doi:10.1093/dnares/dsr009

Draft Genome Sequencing and Comparative Analysis of *Aspergillus sojae* NBRC4239

Atsushi Sato[1,2], Kenshiro Oshima[3], Hideki Noguchi[4], Masahiro Ogawa[1], Tadashi Takahashi[1], Tetsuya Oguma[1], Yasuji Koyama[2], Takehiko Itoh[4], Masahira Hattori[3], and Yoshiki Hanya[1,]*

Research and Development Division, Kikkoman Corporation, 399 Noda, Noda City, Chiba 278-0037, Japan[1]; Noda Institute for Scientific Research, 399 Noda, Noda, Chiba 278-0037, Japan[2]; Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan[3] and Department of Biological Information, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259 Nagatsutacho, Midoriku, Yokohama 226-8501, Japan[4]

*To whom correspondence should be addressed. Tel. +81 4-7123-5515. Fax. +81 4-7123-5959. E-mail: yhanya@mail.kikkoman.co.jp

Edited by Katsumi Isono
(Received 31 December 2010; accepted 13 April 2011)

Abstract

We conducted genome sequencing of the filamentous fungus *Aspergillus sojae* NBRC4239 isolated from the *koji* used to prepare Japanese soy sauce. We used the 454 pyrosequencing technology and investigated the genome with respect to enzymes and secondary metabolites in comparison with other *Aspergilli* sequenced. Assembly of 454 reads generated a non-redundant sequence of 39.5-Mb possessing 13 033 putative genes and 65 scaffolds composed of 557 contigs. Of the 2847 open reading frames with Pfam domain scores of >150 found in *A. sojae* NBRC4239, 81.7% had a high degree of similarity with the genes of *A. oryzae*. Comparative analysis identified serine carboxypeptidase and aspartic protease genes unique to *A. sojae* NBRC4239. While *A. oryzae* possessed three copies of α-amyalse gene, *A. sojae* NBRC4239 possessed only a single copy. Comparison of 56 gene clusters for secondary metabolites between *A. sojae* NBRC4239 and *A. oryzae* revealed that 24 clusters were conserved, whereas 32 clusters differed between them that included a deletion of 18 508 bp containing *mfs1*, *mao1*, *dmaT*, and *pks-nrps* for the cyclopiazonic acid (CPA) biosynthesis, explaining the no productivity of CPA in *A. sojae*. The *A. sojae* NBRC4239 genome data will be useful to characterize functional features of the *koji* moulds used in Japanese industries.

Key words: *Aspergillus sojae*; *Aspergillus oryzae*; comparative genomics; genome sequencing

1. Introduction

Koji moulds are widely used in the production of traditional fermented foods and beverages such as Japanese *miso*, soy sauce, and *sake*. Two typical *koji* moulds, *Aspergillus sojae* and *A. oryzae*, are used. During the fermentation process, *koji* moulds act by breaking down the ingredients. Each species of *koji* moulds reacts differently to the ingredients used and must therefore be selected based on the desired product. For example, *A. sojae* is selected to produce *miso* and soy sauce due to its high proteolytic ability, and *A. oryzae* is used widely in *sake*, *miso*, and soy sauce production for its high amylolytic ability. Among *Aspergillus* strains deposited in the RIKEN Bioresource Center Japan Collection of Microorganisms (http://www.jcm.riken.jp/JCM/JCM_DB.shtm), 15 strains out of 53 in *A. oryzae* strains were derived from *sake koji*, 6 from *miso*, and

17 from soy sauce *koji*, while 16 strains out of 20 strains of *A. sojae* were derived from soy sauce *koji*.

For taxonomic classification of *Aspergillus* species, molecular strategies have been developed to discriminate several *Aspergillus* species.[1] *Aspergillus oryzae* and *A. sojae* are classified in the *Aspergillus* section *Flavi*, which also includes plant pathogen *A. flavus* and *A. parasiticus* that produces aflatoxins known to be carcinogenic substances. The analysis based on restriction-site polymorphisms of genes coding for 11 proteins and sequences of five of those genes suggested that *A. oryzae* is a species derived from *A. flavus* through human handling.[2] From a viewpoint of evolutionarily close relation of *A. oryzae* and *A. sojae* with pathogenic *Aspergillus* species, it is vital to distinguish between aflatoxin productive and non-productive moulds and to select the latter for industrial use. It has been reported that *A. oryzae* does not produce these substances from expressed sequence tag (EST) analysis of *A. oryzae* RIB40, in which many of the aflatoxin biosynthesis gene clusters were found to be unexpressed.[3] For *A. sojae,* a termination point mutation in *aflR* which controls transcription of aflatoxin biosynthesis gene clusters and lack of the polyketide synthase (PKS) gene are correlated with its aflatoxin non-productivity.[4]

The genome of *A. oryzae* RIB40 was recently completely sequenced and the comprehensive analysis showed that this strain possesses 134 protease genes including many paralogous genes and multiple copies of α-amylase and α-glucosidase genes.[5] These genetic features may account for its high proteolytic and amylolytic abilities in *A. oryzae* RIB40.

Many of the moulds classified in *Aspergillus* section *Flavi* are known to produce various secondary metabolites. The *A. oryzae* RIB40 genome encodes genes for numerous secondary metabolites other than aflatoxins,[6] although EST and microarray analysis of *A. oryzae* RIB40 suggested that it has almost no productivity of secondary metabolites.[7] These features of *A. oryzae* on the basis of quality, productivity, and safety may be one of the reasons for that *A. oryzae* strains have gradually been selected as industrially useful strains.[8]

Though similar studies in *A. sojae* have not been conducted so much as *A. oryzae*, it is thought to be a domesticated strain selectively bred from natural strains as well as *A. oryzae* above.

However, the whole genetic information of *A. sojae* is insufficient to investigate the functional features important for its industrial use including the protease and amylase activities as well as safety. Therefore, we conducted the whole-genome sequencing of the practical strain *A. sojae* NBRC4239 isolated from Japanese soy sauce *koji* by using the next generation sequencer 454 pyrosequencer. The genetic information of *A. sojae* NBRC4239, combined with that of *A. oryzae*, will synergistically provide the knowledge for deep understanding of the biological nature of industrially important *koji* mould and for its further development of usefulness in food science field.

2. Materials and methods

2.1. Strain and DNA preparation

Aspergillus sojae NBRC4239 was obtained from NBRC (http://www.nbrc.nite.go.jp/). This strain is a practical strain isolated from Japanese soy sauce *koji*.

Aspergillus sojae NBRC4239 was incubated in PD liquid media (1% peptone, 2% dextrin, 0.5% KH_2PO_4, 0.1% $NaNO_3$, 0.05% $MgSO_4$, and 0.1% casamino acids, pH 6.0) on a shaker at 150 rpm at 30°C for 24 h. After collection on a mortar, mould was frozen in liquid nitrogen and then crushed with a pestle. The genome was extracted from this mould using a Wizard Genomic DNA Purification Kit (Promega Corporation, USA) and purified using a DNeasy Blood & Tissue Kit (QIAGEN Sciences, USA), according to the respective manufacturer's protocols.

2.2. Genome sequencing and data assembly

For GS FLX Titanium fragment sequencing, 500 ng of genomic DNA was sheared into DNA fragments ranging from 300 to 800 bp by nebulization. After both ends of the DNA fragments were repaired and phosphorylated, two types of adaptors (A and B) were ligated to the DNA fragments. Next, the DNA fragments carrying the 5′-biotin of adaptor B from the ligation mixture were immobilized onto magnetic streptavidin-coated beads. The single-stranded template DNA (ssDNA) molecules carrying Adaptor A at 5′-end and Adaptor B at 3′-end were isolated by alkaline denaturation. These purified ssDNAs were then hybridized to DNA capture beads and clonally amplified by an emulsion polymerase chain reaction (PCR) method. After denaturation of the amplified double-stranded DNAs on the capture beads, these beads with single-stranded molecules were spread onto each well of a pico titre plate. For GS FLX Titanium Paired-end sequencing, 15 μg of genomic DNA was sheared into DNA fragments ranging from ~8 kb by fragmentation. After the ends of the DNA fragments were repaired and internal adaptors were ligated to the DNA fragments, each DNA fragment was circularized and self-ligated. The circular DNA was sheared into DNA fragments ranging from 300 to 800 bp by nebulization. The DNA fragments carrying the 5′-biotin of internal adaptor from the ligation mixture were immobilized onto magnetic streptavidin-coated beads. Paired-end sequencing was carried out similar to the method described above. Two sequencing runs in total were carried out. The GS FLX sequence data were assembled using Newbler assembly software.

2.3. *Gene prediction and annotation*

GlimmerHMM,[9] AUGUSTUS,[10,11] SNAP,[12] and GeneMark + ES[13] were used as *ab initio* predictors, and Genewise[14] was used as the evidence-based predictor. The *ab initio* GlimmerHMM, AUGUSTUS, and SNAP parameters were trained on all *A. oryzae* RIB40 gene models, while GeneMark + ES performed an iterative self-training procedure. The amino acid sequence of *A. oryzae* RIB40 was used for alignment by Genewise. GFF files obtained from these prediction programmes were incorporated by Evigan[15] to produce prediction results. Out of the predicted open reading frames (ORFs), those with more than 100 amino acid residues were selected as predictor genes.

Amino acid sequences of the predictor genes were matched with non-redundant protein database (nr, NCBI) by BLAST[16] and were annotated based on identity. tRNA were predicted by tRNAscan-SE.[17]

2.4. *Comparative genomics*

Nucleotide and amino acid sequences of *A. oryzae* RIB40 were obtained from DOGAN (http://www.nbrc.nite.go.jp/dogan/). Nucleotide and amino acid sequences of *A. flavus* NRRL3357, *A. fumigatus* Af293, and *A. nidulans* FGSC A4 were obtained from the *Aspergillus* Genome Database (AspGD).[18] Putative domains were predicted with the HMMER[19] programme hmmscan using the hidden Markov models from the pfam database.[20] For *A. oryzae* RIB40, *A. flavus* NRRL3357, *A. fumigatus* Af293, and *A. nidulans* FGSC A4, ORFs with Pfam domain scores of >150 were subject to comparison.

2.4.1. Protease A domain list for protease was created by matching amino acid sequences registered in MEROPS[21] with the HMMER from the Pfam database. Based on this list, amino acid sequences with protease domain scores >150 were compared between *A. sojae* NBRC4239 and *A. oryzae* RIB40. For phylogenetic analysis, multiple alignments were carried out by ClustalX[22] and phylogenic trees were drawn with TreeView.[23]

2.4.2. Amylolytic enzymes It is known that out of the glycoside hydrolases, families 13, 15, and 31 are involved in amylolysis. Entries of glycoside hydrolase belonging to these three families in relation to *A. oryzae* RIB40 were extracted from the Carbohydrate Active Enzyme (CAZy) database.[24] From these data, amylolytic enzymes possessed by *A. sojae* NBRC4239 strains were predicted. Furthermore, for these predicted amylolytic enzymes, checks were made to confirm the presence of active centre residues (data not shown). Alignment of nucleotide and amino acid sequences were carried out by GENETYX, and BLAST was used for homology searches. PCR primer sequences used for partial nucleotide sequence checks are shown in Supplementary Table S1.

2.4.3. Secondary metabolism Sequences of secondary metabolite gene clusters in *A. oryzae* RIB40 were obtained from the Secondary Metabolite Unknown Region Finder (SMURF) database.[25] Cyclopiazonic acid (CPA) biosynthesis gene cluster sequences of *A. flavus* were obtained from Broad Institute (http://www.broadinstitute.org/annotation/genome/Aspergillus_group) with reference to reports by Chang *et al*.[26] Sequences near the aflatrem gene clusters in *A. flavus* and *A. oryzae* were obtained from reports by Nicholson *et al*.,[27] where the sequences for *A. flavus* NRRL3357 and *A. oryzae* RIB40 were obtained from Broad Institute and DOGAN, respectively. For *A. flavus* NRRL6541, sequences ATM1 (AY559849.2 GI:161621808) and ATM2 (AM921700.1 GI:162286818) entered in GenBank (http://www.ncbi.nlm.nih.gov/genbank) were used. Harr plots were generated by *In silico* Molecular Cloning Series IMC, genomics edition (In Silico Biology, Inc.). PCR primer sequences used for partial nucleotide sequence checks are shown in Supplementary Table S2.

2.5. *Accession numbers*

Nucleotide sequence data were entered into the DDBJ/EMBL/GenBank DNA databases. Accession numbers for the 65 scaffold sequences are DF093557–DF093585 and for the 1034 contig sequences are BACA01000001– BACA01001034.

3. Results and discussion

3.1. *Sequencing and assembly*

We obtained 1034 contigs (>100 bp) and 65 scaffolds by assembling the reads obtained from sequencing (Supplementary Table S3). The 65 scaffolds are composed of 557 contigs, thus 477 contigs did not make up scaffolds. Out of the 1034 contigs, 707 were >500 bp. Total length of the contigs and scaffolds each exceeded 39 Mb. As the genome size reported for *A. oryzae* RIB40 is 37.6 Mb,[5] the genome of *A. sojae* NBRC4239 was predicted to exceed that of *A. oryzae* RIB40.

3.2. *Gene prediction and annotation*

ORF prediction was carried out on the 65 scaffolds and the 477 contigs that did not make up scaffolds. As a result, we obtained 13 033 ORFs with amino acid residues >100. Also, 275 tRNAs were predicted by tRNAscan-SE. Table 1 showed the comparison of ORFs of *A. oryzae* RIB40 and *A. sojae* NBRC4239.

3.3. *Comparative genomics*

The proportion of ORFs with Pfam domain scores of >150 was 2847/13 033 (21.8%) for *A. sojae* NBRC4239, 2868/12 074 (23.8%) for *A. oryzae* RIB40, 2880/12 604 (22.9%) for *A. flavus* NRRL3357, 2517/9887 (25.5%) for *A. fumigatus* Af293, and 2641/11 272 (23.4%) for *A. nidulans* FGSC A4.

ORFs with Pfam domain scores of >150 were compared between *A. sojae* NBRC4239 and *A. oryzae* RIB40 by BLASTP, and it was found that 2326/2847 (81.7%) had >90% identity (Supplementary Fig. S1a). Next, 397/2847 ORFs from *A. sojae* NBRC4239 were found to have <70% identity with those from *A. oryzae* RIB40. Of those, 134 ORFs had >90% identity and 192 ORFs had <70% identity in the nucleotide sequence to the corresponding regions in *A. oryzae* RIB40 by tBLASTN. Therefore, the 134 ORFs found in *A. sojae* NBRC4239 may have been missed in the gene prediction of *A. oryzae* RIB40, and the orthologous ORFs to the 192 ORFs may be absent in *A. oryzae* RIB40. These 192 ORFs in *A. sojae* NBRC4239 were matched with *A. flavus* NRRL3357 genes with domain scores >150 using BLASTP. As a result, 22 had >90% identity and 170 had <70% identity. These results indicated that the 22 ORFs are present in both *A. sojae* NBRC4239 and *A. flavus* NRRL3357 but absent in *A. oryzae* RIB40. The 170 ORFs with <70% identity were matched with nr of the NCBI database by BLASTP, and 132 ORFs were found to have <70% identity. Thus, these 132 ORFs may be unique to *A. sojae* NBRC4239 (Supplementary Fig. S1b).

Table 1. Comparison of ORFs of *A. oryzae* RIB40 and *A. sojae* NBRC4239

	A. oryzae RIB40	*A. sojae* NBRC4239
Size of assembly (MB)	37.6	39.5
GC content (%)	48.2	48.1
tRNA genes	270	275
Number of ORFs	12, 074	13, 033
Average ORF size	449.8	455.9
Min ORF size	101	101
Max ORF size	6,886	7,566

The methods of the sequencing and the ORF prediction procedure were different in *A. oryzae* and *A. sojae*, and the value 101, the minimum size, just indicated the artificial value for cutoff.

3.3.1. Protease Out of the 2847 ORFs of *A. sojae* NBRC4239 and 2868 of *A. oryzae* RIB40, 83 ORFs in *A. oryzae* RIB40 and 76 in *A. sojae* NBRC4239 had protease domain scores >150; thus, *A. sojae* NBRC4239 had seven fewer ORFs. The total number of predicted protease genes in *A. oryzae* RIB40 was considerably less than the reported 134,[5] and this difference is likely due to the strict domain score set at >150.

ORFs with domain scores >150 from *A. sojae* NBRC4239 and *A. oryzae* RIB40 were sorted and compared by domains. The number of proteases with specific domains was not very different in either species. Four types of proteases in *A. sojae* NBRC4239 had one more gene than those in *A. oryzae* RIB40, respectively. Eleven types of proteases in *A. sojae* NBRC4239 had one gene less than those in *A. oryzae* RIB40, respectively. In both *A. oryzae* RIB40 and *A. sojae* NBRC4239, serine carboxypeptidases were most abundant, followed by aspartic proteases. *Aspergillus sojae* NBRC4239 had 13 serine carboxypeptidases, which was one more than *A. oryzae* RIB40, and it had seven aspartic proteases which was two less than *A. oryzae* RIB40.

Under the strict condition of protease domain scores >150, we found no significant difference in the number of protease genes between the two species. However, by using a less strict condition, a difference in the number of protease genes may be observed between *A. sojae* and *A. oryzae*.

3.3.1.1. Serine carboxypeptidase

Phylogenic tree for serine carboxypeptidases based on sequences with domain scores >150 was constructed and then compared for *A. sojae* NBRC4239, *A. oryzae* RIB40, *A. flavus* NRRL3357, *A. fumigatus* Af293, and *A. nidulans* FGSC A4 (Fig. 1). We found that *A. sojae* NBRC4239 possesses a serine carboxypeptidase gene (scaffold00048.369) that has low sequence similarity with the other four species. This gene was also included in the 132 ORFs that had <70% identity against nr (refer to the 'Comparative genomics' section). The similarity search of scaffold00048.369 against the ORFs of *A. oryzae* RIB40 by BLASTP identified a gene AO090103000026 with the closest match of 56% identity. AO090103000026 was annotated as a serine carboxypeptidase in *A. oryzae* RIB40 (Fig. 1). In addition, Scaffold00048.369 was found to have the highest similarity to carboxypeptidase S1 in *Neosartorya fischeri* NRRL181 with 58% identity by searching

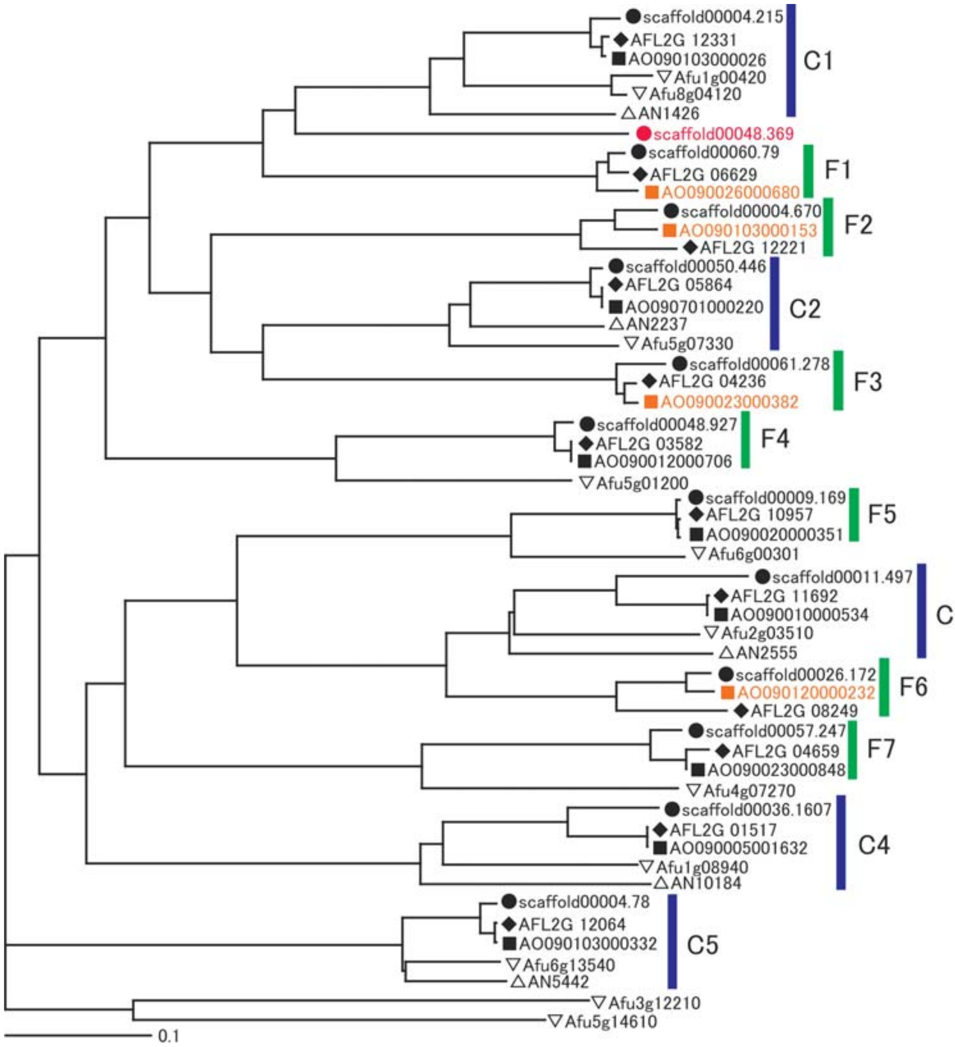

Figure 1. Phylogenic analysis of serine carboxypeptidase in five *Aspergillus* species. ORFs of serine carboxypeptidase with Pfam domain scores of >150 were analysed by ClustalW and were drawn by TreeView for five *Aspergillus* species. Filled circles, *A. sojae*; filled squares, *A. oryzae*; filled diamonds, *A. flavus*; triangle, *A. nidulans*; inverted triangles, *A. fumigatus*. Blue lines indicate the genes in common with all five *Aspergillus* species. Green lines indicate the genes in common with three species in *Aspergillus* section *Flavi*. Red shows the gene unique to *A. sojae*. Oranges show extra homologues in *A. oryzae* reported in Machida *et al.*[6]

against the nr NCBI database by BLASTP. From these results, scaffold00048.369 is likely to be a serine carboxypeptidase unique to *A. sojae* NBRC4239. We confirmed the expression of scaffold00048.369 by RT–PCR for mRNA isolated from *A. sojae* NBRC4239 incubated in wheat bran media. The sequencing of the RT–PCR product revealed a sequence identical to the predicted ORF, indicating that our ORF prediction for scaffold00048.36 was correct, and this gene is expressed in wheat bran media (data not shown).

Phylogenic analysis showed that serine carboxypeptidase genes are classified into five clusters that are in common with the five *Aspergillus* species (C1–5), and seven clusters that are in common only with three species of *Aspergillus* section *Flavi* (F1–7). Three of the five common clusters contained serine carboxypeptidase genes that are unique to *Aspergillus* section *Flavi*, in addition to the putative orthologous genes. The serine carboxypeptidases unique to *A. oryzae* RIB40 previously reported[6] are therefore considered to be in common with the *Aspergillus* section *Flavi*. It is conceivable that *A. sojae* and *A. oryzae* have been used widely in *miso* and soy sauce fermentation because of their possession of highly similar protease genes.

3.3.1.2. Aspartic protease

Phylogenic tree for aspartic proteases based on sequences with domain scores of >150 was constructed and then compared for *A. sojae* NBRC4239, *A. oryzae* RIB40, *A. flavus* NRRL3357, *A. fumigatus* Af293, and *A. nidulans* FGSC A4 (Fig. 2). We found

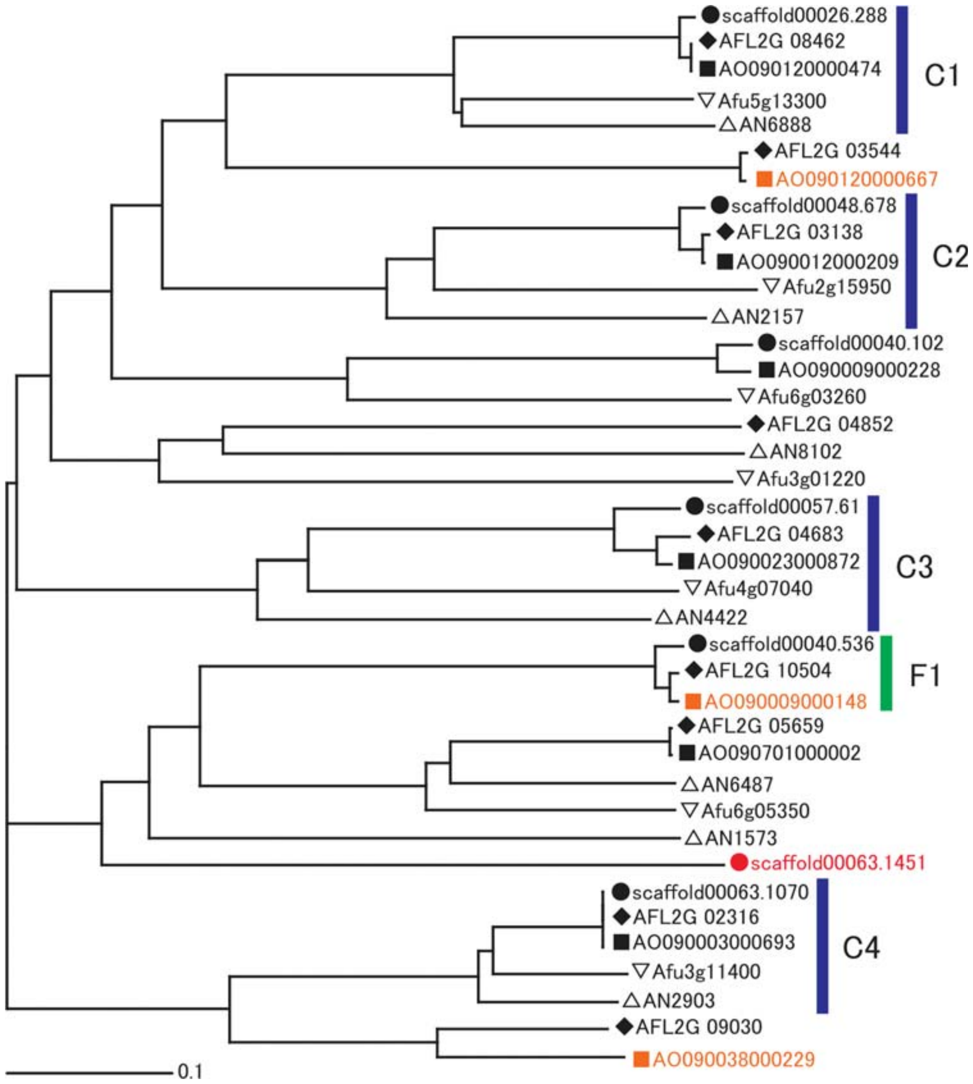

Figure 2. Phylogenic analysis of aspartic protease in five *Aspergillus* species. ORFs of aspartic protease with Pfam domain scores of >150 were analysed by ClustalW and were drawn by TreeView for 5 *Aspergillus* species. Filled circles, *A. sojae*; filled squares, *A. oryzae*; filled diamonds, *A. flavus*; triangle, *A. nidulans*; inverted triangles, *A. fumigatus.* Blue lines indicate the genes in common with all five *Aspergillus* species. Green line indicates the genes in common with three species in *Aspergillus* section *Flavi*. Red shows the gene unique to *A. sojae*. Oranges show extra homologues in *A. oryzae* reported in Machida *et al*.[6]

that *A. sojae* NBRC4239 possesses an aspartic protease gene (scaffold00063.1451) that showed low in sequence similarity with those in the other four species. This gene was also included in the 132 ORFs that had <70% identity against nr (refer to the 'Comparative genomics' section). Scaffold00063.1451 was searched against the ORFs of *A. oryzae* RIB40 by BLASTP, and AO090701000002 was found to be the closest gene with 33% identity. AO090701000002 was annotated as an aspartic protease in *A. oryzae* RIB40 (Fig. 2). Similarity search of scaffold00063.1451 identified yapsin of *Penicillium marneffei* ATCC18224 with the highest similarity of 46% identity against the nr NCBI database by BLASTP. From these results, scaffold00063.1451 is likely to be an aspartic protease unique to *A. sojae* NBRC4239. We confirmed the expression of scaffold00063.1451 by RT–PCR followed by sequencing, as described in the 'Serine carboxypeptidase' section, indicating that our ORF prediction for scaffold00063.1451 was correct, and this gene is expressed in wheat bran media (data not shown).

Phylogenic analysis showed that aspartic protease genes are classified into four clusters that are in common with the five *Aspergillus* species (C1–4). Only one cluster was in common with only the three species belonging to *Aspergillus* section *Flavi* (F1), and other three clusters were shared with two of the three species. These data suggested that aspartic protease genes are less conserved among *Aspergillus* section *Flavi*, in contrast to serine carboxypeptidase genes.

3.3.2. Amylolytic enzymes

3.3.2.1. Amylolytic enzymes in A. sojae NBRC4239 and A. oryzae RIB40

It is known in general that *A. sojae* has lower amylolytic activity compared with *A. oryzae.* We studied the genes for amylolytic enzymes in *A. oryzae* RIB40 and *A. sojae* NBRC4239 to analyse this difference. First, we compared the number of glycoside hydrolases belonging to Family 13, 15, and 31 in both strains, respectively (Supplementary Table S4). We found no difference in gene numbers of glycoside hydrolases between *A. sojae* and *A. oryzae* for Family 31 including α-glucosidase (EC.3.2.1.20), and for Family 15 including glucoamylase (EC.3.2.1.3). Thus, there is unlikely to be a difference in these enzymatic activities between the two *Aspergillus* strains. In contrast, we found that *A. sojae* has two copies less glycoside hydrolases in Family 13 including α-amylase (EC.3.2.1.1) than those in *A. oryzae.* Missing of the two genes in *A. sojae* was due to the copy number variation between the two strains; *A. sojae* only has one copy of *amyB* compared with the three copies (AO090023000944: *amy1*, AO090120000196: *amy2*, and AO090003001210: *amy3*) in *A. oryzae.*[28] In *A. oryzae*, *amyB* codes for so-called Taka-amylase, which is important for amylolysis. Therefore, a decreased copy number of *amyB* orthologues in *A. sojae* likely accounts for the lower amylolytic ability of *A. sojae* than that of *A. oryzae*.

3.3.2.2. α-Amylase genes and their flanking regions

The above-mentioned three α-amylase genes and their flanking 20-kb regions of *A. oryzae* were further compared with the corresponding regions in *A. sojae* NBRC4239 scaffolds. We investigated whether the difference in α-amylase gene copy numbers results from the difference in genomic structures. The results are shown in Fig. 3. In the *A. sojae amy1* region, the 12.5-kb sequence including 2.2 kb of *amy1* ORF, and its promoter and terminator regions were absent. Instead of the 12.5-kb region, a unique 2.9-kb sequence excluding *amy1* was present. This was also the case for *amy2*, where the 12.4-kb region including *amy2* ORF, and its promoter and terminator regions were absent, but a sequence unique as observed in the *amy1* region was not present in *amy2*. Furthermore, a 7.2-kb region was also absent and an inverted region was observed near the missing *amy2* regions in *A. sojae*. These results indicated that this region of the *A. sojae* genome was structurally rearranged. In the *amy3* region, we found that *amy3* structural genes and its terminator regions were conserved between *A. oryzae* and *A. sojae*. However, a 1.9-kb insertion sequence was found at 0.53 kb upstream of the translation initiation site in the *A. oryzae amy3* promoter.

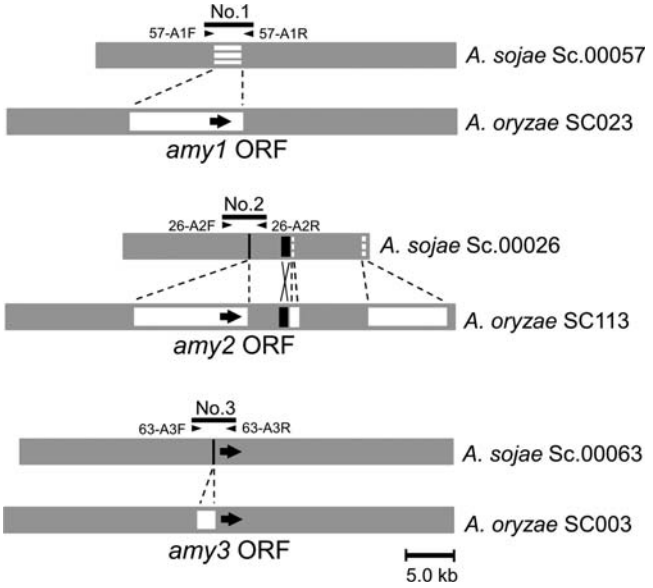

Figure 3. Comparison of α-amylase genes and their flanking regions. Map of α-amylase and flanking regions. Grey box, conserved regions; shaded box, regions unique to *A. sojae*; white box, regions unique to *A. oryzae* ; black box, inversed regions; black arrow, α-amylase gene ORF; black triangle, PCR primer-binding site.

We therefore confirmed that the genomic structures of the α-amylase regions predicted from genome analysis were correct for this strain by PCR (Supplementary Fig. S2). These results indicate that the difference in copy numbers of α-amylase genes between *A. sojae* NBRC4239 and *A. oryzae* RIB40 is a result of a rearrangement in genomic structure.

3.3.2.3. Analysis of transposons surrounding the α-amylase genes

We investigated transposons existing near the α-amylase genes in *A. oryzae* and *A. sojae* (Fig. 4A). We found that the ∼1.9-kb insertion sequence locating upstream the *A. oryzae amy3* promoter is a transposon *Tao1* (DDBJ/EMBL/GenBank accession number: AB021710.1). This transposon was flanked by inverted repeat sequences characteristic to ClassII DNA transposons (Fig. 4A).[29] The *Tao1* insertion site at the *A. oryzae amy3* promoter region was found to correspond to the ‘TA’ sequence in the −533 to −532 region upstream the *A. sojae amy3* promoter.

This is consistent with that ClassII transposons tend to be preferentially integrated at a TA sequence, resulting in target site TA duplication on both flanking of the integrated transposon.[29] The *Tao1* is located at the further upstream of the amylase transcription factor AmyR recognition site[30] and the CreA recognition sites[31] involved in carbon catabolite repression (Fig. 4B). It is not clear whether *Tao1* insertion affects the expression of the *A. oryzae amy3* gene or not, and further study will be needed to clarify the effect of *Tao1* insertion.

The *Tao1* transposon was also present within the promoters of *amy1* and *amy2* of *A. oryzae*. The insertion sites were identical to *amy3* insertion (Fig. 4A). However, *Tao1* inserted in the *amy1* and *amy2* promoters were largely truncated and only partial sequences of 575 bp at the 3′-end of the total 1.9 kb sequence were left. The truncation of *Tao1* in *amy1* and *amy2* have occurred after triplication of the single *amy* gene having *Tao1* transposon, which inserted in the *A. oryzae* lineage after divergence of *A. sojae* from the common ancestor.

We also found a 981-bp ORF similar to *Ant1* transposase near *amy1* and *amy2* of *A. oryzae*. *Ant1* transposon is a member of ClassII DNA transposon classified in the *Tc1/mariner* group,[29,32] and *Ant1* in *A. niger* is reported to have transfer activity.[33] The corresponding *Ant1* transposase homologues were not found in the downstream regions of *amy3* of *A. oryzae* or *amyA* of *A. sojae*.

As shown above, several transposons, such as *Tao1* and *Ant1* transposase homologue, were found near the three α-amylase genes of *A. oryzae,* but were absent around *amyA* of *A. sojae*. The mechanism for the multiplication of α-amylase gene in *A. oryzae* is unclear, but these transposons might have a crucial role for the amylase gene multiplication in *A. oryzae*. The difference in α-amylase gene copy numbers might result in the difference in amylolytic activity between the two strains. This is likely to be a major factor for why *A. oryzae* became widely used in industry, such as in fermentation of *sake*, soy sauce, and *miso*, whereas *A. sojae* became used solely for soy sauce fermentation.

3.3.3. *Analysis of secondary metabolism-related genes*

3.3.3.1. Comparison with secondary metabolite gene clusters in A. oryzae

The 56 secondary metabolite cluster sequences were predicted using SMURF in the *A. oryzae* RIB40 genome.[25] We analysed these clusters for the *A. sojae* NBRC4239 genome by Harr plots. Out of the 56 predicted secondary metabolite clusters, 24 clusters were found to be almost identical and the remaining 32 clusters differed from those in *A. oryzae* RIB40 (Supplementary Table S5). *Aspergillus sojae* NBRC4239 had no sequence homologous to Cluster 51 located on the end of chromosome 5 in *A. oryzae*. In addition, large portions of Clusters 38 (non-ribosomal peptide synthetase: NRPS) and 47 (NRPS) were missing in *A. sojae* NBRC4239 (Fig. 5). For Cluster 38, *A. sojae* NBRC4239 had a replaced 363-bp sequence and a sequence containing the 5′ portion of 460 bp and the 3′ portion of 2.9 kb generated by a deletion of

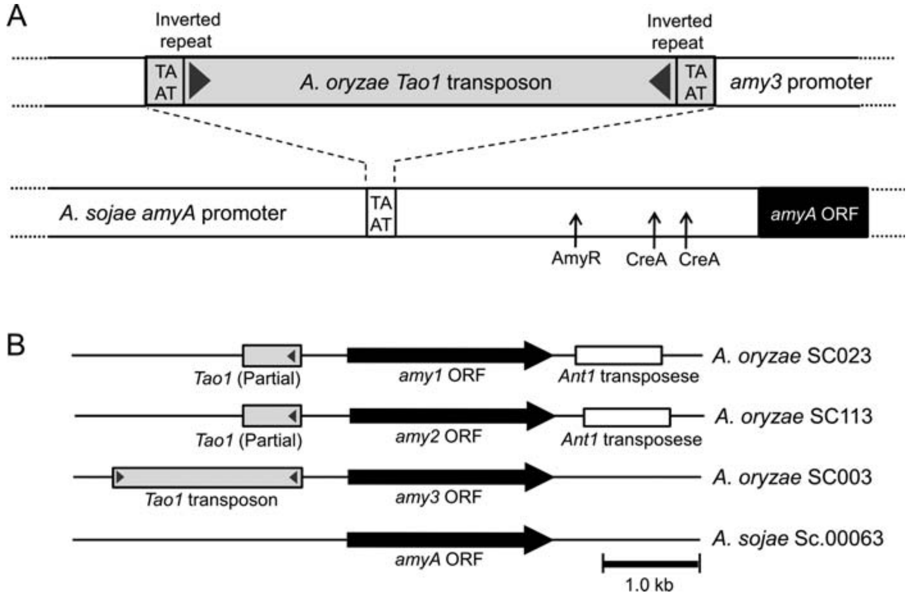

Figure 4. Analysis of transposons surrounding α-amylase genes in *A. oryzae* RIB40 and *A. sojae* NBRC4239. (**A**) Expected insertion site of *A. oryzae Tao1* transposon. (**B**) Map of transposons surrounding *A. oryzae* and *A. sojae* α-amylase genes.

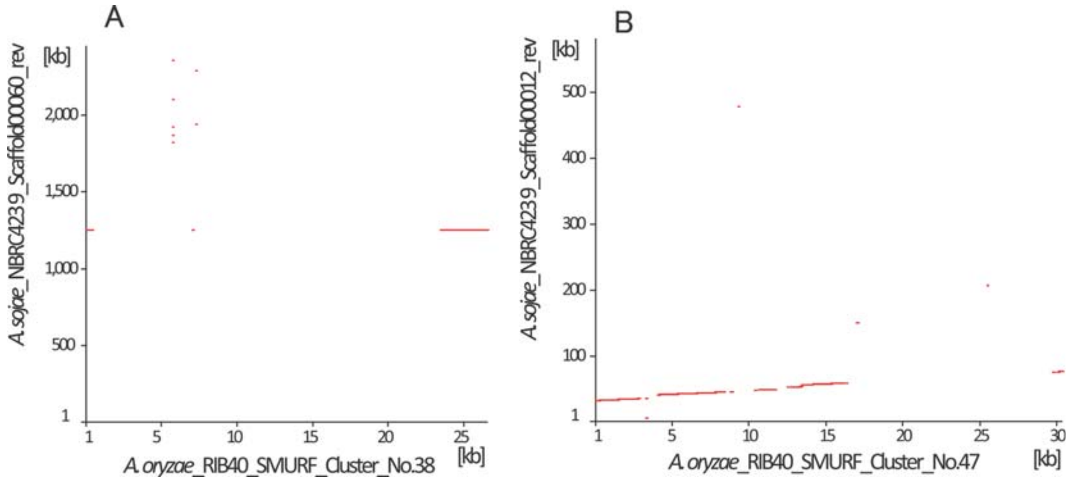

Figure 5. Comparison of *A. oryzae* predicted secondary metabolite clusters with *A. sojae*. Secondary metabolite Clusters 38 (**A**) and 47 (**B**) which were predicted from *A. oryzae* RIB40 by SMURF were compared with sequences in *A. sojae*NBRC4239 by Harr plots.

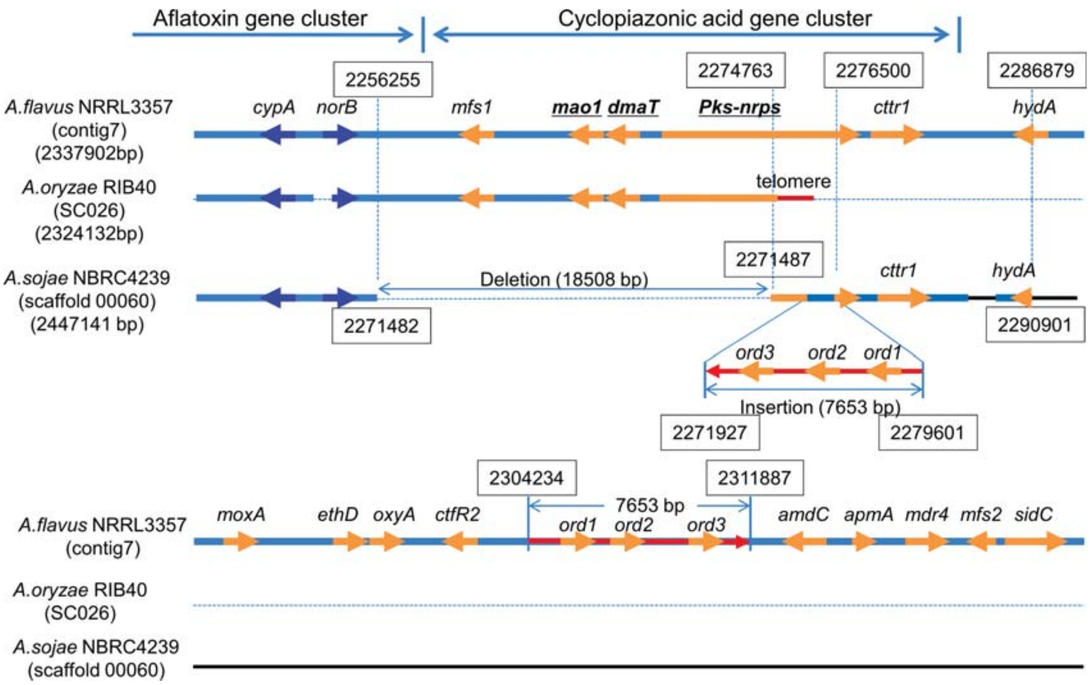

Figure 6. Comparison of CPA biosynthesis gene cluster regions between *A. sojae* NBRC4239, *A. flavus* NRRL3357, and *A. oryzae* RIB40. Nucleotide sequences of CPA biosynthesis gene cluster regions were compared between *A. sojae* NBRC4239, *A. flavus* NRRL3357, and *A. oryzae* RIB40 by BLASTN and are presented diagrammatically.

23 188 bp from the 26 571 bp (SC026: 1081838–1108408) predicted in *A. oryzae* RIB40 (Fig. 5A). For Cluster 47, the 3′ portion of 14 kb in 30 338 bp (SC102: 1249352–1279689) predicted in *A. oryzae* RIB40 was replaced by unrelated 17-kb sequence (Fig. 5B). Also, small deletions and insertions were observed in the remaining 29 clusters (data not shown).

Unlike Cluster 51 (PKS) located near the end of chromosome 5 (SC113: 1828103-1841684) in *A. oryzae* RIB40, missing of the equivalent cluster to Cluster 51 in *A. sojae* NBRC4239 may be partly explained by the instability of the region near the chromosome end in *A. sojae* NBRC4239. On the other hand, Clusters 38 and 47 in *A. oryzae* RIB40, of which equivalent clusters that had large deleted portions in *A. sojae* NBRC4239 are located near the centre of arm of chromosome 3 (SC26: 1081838–1108408 of 2 324 132 bp) and near the centromere of chromosome 4 (SC102: 1249352–1279689 of 1 779 707 bp), respectively. Therefore, the reasons for the depletion of these secondary metabolite

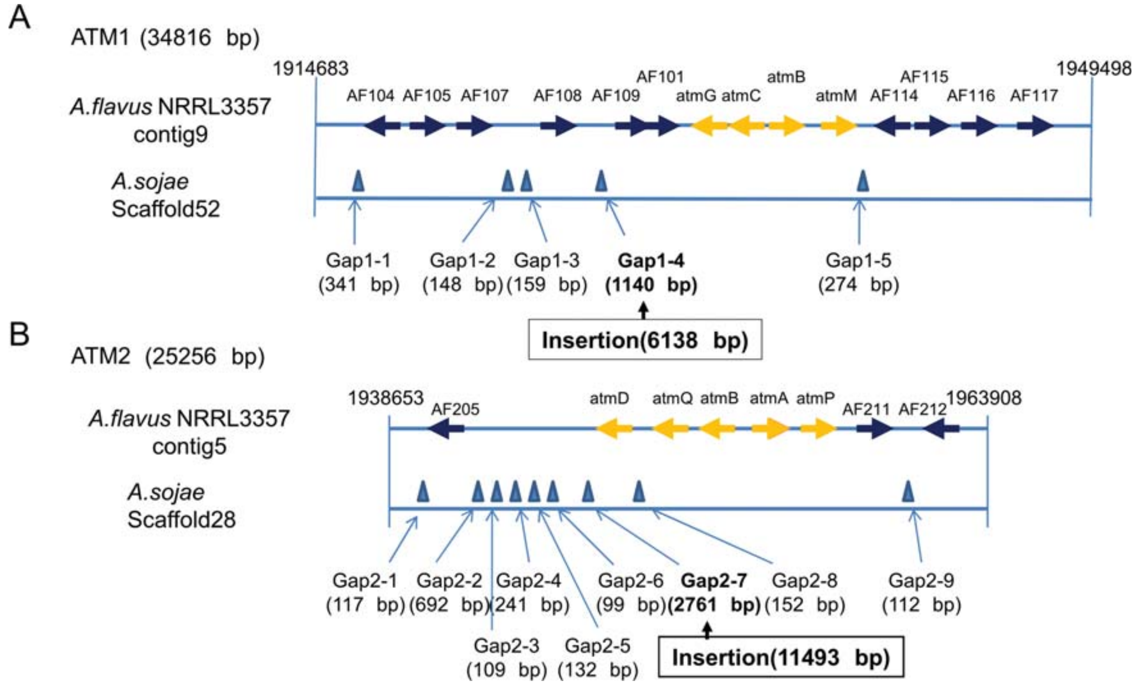

Figure 7. Comparison of aflatrem cluster regions in *A. sojae* NBRC4239 and *A. flavus*. Diagram of aflatrem cluster sequence comparison between *A. flavus* NRRL3357 and *A. sojae* NBRC4239 strains. (**A**) ATM1 region. (**B**) ATM2 region.

gene clusters may be different from that for missing of the equivalent cluster to Cluster 51 at the chromosomal end.

3.3.3.2. Analysis of CPA gene cluster regions

Gene cluster regions for CPA biosynthesis in *A. flavus* and *A. oryzae* were analysed for genomes of *A. sojae* NBRC4239, *A. flavus* NRRL3357, and *A. oryzae* RIB40 by BLASTN.[34] The results are shown in Fig. 6. In *A. flavus* and *A. oryzae*, CPA clusters were found at the end of chromosome 3, next to the aflatoxin biosynthesis gene clusters (Fig. 6). Genes *mfs1*, *mao1*, *dmaT*, *Pks-nrps*, and *cttr1* shown in the figure are considered to be involved in CPA biosynthesis. Since the large portion of *Pks-nrps* at the telomere side is deleted, CPA cannot be synthesized in *A. oryzae* RIB40.[34] On the other hand, 18 508 bp of the CPA biosynthesis cluster region was found to be deleted in *A. sojae* NBRC4239, thus most of the *mfs1*, *mao1*, *dmaT*, and *Pks-nrps* sequences were missing (Fig. 6). Furthermore, a 7653-bp sequence including *ord1*, *ord2*, and *ord3* is present 25 kb distant from the CPA cluster toward the telomere side in *A. flavus*. This sequence was found to be inserted inversely next to the missing CPA gene cluster in *A. sojae*.

PCR was carried out to confirm the missing *A. sojae* CPA biosynthesis gene cluster region and the inverted insert (Supplementary Fig. S3). These results confirmed the 18.5-kb deletion and the inverted 7.6-kb insertion found in the *A. sojae* NBRC4239 genome.

In addition to the finding of complete deletion of *mfs1*, *mao1*, and *dmaT*, the present analysis also found the deletion of a promoter and the half of the ORF containing the ketoacyl synthase (KR) domain and the acyltransferase (AT) domain for *Pks-nrps* in *A. sojae*. The genes *mao1*, *dmaT*, and *Pks-nrps* are essential for CPA biosynthesis in *A. flavus*.[26] Therefore, the present data lead to the conclusion that *A. sojae* is unable to produce CPA, which also verifies the safety of *A. sojae* for use in industry.

3.3.3.3. Analysis of aflatrem biosynthesis gene cluster regions

The *A. sojae* genome was analysed for the aflatrem biosynthesis gene cluster found in *A. flavus*.[27] Aflatrem biosynthesis genes in *A. flavus* are known to consist of genes required to synthesize the intermediate paspaline (*atmG*, a*tmC*, a*tmM*, and a*tmB*) as well as genes required to convert paspaline to aflatrem (*atmP*, *atmQ*, and *atmD*), which are encoded at two separate loci ATM1 and ATM2 in *A. flavus*, respectively.[27] ATM1 (34 816 bp) and ATM2 (25 256 bp) were analysed for the *A. oryzae* and *A. sojae* genomes by BLASTN. We found almost identical sequences to ATM1 and ATM2 with a few base substitutions in *A. oryzae*. In contrast, five gaps with >100 bp were found in the corresponding region to ATM1 locus in *A. sojae* (Fig. 7A). In Gap 4, a 1140-bp sequence in *A. flavus* was replaced by unrelated 6138-bp sequence in *A. sojae* (Fig. 7A). Nine gaps with >100 bp were also observed in the corresponding region to ATM2 locus in *A. sojae* (Fig. 7B). In Gap 7, a 2761-bp sequence in *A. flavus* was replaced by unrelated 11 493-bp sequence in *A. sojae* (Fig. 7B). All these gaps found in *A. sojae* were present in the non-coding regions. A frameshift due to single base insertion in exon 7 of *atmQ* was reported to account for the non-productivity of aflatrem in *A. oryzae*,[27] but such mutation was not found in *A. sojae* in this study. The presence of insertions in *A. sojae* NBRC4239 was confirmed by PCR (Supplementary Fig. S4).

In this study, we showed that *A. sojae* NBRC4239 had many differences in ATM1 and ATM2 loci including deletions and insertion of unrelated sequences in comparison with those in *A. oryzae and A. flavus*, where both loci are well conserved. In addition to these differences, more than 10 gaps <100 bp were also observed in the corresponding loci in *A. sojae* NBRC4239 (data not shown). As described above, all the 14 gaps were present in the non-coding regions in *A. sojae*. To date, production of aflatrem in *A. sojae* has not been reported but the present study did not provide the evidence for aflatrem non-productivity from the sequence information. Further analysis will be needed to solve this discrepancy on the aflatrem production in *A. sojae*.

Acknowledgements: We thank S.W. Kim, M. Kikuchi, K. Furuya, C. Shindo, H. Inaba, E. Iioka, Y. Takayama, W. Suda, and Y. Hattori from the University of Tokyo, S. Hara, K. Matsushima, K. Ito, and Y. Shinohara from Noda Institute for Scientific Research, and T. Yoshimura, A. Komura, J. Manaka, and M. Yuzuki from Kikkoman Corporation for technical assistance and fruitful discussion.

Supplementary data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

References

1. Godet, M. and Munaut, F. 2010, Molecular strategy for identification in *Aspergillus* section *Flavi*, *FEMS Microbiol. Lett.*, **304**, 157–68.
2. Geiser, D.M., Pitt, J.I. and Taylor, J.W. 1998, Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*, *Proc. Natl Acad. Sci. USA*, **95**, 388–93.
3. Akao, T., Sano, M., Yamada, O., et al. 2007, Analysis of expressed sequence tags from the fungus *Aspergillus oryzae* cultured under different conditions., *DNA Res.*, **14**, 47–57.
4. Chang, P.K., Matsushima, K., Takahashi, T., et al. 2007, Understanding nonaflatoxigenicity of *Aspergillus sojae*: a windfall of aflatoxin biosynthesis research, *Appl. Microbiol. Biotechnol.*, **76**, 977–84.
5. Kobayashi, T., Abe, K., Asai, K., et al. 2007, Genomics of *Aspergillus oryzae*, *Biosci. Biotechnol. Biochem.*, **71**, 646–70.
6. Machida, M., Asai, K., Sano, M., et al. 2005, Genome sequencing and analysis of *Aspergillus oryzae*, *Nature*, **438**, 1157–61.
7. Machida, M., Terabayashi, Y., Sano, M., et al. 2008, Genomics of industrial *Aspergilli* and comparison with toxigenic relatives, *Food Addit. Contam. Part A Chem. Anal. Control Expo Risk Assess.*, **25**, 1147–51.
8. Machida, M., Yamada, O. and Gomi, K. 2008, Genomics of *Aspergillus oryzae*: learning from the history of Koji mold and exploration of its future, *DNA Res.*, **15**, 173–83.
9. Majoros, W.H., Pertea, M. and Salzberg, S.L. 2004, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics*, **20**, 2878–79.
10. Stanke, M., Schöffmann, O., Morgenstern, B. and Waack, S. 2006, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources, *BMC Bioinformatics*, **7**, 62.
11. Stanke, M. and Waack, S. 2003, Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, **19**, ii215–25.
12. Korf, I. 2004, Gene finding in novel genomes, *BMC Bioinformatics*, **5**, 59.
13. Besemer, J. and Borodovsky, M. 2005, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses, *Nucleic Acids Res.*, **33**, W451–4.
14. Birney, E., Clamp, M. and Durbin, R. 2004, GeneWise and Genomewise, *Genome Res.*, **14**, 988–5.
15. Liu, Q., Mackey, A.J., Roos, D.S. and Pereira, F.C.N. 2008, Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene predictions, *Bioinformatics*, **24**, 597–605.
16. Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
17. Lowe, T.M. and Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–4.
18. Arnaud, M.B., Chibucos, M.C., Costanzo, M.C., et al. 2010, The *Aspergillus* Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the *Aspergillus* research community, *Nucleic Acids Res.*, **38**, D420–27.
19. Eddy, S.R. 1998, Profile hidden Markov models, *Bioinformatics*, **14**, 755–63.
20. Finn, R.D., Tate, J., Mistry, J., et al. 2008, The Pfam protein families database, *Nucleic Acids Res.*, **36**, D281–88.
21. Rawlings, N.D., Barrett, A.J. and Bateman, A. 2010, MEROPS: the peptidase database, *Nucleic Acids Res.*, **38**, D227–33.

22. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. 1997, The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res.*, **24**, 4876–82.
23. Page, R. 1996, TREEVIEW: an application to display phylogenetic trees on personal computers, *Appl. Biosci.*, **12**, 357–8.
24. Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. 2009, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics, *Nucleic Acids Res.*, **37**, D233–38.
25. Khaldi, N., Seifuddin, F.T., Turner, G., et al. 2010, SMURF: genomic mapping of fungal secondary metabolite clusters, *Fungal Genet. Biol.*, **47**, 736–41.
26. Chang, P.K., Horn, B.W. and Dorner, J.W. 2009, Clustered genes involved in cyclopiazonic acid production are next to the aflatoxin biosynthesis gene cluster, *Fungal Genet. Biol.*, **46**, 176–82.
27. Nicholson, M.J., Koulman, A., Monahan, B.J., Pritchard, B.L., Payne, G.A. and Scott, B. 2009, Identification of two aflatrem biosynthesis gene loci in *Aspergillus flavus* and metabolic engineering of *Penicillium paxilli* to elucidate their function, *Appl. Environ. Microbiol.*, **75**, 7469–81.
28. Wirsel, S., Lachmund, A., Wildhardt, G. and Ruttkowski, E. 1989, Three alpha-amylase genes of *Aspergillus oryzae* exhibit identical intron–exon organization, *Mol. Microbiol.*, **3**, 3–14.
29. Muñoz-López, M. and García-Pérez, J.L. 2010, DNA transposons: nature and applications in genomics, *Curr. Genomics*, **11**, 115–8.
30. Ito, T., Tani, S., Itoh, T., Tsukagoshi, N., Kato, M. and Kobayashi, T. 2004, Mode of AmyR binding to the CGGN$_8$AGG sequence in the *Aspergillus oryzae taaG2* promoter, *Biosci. Biotechnol. Biochem.*, **68**, 1906–11.
31. Ruijter, G.J. and Visser, J. 1997, Carbon repression in *Aspergilli*, *FEMS Microbiol. Lett.*, **151**, 103–4.
32. Moerman, D.G. and Waterston, R.H. 1989, Mobile elements in *Caenorhabditis elegans* and other nematodes. In: Berg, D.E. and Howe, M.M. (eds.), *Mobile DNA*. American Society of Microbiology: Washington, DC, pp. 537–56.
33. Glayzer, D., Roberts, I., Archer, D.B. and Oliver, R.P. 1995, *Ant-1*, an active transposon from the fungus *Aspergillus niger*, *Mol. Gen. Genet.*, **249**, 432–8.
34. Tokuoka, M., Seshime, Y., Fujii, I., Kitamoto, K., Takahashi, T. and Koyama, Y. 2008, Identification of a novel polyketide synthase-nonribosomal peptide synthetase (PKS-NRPS) gene required for the biosynthesis of cyclopiazonin acid in *Aspergillus oryzae*, *Fungal Genet. Biol.*, **45**, 1608–15.