MIDDAS-M: Motif-Independent De Novo Detection of Secondary Metabolite Gene Clusters through the Integration of Genome Sequencing and Transcriptome Data

Myco Umemura¹, Hideaki Koike², Nozomi Nagano³, Tomoko Ishii¹, Jin Kawano¹, Noriko Yamane², Ikuko Kozone⁴, Katsuhisa Horimoto⁵, Kazuo Shin-ya⁴,⁶, Kiyoshi Asai³, Jiujiang Yu⁷, Joan W. Bennett⁸, Masayuki Machida¹*¹

¹Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Sapporo, Hokkaido, Japan, ²Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan, ³Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo, Japan, ⁴Japan Biological Informatics Consortium, Koto-ku, Tokyo, Japan, ⁵Molecular Profiling Research Center for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo, Japan, ⁶Biomedical Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo, Japan, ⁷Beltsville Agricultural Regional Research Center, Agricultural Research Service, United States Department of Agriculture, Beltsville, Maryland, United States of America, ⁸Department of Plant Biology and Pathology, Rutgers University, New Brunswick, New Jersey, United States of America

Abstract

Many bioactive natural products are produced as “secondary metabolites” by plants, bacteria, and fungi. During the middle of the 20th century, several secondary metabolites from fungi revolutionized the pharmaceutical industry, for example, penicillin, lovastatin, and cyclosporin. They are generally biosynthesized by enzymes encoded by clusters of coordinately regulated genes, and several motif-based methods have been developed to detect secondary metabolite biosynthetic (SMB) gene clusters using the sequence information of typical SMB core genes such as polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs). However, no detection method exists for SMB gene clusters that are functional and do not include core SMB genes at present. To advance the exploration of SMB gene clusters, especially those without known core genes, we developed MIDDAS-M, a motif-independent de novo detection algorithm for SMB gene clusters. We integrated virtual gene cluster generation in an annotated genome sequence with highly sensitive scoring of the cooperative transcriptional regulation of cluster member genes. MIDDAS-M accurately predicted 38 SMB gene clusters that have been experimentally confirmed and/or predicted by other motif-based methods in 3 fungal strains. MIDDAS-M further identified a new SMB gene cluster for ustiloxin B, which was experimentally validated. Sequence analysis of the cluster genes indicated a novel mechanism for peptide biosynthesis independent of NRPS. Because it is fully computational and independent of empirical knowledge about SMB core genes, MIDDAS-M allows a large-scale, comprehensive analysis of SMB gene clusters, including those with novel biosynthetic mechanisms that do not contain any functionally characterized genes.

Introduction

Chemists have been deciphering the chemical structures of natural products for a century and a half. Many of these natural products are produced as “secondary metabolites” by plants, bacteria, and fungi. During the middle of the 20th century, several secondary metabolites from fungi revolutionized the pharmaceutical industry. These include the antibiotic, penicillin; the cholesterol-level lowering compound, lovastatin; and the immune suppressor, cyclosporin. Other fungal secondary metabolites have achieved notoriety, such as aflatoxin [1]. In the late 20th century, with the advent of gene cloning, it became apparent that fungal secondary metabolites are biosynthesized by clusters of coordinately regulated genes. Such gene clustering is rare in eukaryotes.

In spite of limited number of secondary metabolites identified from a single species, sequencing the genomes of filamentous fungi has revealed far more than the expected numbers of secondary metabolite biosynthetic (SMB) genes. The numbers of SMB genes encoding polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) range from 17–35 and 14–24, respectively, in many lawfully species [2]. To identify potential secondary metabolites (SMs) in filamentous fungi,
various bioinformatics tools, including SMURF [3], antiSMASH [4,5], CLUSEAN [6], and the method described by Andersen et al. [7], have been developed and successfully applied. The basic concept underlying these tools is the existence of SMB gene clusters, which typically contain approximately 20 genes, including the so-called core genes of PKS, NRPS, or dimethylallyl tryptophan synthases (DMATs). These methods are completely dependent on the known sequence motifs of the core genes; therefore, they can only be used to detect SMB gene clusters that include these core genes. In addition, they cannot distinguish functional clusters from silent or cryptic clusters in fungi [8] because they do not incorporate transcriptomics data.

Many secondary metabolites with important medicinal activities have scaffold structures that are mostly synthesized by the core genes of PKS or NRPS, but there are also others independent of those core genes such as oxylipins, a derivative of fatty acids [9]. We recently discovered the SMB gene cluster for kojic acid (KA), which is the representative secondary metabolite of Aspergillus oryzae [10,11]. The KA cluster could not be detected by conventional methods due to the lack of the core genes. KA was discovered in 1907 and has been used industrially [12], but its biosynthetic gene cluster was found only recently. This fact indicates the extreme difficulty in identifying SMB gene clusters without any core genes.

Comparative genomics has shed light on the characteristics of SMB genes that localize to so-called non-syntenic blocks (NSBs) [13–15], NSBs harbor genes that have roles in the transport and metabolism of various compounds [13] and are highly divergent between species [16–18]. Two-thirds of the genes in NSBs are not homologous with any genes with known functions [13]. Considering our limited knowledge regarding SMB genes and their high level of diversity, it can be speculated that the significant accumulation of unknown genes on NSBs is due to the presence of a large number of SMB genes on NSBs. In support of this hypothesis, the KA gene cluster is located in an NSB [11].

To enhance the exploration of SMB gene clusters in fungal genomes, especially those without core genes, we have developed MIDDAS-M, a motif-independent de novo detection algorithm for secondary metabolite gene clusters. We used virtual gene cluster generation on an annotated genome sequence integrated with metabolism of various compounds and bacterial causes that are non-SMB clusters [13,14]. We have experimentally confirmed and/or predicted by other methods. MIDDAS-M also automatically generates the candidate clusters from all possible pairwise comparisons of transcriptomes from several or more culture conditions. This allows comprehensive de novo predictions using large-scale transcriptome datasets based on a variety of culture conditions. See Supplementary Method, the “MIDDAS-M computation” section in the Supplementary Method in Appendix S1). MIDDAS-M also automatically generates the candidate clusters from all possible pairwise comparisons of transcriptomes from several or more culture conditions. This allows comprehensive de novo predictions using large-scale transcriptome datasets based on a variety of culture conditions. See Supplementary Method, the “MIDDAS-M computation” section in the Supplementary Method in Appendix S1 for further details. MIDDAS-M is available for use at the following server (http://133.242.13.217/MIDDAS-M).
Accurate detection of experimentally validated SMB gene clusters

MIDDAS-M was applied to the filamentous fungus *A. oryzae* for the detection of the KA gene cluster. This metabolite is an inhibitor of pigment formation in animal tissues and is therefore used as a skin-whitening compound in cosmetics [19,20]. The KA cluster was recently found to be composed of only three genes, none of which encodes a PKS, NRPS, or other core SMB enzyme. Instead, the three genes encode an oxidoreductase, a Zn(II)$_2$-Cys$_6$ (C6)-type transcription factor, and a major facilitator superfamily transporter [10,11]. KA production is typically observed after 3 to 4 days of inoculation of *A. oryzae* in liquid growth media, and can be stopped by adding a small amount of sodium nitrate to the medium [21,22].

Figure 2 shows the results of MIDDAS-M analysis for three *A. oryzae* transcriptomes in the relative transcription observed under KA-inducing vs. KA-non-inducing conditions in two-color DNA microarray experiments; 4 vs. 2 days, 7 vs. 4 days, and without vs. with nitrate. Among the 12,084 genes of *A. oryzae* [13], 5,046 genes with expression in all three datasets were used for the analysis. The $M$ scores for the 7/4-day dataset are normally distributed when the cluster size $ncl = 1$, but the symmetry was lost, and the top of the distribution slid to the left, when $ncl = 3$ and 5, accompanied by the emergence of large $M$ scores outside of the normal distribution (Fig. 2A). MIDDAS-M emphasizes this deviation of the SMB cluster candidates through Equation 2, enabling their sensitive detection. In the 7/4-day dataset, a distinct single peak emerged in the $v_{max}$ score from the gene induction ratio ($m$ value) as designated by a red arrow in Fig. 2B. The gene cluster corresponding to this peak was composed of three genes, AO090113000136, AO090113000137, and AO090113000138, which were exact matches to the three KA biosynthetic genes [10,11]. The highly sensitive and specific detection of the KA gene cluster, which has a small cluster size of 3 and does not include any core genes, indicates that MIDDAS-M has strong potential as a motif-independent predictor of SMB gene clusters. In the 4/2-day
and without/with nitrate datasets, only small \( \omega_{\text{max}} \) signals were observed, indicating that the increase of KA productivity in the two datasets was not due to the transcriptional induction of the genes responsible for KA biosynthesis.

MIDDAS-M was also tested for *Fusarium verticillioides* using a time series of four transcriptomes obtained from mycelia grown in the liquid medium used to induce fumonisin production [23]. This fungus is a plant pathogen that produces mycotoxins and is phylogenetically distantly related to *Aspergillus*. A comprehensive comparison of the 4 transcriptomes followed by the MIDDAS-M prediction yielded several distinct peaks of \( \omega_{\text{max}} \), of which 5 corresponded to the known SMB gene clusters for fumonisin [24], perithecium pigment [23], fusaric acid [23], bikaverin [25], and fusarin [23] (Fig. 3). Although the size of the predicted SMB gene cluster for fusaric acid was three-fold larger than the experimentally validated clusters, the others were almost correct in size (Table 1). This result clearly illustrates the high sensitivity of MIDDAS-M in detecting functional SMB clusters.

The cluster harboring fusaric acid biosynthetic genes (peak c in Fig. 3B) was predicted to have 17 genes (FVEG_12519–FVEG_12535) by MIDDAS-M, whereas the cluster size reported by Brown et al was 5 (FVEG_12519–FVEG_12523) [23] (Table 1, Fig. S2 in Appendix S1). The gene expression profile in this region suggests existence of another cluster adjacent to the fusaric acid gene cluster with a few additional genes in between (Fig. S2 in Appendix S1). One of the remarkable features of MIDDAS-M is the potential to predict a gene cluster even though it includes a small number of genes that are not co-regulated with other cluster member genes. This enables sensitive detection of gene clusters from the dataset containing inaccurate data points due to their low expression levels and/or biological fluctuation under the same condition. It is thought that this characteristic led to the prediction of the above cluster much longer than the actual size by combining the two clusters into one. In addition to detecting the five clusters noted above, this analysis revealed two other VCs with high \( \omega_{\text{max}} \) scores (\( y_1 \) and \( y_2 \) in Fig. 3B). They were not predicted by SMURF, and were composed of 3 and 4 genes, respectively, the latter of which included an NRPS-like enzyme (Fig. 3B, Table S2 in Appendix S1). To assign peaks to their corresponding compounds, detailed analysis of the linkage between the gene cluster expression and compound productivity is necessary.

**Figure 2. Behavior and performance of MIDDAS-M in *A. oryzae*.** (A) Histograms of \( M \) scores at \( ncl = 1, 3, 5, 7, \) and 10 in the transcriptomes at 7 vs. 4 days of cultivation in kojic acid (KA)-production medium. The symmetry broke at a cluster size of 3 because of the emergence of large \( M \) scores due to the induction of the KA cluster genes. Arrows at the termini of the x-axis indicate the smallest and the largest values. (B) Emergence of a \( \omega_{\text{max}} \) peak by MIDDAS-M from the raw induction ratio. The x-axis designates relative position of the genes on the *A. oryzae* RIB40 genome when eight chromosomes are concatenated into one. The y-axis scales are the same for all three datasets in the same raw. The \( \omega_{\text{max}} \) peak indicated by the red arrow corresponds exactly to the three genes responsible for KA production.

doi:10.1371/journal.pone.0084028.g002

Large-scale detection of SMB gene clusters by MIDDAS-M

To demonstrate the fully computational and motif-independent features of MIDDAS-M for the comprehensive analysis of SMB gene clusters, we employed a systematic pairwise comparison of *A. flavus* 28 transcriptome datasets from a variety of cultivation conditions (GSE15435 [26], Fig. 4A). MIDDAS-M detected 240 candidate clusters with the threshold of 0.05 for the statistical
By MIDDAS-M. Certain peaks were detected under only limited combinations of conditions, illustrating the utility of MIDDAS-M for the comprehensive analysis of culture conditions that induce rarely expressed SMB genes (Fig. 4B). For example, the peak circled in Fig. 4B detected only in a limited conditions, composed of AFLA_035680 through AFLA_035720, was not detected either by SMURF or by antiSMASH.

The detected peaks were highly localized to NSBs (702 detected cluster genes out of 969 total; see Table S3 in Appendix S1). This result is in good agreement with the fact that the genes related to secondary metabolite biosynthesis, transport, and catabolism (Q-genes), identified in the EuKaryotic Orthologous Groups (KOG) [28,29] on NSBs [13]. In addition, the detected gene clusters were enriched for Q-genes compared with the whole genome, regardless of their inclusion of core genes (SMURF+/−) (Fig. 5A). Genes annotated as cytochrome P450 enzymes, which constitute a large enzyme family often involved in SMB gene clusters [30], represent 1.1% of the 13,471 genes in the A. flavus genome, and are contained in 9.1% of the 240 unique clusters detected by MIDDAS-M. The P450 gene content in the detected gene clusters increased drastically to >60%, by applying threshold $\omega_{\max} \text{max} \geq 15,800$ (Fig. 5B), although the number of clusters decreased exponentially along with increasing the threshold of $\omega_{\max}$ score (24 clusters when $\omega_{\max} \geq 10,000$, Fig. S3 in Appendix S1). SMB clusters are often regulated by C6-type transcription factors [31], and major facilitator superfamily (MFS) transporters are often present in SMB clusters [32]. These two genes also appear more frequently in the clusters as the threshold increased. Among 240 candidate SMB gene clusters detected by MIDDAS-M with the threshold of 0.05 false positive rate, 89% (213) were detected by SMURF or by antiSMASH. Of 240 candidate SMB gene clusters detected by MIDDAS-M, 60%, by applying threshold $\omega_{\max} \text{max} \geq 10,000$, Fig. S3 in Appendix S1), and this tendency continued when $\omega_{\max} > 10,000$ (71% or 17 in 24). These results strongly suggest that MIDDAS-M detected clusters of SMBs even when the clusters did not include the core genes. Detection of the KA cluster is the typical example. The ustiloxin B biosynthetic gene cluster, which was first detected by MIDDAS-M and experimentally-validated in this study, is another good example. These two clusters are both lacking known core genes, thus have never been predicted by the existing software tools based on sequence information of core genes, such as SMURF and antiSMASH (see detail in the next section). Use of high threshold of $\omega_{\max}$ and gene functional information will increase accuracy of predicting SMB gene clusters, though it may fail to detect novel SMB clusters.

**Figure 3. Clear detection of known SMB gene clusters in F. verticillioides by MIDDAS-M.** (A) Expression levels of each gene on the F. verticillioides genome in 4 samples of a transcriptome time series at 24, 48, 72, 96 h in liquid fumonisin-inducing media. The highest value of the 4 expression levels was plotted for each gene. (B) Absolute maximum cluster scores ($|\omega_{\max}|$) by the comprehensive pair-wise calculation ($C_j$) for each gene detected from the same transcriptome data as A. The step line plot in gray denotes the individual chromosomes. The peaks designated by a through e correspond to the 5 experimentally validated SMB clusters: a. fumonisin; b. perithe- cium pigment; c. fusaric acid; d. bikaverin; e. fusarin. Two peaks to which any known gene clusters do not correspond were designated as y1 and y2.

doi:10.1371/journal.pone.0084028.g003

**Identification of a novel ustiloxin B gene cluster by MIDDAS-M**

The comprehensive analysis of A. flavus transcriptomes by MIDDAS-M revealed a pair of culture conditions (cracked maize at 28°C versus 37°C) that showed 3 distinct peaks: the first peak corresponded to the aflatoxin biosynthetic gene cluster; the second peak to a putative cluster (designated cluster a) consisting of 18 genes (AFLA_094940–AFLA_095110; gene ID interval = 10 in most cases); and the third peak to a putative cluster (cluster b) consisting of 5 genes (AFLA_039200–AFLA_039240). To identify the compounds produced by clusters a and b, we constructed three types of A. flavus deletion mutants for each cluster using pfug as a selectable marker. For cluster a, mutant ΔAF_a had 13 genes (ΔAFLA_094940–AFLA_095060) deleted, mutant ΔAF_a had one gene (ΔAFLA_094960) deleted, and mutant ΔAF_a had one gene (ΔAFLA_095040) deleted. For cluster b, mutant ΔAF_b had five genes (ΔAFLA_039200–AFLA_039240) deleted, mutant ΔAF_b had one gene (ΔAFLA_039210) deleted, and mutant ΔAF_b had one gene (ΔAFLA_039230) deleted (Fig. S1 and Table S1 in Appendix S2). Certain peaks were detected under only limited conditions, illustrating the utility of MIDDAS-M for the comprehensive analysis of culture conditions that induce rarely expressed SMB genes (Fig. 4B). For example, the peak circled in Fig. 4B detected only in a limited conditions, composed of AFLA_035680 through AFLA_035720, was not detected either by SMURF or by antiSMASH.

The detected peaks were highly localized to NSBs (702 detected cluster genes out of 969 total; see Table S3 in Appendix S1). This result is in good agreement with the fact that the genes related to secondary metabolite biosynthesis, transport, and catabolism (Q-genes), identified in the EuKaryotic Orthologous Groups (KOG) [28,29] on NSBs [13]. In addition, the detected gene clusters were enriched for Q-genes compared with the whole genome, regardless of their inclusion of core genes (SMURF+/−) (Fig. 5A). Genes annotated as cytochrome P450 enzymes, which constitute a large enzyme family often involved in SMB gene clusters [30], represent 1.1% of the 13,471 genes in the A. flavus genome, and are contained in 9.1% of the 240 unique clusters detected by MIDDAS-M. The P450 gene content in the detected gene clusters increased drastically to >60%, by applying threshold $\omega_{\max} \text{max} \geq 15,800$ (Fig. 5B), although the number of clusters decreased exponentially along with increasing the threshold of $\omega_{\max}$ score (24 clusters when $\omega_{\max} \geq 10,000$, Fig. S3 in Appendix S1). SMB clusters are often regulated by C6-type transcription factors [31], and major facilitator superfamily (MFS) transporters are often present in SMB clusters [32]. These two genes also appear more frequently in the clusters as the threshold increased. Among 240 candidate SMB gene clusters detected by MIDDAS-M with the threshold of 0.05 false positive rate, 89% (213) were detected by SMURF or by antiSMASH. Of 240 candidate SMB gene clusters detected by MIDDAS-M, 60%, by applying threshold $\omega_{\max} \text{max} \geq 10,000$, Fig. S3 in Appendix S1), and this tendency continued when $\omega_{\max} > 10,000$ (71% or 17 in 24). These results strongly suggest that MIDDAS-M detected clusters of SMBs even when the clusters did not include the core genes. Detection of the KA cluster is the typical example. The ustiloxin B biosynthetic gene cluster, which was first detected by MIDDAS-M and experimentally-validated in this study, is another good example. These two clusters are both lacking known core genes, thus have never been predicted by the existing software tools based on sequence information of core genes, such as SMURF and antiSMASH (see detail in the next section). Use of high threshold of $\omega_{\max}$ and gene functional information will increase accuracy of predicting SMB gene clusters, though it may fail to detect novel SMB clusters.
Table 1. Experimentally-validated or SMURF-annotated SMB gene clusters detected by MIDDAS-M.

| Fungus       | Compound/SMURF          | \( n_{\text{max}} \) | Gene ID               | Cluster size | Otherd | Source              |
|--------------|-------------------------|-----------------------|-----------------------|--------------|--------|---------------------|
| A. oryzae    | Kojic acid              | 9544                  | AO090113000136 - AO090113000138 | 3            | 3      | [10,11]             |
| F. verticillioides | Bikaverin (Cluster 7) | 11708                 | FVEG_03379 – FVEG_03383 | 4            | 6      | [25], SMURF         |
|               | Fumonisin (Cluster 3)   | 9780                  | FVEG_00316 – FVEG_00329 | 14           | 15     | [24], SMURF         |
|               | Fusaric acid (Cluster 27) | 6398              | FVEG_12519 – FVEG_12535 | 17           | 5      | [23], SMURF         |
|               | Fusarin                 | 840                   | FVEG_11078 – FVEG_11086 | 9            | 9      | [23], SMURF         |
|               | Penicillicin pigment (Cluster 9) | 12533              | FVEG_03696 – FVEG_03699 | 6            | 4      | [23], SMURF         |
|               | Cluster 10              | 1700                  | FVEG_05526 – FVEG_05530 | 5            | 10     | SMURF               |
|               | Cluster 24              | 866                   | FVEG_11927 – FVEG_11931 | 5            | 7      | SMURF               |
| A. flavus    | Aflatoxin (Cluster 54)  | 99087                 | AFLA_139150 - AFLA_139320 | 18+5         | 29     | [37], SMURF         |
|              | Aflaxetin               | 24302                 | AFLA_139370 – AFLA_139410 |              |        |                     |
|              | (Cluster 14)            | 8984                  | AFLA_045490 - AFLA_045540 (ATM1) |              |        |                     |
|              | Cyclopiazonic acid      | 36281                 | AFLA_139460 – AFLA_139490 | 4            | 3      | [45]                |
|              | Gliotoxin-like (Cluster 22) | 32872           | AFLA_064380 – AFLA_064590 | 22           | 26     | Annotation, SMURF   |
|              | Kojic acid              | 8273                  | AFLA_096030 – AFLA_096060 | 4            | 3      | [10,11]; Blastp, E0.0 |
|              | Ustiloxin B             | 21857                 | AFLA_094940 – AFLA_095110 | 18           | ?      | This study          |
|              | Cluster 3               | 7369                  | AFLA_005320 – AFLA_005350 | 4            | 8      | SMURF               |
|              | Cluster 5               | 1960                  | AFLA_006170 – AFLA_006190 | 3            | 7      | SMURF               |
|              | Cluster 7               | 5193                  | AFLA_009980 – AFLA_010030 | 6            | 8      | SMURF               |
|              | Cluster 8               | 9341                  | AFLA_010600 – AFLA_010630 | 4            | 10     | SMURF               |
|              | Cluster 10              | 18356                 | AFLA_023000 – AFLA_023040 | 5            | 15     | SMURF               |
|              | Cluster 17              | 1423                  | AFLA_054370 – AFLA_054390 | 3            | 25     | SMURF               |
|              | Cluster 18              | 1072                  | AFLA_060030 – AFLA_060050 | 3            | 15     | SMURF               |
|              | Cluster 19              | 26351                 | AFLA_060660 – AFLA_060700 | 5            | 9      | SMURF               |
|              | Cluster 20              | 2079                  | AFLA_062820 – AFLA_062900 | 9            | 18     | SMURF               |
|              | Cluster 21              | 8227                  | AFLA_064260 – AFLA_064330 | 8            | 21     | SMURF               |
|              | Cluster 23              | 5702                  | AFLA_066690 – AFLA_066720 | 4+6          | 33     | SMURF               |
|              | Cluster 24              | 4508                  | AFLA_069320 – AFLA_069340 | 3            | 10     | SMURF               |
|              | Cluster 25              | 4219                  | AFLA_070860 – AFLA_080890 | 4+4          | 26     | SMURF               |
|              | Cluster 27              | 2012                  | AFLA_082140 – AFLA_082160 | 3            | 14     | SMURF               |
|              | Cluster 33              | 5797                  | AFLA_101700 – AFLA_101770 | 8            | 6      | SMURF               |
|              | Cluster 36              | 1026                  | AFLA_105410 – AFLA_105450 | 5            | 5      | SMURF               |
|              | Cluster 37              | 13236                 | AFLA_108550 – AFLA_108580 | 4            | 18     | SMURF               |
|              | Cluster 41              | 2503                  | AFLA_116130 – AFLA_116150 | 3+3          | 26     | SMURF               |
|              | Cluster 44              | 6277                  | AFLA_118390 – AFLA_118410 | 3            | 11     | SMURF               |
|              | Cluster 45              | 2494                  | AFLA_118940 – AFLA_119000 | 7            | 19     | SMURF               |
|              | Cluster 46              | 4420                  | AFLA_119080 – AFLA_119120 | 5            | 6      | SMURF               |
|              | Cluster 47              | 12300                 | AFLA_121470 – AFLA_121540 | 8            | 8      | SMURF               |
|              | Cluster 49              | 1429                  | AFLA_128030 – AFLA_128110 | 9            | 13     | SMURF               |
|              | Cluster 53              | 2813                  | AFLA_137830 – AFLA_137860 | 4+3          | 15     | SMURF               |
|              |                        | 1844                  | AFLA_137890 – AFLA_137910 |              |        |                     |

The detection threshold is >95th quantile (false positive rate 0.05). The most induced combinations of culture conditions are listed in Appendix S2.

aClusters with numbers are those predicted by SMURF. The list of the predicted gene clusters can be downloaded from http://jcvi.org/smurf/precomputed.php.
bGene IDs are for annotated genome sequences in GenBank (A. oryzae, F. verticillioides, and A. flavus) as described in Appendix S1.
cTwo numbers are described when the predicted clusters are divided into two regions and represent the corresponding clusters.
dCluster size experimentally validated or predicted by SMURF (refer to Source in detail).
doi:10.1371/journal.pone.0084028.t001
Figure 4. SMB gene cluster detection by MIDDAS-M in *A. flavus*. (A) A 3D view of the $\omega_{\text{max}}$ scores for all genes and combinations of culture conditions. Comprehensive detection of SMB gene clusters was performed on all 378 pairwise combinations of culture conditions from 28 transcriptomes. The gray and green areas denote blocks of synteny and non-synteny, respectively, with the *A. nidulans* genome. The positions of gene clusters possessing PKS and NRPS core genes predicted by SMURF are shown in orange and blue, respectively. The chemical structures of four *A. flavus* secondary metabolites are shown at the positions of corresponding SMB gene clusters; the ustiloxin B gene cluster was first identified in this paper. (B) Magnified view of the area on chromosome 2 corresponding to the black square in A. As an example, a yellow circle designates the peak observed specifically at particular positions, from which conditions for producing the corresponding compound were determined.

doi:10.1371/journal.pone.0084028.g004

Figure 5. Frequency of SMB-related genes in clusters detected by MIDDAS-M. (A) Ratios of SMB-related genes (Q-genes) detected by KOG analysis with the cluster genes detected by MIDDAS-M (hatched bars) and all the genes in the corresponding genome (gray bars). (B) The proportion of clusters containing genes annotated as P450 enzymes (pink), C6 transcription factors (blue), and major facilitator superfamily members (green) were calculated for detected clusters with the threshold score of $\omega_{\text{max}}$ in *A. flavus*. The value is plotted to a $\omega_{\text{max}}$ of 18,350, at which 10 clusters remain to be detected.

doi:10.1371/journal.pone.0084028.g005
S1). The deletion mutant lacking the entire aflatoxin cluster and the pyrG revertant were also constructed as positive controls. After solid cultivation of the 7 deletion mutants and the control strain (pyrG revertant) on cracked maize at 20°C for 7 days, water-soluble metabolites were analyzed by high-performance liquid chromatography-mass spectrometry (HPLC-MS). By comparing metabolite profiles between mutants, we found a negative ion spectrum at m/z 644.2 with a retention time (RT) of 8.9 min that was absent only in water extracts from the three deletion mutants corresponding to cluster a (Figs. 6B, 6C). Ultra-performance liquid chromatography-high-resolution mass spectrometry (UPLC-HRMS) showed that the accurate mass of the corresponding ion was 646.240 [M+H]^+ and 644.231 [M-H]^− with UV absorption at 290, 250, and 209 nm. By searching an organic compound database, we found that these measurements corresponded to ustiloxin B (C_{29}H_{39}N_{3}O_{13}S; MW 645.681). Ustiloxin B was first isolated as a water-soluble component of false smut balls on rice panicles infected by the fungus Ustilaginoidea virens [33-35]. The HPLC-purified compound from the water extract of the control strain (pyrG revertant) was compared with a ustiloxin B standard using UPLC-HRMS. The two compounds showed identical mass spectra with an RT of 1.61 min (Fig. 6D) as well as identical peaks in the extracted ion chromatogram at m/z 644.231 [M−H]^− and in the UV spectra at 290, 254, and 220 nm (Fig. 6E). These results provide the first evidence that the genes AFLA_094960 and AFLA_095040 are responsible for ustiloxin B biosynthesis, indicating that cluster a, composed of AFLA_094940 through AFLA_095110, is a ustiloxin B biosynthetic cluster.

Based on its chemical structure, ustiloxin B is likely characterized as a non-ribosomal peptide. One of the genes responsible for producing ustiloxin B, AFLA_095040, was putatively annotated as an NRPS-like enzyme in the NCBI database (gene ID: 7917917). However, the AFLA_095040 gene contains only the catalytic domain of a pyridoxal 5’-phosphate-dependent enzyme from amimotransferase family-5, which must be involved in reactions other than non-ribosomal peptide bond biosynthesis (Fig. S4 in Appendix S1 and the “ust” sheet in Appendix S2). Moreover, none of the NRPS-specific catalytic domains (A, C, PCP, or TE) were found in any genes in or near the cluster (AFLA_094930–AFLA_095170), as determined by a BLAST [36] search against the UniProtKB database [37,38]. Accordingly, the cluster was not detected by SMURF (http://jCVI.org/smurft/precomputed.php), antiSMASH (the “antiSMASH.AF” sheet in Appendix S2), or other currently available conventional SMB gene cluster prediction methods, which use catalytic domain sequence motif information. This result clearly indicates that MIDDAS-M has potential use as a motif-independent predictor of functional SMB gene clusters.

Discussion

In this work, we described the first sequence motif-independent algorithm for the discovery of functional fungal SMB gene clusters based on a combination of whole genome sequence data and transcriptome information. To achieve this novel and fully computational approach, we combined an algorithm to generate comprehensive virtual gene clusters on a genome of interest with a combination of whole genome sequence data and transcriptome information. This result clearly indicates that MIDDAS-M has potential use as a motif-independent predictor of functional SMB gene clusters.}

production of these compounds. In contrast to the former 3 clusters, the KA gene cluster does not include any genes considered as core SMB genes, such as PKSs, NRPSs, DMATs, or terpene cyclases (TCs). The KA gene cluster predicted by MIDDAS-M was the sole candidate with a correct cluster size. Nine gene disruption experiments were required to identify this cluster without MIDDAS-M prediction in our previous work using the same transcriptomes [11].

The fully computational and motif-independent feature of MIDDAS-M allowed for the comprehensive analysis of SMB gene clusters based on expression differences in a given pair of multiple transcriptomes. Because little is known about SMB gene clusters other than those containing PKS, NRPS, TC, and DMATS, the validation of the MIDDAS-M results is extremely difficult. Nonetheless, based on the MIDDAS-M prediction, we identified the first SMB gene cluster for ustiloxin B, the non-ribosomal peptide-like compound that inhibits mitochondrial assembly [35], in A. flavus. Although ustiloxin B was identified more than 20 years ago, the ustiloxin B biosynthetic gene cluster had remained unknown until the present study. The lack of the NRPS catalytic domains A, C, PCP, and TE in all genes both in the cluster and within 10 adjacent genes outside the cluster suggests a novel mechanism for cyclic peptide biosynthesis. Our further deletion experiments and sequence analysis revealed that at least 3 genes with unknown functions (AFLA_094970, AFLA_094900, and AFLA_094990) may be involved in the peptide bond synthesis and cyclization of the compound, supporting the idea above (data not shown). However, there still remains a possibility that additional gene encoding an NRPS for the ustiloxin biosynthesis may be located distantly from the cluster.

MIDDAS-M enables the highly sensitive identification of SMB gene clusters, but the predicted cluster sizes may be smaller than the actual cluster sizes in some cases. For example, the aflatoxin gene cluster of A. flavus was composed of 29 genes from AFLA_139150 through AFLA_139440 [39,40], but MIDDAS-M detected 25 genes, AFLA_139150 through AFLA_139410 (excluding AFLA_139330–AFLA_139360). This discrepancy is likely due to the Z-score transformation at each ncl used to normalize M scores before enhancement. When information from a candidate gene cluster(s) is included at a certain ncl, the standard deviation used for the denominator in Z-score transformation increases. As a result, the M score(s) of the strongly positive gene clusters tend to be smaller at the correct size. This factor does not affect the detection sensitivity of cluster positions but does affect the cluster boundary detection. One potential solution for this problem is to use another algorithm, such as co-expression analysis, for the precise prediction of cluster boundaries after the sensitive detection of cluster candidates by MIDDAS-M.

There are more than 100,000 fungal species in nature [41] that are potential producers of bioactive compounds [31]. Because fungal SMB genes are highly divergent [16,42,43], even fungal species closely related to those that have already been sequenced are worth sequencing to discover new SMB genes. We have confirmed that MIDDAS-M performs equally well when using transcriptomes from RNA-seq data in a comparative performance with DNA microarray for SMB gene cluster detection. MIDDAS-M enables the comprehensive exploration of functional SMB genes in fungal genomes by effectively utilizing the vast amount of available genome and transcriptome information, which will accelerate the discovery of biosynthesis or other functional categories of genes in the future.
Supporting Information

Appendix S1
Experimental details pertaining to the algorithm execution using transcriptome data, gene disruption, and the identification of ustiloxin B. (DOCX)

Appendix S2
The comprehensive MIDDAS-M prediction data for *F. verticillioides* (the “F.verticillioides” sheet) and *A. flavus* (the “A.flavus” sheet), the functional annotations of genes in or near the ustiloxin B cluster found by BLAST against UniProtKB (the “ust” sheet), and the result of antiSMASH prediction for *A. flavus* in comparison to that of MIDDAS-M (the “antiSMASH.AF” sheet). (XLSX)

Acknowledgments
The authors would like to thank Prof. Y. Hashimoto and Dr. H. Kobayashi, the University of Tokyo, for the kind donation of ustiloxin B, Dr. M. Arita, the University of Tokyo, and Dr. K. Tomii, AIST, for their critical review of the equations, and Dr. S. Aburatani, AIST, for effort of building the public URL. We thank Hokuren, an agricultural cooperative association in Hokkaido Prefecture, for their kind donation of cracked maize. We are grateful to Dr. R.H. Proctor, Dr. D.W. Brown, Dr. R.A.E. Butchko, and Dr. M. Busman, USDA-ARS-NCAUR, for valuable discussions, and to our colleagues for their help and advice.

Author Contributions
Conceived and designed the experiments: MU HK MM. Performed the experiments: MU JK TI NY IK. Analyzed the data: MU NN. Wrote the paper: MU MM JWB. Built the algorithm: MU HK. Provided critical discussions: KH KS KA JY JWB.

Figure 6. Identification of the ustiloxin B cluster in *A. flavus* based on the MIDDAS-M prediction. (A) MIDDAS-M results from a combination of culture conditions in maize at 28°C versus 37°C. The leftmost distinct peak corresponds to the aflatoxin gene cluster. The other two peaks were designated as clusters a and b. The step line plot in gray denotes the chromosomes. (B) Peaks at a retention time of 8.9 min detected in the extracted ion chromatograms of m/z 644.2±0.1 in negative ion mode were not observed in the *A. flavus* deletion mutants of the genes in cluster a (red). Chromatograms are for medium only (blue, negative control), the control strain (pyrG revertant, black), the aflatoxin cluster deletion mutant, and three mutants with deletions in cluster b (gray). (C) The mass spectra at the 8.9 min retention peaks in the control strain (above) and the deletion mutant ΔAF_a (below). The MS peak of m/z 644.2 in the control strain was not present in the deletion mutant. (D) Comparison of the mass spectra for ustiloxin B and the compound with m/z 644.2 in negative ion mode isolated from the control strain. (E) Comparison of the mass spectra for ustiloxin B and the compound with m/z 644.23 in negative ion mode and UV chromatograms at 290, 254, and 220 nm are indicated.

doi:10.1371/journal.pone.0084028.g006
References

1. Yu J, Bhatnagar D, Ehrlich KC (2002) Aflatoxin biosynthesis. Rev Iberoam Micol 19: 191–200.
2. Rask C, Larsen TO, Friisad JC (2010) Functional systems biology of Aspergillus. In: Machida M, Gomi K, editors. *Aspergillus*. Molecular Biology and Genomics. Norfolk, UK: Caister Academic Press. pp. 173–198.
3. Khalidi N, Seifeddin FT, Turner G, Haft D, Nierman WC, et al. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol 47: 736–741.
4. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski R, et al. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthetic gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39: W339–346.
5. Blin K, Medema MH, Kazerempour D, Fischbach MA, Breitling R, et al. (2013) antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 41: W204–W212.
6. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, et al. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J Biotechnol* 140: 13–17.
7. Andersen MR, Nielsen JB, Kløgaard A, Petersen LM, Zachariassen M, et al. (2012) Accurate prediction of secondary metabolite gene clusters in filamentous fungi. Proc Natl Acad Sci U S A 109: E99–107.
8. Chang PK, Horn BW, Dorner JW (2005) Sequence breakpoints in the aflatoxin biosynthesis gene cluster and flanking regions in nonaflatoxigenic *Aspergillus* *flavus* isolates. Fungal Genet Biol 42: 914–923.
9. Brodhun F, Feussner I (2011) Oxylipins in fungi. *FLBS* 278: 1047–1063.
10. Maru J, Yamane N, Ohashi-Kunihoro S, Ando T, Terabayashi Y, et al. (2011) Kojic acid biosynthesis in *Aspergillus* *oryzae* is regulated by a Zn(II)(2)Cys(6) transcriptional activator and induced by kojic acid at the transcriptional level. *J Biosci Bioeng* 112: 40–43.
11. Terabayashi Y, Sano M, Yamane N, Maru J, Tamano K, et al. (2010) Identification and characterization of genes responsible for biosynthesis of kojic acid, an industrially important compound from *Aspergillus* *oryzae*. Fungal Genet Biol 47: 953–961.
12. Saito K (1907) Uber die Saurebinding von *Aspergillus* *oryzae*. Botanical Magazine Tokyo 21: 7–11.
13. Machida M, Asai K, Sano M, Tanaka T, Kumagai T, et al. (2005) Genome sequencing and analysis of *Aspergillus* *oryzae*. *Nature* 438: 1157–1161.
14. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, et al. (2005) Sequencing the metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Science* 309: 1353–1357.
15. Koiso Y, Morisaki N, Yamashita Y, Mitsui Y, Shirai R, et al. (1998) Isolation and characterization of genes responsible for biosynthesis of kojic acid, a mycotoxin from false smut balls on rice panicles caused by *Ustilaginoidea virens*. *J Antibiot* (Tokyo) 51: 418–422.
16. Koiso Y, Motomi I, Iwasaki S, Sato S, Sonoda R, et al. (1992) *Ustilago* A: a phytotoxin and a mycotoxin from false smut balls on rice panicles. *Tetrahedron Let*ters 33: 4157–4160.
17. Koiso Y, Morisaki N, Iwashita Y, Misuki Y, Shirai R, et al. (1998) Isolation and structure of an antimicrobial cyclic peptide, ustiloxin F, chemical interrelation with a homologous peptide, ustiloxin B. *Fungal Genet Biol* 29: 316–322.
18. Yu J (2012) Current understanding on aflatoxin biosynthesis and future perspective in reducing aflatoxin contamination. *Toxins (Basel)* 4: 1024–1057.
19. Alachul SF, Gush W, Miller W, Myers EW, Lipman DJ (1998) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
20. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40: 191–195.
21. Sanchez JP, Somoza AD, Keller NP, Wang CC (2012) Advances in *Aspergillus* secondary metabolite research in the post-genomic era. *Nat Prod Rep* 29: 531–571.
22. Wang HM, Chen CC, Giridhar R, Chang TS, Wu WT (2005) Repeated-batch production of kojic acid in a cell-retention fermenter using *Aspergillus* *oryzae* M539. *J Ind Microbiol Biotechnol* 32: 227–233.
23. Brown DW, Butchko RA, Bauman M, Proctor RH (2012) Identification of gene clusters associated with fusaric acid, fusicarpin, and perithecial pigment production in *Fusarium verticillioides*. *Fungal Genet Biol* 49: 521–532.
24. Proctor RH, Brown DW, Platter RD, Desjardins AE (2003) Co-expression of 15 contiguous genes delineates a fumonisin biosynthetic gene cluster in *Gibberella manihotensis*. *Fungal Genet Biol* 38: 237–249.
25. Yu J, Willmann A, Straeten M, Kleigrewe K, Beyer M, et al. (2009) Biosynthesis of the red pigment bikaverin in *Fusarium fujikuroi*: genes, their function and regulation. *Mol Microbiol* 72: 931–946.
26. Georgianna DR, Fedorova ND, Burroughs JL, Dolezal AL, Bok JW, et al. (2010) Beyond aflatoxin: four distinct expression patterns and functional roles associated with *Aspergillus* *flavus* secondary metabolism gene clusters. *Mol Plant Pathology* 11: 213–226.
27. Kusumoto K, Nogata O, Ohta H (2000) Directed deletions in the aflatoxin biosynthesis gene homolog cluster of *Aspergillus* *oryzae*. *Curr Genet* 37: 104–111.
28. Takasawa RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.
29. Tanoue RL, Kimori K, editors. *Fungal Genomics*. Norfolk, UK: Caister Academic Press. pp. 173–198.
30. Weitzel C, Simonson HT (2013) Cytochrome P450-enzymes involved in the biosynthesis of monoo- and sesquiterpenes. *Phytochemistry Reviews*.
31. Keller NP, Turner G, Bennett JW (2005) Fungal secondary metabolism - from biotechnology to genomics. *Nat Rev Microbiol* 3: 937–947.
32. Coleman JJ, Mylonakis E (2009) Efflux in fungi: la piece de resistance. *PLoS Pathog* 5: e1000486.
33. Koiso Y, Li Y, Iwasaki S, Hanaoka K, Kobayashi T, et al. (1994) Ustiloxins, antimicrobial cyclic peptides from false smut balls on rice panicles caused by *Ustilaginoidea virens*. *J Antibiot* (Tokyo) 47: 765–773.
34. Koiso Y, Natori M, Iwasaki S, Sato S, Sonoda R, et al. (1992) *Ustilago* A: a phytotoxin and a mycotoxin from false smut balls on rice panicles. *Tetrahedron* 48: 418–422.
35. Koiso Y, Morisaki N, Yamashita Y, Misuki Y, Shirai R, et al. (1998) Isolation and structure of an antimicrobial cyclic peptide, ustiloxin F, chemical interrelation with a homologous peptide, ustiloxin B. *J Antibiot* (Tokyo) 51: 418–422.
36. Alachul SF, Gush W, Miller W, Myers EW, Lipman DJ (1998) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
37. Tatusov RL, Koonin EV, Lipman DJ (2002) COG: a tool for genome-wide analysis of protein families. *Nucleic Acids Res* 30: 52–60.