnological and industrial applications.

**Keywords:**

*Trichoderma reesei* − Complete genome sequence − Repeat-induced point mutation − AT-rich block – Centromere − Centromere-encoded gene − CAZyme; Gene cluster.

## Background

*Trichoderma* is a fungal genus in soils and many other natural environments. *Trichoderma reesei* (syn. *Hypocrea jecorina*) is a widely used model organism for plant cell wall degradation and industrial enzyme production. The natural strain QM6a (ATCC13631) was first isolated from the Solomon Islands during the Second World War [1]. High enzyme producers (e.g., QM9414 and RUT-C30) were artificially generated from QM6a for industrial applications over the last 70 years [2-6].

*Trichoderma reesei* undergoes a heterothallic reproductive cycle and generates fruiting bodies (stromata) containing linear asci with 16 ascospores [7]. These 16 ascospores are generated via meiosis and two rounds of postmeiotic mitosis [8]. When placed under favorable conditions, ascospores germinate to form vegetative mycelia and produce asexual spores (i.e., conidia). Sexual development of the *Trichoderma reesei* CBS999.97 wild-isolate strain produced two haploid strains, CBS999.97(*MAT1-1*, F/X) and CBS999.97(*MAT1-2*, M/33). The two ancestral scaffolds (M and 33) in CBS999.97(*MAT1-2*, M/33) underwent an unequal translocation to form two new scaffolds (F and X) in CBS999.97(*MAT1-1*, F/X) [8]. Like CBS999.97(*MAT1-2*, M/33), QM6a has a *MAT1-2* mating type locus and two ancestral scaffolds M and 33. Due to chromosome heterozygosity and meiotic recombination, sexual crossing of CBS999.97(*MAT1-1*, F/X) with CBS999.97(*MAT1-2*, M/33) or QM6a often (>90%) generates segmentally aneuploid (SAN) progeny [8]. The CBS999.97 wild-isolate strain is also an excellent fungal model for light responses since light variation greatly affects its sexual development and conidiation [7, 9, 10]. Constant light promotes conidiation and completely inhibits stromata formation, whereas total darkness causes a slowdown of the growth of stromata [10].

The genomes of several *Trichoderma* species have been sequenced and are publicly available from the Joint Genome Institute of the US Department of Energy. The QM6a-v2.0 draft genome (33.4 Mb) contains 87 scaffolds and 9129 predicted genes [11]. The RUT-C30-v1.0 draft genome (32.7 Mb) has 182 scaffolds and 9852 predicted genes [12]. The genomes of *Trichoderma atroviride* and *Trichoderma virens* were thought to be larger than QM6a-v2.0, with sizes of 36.1 and 38.8 Mb, respectively, versus 34.1 MB for QM6a, both encoding more than 2,000 additional predicted genes [4, 13]. These genomes have been used to identify key genes involved in some important biological processes [4, 13-15], e.g., the transcriptional factors and transporters that control induction and expression of carbohydrate-active enzymes (CAZymes) and the plant cell wall degradation enzymes. Using QM6a as a reference, it has also been reported that QM9414 and RUT-C30 might carry multiple alternations, including rearrangements, point mutations, insertions and deletions [12, 16]. Recently, a group of *Trichoderma* researchers have collectively annotated and compared ~30% genes in the JGI genomes of *Trichoderma reesei*, *Trichoderma atroviride* and *Trichoderma virens* [4].

To reveal gene order and dynamic gene expression at the chromosome level, a genome-wide chromosome conformation capture method (referred herein as “HiC”) had been applied to close the gaps between the QM6a-v2.0 scaffolds. The HiC draft genome revealed seven super-scaffolds and four short contigs [17]. Druzhinina et al. then applied the QM6a-HiC draft to annotate 9151 (not 9194) predicted genes. A third of the putative CAZyme genes occurred in loose clusters that also contained a high number of genes encoding small secreted cysteine-rich proteins (SSCPs). Five CAZyme gene clusters are located close to chromosomal ends. These subtelomeric areas are also enriched in genes involved in conidiation, iron scavenging and interactions with other fungi, such as secreted protease genes, amino acid transporter genes, gene clusters for polyketide synthases (PKS), non-ribosomal peptide synthase (NRPS) and PKS-NRPS fusion proteins [18]. The QM6a-HiC annotation (http://trichocode.com/index.php/t-reesei) [18] became publicly available in January, 2017. Because it does not provide any expectation values (E) for the BLAST sequence alignments, its reliability needs to be further confirmed.

Strictly speaking, the HiC draft genome is far from equivalent to a complete genome sequence. We noticed that both the QM6a-v2.0 and QM6a-HiC drafts lack several evolutionarily-conserved genes that are ubiquitiously expressed in almost all studied eukaryotic organisms, including the recA family protein Rad51 and the DNA repair protein Rad50. Therefore, we applied both second and third generation sequencing (SGS and TGS) platforms to resequence the QM6a genome. The Single Molecule Real Time (SMRT) sequencing method developed by Pacific BioSciences (PacBio) offers much longer reads of up to 60 kb [19]. After error correction with short and high-quality Illumina-MiSeq sequencing reads, the PacBio long reads were assembled into seven telomere-to-telomere chromosomes. Our high-quality genome sequence provides a large quantity of new information to facilitate functional and comparative studies of this industrially important workhorse fungus.

## Results

### Resequencing the QM6a genome

The QM6a genome was resequenced using seven SMRT cells on the PacBio RSII platform. Following extraction of reads with *Trichoderma reesei*-only sequences, we recovered the longest raw reads (≥16□kb) with up to 80x coverage, totaling 3,397,762,180 bp. The hierarchical genome assembly process program [20] was used to generate a preliminary PacBio draft with seven superscaffolds, a short unitig and 1.8 kb of contaminating DNA (Additional File 1: Table A1). This short unitig was completely identical in nucleotide sequence to the QM9414 mitochondrial genome (42,139bp; NC_003388.1) [21], indicating that the error rate for the preliminary PacBio draft was extremely low (<0.0024%).

For error correction, the Ilumina MiSeq 300 bp paired-end reads (6.8 Gb) were collected and trimmed. Reads (67.85%) with a quality score threshold (Q) greater than 30 were retained (Additional File 1: Table A2 and A3). The final assembly data contains a circular mitochondrial genome (42,139 bp) and seven unitigs (34,922,528 bp) (Additional File 1: Table A3). We highlight that there were no sequence ambiguities or unidentified bases (Ns) (Table 1). In contrast, the QM6a-v2.0 draft genome was 33,453,791 bp and had 48,252 Ns, whereas the HiC draft was 33,395,328 bp and had 42,879 Ns.

The seven superscaffolds closely match (if not being completely identical to) the full-length chromosomes because all of their termini capture typical telomeric sequences (i.e., TTAGGG at 3′-termini and the reverse complement CCCTAA at 5′-termini) [11] with up to 14 repeats (Additional File 1: Table A3). We categorized these telomere-to-telomere chromosomes with Roman numerals (ChI-ChVII), from largest to smallest. Genetically-defined linkage groups were designated alphabetically (A-G), and linkage group arms are designated L for the short (left) arm at 5′ termini and R for the long (right) arm at 3′ termini (Figure 1). The complete QM6a genome sequences have been submitted to NCBI (accession number CP016232-CP016238).

## The complete QM6a genome compared to two earlier draft genomes

By mapping the MiSeq reads to the QM6a-v2.0 draft, we found that ten scaffolds in the QM6a-v2.0 draft were not genuine *Trichoderma reesei* sequences (Additional File 1: Tables A5). For this reason, 16 previously annotated genes (Additional File 1: Tables A6) in QM6a-v2.0 and QM6a-HiC might not be authentic QM6a genes. In addition, there were numerous sequencing and assembly errors in QM6a-v2.0. The most prominent assembly error was the first and longest scaffold, wherein the 5′ portion (~2.46 Mb) was mapped to ChV and the 3′ portion (~1.79 Mb) to ChIV (Additional File 1: Table A7).

Our complete QM6a genome sequence also covers all seven superscaffolds and the four short contigs of QM6a-HiC [17]. QM6a-HiC contains a large quantity of sequence and/or assembly errors, particularly those sequences (e.g., telomeric repeats) close to both termini of the seven superscaffolds. The four short HiC contigs are all located at chromosome regions with low guanine(G)-cytosine(C) contents. HiC0 is located close to *tel6L* (the left telomere of ChVI), HiC1 and HiC2 at *cen3* (the centromere of ChIII), and HiC3 in an interspersed AT-rich block on the right arm of ChVI. There were at least 18 inversion errors in the HiC draft genome (Figure 1 and Additional File 1: Table A8). Eight of these inversion errors could account for the failure to connect these four short contigs to the corresponding superscaffolds during the HiC experiments [17].

The HiC experiments also resulted in incorrect assembly at the rDNA locus. Using Southern hybridization, we confirmed that the right arm of ChVI harbors the large rDNA locus with nine tandem “head-to-tail” repeats. Each repeat contains an 18S–5.8S–26S rRNA gene cluster and a non-transcribed intergenic spacer (IGS) (Figure 2 and Additional File 1: Table A9). This result approaches the theoretical limit for mapping results using Illumina Miseq short reads, i.e., 200-260x coverage at the rDNA locus versus 25-30x coverage along the entire chromosome (Figure 3). It is worth noting that there are 175-200 copies of the large rDNA tandem repeats in *Neurospora crassa* [22].

Thus, the complete QM6a genome sequence has uncovered many sequencing and assembly errors in QM6a-v2.0 and QM6a-HiC. We suggest that caution should be exercised in applying HiC or the chromosome conformation capture method for gap-closing draft genome sequences produced by SGS technology and for other genomic analyses involved in AT-rich and repetitive sequences.

**QM6a compared to high enzyme producers**

The high-quality QM6a genome sequence we provide here is a better scaffold to order and orient contigs of other *Trichoderma* draft genomes previously generated by SGS technologies. It has been reported that there might be five or eleven potential translocations in RUT-C30 [12, 16]. A BLAST search revealed that there are only three promising translocations in RUT-C30: ChV to ChII (the first scaffold of RUT-C30), ChIII to ChI (the second scaffold of RUT-C30), and ChIII to ChIV (the fifth scaffold of RUT-C30). Theoretically, these three translocations in RUT-C30, together with the one translocation we identified in CBS999.97(*MAT1-1*, F/X), are sufficient to account for the large quantity of inviable SAN ascospores generated from sexually crossing RUT-C30 with CBS999.97(*MAT1-1*, F/X) [8, 23]. There are also three short ectopic insertions in RUT-C30 including the second scaffold (34 bp), the fifth scaffold (231 bp) and the ninth scaffold (2,655 bp) (Additional File 1: Table A10). These three ectopic insertions are more likely due to sequence duplication or assembly errors.

**Genome reannotation**

We applied four different approaches to genome reannotation (Materials and Methods), including the use of all (28748) *Trichoderma reesei* proteins from National Center for Biotechnology Information (NCBI), QM6a-v2.0, RUT-C30-v1.0 and two publicly-available transcriptome datasets [24, 25]. We annotated 1630 newly-predicted QM6a genes, including 70 tRNA genes and 23 5S-rDNA genes (Additional File 2: Tables B1 and B2). The average length of all 10876 QM6a genes is 1579 bp. Their average GC content (56.5%) is ~5.5% higher than that of the entire QM6a genome. Among 1515 new protein-encoding genes, 679 have been annotated in RUT-C30-v1.0. Most of them encode novel or hypothetical proteins, and only 120 and 285 newly-predicted genes encode protein products that have homologs in *Saccharomyces cerevisiae* and *Neurospora crassa*, respectively. It is worth noting that we annotated several essential or biologically important genes, including six DNA repair genes (*rad50*, *sae2*/*com1*, *rad51*, *rad57*, *srs2*, *rrm3* and *pif1*), an essential component of sister chromatin cohesion complex (*smc1*), a key autophagy gene (*atg11*), two cell division cycle genes (*cdc4* and *cdc15*) and two mitochondrial genes (*sod2* and *tom7*) (Additional File 1: Table A11 and Additional File 2: Tables B2 and B3). These evolutionarily conserved genes had never been annotated in QM6a-v2.0 or QM6a-HiC [4, 11, 18]. The complete genome sequence and set of genes we provide here can serve as a better guide for further experiments, especially for global approaches to evolution, biological functions and industrial applications.

We categorized all QM6a genes with a non-italicized uppercase letter, a number and a letter: Tr (for *Trichoderma reesei*); A, B to G (for chromosome I, II through VII); a number corresponding to the order of the transcripts (counting from the left telomere); and W or C to designate the Watson or Crick strand (the Watson strand is 5′→3′ left telomere to right telomere); for example, the 100th gene from the left telomere of chromosome I is TrA0100C (Additional File 2).

## Comparative transcriptome analysis

Next, we applied TopHat—a bioinformatic sequence analysis package tool—to map and count the SGS reads of all annotated genes and then to determine the values of Transcripts Per kilobase Million (TPM) [26]. Compared to RPKM (Reads Per Kilobase Million) and FPKM (Fragments Per Kilobase Million), TPM is a stable and reliable RNA-seq expression unit across experiments [27]. All reads mapped to the rRNA and tRNA genes were excluded before calculating TPMs (Additional File 2: Tables B1).

Our results confirmed those reported by [24] that 35 CAZyme genes and 27 non-CAZyme genes were highly induced (≥ 20-fold) by straw substrate in QM6a (Additional File 3: Table C1). In addition, straw also upregulated (≥ 20-fold) 210 previously annotated genes in QM6a-v2.0, including *xyr1* (xylanase regulator 1; ≥ 100-fold), 38 CAZyme genes, 1 NRPS gene, the mating type gene *mat1-2-1* [7] and hybrid-type peptide pheromone precursor 1 (*hpp1*) [28] (Additional File 3: Table C2). Many straw-induced genes in QM6a [24] were shown to be differentially regulated in response to cellulose and sophorose in QM9414 [25] (Additional File 3: Table C1 and C2).

Among the 1535 new protein-encoding genes, there are 39 straw-induced (≥ 20-fold) genes in QM6a and 2 cellulose-induced genes (≥ 20-fold) in QM9414 (Additional File 3: Table C3). One hundred and forty four new QM6a genes are also significantly upregulated (5-20 fold) by straw (Additional File 3: Table C4). For example, TrC1345C, a new straw-induced gene, encodes the homolog of *S. cerevisiae* Cat8 zinc-cluster transcription factor. Cat8 is involved in gluconeogenesis, the glyoxylate cycle and ethanol utilization, and it is necessary for derepression of a variety of genes under non-fermentative growth conditions (e.g., diauxic shift and sporulation) [29]. Further research on this transcriptional factor might help to enhance cellulolytic enzyme production. We also identified nine new genes with TPM values in QM6a that are much higher than those in QM9414 (Additional File 3: Table C5). These genes are likely mutated or even deleted in QM9414.

## Wide distribution of AT-rich blocks along seven QM6a chromosomes

To establish the link between DNA sequence and chromatin architecture, the local GC contents along each chromosome were calculated using a 0.5 kb sliding window. We identified 2349 AT-rich chromosome blocks with GC contents ≥12% and ≥6% lower than the average GC content of the all predicted genes (56.5%) and the entire QM6a genome (51.1%), respectively (Table 2).

The most prominent or longest AT-rich blocks in each chromosome are the centromeres, ranging from 162.5 kb (*cen2*) to 208.5 kb (*cen6*) (Figure 4 and Additional File 3: Table A2). The longest AT-rich blocks in each *Neurospora crassa* chromosome are also centromeres [30]. A BLAST search revealed that the seven QM6a centromeres collectively harbored 24 conserved sequences (≥90% identity; maximum length 8625 bp and minimum length 4847 bp) with a copy number per chromosome ranging from one to five (Additional File 4). These conserved sequences are centromere-specific and highly AT-rich, perhaps representing centromeric repeats.

The seven QM6a centromeres also separately encode 13 previously-annotated genes and 24 newly-predicted genes (Additional File 3: Table C6). Centromere-encoded transcripts are known to be integral components of the genomes of mammals, higher plants and the fission yeast *Schizosaccharomyces pombe* [31-34]. Copy numbers of these 37 centromere-encoded transcripts were relatively low; their expression (TPM > 0) could only be detected in QM9414, but not in QM6a (Additional File 3: Table C6). The QM9414 transcriptomic data from the Illumina HiSeq 2000 platform [25] apparently had better sequencing depth than the QM6a transcriptomic data from the SOLiD platform [24]. It will be of importance to investigate whether these 24 centromeric repeats and 37 centromere-encoded genes are involved in centromere integrity and chromosome segregation fidelity in *Trichoderma reesei*.

The AT-rich chromosomal blocks next to the 14 telomeres are subtelomeres, with the shortest being ~1 kb and the longest up to 87 kb (Additional File 1: Table A3). Other than centromeres and subtelomeres, there are 2328 interspersed AT-rich blocks (Table 2). On average, there are only five genes between two neighboring interspersed AT-rich blocks. The biological relevance of these interspersed AT-rich blocks remains to be elucidated (see below). It is worth noting that both the 5′ and 3′ flanking sequences of the rDNA locus contain a 2 kb AT-rich block (Figure 3). We postulate that these two AT-rich blocks might be involved in regulating nucleolar organization, rRNA transcription, rDNA copy homeostasis and prevention of repeat-induced point mutation (RIP) (see below).

## Comparative genomic analyses of chromosome architectures

Next, we compared the QM6a genome with 11 publicly-available and well-assembled fungal genomes for their AT-rich blocks. We highlight that all these fungal genomes contain ≤ 0.11% unresolved or unknown bases (Ns) (Additional File 1: Table A12). Unresolved bases jeopardize the integrity of *in silico* genomic analysis.

All of these 12 fungal genomes have many short AT-rich blocks with lengths of

0.5-3 kb (Table 2). In *Saccharomyces cerevisiae*, these short AT-rich blocks correlate with the convergent intergenic regions and the pericentromeres that associate with cohesin (an evolutionarily-conserved protein complex that functions to hold a pair of sister chromatids during mitosis and meiosis). The average distance between neighboring cohesin-binding sites along yeast chromosome arms is 10-15 kb, which is compatible with the observed localized oscillations in base composition [35-37]. It has been suggested that sister chromatid connections via cohesin complexes occur preferentially at the chromosome axes, the bases of intrachromosomal loops or topologically-associated domains (TADs) [35, 38-40]. Expression of genes within TADs is somewhat correlated [41-45]. Intriguingly, the average distance between neighboring AT-rich blocks along QM6a chromosomal arms is ~13 kb. We suggest that these AT-rich blocks might be functionally associated with chromosomal loading of cohesin in *Trichoderma reesei*.

We were able to categorize these 12 fungal genomes into three different groups according to: (1) the average AT content, (2) the genome-wide distribution of local AT content, and (3) the number of long (≥ 3kb) AT-rich blocks. For example, in the QM6a genome, there were 167 AT-rich blocks with length ≥ 3 kb (Table 2 and Figure 6).

The Group I fungi consist of six filamentous ascomycetes (Pezizomycotina), including QM6*a*, *Neurospora crassa* (OR74A), *Penicillium chrysogenum* (P2niaD18), *Mycosphaerella graminicola* (IPO323), *Fusarium fujikuroi* (IMI 58289), and *Aspergillus nidulans* (FGSC A4). Their genomes not only have similar average AT contents (~50%) but also display biphasic local AT content distributions due to the presence of many longer (≥ 3kb) AT-rich blocks (Table 2 and Figure 6A, top two panels). In *Neurospora crassa*, AT-rich blocks are DIM-3 (importin α)-dependent constitutive heterochromatins with transposon relicts and trimethylated histone 3 at lysine 9 (H3K9me3). Genome organization in *Neurospora crassa* nuclei is largely defined by constitutive heterochromatins via strong intra- and inter-chromosomal contacts [46, 47].

Group II are three basidiomycetes, i.e., *Ustilago maydis* (521), *Coprinopsis cinerea* (Okayama 7#130) and *Cryptococcus neoformans* (JEC21). Their average AT contents are ~50%, similar to those of Group I fungal genomes (Table 2). The Group II fungal genomes display relatively normal local AT content distribution (Figure 5A, the lowest panel) due to the absence of longer AT-rich blocks (Figure 5B).

Group III are three hemiascomycete yeasts, *Saccharomyces cerevisiae* (S288C), *Schizosaccharomyces pombe* (972h-) and *Candida glabrata* (CBS138). Their average AT contents (62-64%) are 12-14% higher than those of Group I and Group II fungi (Table 2). These three yeast genomes also display a relatively normal distribution of AT local content (Figure 5A, the lowest panel) and have no or very few longer AT-rich blocks (Figure 5B).

## Repeat-Induced Point mutation (RIP) is identified in *Trichoderma reesei*

The high number of long AT-rich blocks in the Group I fungal genomes might be correlated with RIP: a phenomena originally discovered in *Neurospora crassa* at a premeiotic stage during sexual development [46]. The process of RIP requires a specialized cytosine methyltransferase gene *rid-1* (RIP-defective) and induces cytosine-to-thymine (C-to-T) point mutations in a homology-dependent manner [48, 49]. Several previous studies found no evidence of RIP in any of the Group II or Group III fungi we investigated here [50-52]. RIP has been documented in *Penicillium chrysogenum* [53], *Fusarium fujikuroi* [54], *Mycosphaerella graminicola*

[55] and *Aspergillus nidulans* [56], but it has never been experimentally demonstrated in any *Trichoderma* species [18, 57]. Intriguingly, the QM6a complete genome encodes almost all proteins known to be involved in RIP and DNA methylation in *Neurospora crassa*, including *rid1* (TrC1298W) [57], *dim2* (TrB0908W; DNA methyltransferase), *dim3* (TrE0517W; importin α), *dim5* (TrB0159C; H3K9 methyltransferase) [13], *dim7* (TrD0784W), *dim8* (TrG0915W), *dim9* (Trc0267C) and *hpo* (TrB1406W; HP1) (Additional File 2: Tables B1 and B2).

We carried out both bioinformatic and molecular genetic analyses to determine whether RIP could operate in *Trichoderma reesei*. The RIPCAL software tool was applied to compare differences in the extent of RIP mutations of different sequences by determining two widely-used RIP indices: TpA/ApT and [(CpA + TpG) / (ApC + GpT)] [58, 59]. Higher values of TpA/ApT and lower values of [(CpA + TpG) / (ApC + GpT)] indicate stronger RIP responses [60, 61]. We found that the hierarchy for RIP in QM6a is the mating type gene *MAT1-2-1* < all predicted genes < the whole QM6a genome < the large rDNA tandem repeats (18S-5.8S-26S) < 5S rDNAs. It has been reported that the large rDNA tandem repeats and the 5S rDNAs in *Neurospora crassa* survived RIPs due to either nucleolar sequestration [62] or their smaller size [60], respectively. The RIP indices of the mating gene (*MAT1-2-1 or MFa*), all predicted genes and the whole genome suggest that the RIP respones in QM6a are as robust as those in *Neurospora crassa* (Additional File 1: Table A13). Thus, we suggest that *Trichoderma reesei* has evolved an RIP system similar to that of *Neurospora crassa*. Our results also reveal that the longer the AT-rich blocks the greater the extent of RIP mutations in QM6a (Additional File 1: Table A13), consistent with RIP mutations accumulating in AT-rich sequences.

Next, we carried out sexual crossing tests between four parental (F0) strains, including wild-type CBS999.97 [7], *blr1*Δ, *env1*Δ [9, 10] and *ku70*Δ [63], and then tested their F1 progeny for mutations in the full-length hygromycin-resistant (*hph*) genes present in the selection marker construct used for deletion of the corresponding *Trichoderma reesei* genes [10]. Progeny were isolated using the hexadecad dissection technique [8]. No *hph* sequence is present in the wild-type CBS999.97 F0 strains, whereas *env1*Δ and *ku70*Δ each carries one copy of full-length *hph*. In contrast, the *blr1*Δ mutant contains two tandem head-to-tail *hph* sequences resulting in repeats; one is a full-length *hph* and the other an N-terminal truncated *hph*-ΔN (Figure 6A). All of the *hph* and *hph*-ΔN alleles in the corresponding parental strains were confirmed first by genomic PCR and Sanger sequencing (see Additional File 5 for their nucleotide sequences) and then by Southern hybridization (Figure 6B).

After sexual crossing (n = 10), the F1 progeny displayed numerous C-to-T point mutations in all cases where two similar sequences were present in one mating partner before crossing, i.e., in progeny of *blr1*Δ (*hph* and *hph*-ΔN), but not in progeny of *ku70*Δ or *env1*Δ that comprise only one copy of full-length *hph* in their genome (Figure 6C and Additional File 1: Table A14).

To determine whether RIP could operate between two genetically unlinked *hph* alleles, we generated a *ku70*Δ*env1*Δ double mutant that had one deletion on ChII (*env1*Δ) and the other on ChIII (*ku70*Δ). Our data revealed that sexual crossing (n = 10) between *ku70*Δ*env1*Δ and a wild-type mating partner also resulted in C-to-T point mutations in progeny of these crosses (Figure 6C and Additional File 5).

In *Neurospora crassa*, sequences mutated by RIP showed skewed dinucleotide frequencies because of the sequence preference of RIP (CpA > CpT > CpG > CpC) [64]. By comparing the nucleotide sequences of the *hph* alleles in all F1 progeny (Appendix A4), we found that *Trichoderma reesei* displayed a different sequence preference of RIP, i.e., CpA ≈ CpG >> CpT > CpC (Additional File 1: Table A14).

We conclude that *Trichoderma reesei*, like *Neurospora crassa*, exhibits high homology pairing and RIP activities at a premeiotic stage before premeiotic DNA synthesis. Our results are thus of tremendous importance for industrial strain improvement.

## Repetitive features

The completeness of the QM6a and other 11 high-quality fungal genomes also allowed us to accurately survey genome-wide repetitive features and their correlation to RIP using the RepeatMasker search program (http://www.repeatmasker.org/). We were able to identify almost all Ty elements along the 16 chromosomes of *Saccharomyces cerevisiae* [65, 66] (Additional File 1: Table A15). Our results also confirm that the genome of *Neurospora crassa* accumulates fragmented and rearranged transposon relics, in particular *gypsy*-LTRs (long terminal repeats) and Tad-LINEs (long interspersed repeat elements) [30]. The fungal wheat pathogen *Mycophaerella graminicola* has a high copy number of *gypsy*-LTRs, *copia*-LTRs and Tad-LINEs [55]. *gypsy*-LTRs are highly overrepresented in the genomes of *Schizosaccharomyces pombe* [67], *Cryptococcus neoformans* [68] and *Coprinopsis cinerea* [69]. The *Candida glabrata* CBS138 genome contains very few (~0.4%) repetitive sequences (http://www.candidagenome.org/) (Additional File 1: Table A16).

Intriguingly, the QM6a genome has the fewest transposon sequences among the six Group I fungal genomes (Additional File 1: Table A16). The majority of *copia*-LTRs (6/8), *gypsy*-LTRs (9/10), *CMC-EnSpm* (6/6) and *MULE-MuDR* (13/21) are located in longer AT-rich blocks. In contrast, most LINEs (16/18) are located in non-AT-rich regions (Additional File 1: Table A17). Neither centromeres nor telomeres show a preponderance of any particular type of transposable element (Additional File 1: Table A17). According to their RIP indices (Additional File 1: A13) and smaller size (Additional File 1: Tables A17 and A18), we conclude that almost all transposon sequences in QM6a are fragmented or rearranged transposon relics.

The copy numbers of transposon sequences in QM6a we report here (Additional File 1: Table A16) are much lower than those reported by Kubicek et al. using the QM6a-v2.0 draft [13]. Both we and Kubicek *et al.* used the RepeatMasker search program to search for repetitive sequences. It is important to point out that the limitation of this widely-used program is that it often generates many false-positive results. This is why we first applied it to the *Saccharomyces cerevisiae* genome to determine the optimum parameters for filtering the preliminary RepeatMasker data (see Materials and Methods). The number and locations of five different Ty elements in all of that 16 yeast's chromosomes had been determined before [65, 66].

## Partitioning of gene clusters by AT-rich blocks

A hallmark of the QM6a-v2.0 draft genome is that a third of the 228 CAZyme genes are non-randomly distributed and form several CAZyme gene clusters. Several of the regions of high CAZyme gene density also contain genes encoding proteins involved in secondary metabolism. Accordingly, it has been proposed that gene clustering and/or coexpression might be evolutionarily advantageous for *Trichoderma reesei* in its competitive soil habitat or other natural environments [11, 70]. Using the QM6a-HiC draft genome as a reference, 20 CAZyme gene clusters and 42 SSCP gene clusters were identified in the QM6a-HiC genome draft [18]. All these gene clusters consisted of only 3-6 CAZyme and/or SSCP genes but, surprisingly, gene clustering did not influence gene expression.

To gain a better insight, we reexamined all these 62 CAZyme and SSCP gene clusters in QM6a-HiC [18]. Due to sequencing and assembly errors, some gene clusters are overlapped or even duplicated. The original 62 gene clusters locate to 46 chromosomal blocks in the complete QM6a genome. The majority of these gene clusters also contain new QM6a genes we annotate in this study (Additional File 2: Table B4). One gene cluster even contains a counterfeit gene (QM6a-v2.0 gene number 71245) [18].

Intriguingly, except for 5 short gene clusters (≤ 4 contiguous genes), all the other 41 gene clusters in the complete QM6a genome are divided into smaller compartments by AT-rich blocks (Additional File 2: Table B4). The first example of this phenomenon is a nitrate assimilation gene cluster and an annotated CAZyme gene cluster [18] immediately adjacent to *tel2R*. These two gene clusters were divided by *tel2R* and five interspersed AT-rich blocks into five smaller compartments (Figure 7A). The first (or rightmost) compartment consists of one hypothetical protein gene (TrB1976W) and three nitrate reductase genes (*nit3*, TrB1975W; *nit2*, TrB1974C; *nit6*, TrB1973C); transcripts of these genes were barely detectable (TPMs < 0.5) under glucose for 48 hours, then in straw for 24 hours and finally with the addition of glucose for 5 hours. The second compartment comprises a sole nitrate transporter gene (*nit10*, TrB1972C); *nit10* was slightly induced by straw and then repressed by the addition of glucose. The third compartment contains a β-mannosidase gene, a member of glycoside hydrolase family 2 (GH2, TrB1971C) and a hypothetical gene (TrB1970W). Compared to *nit10,* this β-mannosidase gene exhibited ~ ~5-fold greater induction by straw. The fourth compartment has only a β-1,4-glucuronan lyase gene (*trgL*, TrB1969WW). Like the nitrate reductase genes, *trgL* did not respond to glucose or straw substrate. Finally, the last (or leftmost) compartment harbors two CAZyme genes and a hypothetical protein gene (TrB1966). The two CAZyme genes are the glucuronoyl esterase *cip2* (TrB2050W) and a GH30 endo-β-1,4-xylanase (TrB2049C) (Figure 7A). These two genes were highly induced by straw and then repressed by the addition of glucose (Additional File 1: Table A19A).

The second example is a CAZyme gene cluster on ChIV with 22 genes. It is divided by eight AT-rich blocks into seven smaller compartments (C1-C7). The four CAZyme genes are located in different compartments; a GH31 α-glucosidase (TrD1393W) in C1, *cel3c* (TrD1398C) in C3, and a GH28 polygalacturonase (TrD1411C) and *pgx1* (TrD1412C) in C6 (Figure 7B). These four CAZyme genes were differentially regulated by straw and the addition of glucose (Additional File 1: Table A19B).

The third example is the CAZyme gene cluster close to *tel7R*. Six AT-rich blocks partitioned this gene cluster into four smaller compartments (C1-C4). The four CAZyme genes were allocated to three different compartments; the *egl2/cel5* endo-β-1,4-glucanase (TrG1193W) in C1, GH79 ß-glucuronidase (TrG1202W) and rhamnogalacturonyl hydrolase (TrG1204W) in C2, and the *cbh* cellobiohydrolase (TrG1206C) in C4. Only *cbh* was highly induced by straw in QM6a (Additional File 1: Table A19C).

Together, these results might explain why it was previously reported that gene clustering did not influence gene expression [18]. We propose that these smaller compartments are structurally and functionally similar to intrachromosomal loops or TADs.

It should be noted that occurrence of intergenic AT-rich blocks did not always result in differential gene expression. For example, a gene cluster with three contiguous CAZyme genes—the acetyl xylan estrease *gene axe1* (TrE0669C; acetyl xylan estrease), the cellulose-induced protein *cip1* (TrE0670C) and the β-1,6-N-acetylglucosaminyl transferase gene *egl4* (TrE0671C)—resides in the middle of ChV. These three CAZyme genes are separated by four AT-rich blocks into three smaller compartments, but they were all highly induced by straw and then repressed by the addition of glucose (Additional File 1: Table A19D). In this case, each gene becomes a partitioned functional unit (see Discussion). Simultaneous expression of the *axe1-cip1-egl4* triad might be independently controlled by other determinants, e.g., common transcription factors and/or similar chromosomal conformation.

## Discussion

*Trichoderma reesei* QM6a and its derivatives have been widely used for nearly four decades to produce plant cell wall-degrading enzymes and heterologous recombinant proteins. In this study, we have obtained a high-quality complete genome sequence of QM6a. We readily uncovered many sequencing, assembly and gap-closing errors in earlier draft genomes. The seven telomere-to-telomere QM6a chromosomes can be used as better scaffolds for comparative genomic analyses, not only with industrial strains but also other *Trichoderma reesei* wild isolates (e.g., CBS999.97) and other species in the same genus (e.g., *Trichoderma atroviride* and *Trichoderma virens*). Our results also revealed much new genomic information never provided by earlier draft genomes, including 7 centromeres, 14 telomeres, 2328 AT-rich blocks, 1630 newly-predicted genes, 37 centromere-encoded genes and 24 centromeric repeats. Therefore, our complete QM6a genome sequence provides a comprehensive roadmap for further studies of this economically-important fungus, including industrial strain improvements and elucidation of the functional relationships between sequences, gene products and genome organization.

The central finding of this study is that *Trichoderma reesei* has evolved a robust RIP system. Firstly, the QM6a genome contains the lowest overall copy number of transposons among six studied filamentous ascomycetes (including *Neurospora crassa*). Secondly, as in *Neurospora crassa*, sexual crossing readily induced C-to-T point mutations on both tandem and unlinked duplicated sequences in *Trichoderma reesei*. Thirdly, almost all 2349 AT-rich blocks in QM6a were predicted by the RIPCAL software program to be affected by RIP. Considerable evidence suggests that AT-rich blocks establish a link between DNA sequence and chromatin architecture. In *Neurospora crassa*, AT-rich blocks form constitutive heterochromatins and mediate intra- and inter-chromosomal contacts [46, 47]. In *Saccharomyces cerevisiae*, AT-rich blocks constitute the chromosome axes or the bases of chromosomal loops or TADs [35, 38-40]. In mammalian interphase chromosomes, the spatial distribution of AT-rich blocks (e.g., lamina-associated domains, LADs) and GC-rich blocks (e.g., TADs or chromosomal loops) are evolutionarily conserved. LADs preferentially interact with other LADs, whereas TADs exhibit more localized chromosomal domains [71]. Intriguingly, our results reveal that the rDNA locus of QM6a is surrounded by two interspersed AT-rich blocks. It would be of interest to further investigate whether and how these two interspersed AT-rich blocks are involved in preventing rDNA from being affected by RIP as well as in regulating nucleolar organization, rRNA transcription and rDNA copy homeostasis.

From the results of this study, we postulate that RIP does not function solely as a genome defense mechanism to diminish the potentially deleterious effects caused by the spread of transposable elements. It may also have important roles in reshaping the *Trichoderma reesei* genome. We demonstrate that the widespread interspersed AT-rich blocks lead to genome-wide partitioning of the gene clusters in QM6a. Our findings can readily account for why gene clustering does not affect gene expression in *Trichoderma reesei*. Mechanistically, RIP-mediated C-to-T mutations presumably can transform duplicated sequences in a CAZyme gene cluster into interspersed AT-rich blocks, thus dividing an ancestral gene cluster with multiple CAZyme or SSCP genes into multiple smaller compartments or TADs. Intriguingly, it has been reported previously that many pathogenic fungi (*Leptosphaeria maculans*, *Magnaporthe oryzae*, *Fusarium* spp.) comprise AT-rich blocks with RIP affected effector genes and transposable elements [72, 73] and that RIP is a potential factor in *Leptosphaeria maculans* in creating the rapid sequence diversification (i.e., of the effector genes) needed for selection pressure [74] and concerted epigenetic regulation of their expression [75]. Partitioning of gene clusters by AT-rich blocks may also help to control simultaneous expression of the rDNA locus (Figure 3) and of the three partitioned functional units in the *axe1*-*cip1*-*egl4* triad (Additional file 1: Table A19D). Further research will reveal how RIP provides evolutionary advantages to *Trichoderma reesei* and other filamentous ascomycetes (Pezizomycotina) to survive in natural environments and pathogenic conditions.

## Conclusion

The earlier genome drafts of QM6a and other *T. reesei* industrial strains have been useful in identifying key genes involved in several important biological processes. However, due to numerous sequencing and assembly errors, they are not suitable for several other studies, e.g., genome-wide determination of gene order, chromosome architecture and expression dynamics as well as chromosome engineering for genetic modification(s). The complete QM6a genome sequence provides an unprecedented opportunity to overcome those obstacles associated with earlier draft genomes. To avoid the limitations of working with incomplete datasets and the false leads that can come from trying to work with imperfect data, more caution should be exercised in utilizing the draft genome sequences solely determined by SGS and/or HiC for functional and comparative analyses.

## Materials and Methods

### Fungal growth, DNA preparation and pulsed-field gel electrophoresis (PFGE)

QM6a was inoculated on petri-plates with malt extract agar (MEA) medium at 25 °C until full asexual sporulation was observed (~5 days). $2 \times 10^8$ conidial spores were collected, and then inoculated in 50 mL potato dextrose medium (PDB) at 30 °C for 6 h. The germinate hyphae were harvested by centrifugation at 3000*g* for 5 min at room temperature and incubated in 2 mL lysing enzyme buffer [0.1 M $KH_2PO_4$ (pH 5), 1.2M Sorbitol, 5% lysing enzyme (Sigma, USA)] at 30 °C for 1.5 h. The protoplasts were harvested by centrifugation at 600*g* for 10 min at 4 °C, dissolved in 1.2 ml GUHCl solution (43% guanidine-HCl, 0.1M EDTA pH8.0, 0.15M NaCl, 0.05% Sarkosyl) at 65 °C for 20 min and then mixed with 6.4 mL ice-cold ethanol to precipitate the genomic DNA. The pellet was dissolved in 10X TE with 0.6 mg/ml RNaseH at 37 °C for 1hr and then in 0.4 mg/mL proteinase K at 65 °C for 1hr. The genomic DNA was purified with the phenol:chloroform:isoamyl alcohol (25:24:1) method and then recovered by standard precipitation with ethanol. Next, the quality of high molecular weight genomic DNA for Illumina MiSeq and PacBio sequencing was validated by PFGE. The genomic DNA was separated in a 1% agarose gel in 0.5x TBE buffer, using a CHEF DR II (Biorad) with 0.5x TBE running buffer, continuously refrigerated at 14 °C and 6 V/cm (current 110-125 mA) for 18 h. The

Lambda DNA-MonoCut Mix (New England Biolabs, N3019S) was used as size marker. Visualization was performed after staining with ethidium bromide after the electrophoresis.

### DNA sequencing and *de novo* assembly

Illumina MiSeq sequencing was carried out at the DNA sequencing facility in the Institute of Molecular Biology, Academia Sinica (Taipei, Taiwan) in May 2016 (sequencing was coordinated by the author SYT). A shotgun paired-end library (average size of 550 bp) was prepared using the Illumina TruSeq DNA nano Sample Prep Kit. The Illumina library (including PCR amplification and quantification) was prepared automatically by the NeoPrep Library Prep System, and then sequenced on the Miseq platform (300 cycles, paired-end sequencing) with the Miseq Control software version 2.5.1 and Sequencing Analysis Viewer version 1.8.20. Sequencing data were sent to Illumina BaseSpace automatic analysis during running. For report version 2.2.9 with MiSeq 600 cycler V3 chemistry, 96.16% of clusters passed the filter and 78.32% of bases qualified higher than Q30 at 2x300 n.

For Pacbio continuous long read sequencing, high-quality genomic DNA was submitted to the Ramaciotti Centre for Genomics (University of New South Wales, Sydney, Australia). Sequencing was coordinated by Carolina Correa and Tonia Russel. High molecular weight DNA was sheared with g-TUBE (Covaris PN 520079), aiming at DNA fragments of about 20 kb. The library was constructed with a 20kb size-selected protocol using DNA Template Prep Kit 2.0 (PN 001-540-726), purified and further selected for long insert size with a 0.35X AMPure (AMPure PB PN 100-265-900) bead size selection. The library was sequenced on a PacBio RSII device using the reagents DNA/Polymerase Binding Kit P4 (PN 100-236-500), DNA Sequencing

Reagent 2.0 PN (100-216-400), SMRT®Cell V3 (PN 100-171-800) and Magbead (PN 100-133-600), loading at 200 pM on the plate. Data were collected with Stage Start and 180 minute movies. Seven SMRT®Cell cells were used to generate 263,312 reads and 3,397,762,180 bases.

The Pacbio data were assembled with the SMRT Analysis Software v2.3.0 (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/). The assembly was performed with HGAP (3.0 protocol) with the following parameters: (1) PreAssembler Filter v1 (minimum sub-read length = 500bp, minimum polymerase read quality = 0.80, minimum polymerase read length = 100bp); (2) PreAssembler v2 (minimum seed length = 10,000bp, number of seed read chunks = 6, alignment candidates per chunk = 10, total alignment candidates = 24, min coverage for correction = 6); (3) AssembleUnitig v1 (Genome Size 34000000bp, target genome coverage = 30, overlap error rate = 0.06, minimum overlap = 100bp and overlap k-mer = 22); (4) BLASR v1 mapping of reads for genome polishing with Quiver (max divergence percentage = 30, minimum anchor size = 12). The final assembly contained nine unitigs (seven telomere-to-telomere chromosomes, a circular mitochondrial genome and a 1870bp short sequence) with a total 34,993,035 bp and approximately 87x genome coverage. N50 was 5,311,312 bp and max was 6,835,650 bp. The per-base Quality Value (QV) is higher than 45 and a little less than 50 (99.999% accuracy; average 48.8). A BLASTN search revealed that the circular mitochondrial genome (42,130 bp) was completely identical in nucleotide sequence with the QM9414 mitochondrial genome (accession: NC_003388.1) [21]. The shortest unitig (1870 bp) is likely contaminating DNA because no Illumina MiSeq reads were mapped to its sequence.

## Mapping of Illumina reads over three different genomic drafts

The two runs of Illumina MiSeq with 300-bp paired-end reads were combined into one. The forward and reverse data were 6.8 Gb. Reads were preprocessed with Trimmomatic V0.32 (http://www.usadellab.org/cms/?page=trimmomatic) to trim and remove reads (33.15%) that fell below a quality score threshold of 30 (Q30) and were shorter than 30bp. Next, CLC Genomics Workbench 7.5 (http://www.clcbio.com/blog/clc-genomics-workbench-7-5/) was used to map the qualified reads to three different QM6a genomic drafts: (1) QM6a-v2.0 (http://genome.jgi.doe.gov/Trire2/Trire2.home.html) plus the complete mitochondrial genome [20]; (2) the HiC draft genome [16] plus the complete mitochondrial genome [20]; (3) the PacBio *de novo* assembly (this study). The parameters used for mapping were (1) Mismatch cost = 3; (2) Insertion cost = 2; (3) Deletion cost = 2; (4) Length fraction = 0.8; (5) Similarity fraction = 0.8. The results are listed in Table S4. There were some differences between the PacBio and Illumina platforms, so the mapping results were used to extract a consensus sequence to adjust the bases to a final version of the chromosome sequences. We defined a threshold=2 to identify low coverage regions. For low coverage regions (i.e. threshold ≥2), sequences from the Illumina Miseq platform were used to construct the consensus sequence.

The seven telomere-to-telomere chromosomes were categorized with Roman numerals (ChI-ChVII), from the largest to the smallest. Artemis was used to determine the percentage of G+C in 500 bp non-overlapping windows. The centromere of each chromosome was their longest interspersed AT-rich block. The EMBOSS revseq software tool (v6.5.7; http://www.bioinformatics.nl/cgi-bin/emboss/revseq) was used to reverse and complement the nucleotide sequences of ChII, ChIV and ChVI, respectively, so that all seven chromosomes had a shorter (left) arm at 5′ termini and a longer (right) arm at 3′ termini. The final sequences were first compared to the original version using BLAST searches (Additional File 1: Table A1) before being submitted to NCBI (accession numbers CP016232-CP016238).

### Genome reannotation

The MAKER2 v2.31.8 (http://www.yandell-lab.org/software/maker.html) genome annotation pipeline [74] was applied for genome reannotation. Firstly, all (28748) *T. reesei* protein sequences from NCBI were used for *ab initio* gene predictions. Secondly, the Augustus v3.0.3 gene prediction software (http://augustus.gobics.de/) [75] was also used to predict new genes. All proteins from *Neurospora spp.* and *Fusarium spp.* were used for Augustus training to impose constraints on the predicted gene structure, including splice sites, translation initiation sites and stop codons. Thirdly, we isolated QM6a poly(A) RNA and performed a MiSeq RNA sequencing experiment. The resulting mRNA reads (trimming quality=30, min length=20bp) were applied to Trinity (v2.0.6; https://github.com/trinityrnaseq/trinityrnaseq/wiki) for *de novo* transcriptome assembly. The expressed sequence tag (EST) results predicted by “Maker” were used to identify new genes. Finally, we integrated all the ESTs and protein sequences from the two publicly-available databases (QM6a-v2.0 and RUT-C30-v1.0; at JGI) as well as all the ESTs assembled from two published transcriptome datasets [24, 25].

**A. QM6a-v2.0**

(http://genome.jgi.doe.gov/Trire2/Trire2.home.html)

TreeseiV2_FilteredModelsv2.0.transcripts.fasta

TreeseiV2_FilteredModelsv2.0.proteins.fasta

**B. RUT-C30-v1.0**

(http://genome.jgi.doe.gov/TrireRUTC30_1/TrireRUTC30_1.home.html)

TrireRUTC30_1_GeneCatalog_transcripts_20110526.nt.fasta

TrireRUTC30_1_GeneCatalog_proteins_20110526.aa.fasta.

### C. The Illumina HiSeq 2000 sequencing reads from QM9414 treated with cellulose (24, 48 and 72 hours), sophorose (2, 4 and 6 hours) and glucose (24 and 48 hours) [25].

GSE53629 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53629)

SRR1057947: QM9414 Cellulose replication 1

SRR1057948: QM9414 Cellulose replication 2

SRR1057949: QM9414 Cellulose replication 3

SRR1057950: QM9414 Sophorose replication 1

SRR1057951: QM9414 Sophorose replication 2

SRR1057952: QM9414 Sophorose replication 3

SRR1057953: QM9414 Glucose replication 1

SRR1057954: QM9414 Glucose replication 2

SRR1057955: QM9414 Glucose replication 3

The Illumina-HiSeq .sra files were converted into fastq using the fastq-dump program of the SRA Toolkit. Trimmomatic v0.32 [76] was used to trim and remove reads that fell below a quality score threshold of 30 (Q30) and that were shorter than 10bp. The Cufflinks pipeline tools (http://cole-trapnell-lab.github.io/cufflinks/) were downloaded for transcriptome assembly and differential expression analysis, including bowtie2-build (bowtie v2.2.3), tophat2 (v2.0.13) and cufflinks (v2.2.1). For read alignment, the parameters of tophat2 were: (1) min-intron-length: 20 bp; (2) max-intron-length: 5000bp (to reflect introns and splicing elements of five diverse fungi); and (3) No transcript GTF (Gene Transfer Format) file was provided to guide assembly. There were ~95% trimmed paired reads that could be aligned to the QM6a complete genome. After each QM9414 RNA-seq raw dataset was assembled into transcripts, the cuffmerge program was used to merge all GTF files into one. The gffread program was used to extract transcript sequences.

**D. The SOLiD sequencing reads from QM9a grown first in glucose for 48 hours, then in straw for 24 hours and finally with the addition of glucose for 5 hours [24].**

GSE44648 (http://www.ncbi. nlm.nih.gov/geo/query/acc.cgi?acc=GSE44648)

SRR764963: QM6a Glucose 48 h replication 1

SRR764964: QM6a Glucose 48 h replication 2

SRR764965: QM6a Glucose 48 h replication 3

SRR764966: QM6a Straw 24 h replication 1

SRR764967: QM6a Straw 24 h replication 2

SRR764968: QM6a Straw 24 h replication 3

SRR764969: QM6a Straw + Glucose 5 h replication 1

SRR764970: QM6a Straw + Glucose 5 h replication 2

The SOLID .sra files were converted into csfasta and QV.qual using the abi-dump program of the SRA Toolkit. For transcriptome assembly and quantification, bowtie-build (bowtie v1.1.0), tophat2 and cufflinks were used. All parameters were as for the QM9414 dataset, except the colorspace option was used for the bowtie1 program. There were ~33% reads that could be aligned to the QM6a complete genome.

Next, we performed gene filtering to finalize all predicted protein sequences, the filtering order was all *Trichoderma reesei* protein sequences from NCBI > *Neurospora spp.* and *Fusarium spp.* (Augustus v3.0.3) > *de novo* assembly QM6a-

RNA (Trinity v2.0.6) > QM6a v2.0 + RutC-30 v1.0 + QM6a-RNA (SOLiD) + QM9414-RNA (Illumina HiSeq 2000).

We also manually integrated almost all the annotation results reported by the *Trichoderma* research community [4, 18, 24, 25]. We annotated 10786 predicted genes that covered almost all previously annotated genes in QM6a-v2.0 (9105 out of 9129) and RUT-C30-v1.0 (9717 out of 9852), respectively. Of these predicted QM6a genes, 904 were previously considered to be RUT-C30-specific. We successfully annotated several evolutionarily-conserved DNA repair genes, including *rad50*, *sae2*/*com1*/CtIP, *rad51*, *rad57*, *srs2* and *pif1* (Additional File 2: Tables B1 and B2). These genes were not annotated in QM6a-v2.0 (JGI), RUT-C30-v1.0 (JGI) and QM6a-HiC [18].

Finally, the BLASTP program was applied to compare our final predicted protein sequences to several publicly-available protein databases, including NCBI non-redundant (nr) database, Universal Protein Resource (Uniprot) Prot v2.0, *Neurospora crassa* database (Broad Institute), *Fusarium fujikuroi* (JGI), *Sordaria macrospora* (NCBI), *Saccharomyces cerevisiae* (SGD), and *Schizosaccharomyces pombe* (PomBase and NCBI). The e-values of all BLAST results were $< 1.0 \times 10^{-5}$. The Retrieve/ID mapping tool (Uniprot) was used to map gene ID (Uniprot sprot v2.0 and NCBI BLAST result) and to determine gene ontology (GO) (Additional File 2: Tables B2).

Next, the bowtie-tophat program was used to align all QM6a (SOLiD) and QM9414 (Illumina MiSeq) RNA reads with the following parameters: (1) min-intron-length: 20bp; (2) max-intron-length: 5000 bp (because it was reported previously that the ranges of intron length in *Saccharomyces cerevisiae*, *Aspergillus nidulans* and *Neurospora crassa* are 52-1002 bp, 27-1903 bp and 46-1740 bp, respectively [76]);

and (3) a transcript GTF (Gene Transfer Format) file was provided. To calculate read counts for each transcript, we applied featureCounts [77]. The values of Transcripts Per kilobase Million (TPM) [26] were used as expression values: TPM = (individual gene RPK/the sum of all RPKs) x $10^6$, whereas RPK (Reads per kilobase) = (read counts/transcription length) (Additional File 2: Tables B1 and Additional File 3).

## Repeat Modeler and RepeatMasker

Novel repeat elements were identified by Repeat Modeler-1.0.4 (http://www.repeatmasker.org/RepeatModeler.html) with default parameters. RepeatMasker (version 4.0.6) and the Repbase Library (http://www.repeatmasker.org) were used to scan 12 different fungal genomes for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence, as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). To obtain high-confidence data, we first analyzed the genome sequences of *Saccharomyces cerevisiae* because the number and location of five different Ty elements in all its 16 yeast chromosomes had been reported previously [65, 66]. When the preliminary RepeatMasker data were filtered with two parameters (length >= 140, Smith–Waterman local similarity scores >= 450), the final data (Additional File 1: Table A15) were quite consistent with the known results [65]. The same parameters were then applied to scan all other studied fungal genomes (Additional file 1: Table A16).

## RIP and RIPCAL

The latest RICAL program (https://sourceforge.net/projects/ripcal/) was downloaded

[58, 59]. An initial analysis was done to predict if the complete QM6a genome sequence had been mutated by RIP. All strains used in this study had been previously described [10]. The methods for sexual crossing, single ascospore isolation and preparation of genomic DNA have also been described before [8, 10]. The full-length *hgh* cassettes in all F0 parental strains and representative F1 progeny were amplified by PCR and analyzed by Sanger sequencing technology. All the nucleotide sequences of primers and the *hgh* cassettes are listed in Additional file 1: Table A9 and Additional file 5, respectively. The RIPCAL dinucleotide frequency and EMBOSS compseq tool were used to determine the values of two RIP indices, TpA/ApT and [(CpA + TpG) / (ApC + GpT)], respectively.

## Declarations

### Authors' contributions

WCL and CLC isolated the fungal genomic DNA for sequencing and conducted the sexual crossing experiments. CHH, WCL and YCC carried out bioinformatic analyses. SYT performed Illumina Miseq sequencing experiments. TFW conceived and designed the experiments, and analyzed the data. TFW and WCL wrote the paper. All authors read and approved the final manuscript.

### Acknowledgments

We thank Paul Hsu at the IMB Bioinformatics Service Core, Yu-Tang Huang at the IMB Computer Room and the IMB Genomics Core for technical support, and John O′Brien for English editing. TFW would like to thank Frédérique Bidard-Michelot (IFP Energies Nouvelles, Paris, France) for personal communications on unpublished

results of RUT-C30 genome assembly and a common consensus for QM6a chromosome numbering, ordering and orientation.

## Competing interests

The authors have declared that no competing interests exist.

## Availability of data and materials

- All data generated or analyzed during this study are included in this published article and its supplementary information files. Data deposition: BioProject (PRJNA325840), BioSample (SAMN05250858), QM6a genomic DNA illumina Mi-seq (SRR4417032), QM6a Poly(A) RNA illumina Mi-seq (SRR5229930) and the nucleotide sequences of seven complete QM6a chromosomes (CP016232-CP016238).

## Funding

This work received supports from Academia Sinica, Taipei, Taiwan to TFW. The funding body was not involved in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

## Ethical approval

This study needed no ethical approval and had no experimental research on humans.

## References

1. Reese ET: **History of the cellulase program at the U.S. army Natick Development Center**. *Biotechnol Bioeng Symp* 1976:9-20.

2. Peterson R, Nevalainen H: ***Trichoderma reesei*** **RUT-C30--thirty years of strain improvement**. *Microbiology* 2012, **158**(:58-68.

3. Glass NL, Schmoll M, Cate JH, Coradetti S: **Plant cell wall deconstruction by ascomycete fungi**. *Annu Rev Microbiol* 2013, **67**:477-498.

4. Schmoll M, Dattenbock C, Carreras-Villasenor N, Mendoza-Mendoza A, Tisch D, Aleman MI, Baker SE, Brown C, Cervantes-Badillo MG, Cetz-Chel J *et al*: **The Genomes of Three Uneven Siblings: Footprints of the Lifestyles of Three** ***Trichoderma*** **Species**. *Microbiol Mol Biol Rev* 2016, **80**:205-327.

5. Bischof RH, Ramoni J, Seiboth B: **Cellulases and beyond: the first 70 years of the enzyme producer** ***Trichoderma reesei***. *Microbial Cell Fact* 2016, **15**(1):106.

6. Druzhinina IS, Kubicek CP: **Familiar stranger: ecological genomics of the model saprotroph and industrial enzyme producer** ***Trichoderma reesei*** **breaks the stereotypes**. *Adv Appl Microbiol* 2016, **95**:69-147.

7. Seidl V, Seibel C, Kubicek CP, Schmoll M: **Sexual development in the industrial workhorse** ***Trichoderma reesei***. *Proc Nat Acad Sci* 2009, **106**:13909-13914.

8. Chuang YC, Li WC, Chen CL, Hsu PW, Tung SY, Kuo HC, Schmoll M, Wang TF: ***Trichoderma reesei*** **meiosis generates segmentally aneuploid progeny with higher xylanase-producing capability**. *Biotech for Biofuels* 2015, **8**:30.

9. Seibel C, Tisch D, Kubicek CP, Schmoll M: **ENVOY is a major determinant in regulation of sexual development in** ***Hypocrea jecorina*** **(*****Trichoderma***

***reesei*)**. *Eukaryotic Cell* 2012, **11**:885-895.

10. Chen CL, Kuo HC, Tung SY, Hsu PW, Wang CL, Seibel C, Schmoll M, Chen RS, Wang TF: **Blue light acts as a double-edged sword in regulating sexual development of *Hypocrea jecorina* (*Trichoderma reesei*)**. *PloS One* 2012, **7**:e44969.

11. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D *et al*: **Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*)**. *Nat Biotechnol* 2008, **26**:553-560.

12. Koike H, Aerts A, LaButti K, Grigoriev IV, Baker SE: **Comparative Genomics Analysis of *Trichoderma reesei* Strains**. *Indust Biotech* 2013, **9**:352-367.

13. Kubicek CP, Herrera-Estrella A, Seidl-Seiboth V, Martinez DA, Druzhinina IS, Thon M, Zeilinger S, Casas-Flores S, Horwitz BA, Mukherjee PK *et al*: **Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma***. *Genome Biol* 2011, **12**:R40.

14. Arvas M, Pakula T, Smit B, Rautio J, Koivistoinen H, Jouhten P, Lindfors E, Wiebe M, Penttila M, Saloheimo M: **Correlation of gene expression and protein production rate - a system wide study**. *BMC Genomics* 2011, **12**:616.

15. Hakkinen M, Valkonen MJ, Westerholm-Parvinen A, Aro N, Arvas M, Vitikainen M, Penttila M, Saloheimo M, Pakula TM: **Screening of candidate regulators for cellulase and hemicellulase production in *Trichoderma reesei* and identification of a factor essential for cellulase production**. *Biotech for Biofuels* 2014, **7**:14.

16. Vitikainen M, Arvas M, Pakula T, Oja M, Penttila M, Saloheimo M: **Array comparative genomic hybridization analysis of *Trichoderma reesei* strains with enhanced cellulase production properties**. *BMC genomics* 2010, **11**:441.

17. Marie-Nelly H, Marbouty M, Cournac A, Flot JF, Liti G, Parodi DP, Syan S, Guillen N, Margeot A, Zimmer C *et al*: **High-quality genome (re)assembly using chromosomal contact data**. *Nat Commun* 2014, **5**:5695.

18. Druzhinina IS, Kopchinskiy AG, Kubicek EM, Kubicek CP: **A complete annotation of the chromosomes of the cellulase producer *Trichoderma reesei* provides insights in gene clusters, their expression and reveals genes required for fitness**. *Biotech for Biofuels* 2016, **9**:75.

19. Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin CS, Rapicavoli NA, Rank DR, Li J *et al*: **Long-read, whole-genome shotgun sequence data for five model organisms**. *Sci Data* 2014, **1**:140045.

20. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE *et al*: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**. *Nat Methods* 2013, **10**:563-569.

21. Chambergo FS, Bonaccorsi ED, Ferreira AJ, Ramos AS, Ferreira Junior JR, Abrahao-Neto J, Farah JP, El-Dorry H: **Elucidation of the metabolic fate of glucose in the filamentous fungus *Trichoderma reesei* using expressed sequence tag (EST) analysis and cDNA microarrays**. *J Biol Chem* 2002, **277**:13983-13988.

22. Aign V, Schulte U, Hoheisel JD: **Hybridization-based mapping of *Neurospora crassa* linkage groups II and V**. *Genetics* 2001, **157**:1015-1020.

23. Li W-C, Chuang Y-C, Chen C-L, Wang T-F: **Hybrid infertility: The dilemma or opportunity of applying sexual development to improve *Trichoderma reesei* industrial strains**. In: *Gene Expression Systems in Fungi: Advancements and Applications.* Edited by Schmoll M, Dattenböck C. Cham: Springer International Publishing; 2016: 351-359.

24. Ries L, Pullan ST, Delmas S, Malla S, Blythe MJ, Archer DB: **Genome-wide transcriptional response of *Trichoderma reesei* to lignocellulose using RNA sequencing and comparison with *Aspergillus niger***. *BMC genomics* 2013, **14**:541.

25. Dos Santos Castro L, Pedersoli WR, Antonieto AC, Steindorff AS, Silva-Rocha R, Martinez-Rossi NM, Rossi A, Brown NA, Goldman GH, Faca VM *et al*: **Comparative metabolism of cellulose, sophorose and glucose in *Trichoderma reesei* using high-throughput genomic and proteomic analyses**. *Biotech for Biofuels* 2014, **7**:41.

26. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**:493-500.

27. Wagner GP, Kin K, Lynch VJ: **Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples**. *Theory in Biosciences* (*Theorie in den Biowissenschaften*) 2012, **131**:281-285.

28. Schmoll M, Seibel C, Tisch D, Dorrer M, Kubicek CP: **A novel class of peptide pheromone precursors in ascomycetous fungi**. *Mol Microbiol* 2010, **77**:1483-1501.

29. Tachibana C, Yoo JY, Tagne JB, Kacherovsky N, Lee TI, Young ET: **Combined global localization analysis and transcriptome data identify**

**genes that are directly coregulated by Adr1 and Cat8**. *Mol Cell Biol* 2005, **25**:2138-2146.

30. Borkovich KA, Alex LA, Yarden O, Freitag M, Turner GE, Read ND, Seiler S, Bell-Pedersen D, Paietta J, Plesofsky N *et al*: **Lessons from the genome sequence of *Neurospora crassa*: tracing the path from genomic blueprint to multicellular organism**. *Microbiol Mol Biol Rev* 2004, **68**:1-108.

31. Topp CN, Zhong CX, Dawe RK: **Centromere-encoded RNAs are integral components of the maize kinetochore**. *Proc Nat Acad Sci* 2004, **101**:15986-15991.

32. Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KH, Wong LH: **Active transcription and essential role of RNA polymerase II at the centromere during mitosis**. *Proc Nat Acad Sci* 2012, **109**:1979-1984.

33. Choi ES, Stralfors A, Castillo AG, Durand-Dubief M, Ekwall K, Allshire RC: **Identification of noncoding transcripts from within CENP-A chromatin at fission yeast centromeres**. *J Biol Chem* 2011, **286**:23600-23607.

34. Gent JI, Dawe RK: **RNA as a structural and regulatory component of the centromere**. *Annu Rev Genet* 2012, **46**:443-453.

35. Blat Y, Kleckner N: **Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region**. *Cell* 1999, **98**:249-259.

36. Laloraya S, Guacci V, Koshland D: **Chromosomal addresses of the cohesin component Mcd1p**. *J Cell Biol* 2000, **151**:1047-1056.

37. Glynn EF, Megee PC, Yu HG, Mistrot C, Unal E, Koshland DE, DeRisi JL, Gerton JL: **Genome-wide mapping of the cohesin complex in the yeast *Saccharomyces cerevisiae***. *PLoS Biol* 2004, **2**:E259.

38. Gasser SM: **Chromosome structure. Coiling up chromosomes**. *Curr Biol* 1995, **5**:357-360.

39. Kleckner N: **Meiosis: how could it work?** *Proc Nat Acad Sci* 1996, **93**:8167-8174.

40. Guacci V, Koshland D, Strunnikov A: **A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of *MCD1* in *S. cerevisiae***. *Cell* 1997, **91**:47-57.

41. Hou C, Li L, Qin ZS, Corces VG: **Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains**. *Mol cell* 2012, **48**:471-484.

42. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression**. *Nature Genet* 2000, **26**:183-186.

43. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J *et al*: **Spatial partitioning of the regulatory landscape of the X-inactivation centre**. *Nature* 2012, **485**:381-385.

44. Cavalli G, Misteli T: **Functional implications of genome topology**. *Nature Stru Mol Bio* 2013, **20**:290-299.

45. Gibcus JH, Dekker J: **The hierarchy of the 3D genome**. *Mol cell* 2013, **49**:773-782.

46. Aramayo R, Selker EU: ***Neurospora crassa*, a model system for epigenetics research**. *Cold Spring Harb Perspect Biol* 2013, **5**:a017921.

47. Galazka JM, Klocko AD, Uesaka M, Honda S, Selker EU, Freitag M: ***Neurospora* chromosomes are organized by blocks of importin alpha-**

**dependent heterochromatin that are largely independent of H3K9me3**. *Genome Res* 2016, **26**:1069-1080.

48. Cambareri EB, Jensen BC, Schabtach E, Selker EU: **Repeat-induced G-C to A-T mutations in *Neurospora***. *Science* 1989, **244**:1571-1575.

49. Freitag M, Williams RL, Kothe GO, Selker EU: **A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa***. *Proc Nat Acad Sci* 2002, **99**:8802-8807.

50. Clutterbuck AJ: **Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes**. *Fungal Genetics and Biology* 2011, **48**:306-326.

51. Horns F, Petit E, Yockteng R, Hood ME: **Patterns of repeat-induced point mutation in transposable elements of basidiomycete fungi**. *Genome Biol Evol* 2012, **4**:240-247.

52. Wang X, Hsueh YP, Li W, Floyd A, Skalsky R, Heitman J: **Sex-induced silencing defends the genome of *Cryptococcus neoformans* via RNAi**. *Genes Dev* 2010, **24**:2566-2582.

53. Braumann I, van den Berg M, Kempken F: **Repeat induced point mutation in two asexual fungi, *Aspergillus niger* and *Penicillium chrysogenum***. *Curr Genet* 2008, **53**:287-297.

54. Chiara M, Fanelli F, Mule G, Logrieco AF, Pesole G, Leslie JF, Horner DS, Toomajian C: **Genome sequencing of multiple isolates highlights subtelomeric genomic diversity within *Fusarium fujikuroi***. *Genome Biol Evol* 2015, **7**:3062-3069.

55. Dhillon B, Gill N, Hamelin RC, Goodwin SB: **The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola***. *BMC genomics* 2014, **15**:1132.

56. Clutterbuck AJ: **MATE transposable elements in *Aspergillus nidulans*: evidence of repeat-induced point mutation**. *Fungal Genet Biol* 2004, **41**:308-316.

57. Schuster A, Bruno KS, Collett JR, Baker SE, Seiboth B, Kubicek CP, Schmoll M: **A versatile toolkit for high throughput functional genomics with *Trichoderma reesei***. *Biotech for Biofuels* 2012, **5**:1.

58. Hane JK, Oliver RP: **RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences**. *BMC Bioinformatics* 2008, **9**:478.

59. Hane JK, Oliver RP: **In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes**. *BMC Genomics* 2010, **11**:655.

60. Watters MK, Randall TA, Margolin BS, Selker EU, Stadler DR: **Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in *Neurospora***. *Genetics* 1999, **153**:705-714.

61. Selker EU, Tountas NA, Cross SH, Margolin BS, Murphy JG, Bird AP, Freitag M: **The methylated component of the *Neurospora crassa* genome**. *Nature* 2003, **422**:893-897.

62. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S *et al*: **The genome sequence of the filamentous fungus *Neurospora crassa***. *Nature* 2003, **422**:859-868.

63. Guangtao Z, Hartl L, Schuster A, Polak S, Schmoll M, Wang T, Seidl V, Seiboth B: **Gene targeting in a nonhomologous end joining deficient *Hypocrea jecorina***. *J Biotech* 2009, **139**:146-151.

64. Margolin BS, Garrett-Engele PW, Stevens JN, Fritz DY, Garrett-Engele C,

Metzenberg RL, Selker EU: **A methylated *Neurospora* 5S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation**. *Genetics* 1998, **149**:1787-1797.

65. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF: **Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence**. *Genome Res* 1998, **8**:464-478.

66. Jordan IK, McDonald JF: **Comparative genomics and evolutionary dynamics of *Saccharomyces cerevisiae* Ty elements**. *Genetica* 1999, **107**:3-13.

67. Bowen NJ, Jordan IK, Epstein JA, Wood V, Levin HL: **Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe***. *Genome Res* 2003, **13**:1984-1997.

68. Goodwin TJ, Poulter RT: **The diversity of retrotransposons in the yeast *Cryptococcus neoformans***. *Yeast* 2001, **18**:865-880.

69. Labbe J, Murat C, Morin E, Tuskan GA, Le Tacon F, Martin F: **Characterization of transposable elements in the ectomycorrhizal fungus *Laccaria bicolor***. *PloS One* 2012, **7**:e40197.

70. Hakkinen M, Arvas M, Oja M, Aro N, Penttila M, Saloheimo M, Pakula TM: **Re-annotation of the CAZy genes of *Trichoderma reesei* and transcription in the presence of lignocellulosic substrates**. *Microbial Cell Fact* 2012, **11**:134.

71. Solovei I, Thanisch K, Feodorova Y: **How to rule the nucleus: divide et impera**. *Curr Opin Cell Biol* 2016, **40**:47-59.

72. Raffaele S, Kamoun S: **Genome evolution in filamentous plant pathogens: why bigger can be better**. *Nat Rev Microbiol* 2012, **10**:417-430.

73. Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, Zuccaro A, Reissmann S, Kahmann R: **Fungal effectors and plant susceptibility**. *Annu Rev Plant Biol* 2015, **66**:513-545.

74. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S *et al*: **Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations**. *Nature Commun* 2011, **2**:202.

75. Soyer JL, El Ghalid M, Glaser N, Ollivier B, Linglin J, Grandaubert J, Balesdent MH, Connolly LR, Freitag M, Rouxel T *et al*: **Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans***. *PLoS Genet* 2014, **10**:e1004227.

76. Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW: **Introns and splicing elements of five diverse fungi**. *Eukaryotic Cell* 2004, **3**:1088-1100.

77. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics* 2014, **30**:923-930.

78. Steiger MG, Vitikainen M, Uskonen P, Brunner K, Adam G, Pakula T, Penttila M, Saloheimo M, Mach RL, Mach-Aigner AR: **Transformation system for *Hypocrea jecorina* (*Trichoderma reesei*) that favors homologous integration and employs reusable bidirectionally selectable markers**. *Appl Environ Microbiol* 2011, **77**:114-121.

# Figure Legends

**Figure 1. The complete QM6a genome compared to the HiC draft genome.** The top tracks represent the graphs of GC contents (window size 5000 bp) of seven telomere-to-telomere chromosomes (ChI-ChVII). The seven centromeres (*cen1*-*cen7*; in blue) are located at the longest AT-rich blocks in each chromosome. The telomeres (*tel*) of the right (*R*) and left (*L*) arms in each chromosome are indicated (in grey). The CAZyme genes (in red) and several genes involved in DNA repair, light response and sexual development (in black) are indicated. The rDNA (18S-5.8S-26S) locus on ChVI is indicated in orange. The bottom tracks represent the seven superscaffolds (in green) and four short contigs (in blue) of the HiC draft genome. Chromosomal regions with incorrect orientation (i.e., inversion errors) are indicated in pink.

**Figure 2. The rDNA locus.** (**A**) Organization of the rDNA locus. The top panel illustrates the nine tandem "head-to-tail" repeats revealed by the PacBio RSII platform. Each repeat contains an 18S–5.8S–26S rRNA gene cluster and a full-length non-transcribed intergenic spacer (IGS). The bottom panel shows the five repeats in the HiC draft genome. Each repeat has an 18S–5.8S–26S rRNA gene cluster and two truncated IGSs (IGS-5′Δ and IGS-3′Δ). The location of restriction enzymes *XbaI* (X) and *NheI* (N) are indicted. (**B**) Southern hybridization. Genomic DNA (1 μg) was isolated from three different wild isolate strains: QM6a, CBS999.97(1-1) and CBS999.97(1-2). After digestion with *XbaI* (X) or *NheI* (N), the genomic DNA was subjected to agarose gel electrophoresis, Southern blotting and hybridization with a 28S rDNA probe (**A**) or a *mus53* probe (as the DNA loading control). The *mus53* gene encodes the DNA ligase IV protein and there is only one copy of the *mus53* gene in

the *Trichoderma reesei* genome [78].

**Figure 3. The repetitive features of a representative QM6a chromosome (ChIV).** The top tracks represent the graphs of GC contents (in red; window size 100 bp), the mapping coverage of Illumina-MiSeq reads (in black), predicted genes (in blue) and the sequences not affected by RIP (No RIP; in pink) along the entire ChIV (**A**) and the rDNA locus (**B**). The mapping coverage of Illumina-MiSeq reads for overall genomic DNA is 25-30X, whereas that specifically for the rDNA locus is 200-260X. The RICAL program was used to predict the sequences mutated by RIP using default settings, and the sequences not affected by RIP are shown (No RIP; in pink).

**Figure 4. QM6a centromeres.** The tracks represent the graphs of GC contents of seven centromeres (*cen1*-*cen7*) and the corresponding pericentromeric regions. The centromere-encoded genes are indicated by red bars and the conserved centromeric repeats are indicated by green arrows.

**Figure 5. Comparative analysis of chromosome architecture for 12 different fungal genomes.** (**A**) The AT content for each array element (500 bp) was calculated and put into bins of 5% intervals (black or gray bars for QM6a and red for other fungi, y-axis). The average AT content of each fungal genome is shown. (**B**) Numbers of AT-rich blocks of different lengths are shown. We were able to categorize these 12 fungal genomes into three different groups according to: (1) their average AT contents, (2) the genome-wide distribution of local AT content, and (3) the number of long (≥ 3kb) AT-rich blocks.

**Figure 6. RIP in *Trichoderma reesei*.** (**A**) Schematic diagram of the gene deletion cassettes in the three mutants. Each cassette consists of three components; the dominant hygromycin B-resistant marker open reading frame (*hph*; green box), flanked by an upstream promoter ($P_{pki}$ or $P_{trpC}$) and/or a downstream terminator ($T_{cbh2}$). The full-length *hph* gene has 1026 bps. A truncated *hph*-Δ$N_{304\text{-}1026}$-$T_{cbh2}$ cassette was spontaneously generated in *blr1*Δ during transformation. The dark line underneath *hph* represents the DNA probe used for Southern hybridization. The neighboring genes upstream and downstream of the deleted gene are indicated by white boxes and their protein identity numbers. In *blr1*Δ, there are two identical 12960 fragments (in yellow) and two copies of $T_{cbh2}$ (in white). (**B**) Southern hybridization. Genomic DNA of indicated strains was digested by *SalI* or *Sfo1*, and then subjected to Southern blot analysis using the *hph* DNA probe shown in (A). (**C**) Occurrence of RIP in the F1 progeny. The C-to-T point mutations in the respective full-length *hph* cassesttes (in green) are indicted by vertical black bars. Only the results of one representative sexual crossing experiment (n = 10) are shown. The full-length *hph* cassette in each progeny was amplified by polymerase chain reaction (PCR) and then sequenced by Sanger's method.

**Figure 7. Partitioning of gene clusters by AT-rich blocks.** The tracks represent the graphs of GC contents (window size 500 bp) of a gene cluster next to *tel2R* (A) and pericentromeric regions (B). All exons of all the predicted genes are indicated with blue squares, short AT-rich blocks by smaller green arrows, *tel2R* and long AT-rich blocks by larger green arrows, the CAZyme genes in red, the nitrate assimilation genes in dark green and the systematic names of representative QM6a genes in black.

## Additional files

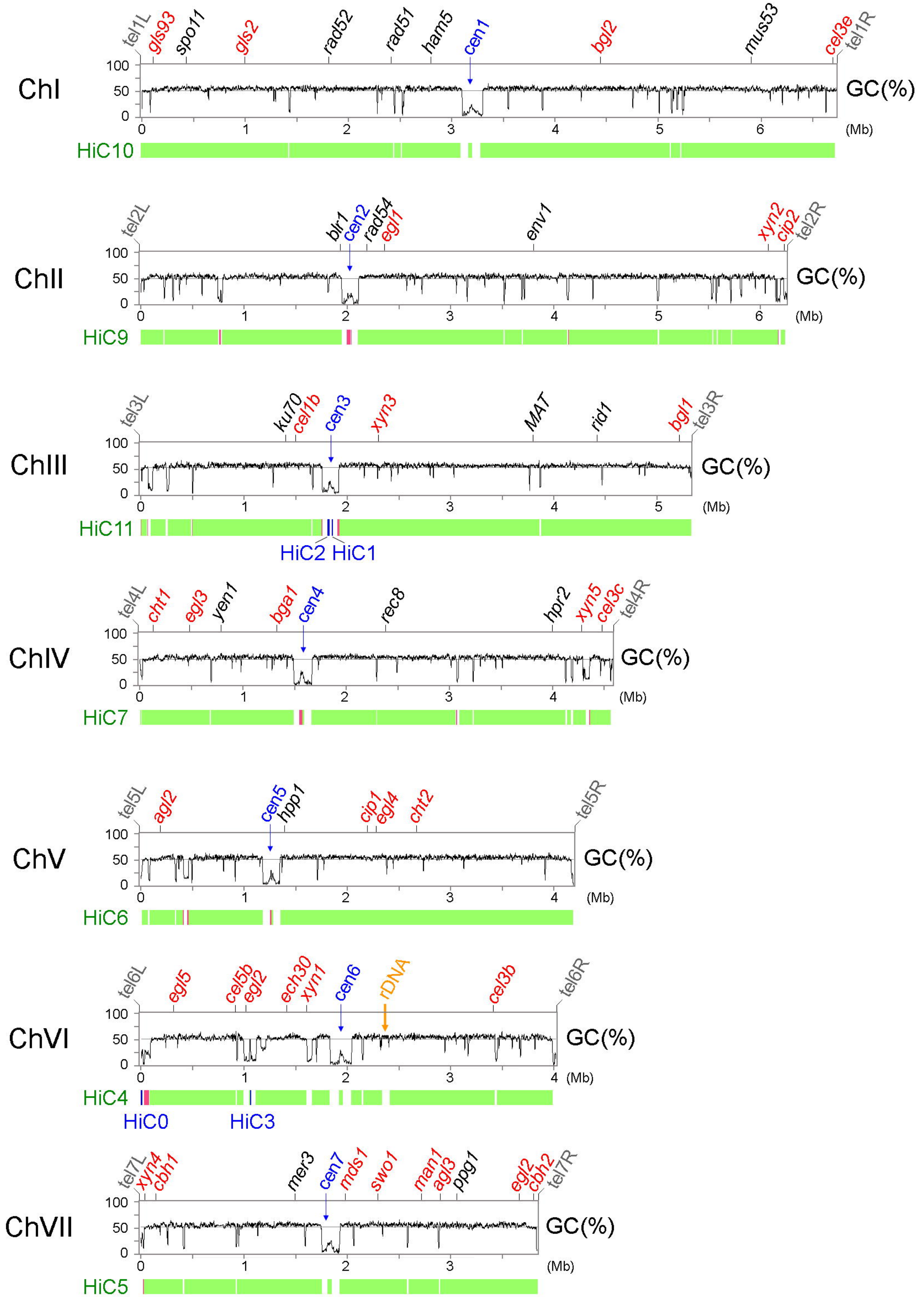
**Li et al. (TFW) Additional file 1.pfd Additional File 1: Table A1.** Preliminary assembly results obtained by the Hierarchical Genome Assembly Process (HGAP 3.0). **Table A2.** Error corrections of the seven PacBio unitigs using the Illimina-MiSeq reads. **Table A3.** Characteristics and assembly of the seven QM6a chromosomes. **Table A4.** Mapping of all trimmed paired-end Illumina MiSeq reads to three QM6a genome drafts. **Table A5.** The MiSeq reads mapped to the QM6a-v2.0 draft genome. **Table A6.** Sixteen false predicted genes in QM6a-v2.0. **Table A7.** The QM6a-v2.0 draft genome compared to the complete QM6a genome sequence. **Table A8.** The HiC draft genome compared to the complete QM6a genome sequence. **Table A9.** PCR primers. **Table A10.** The complete QM6a genome sequence versus the RUT-C30-v1.0 draft genome. **Table A11.** Gene Ontology of some newly-predicted genes**. Table A12.** Sequencing quality of 18 different fungal genomes. **Table A13.** RIP indices of various sequences in QM6a and *Neurospora crassa*. **Table A14.** Repeat-induced C-to-T mutations observed in the *hph* alleles in all F1 progeny. **Table A15.** *Saccharomyces cerevisiae* Ty elements by chromosome. **Table A16.** Transposable elements in 12 well-assembled fungal genomes. **Table A17.** Repetitive sequences in different chromosomal regions. **Table A18.** Repetitive sequences in seven QM6a chromosomes. **Table A19.** Partitioning of four gene clusters by the AT-rich islands. The TPM values in glucose (48 hrs), in straw (24 hrs) and in straw (24hrs) then glucose (5 hrs) are shown [24].

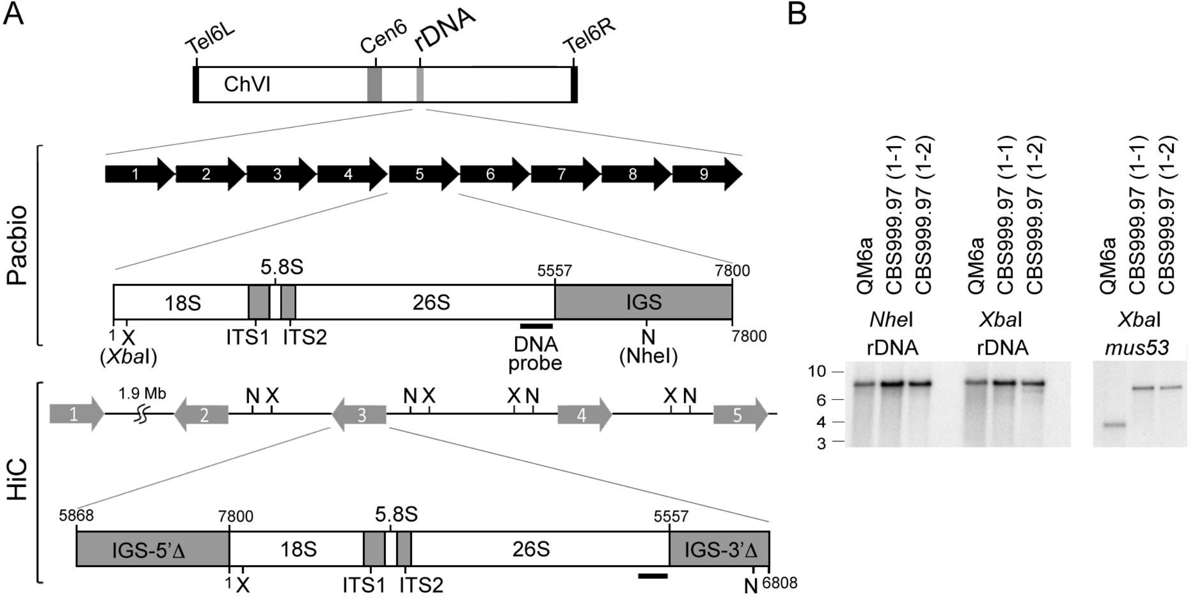
**Li et al. (TFW) Additional file 2.xlsx Additional File 2: Table B1:** Genome annotation of the complete QM6a genomes, including AT-rich blocks, predicted genes, repetitive features, gene ontology (GO), gene names in *Trichoderma reesei*, *Saccharomycetes cerevisiae and Neurospora crassa,* as well as gene identity numbers in QM6a-v2.0 and RUT-C30-v1.0. (**Columns A-H**) All the genes that were not annotated in QM6a-v.20 and Rut-C30-v1.0 are highlighted with yellow background, whereas those only annotated in Rut-C30-v1.0 are highlighted with yellow-green background. (**Columns O-AE**) Comparative transcriptomic analysis of all QM6a genes in different carbon sources using two published transcriptome datasets: (1) The SOLiD sequencing reads from QM9a grown first in glucose for 48 hours, then in straw for 24 hours and finally with the addition of glucose for 5 hours [24]; (2) The Illumina HiSeq 2000 sequencing reads from QM9414 treated with cellulose (24, 48 and 72 hours), sophorose (2, 4 and 6 hours) and glucose (24 and 48 hours) [25]. The TPM values are shown. **Table B2:** Annotation of the complete QM6a genome. All the annotated QM6a genes are compared to those in QM6a-v2.0, Rut-C30-v1.0, all *Trichoderma reesei* proteins from NCBI, the QM6a-HiC annotation results by Druzhinina et al. [18], the annotation results by Schmoll et al. [4], as well as the BLAST results from Uniprot, NCBI non-redundant (nr) database, *Fusarium fujikuroi* (JGI), *Sordaria macrospora* (NCBI)、*Saccharomyces cerevisiae* (SGD) and *Schizosaccharomyces pombe* (PomBase and NCBI). The e-values of all BLASTP results were $< 1.0 \times 10^{-5}$. **Table B3:** TPM values of all new QM6a genes revealed by two publicly-available transcriptome datasets [24, 25]. **Table B4:** Locations of 62 previously identified CAZyme and SSCP gene clusters [18] in 42 chromosomal blocks of the complete QM6a genome. All previously annotated QM6a-v2.0 or QM6a-HiC genes in these gene clusters are indicated in blue. AT-rich islands in these 42 chromosomal blocks are indicated in red. All new QM6a genes we identified in the complete QM6a genome are indicated in darkgreen.

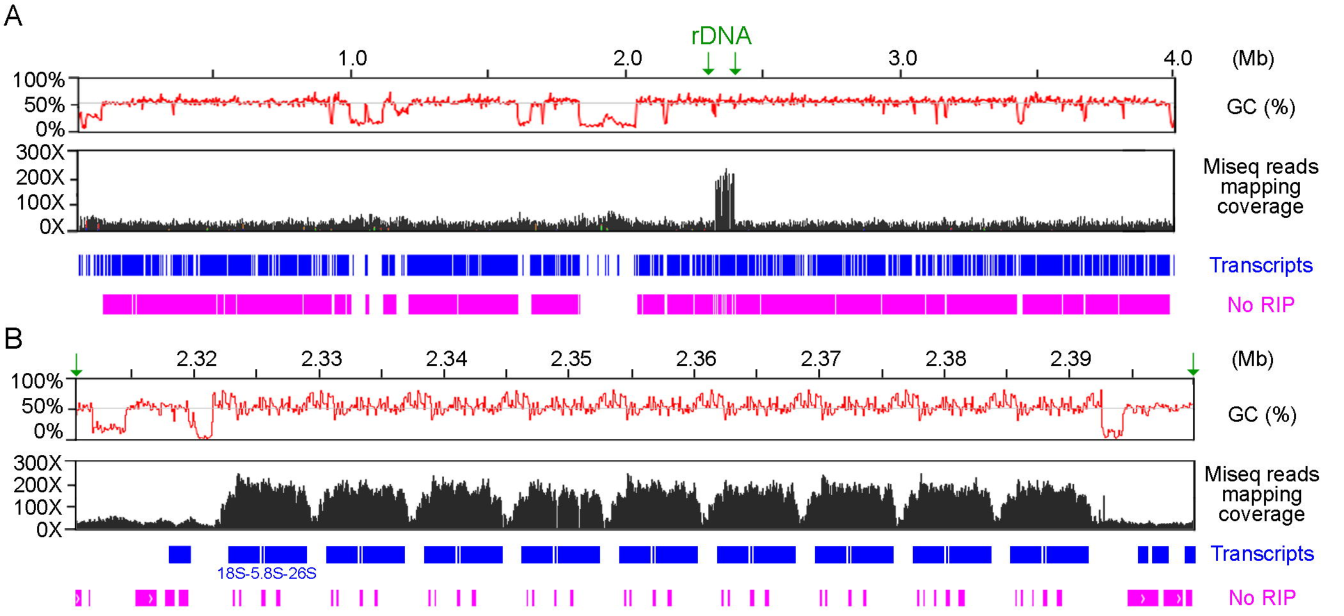
**Li et al. (TFW) Additional file 3.xlsx Additional File 3: Table C1.** Straw-induced QM6a-v2.0 genes identified by Ries et al. [24] and their TPM values. **Table C2.** Straw-induced QM6a-v2.0 genes not identified by Ries et al. [24] and their TPM values. **Table C3.** New QM6a genes that are highly induced (≥ 20-fold) by straw and their TPM values. **Table C4.** New QM6a genes that are significantly induced (5-20-fold) by straw and their TPM values. **Table C4.** Two new QM6a genes that are highly induced by cellulose and their TPM values. **Table C5.** Nine new genes having higher TPM values in QM6a than in QM9414. **Table C6.** Centromere-encoded genes and their TPMs in QM9414.

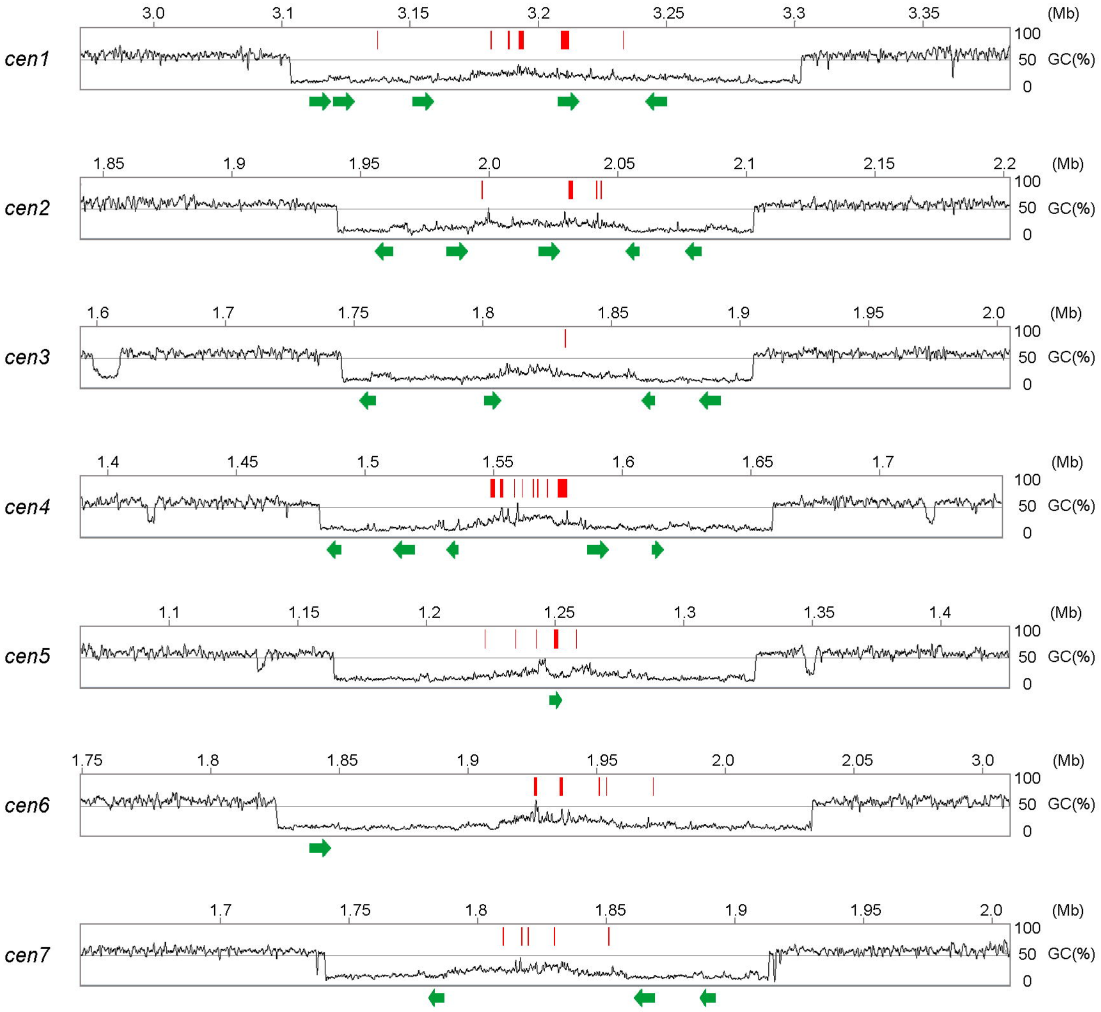
**Li et al. (TFW) Additional file 4.pdf  Additional File 4:** Sequence alignments of the 24 centromeric repeats

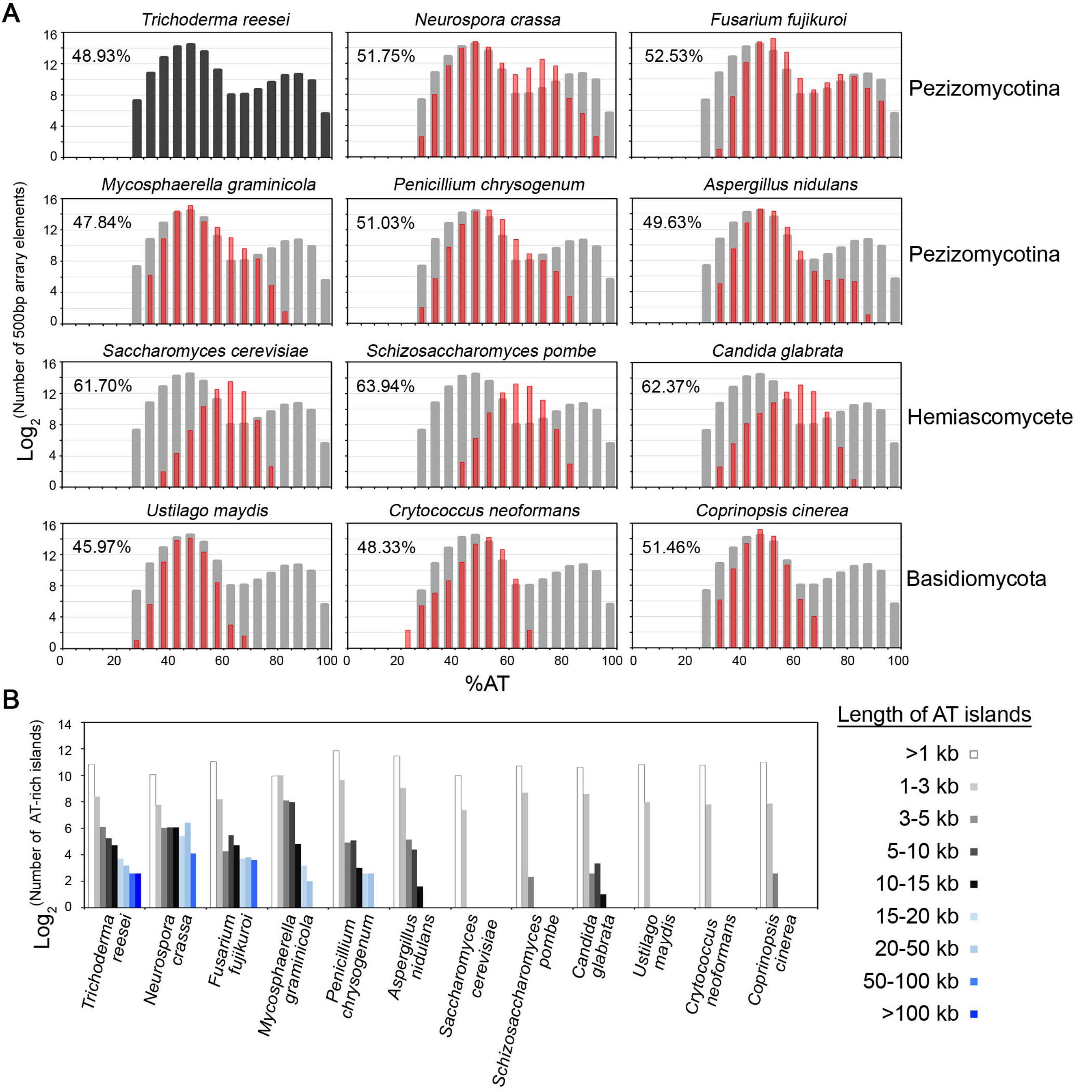
**Li et al. (TFW) Additional file 5.pdf  Additional File 5:** Nucleotide sequences of the *hgh* alleles from all strains listed in Figure 7.

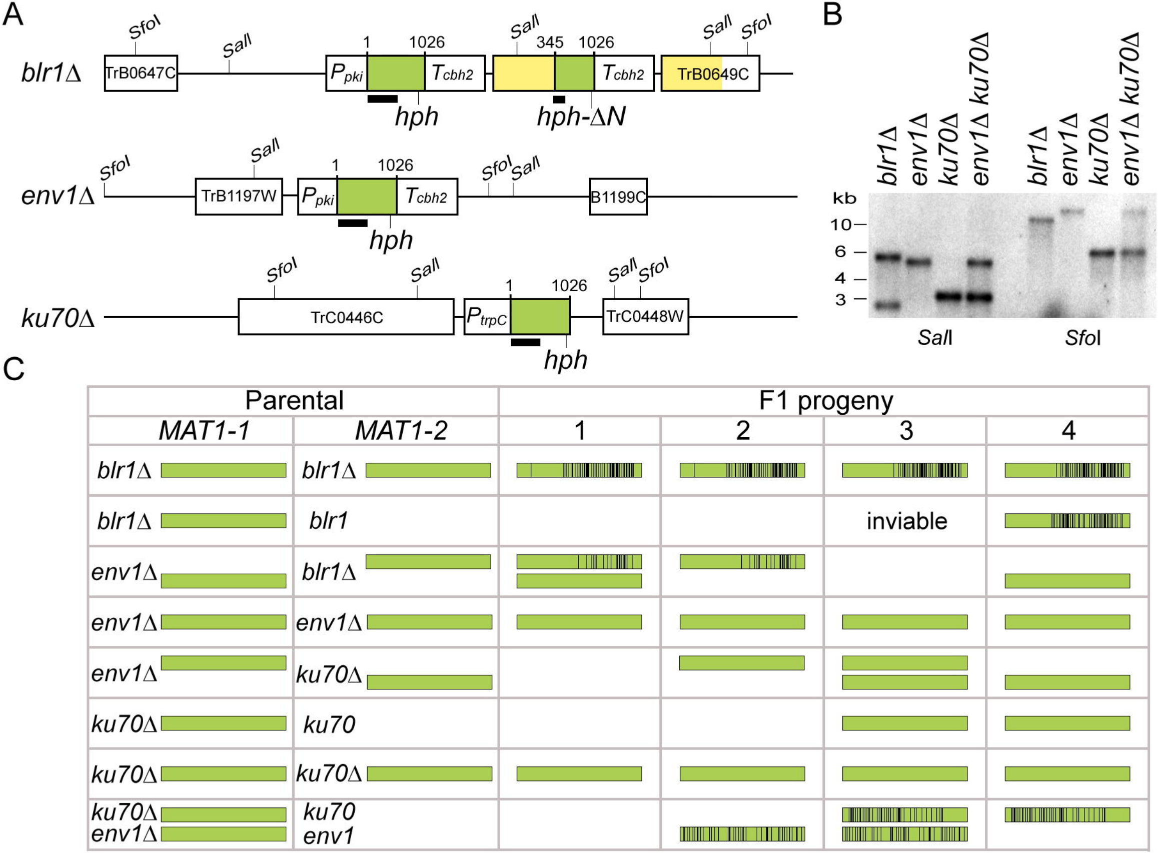

ChI
tel1L
gls93
spo11
gls2
rad52
rad51
ham5
cen1
bgl2
mus53
cel3e
tel1R
100
50
0
GC(%)
0
1
2
3
4
5
6
(Mb)
HiC10
ChII
tel2L
blr1
cen2
rad54
egl1
env1
xyn2
cip2
tel2R
HiC9
ChIII
tel3L
ku70
cel1b
cen3
xyn3
MAT
rid1
bgl1
tel3R
HiC11
HiC2
HiC1
ChIV
tel4L
cht1
egl3
yen1
bga1
cen4
rec8
hpr2
xyn5
cel3c
tel4R
HiC7
ChV
tel5L
agl2
cen5
hpp1
cip1
egl4
cht2
tel5R
HiC6
ChVI
tel6L
egl5
cel5b
egl2
ech30
xyn1
cen6
rDNA
cel3b
tel6R
HiC4
HiC0
HiC3
ChVII
tel7L
xyn4
cbh1
mer3
cen7
mds1
swo1
man1
agl3
ppg1
egl2
cbh2
tel7R
HiC5

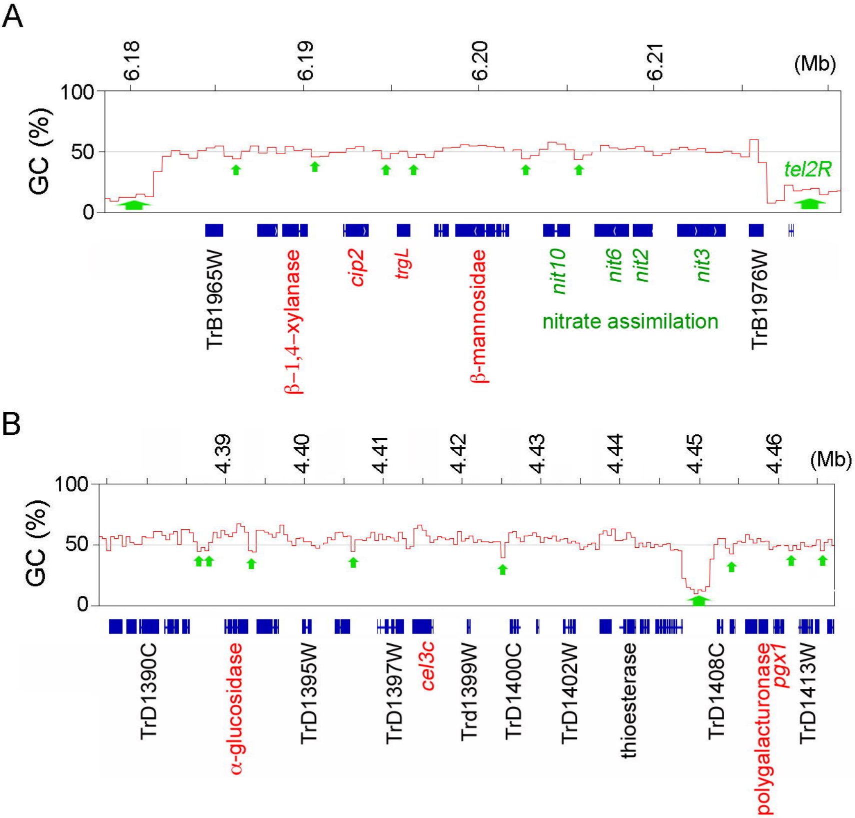

A
Tel6L
Cen6
rDNA
Tel6R
ChVI
Pacbio
1
2
3
4
5
6
7
8
9
5.8S
5557
7800
18S
26S
IGS
1
X
(XbaI)
ITS1
ITS2
DNA
probe
N
(NheI)
7800
HiC
1
1.9 Mb
2
N X
3
N X
X N
4
X N
5
5868
7800
5.8S
5557
IGS-5'Δ
18S
26S
IGS-3'Δ
1
X
ITS1
ITS2
N
6808
B
QM6a
CBS999.97 (1-1)
CBS999.97 (1-2)
QM6a
CBS999.97 (1-1)
CBS999.97 (1-2)
QM6a
CBS999.97 (1-1)
CBS999.97 (1-2)
NheI
rDNA
XbaI
rDNA
XbaI
mus53
10
6
4
3


A
rDNA
1.0
2.0
3.0
4.0
(Mb)
100%
50%
0%
GC (%)
300X
200X
100X
0X
Miseq reads mapping coverage
Transcripts
No RIP
B
2.32
2.33
2.34
2.35
2.36
2.37
2.38
2.39
(Mb)
100%
50%
0%
GC (%)
300X
200X
100X
0X
Miseq reads mapping coverage
Transcripts
18S-5.8S-26S
No RIP


cen1
3.0
3.1
3.15
3.2
3.25
3.3
3.35
(Mb)
100
50
GC(%)
0
cen2
1.85
1.9
1.95
2.0
2.05
2.1
2.15
2.2
(Mb)
100
50
GC(%)
0
cen3
1.6
1.7
1.75
1.8
1.85
1.9
1.95
2.0
(Mb)
100
50
GC(%)
0
cen4
1.4
1.45
1.5
1.55
1.6
1.65
1.7
(Mb)
100
50
GC(%)
0
cen5
1.1
1.15
1.2
1.25
1.3
1.35
1.4
(Mb)
100
50
GC(%)
0
cen6
1.75
1.8
1.85
1.9
1.95
2.0
2.05
3.0
(Mb)
100
50
GC(%)
0
cen7
1.7
1.75
1.8
1.85
1.9
1.95
2.0
(Mb)
100
50
GC(%)
0


A
Trichoderma reesei
48.93%
Neurospora crassa
51.75%
Fusarium fujikuroi
52.53%
Pezizomycotina
Mycosphaerella graminicola
47.84%
Penicillium chrysogenum
51.03%
Aspergillus nidulans
49.63%
Pezizomycotina
Saccharomyces cerevisiae
61.70%
Schizosaccharomyces pombe
63.94%
Candida glabrata
62.37%
Hemiascomycete
Ustilago maydis
45.97%
Crytococcus neoformans
48.33%
Coprinopsis cinerea
51.46%
Basidiomycota
Log2(Number of 500bp array elements)
0
4
8
12
16
0
20
40
60
80
100
%AT
B
Log2(Number of AT-rich islands)
0
2
4
6
8
10
12
14
Trichoderma reesei
Neurospora crassa
Fusarium fujikuroi
Mycosphaerella graminicola
Penicillium chrysogenum
Aspergillus nidulans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Candida glabrata
Ustilago maydis
Crytococcus neoformans
Coprinopsis cinerea
Length of AT islands
>1 kb
1-3 kb
3-5 kb
5-10 kb
10-15 kb
15-20 kb
20-50 kb
50-100 kb
>100 kb


A
blr1Δ
TrB0647C
SfoI
SalI
Ppki
1
1026
hph
Tcbh2
SalI
345
1026
hph-ΔN
Tcbh2
SalI
SfoI
TrB0649C
env1Δ
SfoI
TrB1197W
SalI
Ppki
1
1026
hph
Tcbh2
SfoI
SalI
B1199C
ku70Δ
TrC0446C
SfoI
SalI
PtrpC
1
1026
hph
SalI
SfoI
TrC0448W
B
blr1Δ
env1Δ
ku70Δ
env1Δ ku70Δ
blr1Δ
env1Δ
ku70Δ
env1Δ ku70Δ
kb
10
6
4
3
SalI
SfoI
C
Parental
F1 progeny
MAT1-1
MAT1-2
1
2
3
4
blr1Δ
blr1Δ
blr1Δ
blr1
inviable
env1Δ
blr1Δ
env1Δ
env1Δ
env1Δ
ku70Δ
ku70Δ
ku70
ku70Δ
ku70Δ
ku70Δ
env1Δ
ku70
env1


A
6.18
6.19
6.20
6.21
(Mb)
100
50
0
GC (%)
tel2R
TrB1965W
β−1,4−xylanase
cip2
trgL
β-mannosidae
nit10
nit6
nit2
nit3
nitrate assimilation
TrB1976W
B
4.39
4.40
4.41
4.42
4.43
4.44
4.45
4.46
(Mb)
100
50
0
GC (%)
TrD1390C
α-glucosidase
TrD1395W
TrD1397W
cel3c
Trd1399W
TrD1400C
TrD1402W
thioesterase
TrD1408C
polygalacturonase
pgx1
TrD1413W