

# A scalable platform to identify fungal secondary metabolites and their gene clusters

Kenneth D Clevenger<sup>1,2,7</sup>, Jin Woo Bok<sup>3,7</sup>, Rosa Ye<sup>4,7</sup>, Galen P Miley<sup>1</sup>, Maria H Verdán<sup>1</sup>, Thomas Velk<sup>3</sup>, Cynthia Chen<sup>4</sup>, KaHoua Yang<sup>3</sup>, Matthew T Robey<sup>5</sup>, Peng Gao<sup>2</sup>, Matthew Lamprecht<sup>4</sup>, Paul M Thomas<sup>2,5</sup>, Md Nurul Islam<sup>4</sup>, Jonathan M Palmer<sup>6</sup>, Chengcang C Wu<sup>4\*</sup>, Nancy P Keller<sup>3\*</sup> & Neil L Kelleher<sup>1,2,5\*</sup>

**The genomes of filamentous fungi contain up to 90 biosynthetic gene clusters (BGCs) encoding diverse secondary metabolites—an enormous reservoir of untapped chemical potential. However, the recalcitrant genetics, cryptic expression, and unculturability of these fungi prevent scientists from systematically exploiting these gene clusters and harvesting their products. As heterologous expression of fungal BGCs is largely limited to the expression of single or partial clusters, we established a scalable process for the expression of large numbers of full-length gene clusters, called FAC-MS. Using fungal artificial chromosomes (FACs) and metabolomic scoring (MS), we screened 56 secondary metabolite BGCs from diverse fungal species for expression in *Aspergillus nidulans*. We discovered 15 new metabolites and assigned them with confidence to their BGCs. Using the FAC-MS platform, we extensively characterized a new macrolactone, valactamide A, and its hybrid nonribosomal peptide synthetase–polyketide synthase (NRPS–PKS). The ability to regularize access to fungal secondary metabolites at an unprecedented scale stands to revitalize drug discovery platforms with renewable sources of natural products.**

It is estimated that there are between 500,000 and 3 million ascomycete fungal species on earth<sup>1</sup>, each containing on the order of 50–90 BGCs encoding the production of secondary metabolites<sup>2–4</sup>. However, it has proved challenging to translate this vast biosynthetic potential of fungal genomes into defined and renewable chemical libraries, and most progress has been made in only the handful of fungal species for which well-established genetic tools exist, such as *A. nidulans*<sup>5–7</sup>. Ascomycete fungi represent a tremendous source of natural products; however, the industrial-scale exploitation of their 20–40-Mb genomes awaits a technology that can access their content in a robust, deterministic fashion with established rates of return.

The discovery of new molecules through the expression of secondary metabolite BGCs in heterologous hosts offers many advantages that can complement direct studies of secondary metabolism in wild-type fungi. These advantages include the easy association of mass-spectrometry-detected metabolites with their BGCs, the production of silent metabolites (i.e., metabolites that are not typically observed to be produced by a given species, but whose BGCs are present), facile genetic dissection of metabolite biosyntheses, overproduction of valuable metabolites, and engineering of metabolites through synthetic biology approaches.

The attractiveness of heterologous reconstitution has led to several reports on the reconstitution of individual megasynthase backbone genes<sup>8–11</sup> or entire BGCs<sup>12,13</sup> to produce intermediates or final products of fungal secondary metabolism<sup>13–16</sup>. However, reports on intact clusters have been limited to analyses of only one or two clusters at a time. Because these studies have involved expression vectors such as yeast artificial chromosomes, which are unstable and difficult to handle, or cosmids and fosmids, which are often not large enough to encode complete secondary metabolite BGCs, substantial optimization has been required for each construct<sup>8,17</sup>. For

example, according to antiSMASH<sup>18</sup> gene cluster predictions, the 56 complete clusters analyzed in the present study range from 38,211 bp to 108,634 bp in size, meaning they are all at or beyond the size limit of normal fungal heterologous expression approaches. In investigations of partial clusters or individual megasynthase genes, no more than nine megasynthases have been analyzed in a single study, to our knowledge<sup>14</sup>. One possible reason for this limited scale is the lack of an effective metabolomic strategy. Specifically, there has not been a way to quickly differentiate heterologous expression products from the thousands of compounds in the heterologous host's endogenous metabolome—a critical requirement of a general method for discovering new metabolites in a scalable fashion.

To address the challenge of heterologously expressing large fungal gene clusters, one can use FACs to capture randomly sheared fungal genomic DNA with insert sizes up to 300 kb—large enough to contain full-length BGCs with their accessory genes and regulatory elements. A FAC has been successfully used to produce the known metabolite tereazine D from its BGC through a fungal expression host, thus demonstrating that FACs are able to act as fungal expression vectors<sup>19</sup>. However, this approach has not been used for natural product discovery or for large-scale screening of fungal BGCs because there is not a tool for identifying novel FAC-encoded secondary metabolites of unknown formulae from the thousands of signals generated by untargeted liquid chromatography (LC)–mass spectrometry analysis of a fungal host's metabolome.

Here we report the FAC-MS pipeline and the combination of large-scale FAC generation with a robust screening and scoring pipeline that uses untargeted metabolomics and genetic validation (Fig. 1). Using the FAC-MS approach, we screened a library of 56 FACs harboring uncharacterized fungal BGCs from three diverse species within the *Aspergillus* genus: *Aspergillus terreus*, *Aspergillus*

<sup>1</sup>Department of Chemistry, Northwestern University, Evanston, Illinois, USA. <sup>2</sup>Proteomics Center of Excellence, Northwestern University, Evanston, Illinois, USA. <sup>3</sup>Department of Medical Microbiology and Immunology and Department of Bacteriology, University of Wisconsin–Madison, Madison, Wisconsin, USA. <sup>4</sup>Intact Genomics, Inc., St. Louis, Missouri, USA. <sup>5</sup>Department of Molecular Biosciences, Northwestern University, Evanston, Illinois, USA. <sup>6</sup>Center for Forest Mycology Research, Northern Research Station, US Forest Service, Madison, Wisconsin, USA. <sup>7</sup>These authors contributed equally to this work. \*e-mail: cwu@intactgenomics.com, npkeller@wisc.edu or n-kelleher@northwestern.edu

*aculeatus*, and *Aspergillus wentii*. We detected 17 compounds, each likely to represent different scaffolds, including polyketides, nonribosomal peptides, and terpenoids, and confidently assigned them to their BGCs. These 17 metabolites were produced by 15 different FACs, including one FAC that expressed three distinct scaffolds. Of the 17 metabolites detected and associated with BGCs, 15 were previously unreported in secondary metabolite databases, and thus represent newly identified secondary metabolites. Most of these newly identified metabolites were silent (i.e., not detectable by LC–mass spectrometry analysis) in their native fungal hosts under the conditions used. Also among these 17 secondary metabolites was the benzodiazepine benzomalvin A (**1**), for which the BGC was previously unknown. We demonstrated the flexibility of the FAC–MS pipeline by introducing a series of genetic deletants into a FAC that encodes benzomalvin, which allowed its biosynthesis to be analyzed for the first time, to our knowledge. Furthermore, we show that the FAC–MS process enabled the purification and structural elucidation of new molecules, including the macrolactone valactamide A, with a new scaffold produced by a hybrid NRPS–PKS gene cluster.

## RESULTS

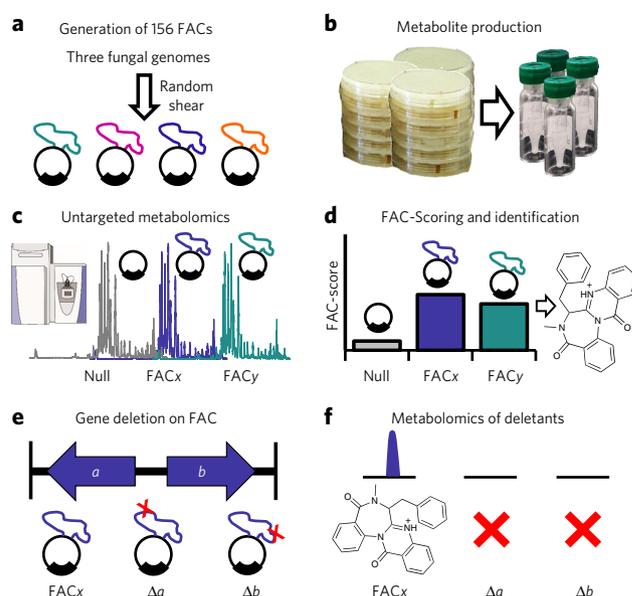
### FAC library creation and screening

We created a library of FACs by capturing ~100-kb inserts of randomly sheared genomic DNA from *A. wentii* (strain DTO 134E9) and *A. aculeatus* (strain ATCC16872), encompassing 95 FACs, each predicted to contain at least one BGC for a secondary metabolite on the basis of insert end-sequencing and alignment to the reference genome. We combined these 95 FACs with our previously reported library of 61 FACs from *A. terreus* (strain ATCC 20542)<sup>19</sup> to produce a final library that contained 156 FACs spanning ~15 Mb of DNA. We chose these three species for the FAC library because they are expected to possess distinct chemical repertoires, as they are phylogenetically distant aspergilli<sup>20</sup>.

Previously, we reported the successful production of the known fungal secondary metabolite tereazine D from a FAC<sup>19</sup>. To test the possibility of using FACs to pursue the high-throughput discovery of new secondary metabolites, we selected 56 FACs predicted to contain uncharacterized BGCs (i.e., BGCs with no known product or well-characterized homolog; Online Methods and Fig. 1a). We transformed each FAC into *A. nidulans* for growth and metabolite extraction (Fig. 1b). We then subjected extracts from the 56 FAC-transformed strains (each grown in biological quadruplicate, repeated from the step of FAC transformation) to LC–mass spectrometry analysis and automated compound detection by XCMS<sup>21</sup> (Fig. 1c).

### Metabolomic scoring of the FAC library

After compound detection, we applied a new scoring system developed for FAC–MS that features a FAC–Score, which can help scientists identify heterologously expressed metabolites from complex mixtures (Online Methods and Fig. 1d) by filtering out signals present in host extracts or in more than one FAC strain. In our previous study<sup>19</sup>, we were unable to identify novel metabolites produced by FAC-encoded gene clusters because of the lack of an effective mechanism for sorting through the ~5,000 LC–mass spectrometry signals detected from each FAC-transformed strain. The FAC–Score examines the abundance of a compound across all controls and FAC-transformed strains simultaneously; high scores are assigned to compounds that are detected in only one FAC strain, and lower scores are assigned to compounds detected in multiple FAC strains or in negative controls. Thus, we were able to use the FAC–Scores to filter the data from 56 FACs so as to eliminate endogenous host metabolites. FAC–Scores ranged from –1.9 to 9.6, with values greater than 0 representing compounds likely to be produced by enzymes encoded on the FAC (Supplementary Results, Supplementary Fig. 1). Application of the FAC–Score to the data set generated from the



**Figure 1 | The FAC–MS platform.** (a) We generated FACs from randomly sheared genomic DNA and cloned them into self-replicating *E. coli*–fungal shuttle vectors. The final library contained 156 FACs, with each FAC putatively containing at least one unique BGC from *A. nidulans*, *A. wentii*, or *A. aculeatus*. (b) We carried out transformation into the heterologous expression host (*A. nidulans*), growth, and metabolite extraction for 56 FACs predicted by bioinformatics to encode previously uncharacterized BGCs. (c) Secreted metabolite extracts from each FAC were subjected to untargeted high-resolution LC–tandem mass spectrometry analysis. (d) Untargeted compound detection and application of the novel FAC–Score systematically identified FAC-encoded secondary metabolites from the complex host metabolic background. (e) Biosynthetic backbone genes for FAC–Score-predicted secondary metabolites were deleted from their FACs, and the deletant FACs were retransformed into *A. nidulans*. ORF, open reading frame. (f) We rescreened FACs with gene deletions via LC–mass spectrometry to validate the role of the deleted gene in the production of each secondary metabolite, which confirmed the elimination of the predicted metabolite by backbone-gene deletion.

library of 56 FACs immediately reduced the data for analysis from thousands of compounds per strain to typically one or a handful of top-scoring compounds per strain. In 29 out of 56 FACs, the top-scoring compound had a score greater than 0. We then manually analyzed each top-scoring compound to determine whether it was unique to a single FAC strain. We found that 16 of the 29 compounds were unique to a single FAC strain. The other 13 could arguably be detected at very low levels in either negative controls or other FAC strains, which prevented a high-confidence assertion that their presence was exclusively FAC dependent; for the sake of conservatism, we excluded these 13 compounds from further analysis in this study.

One possible source of false positives in FAC–MS could be perturbation of the host metabolome by the FAC, leading to compounds with high FAC–Scores that are actually upregulated host metabolites. To empirically validate that the detected metabolites were indeed produced by their assigned BGCs, we deleted a backbone gene (i.e., a PKS or NRPS gene) from each FAC (Fig. 1e, Table 1, Supplementary Fig. 2). We then retransformed the FACs with backbone-gene deletions into the host and analyzed them by LC–mass spectrometry to determine whether the signals from the small-molecule product(s) with high FAC–Scores had indeed been eliminated (Fig. 1f). Importantly, because the engineered FACs were generated and propagated in *Escherichia coli*, this deletion step

**Table 1 | FAC-encoded secondary metabolites detected**

Species	FAC	FAC-Score	Metabolite name	Exact mass	Predicted molecular formula	Gene deleted	Cryptic?
<i>A. aculeatus</i>	AaFAC6A16	8.9	facms0001	402.195	C <sub>25</sub> H <sub>26</sub> N <sub>2</sub> O <sub>3</sub>	NRPS	No
<i>A. aculeatus</i>	AaFAC2P8	4.9	facms0002	558.226	C <sub>29</sub> H <sub>30</sub> N <sub>6</sub> O <sub>6</sub>	NRPS	Yes
<i>A. aculeatus</i>	AaFAC10D7	4.5	facms0003	273.028	C <sub>13</sub> H <sub>7</sub> NO <sub>6</sub>	PKS	Yes
<i>A. terreus</i>	AtFAC9J20	9.6	facms0004/ benzomalvin A	381.147	C <sub>24</sub> H <sub>18</sub> N <sub>3</sub> O <sub>2</sub>	NRPS	No
<i>A. terreus</i>	AtFAC9J20	5.6	facms0005	368.270	C <sub>25</sub> H <sub>36</sub> O <sub>2</sub>	Terpene synthase	No
<i>A. terreus</i>	AtFAC5L7	5	facms0006	673.261	C <sub>36</sub> H <sub>39</sub> N <sub>3</sub> O <sub>10</sub>	NRPS	No
<i>A. terreus</i>	AtFAC9H19	1.5	facms0007	609.918	ND	PKS	Yes
<i>A. terreus</i>	AtFAC5N15	1.2	facms0008	277.204	C <sub>17</sub> H <sub>27</sub> NO <sub>2</sub>	NRPS	Yes
<i>A. wentii</i>	AwFAC1H17	6	facms0009	422.172	C <sub>25</sub> H <sub>26</sub> O <sub>6</sub>	NRPS	Yes
<i>A. wentii</i>	AwFAC1K8	5.6	facms0010	987.550	ND	PKS	Yes
<i>A. wentii</i>	AwFAC4D8	4.8	facms0011	679.349	C <sub>38</sub> H <sub>45</sub> N <sub>7</sub> O <sub>5</sub>	NRPS	Yes
<i>A. wentii</i>	AwFAC4E11	3.8	facms0012	576.310	C <sub>35</sub> H <sub>44</sub> O <sub>7</sub>	PKS	Yes
<i>A. wentii</i>	AwFAC1B1	3.5	facms0013	471.202	C <sub>24</sub> H <sub>29</sub> N <sub>3</sub> O <sub>7</sub>	NRPS	Yes
<i>A. wentii</i>	AwFAC3D3	2.6	facms0014	735.198	C <sub>41</sub> H <sub>29</sub> N <sub>5</sub> O <sub>9</sub>	NRPS	Yes
<i>A. wentii</i>	AwFAC4D17	1.4	facms0015	541.378	C <sub>33</sub> H <sub>51</sub> NO <sub>5</sub>	NRPS	Yes
<i>A. terreus</i>	AtFAC7O19	5.6	facms0016	309.111	C <sub>17</sub> H <sub>15</sub> N <sub>3</sub> O <sub>3</sub>	NRPS	No
<i>A. terreus</i>	AtFAC9J20- ΔbenY	23.6	facms0017/ valactamide A	506.408	C <sub>30</sub> H <sub>54</sub> N <sub>2</sub> O <sub>4</sub>	NRPS/PKS	No

ND, not determined.

was facile and eliminated the need for complex genetic manipulation of the parent fungal strains. Because of the difficulty of genetic manipulation in many fungi, this step is often a major obstacle in studies of secondary metabolite biosynthesis. Of the 16 FAC strains and putative metabolite products identified from different FACs on the basis of FAC-Scores, 15 were supported by the deletion (meaning deletion of the backbone gene led to elimination of the associated metabolite assigned by the FAC-Score), and 1 was refuted.

### Discovery and validation of FAC metabolites

We screened 56 FACs from three species, each encoding a previously uncharacterized BGC (Supplementary Fig. 1). We then generated a gene cluster diagram for each of the 15 validated FACs with an identified secondary metabolite product (Supplementary Fig. 2). The corresponding molecular formulae and specific backbone-gene deletions confirming the metabolite are shown in Table 1, and partial LC–mass spectrometry traces for backbone-gene deletions and parent strains are shown in Supplementary Figure 2.

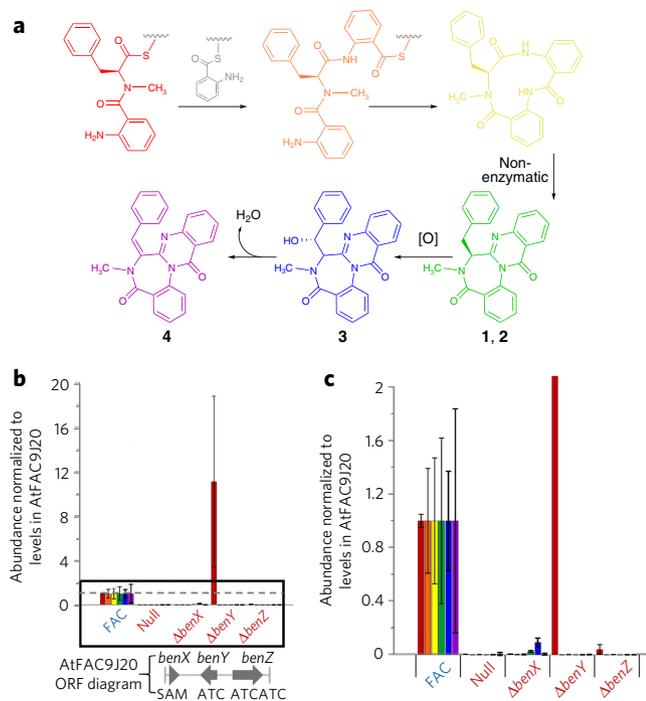
To distinguish known from previously unidentified compounds, we dereplicated the top-scoring metabolites from each FAC strain against the exact masses of ~280,000 known natural products from the Dictionary of Natural Products (<http://dnp.chemnetbase.com>; Online Methods). Also, we compared observed tandem mass spectrometry spectra of metabolites to computationally predicted spectra of compounds in pubchem (using the program MAgMA Online (<https://www.emetabolomics.org>); Online Methods)<sup>22</sup>. Of the 15 compounds, 14 could not be matched to known molecules in this manner, which indicated that they represented novel secondary metabolites. The 15th compound (Supplementary Fig. 2a), which had the highest FAC-Score from the *A. terreus*-derived FAC designated as AtFAC9J20, was predicted to be an atropisomeric mixture of the orphan benzodiazepines benzomalvin A (1) and benzomalvin D (2) (Fig. 2a) on the basis of its molecular formula and tandem mass spectrometry spectra. AtFAC9J20 also expressed several other high-scoring compounds that were unrelated to benzomalvin, leading to two additional compounds

(Supplementary Fig. 2b,c), each from a distinct gene cluster but all encoded on the same FAC (see below).

We derived FAC AtFAC9J20 from an unsequenced strain of *A. terreus* ATCC 20542 (from nucleotide 2,151,734 to the telomeric end of chromosome 4). The FAC contained an insert of 102,722 bp that was sequenced (GenBank [KX449366](https://www.ncbi.nlm.nih.gov/nuclot/KX449366)), aligned to the corresponding region in the reference *A. terreus* strain NIH 2624, and annotated (Supplementary Tables 1 and 2, Supplementary Fig. 3). AtFAC9J20 contained an ~36-kb DNA sequence toward the telomeric end that was missing from the NIH 2624 reference genome, as well as large insertions, deletions, duplications, inversions, and other genetic variations not seen in the same region in NIH 2624 (Supplementary Fig. 3, Supplementary Tables 1 and 2). To map the backbone genes to their metabolite products and demonstrate the utility of FAC-MS for analyses of complex biosynthetic systems, we produced deletants of each predicted backbone gene from AtFAC9J20 in order to readily assign biosynthetic genes to metabolites and dissect their biosynthetic pathways (Supplementary Table 3). Below we describe the dissections of three gene clusters for three distinct scaffolds.

### Identification of benzomalvin A/D from AtFAC9J20

The highest-scoring small molecule from AtFAC9J20, an ion with  $m/z = 382.1547$  and two different retention times of 38 and 41 min, was predicted to be benzomalvin A/D as described above. The chemical formula predicted from the accurate mass was C<sub>24</sub>H<sub>19</sub>N<sub>3</sub>O<sub>2</sub>, which matched that of neutral benzomalvin A/D within 1 part per million (p.p.m.) (observed, 381.1474 Da; theoretical, 381.1477 Da). This identification was also consistent with the observation of two chromatographic peaks, as benzomalvin A is reported to be in conformational equilibrium with its atropisomer, benzomalvin D, which would explain the presence of two LC peaks at different retention times<sup>23</sup> (Supplementary Fig. 4). We unambiguously confirmed the identity of the compound via total synthesis of an LC–mass spectrometry standard for the two atropisomers and direct comparison to the observed products of AtFAC9J20, including



**Figure 2 | Analysis of benzomalvin A/D biosynthesis.** (a) Deletant data suggest that the biosynthesis of benzomalvin A/D depends on the enzymes BenX, BenY, and BenZ, and is followed by the production of benzomalvins E and B at low levels either through a host cytochrome P450 enzyme (benzomalvin E; **3**) or through the spontaneous elimination of water (benzomalvin B; **4**). Wavy lines indicate phosphopantetheine thioesters; [O], oxygen. (b,c) The ion abundances of predicted benzomalvin precursors and products for AtFAC9J20, AtFAC9J20-null (simultaneous deletion of all ORFs from *benX* to *valA* is shown in **Supplementary Table 2**), AtFAC9J20- $\Delta benX$ , AtFAC9J20- $\Delta benY$ , and AtFAC9J20- $\Delta benZ$ . Abundances shown are averages from analyses of biological replicates from the point of fungal transformation. Error bars indicate  $\pm$ s.d. ( $n = 4$ ). Shown are the full chart (b) and a zoomed-in view of the area enclosed by the black box in b (c). The dashed gray lines represent the normalized level of each precursor in AtFAC9J20. Bars are color-coded to represent the intermediates and products shown in a. *benY* and *benZ* deletants both abolished the signal of each intermediate and product, with the exception of the dipeptide Anth-NmPhe in the case of *benY*. Deletion of *benX* substantially decreased, but did not eliminate, the levels of each product and intermediate, consistent with the presence of host methyltransferase activity.

through coinjection of the standard and the AtFAC9J20 extract (**Supplementary Fig. 5**). To our knowledge, the benzomalvin BGC has not been reported previously, and benzomalvins inhibit human indoleamine 2,3-dioxygenase and the neuroreceptor for substance P, called NK1 (refs. 24,25).

Extracts from wild-type *A. terreus* strain 20542 (the strain used to construct the FACs) also contained benzomalvin A/D, as determined on the basis of matching retention times, accurate mass, and tandem mass spectrometry data (**Supplementary Fig. 5**), though at  $\sim$ 350-fold lower abundance than in the AtFAC9J20-containing host (**Supplementary Fig. 6**). In contrast, we did not detect benzomalvin A/D in the sequenced reference strain, *A. terreus* NIH 2624. Additionally, we putatively detected two other known benzomalvins, benzomalvins E (**3**) and B (**4**; **Fig. 2a**), in AtFAC9J20 and *A. terreus* ATCC 20542 on the basis of accurate mass. To analyze benzomalvin biosynthesis in detail, we deleted the newly named genes *benX*, *benY*, and *benZ* from AtFAC9J20 and analyzed the resulting deletants by LC-mass spectrometry (**Fig. 2b,c**). Interestingly, deletion of *benY* led to accumulation of the predicted benzomalvin

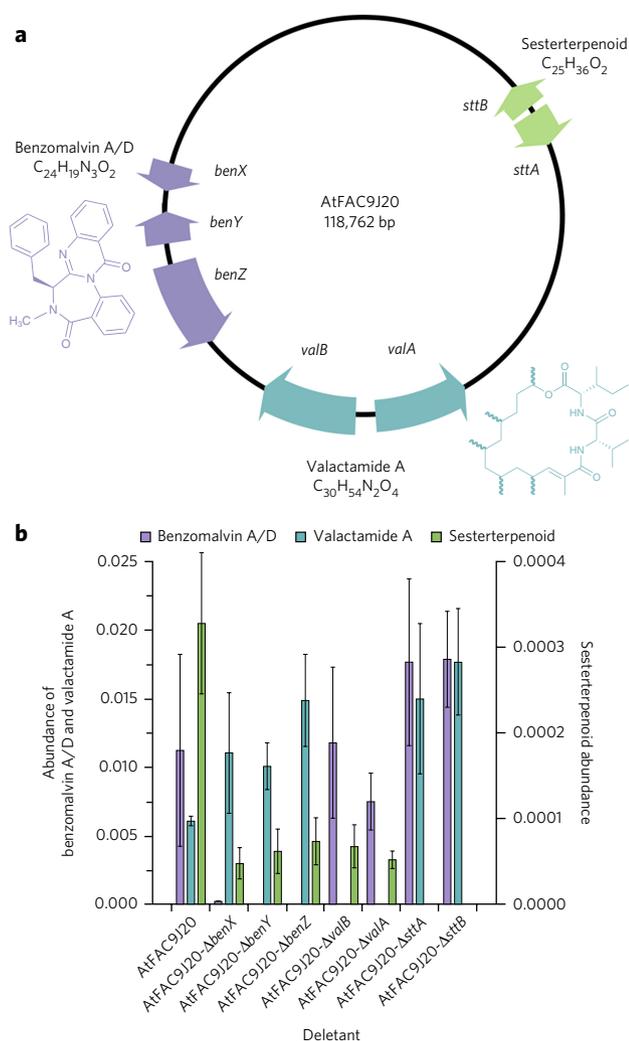
A/D dipeptide precursor, anthranilate-*N*-methylphenylalanine (Anth-NmPhe), whereas deletion of *benZ* eliminated the dipeptide precursor. Deletion of the gene *benX*, predicted to encode a SAM-binding methyltransferase, led to a tenfold decrease in the amount of benzomalvin A/D, but not its complete abolishment (**Fig. 2b,c**), which suggested that a promiscuous host methyltransferase enzyme partially rescued the *benX* deletant. We determined the level of benzomalvin A/D in each deletant, as well as the levels of the following putative biosynthetic precursors and products identified by mass analysis of precursor ions: Anth-NmPhe, Anth-NmPhe-Anth, the 11-member macrocyclic intermediate, benzomalvin E, and benzomalvin B (**Fig. 2b,c**). We used deletant data and bioinformatic analyses to create a proposed model of benzomalvin A/D biosynthesis (**Fig. 2a**, **Supplementary Fig. 7**).

### Identification of a sesterterpenoid from AtFAC9J20

AtFAC9J20 produced many high-scoring compounds, raising the exciting prospect that a single FAC could express multiple intact BGCs, and suggesting that FAC-MS combined with genetic deletants could quickly tease this out. Another compound with a FAC-Score of 5.6 was found to also be unique to AtFAC9J20. That compound had a molecular cation with  $m/z = 369.2783$  and a predicted molecular formula for the neutral compound of  $C_{25}H_{36}O_2$  (observed, 368.2703 Da; theoretical, 368.2715 Da; error,  $-3$  p.p.m.) (**Supplementary Fig. 8**); tandem mass spectrometry also showed fragment ions characteristic of a terpenoid. This molecular formula is reported for four closely related members of the ophiobolin family of sesterterpenoid natural products from *Aspergillus ustus*, as well as the sesterterpenoid variecolin from *Aspergillus variecolor*<sup>26-29</sup> (**Supplementary Fig. 9**). The observed ion was also present in extracts of both *A. terreus* NIH 2624 and *A. terreus* ATCC 20542, which suggested that its biosynthetic machinery would be shared by AtFAC9J20 and NIH 2624. AtFAC9J20 contains an adjacent predicted terpene synthase gene, *sttA*, and the cytochrome P450 gene *sttB*, which are also present in *A. terreus* NIH 2624 (**Fig. 3a**, **Supplementary Fig. 3a**). To test the roles of these genes in sesterterpenoid biosynthesis, we deleted them individually from AtFAC9J20 (**Supplementary Table 2**), which led to elimination of the sesterterpenoid product in both cases (**Fig. 3b**, **Supplementary Fig. 10**). This finding identified *sttA* and *sttB* as a sesterterpenoid BGC that is necessary for the production of a metabolite with the molecular formula  $C_{25}H_{36}O_2$ .

### Identification of a new metabolite named valactamide A

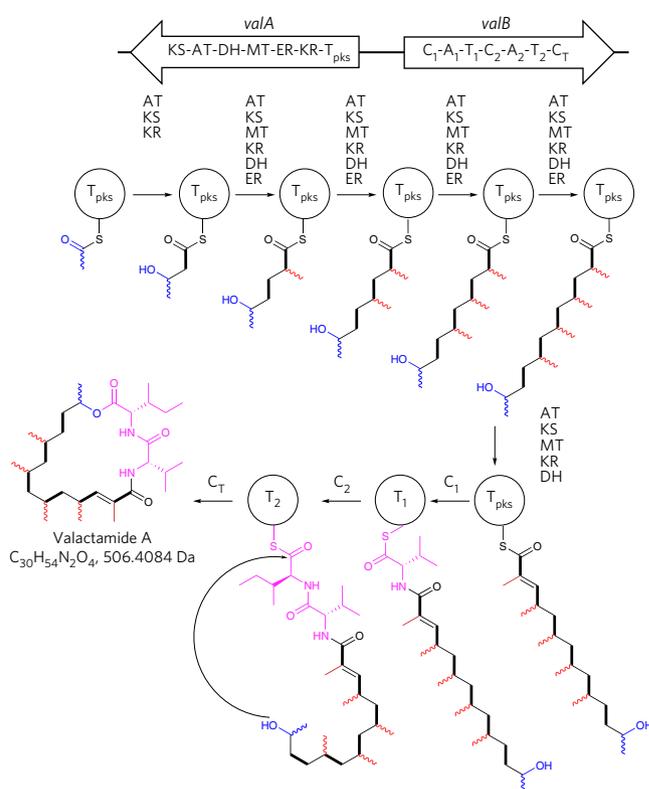
To determine whether the additional highly reducing PKS (HRPKS; *valA*) or NRPS (*valB*) genes on this FAC were able to produce a product, we subjected a host harboring AtFAC9J20- $\Delta benY$  to the FAC-MS pipeline, with the hope that inactivation of the benzomalvin BGC would lead to increased activity of other clusters or improved FAC-Scores for new compounds (or both). Strikingly, the top ion, initially given the platform name facms0017 (**Table 1**), received the highest FAC-Score in our study (23.6); it had an  $m/z$  of 507.4152 and a predicted molecular formula of  $C_{30}H_{54}N_2O_4$  (observed, 506.4079 Da; expected, 506.4084 Da; **Supplementary Fig. 11**). Facms0017 was also present in extracts from *A. terreus* ATCC 20542, as well as AtFAC9J20 (**Supplementary Fig. 12**). Deletion of either *valB* or *valA* eliminated facms0017 completely (**Fig. 3b**). Analysis of tandem mass spectrometry data for facms0017 revealed immonium ions and neutral losses matching Val and either Ile or Leu, consistent with the prediction that the two A domains incorporate branched aliphatic amino acids, based on their Stachelhaus codes<sup>30</sup> (**Supplementary Table 4**, **Supplementary Fig. 11**). The molecular formula also indicated that the protein ValA (the HRPKS) must act iteratively in order to introduce enough ketide units<sup>31,32</sup>. A BLAST analysis of ValA showed that its closest annotated match was the iterative lovastatin diketide synthase from *A. terreus*, LovF, which



**Figure 3 | Analysis of three gene clusters on AtFAC9J20.** (a) A map of the AtFAC9J20 construct (118,762 bp: 102,722 (insert) + 16,040 (vector) bp) and biosynthetic products, showing the genes identified as producing three distinct secondary metabolites, and their corresponding products. Purple, benzomalvin BGC; green, sesterterpenoid BGC; teal, valactamide BGC. (b) Effect of FAC backbone-gene deletants on the ion abundances of biosynthetic products of AtFAC9J20; the key follows the color-coding in a. Deletants confirmed the assignment of BGCs to the three observed products. The abundances shown are averages from an analysis of biological replicates from the point of fungal transformation. Error bars indicate  $\pm$ s.d. ( $n = 4$ ).

uses an acetyl and malonate unit to produce  $\alpha$ -methyl butyrate (31% amino acid identity; E-value = 0.0)<sup>33</sup>.

To determine the structure of facms0017, we purified ~0.2 mg of the compound and analyzed it by <sup>1</sup>H and <sup>13</sup>C NMR spectroscopy, as well as two-dimensional correlation spectroscopy via the COSY, HMBC, and HSQC methods (Online Methods, **Supplementary Figs. 13–18, Supplementary Table 5**). Analysis of the structure of facms0017 revealed a novel scaffold that contained Val and Ile residues, as well as a branched aliphatic chain cyclized at its terminus to form a 20-member macrolactone (**Fig. 4**). Using Marfey's reagent, we confirmed the absolute stereochemical configurations of the Val and Ile residues as the natural L-configuration (**Supplementary Fig. 19**), consistent with the absence of an epimerase domain in the gene cluster. This observed chemical structure led us to replace the platform name "facms0017" with the common name "valactamide A" (5; **Table 1, Fig. 4**).



**Figure 4 | Structure and proposed biosynthesis of valactamide A (facms0017).** The structure of the macrolactone product of the hybrid NRPS-PKS gene cluster found on AtFAC9J20, valactamide A, and its proposed biosynthesis. The atoms proposed to originate from the acetyl starter unit are shown in blue. Methyl groups likely to be incorporated by the ValA methyltransferase (MT) domain are shown in red (methyl substituent stereochemistry not determined). Each backbone ketide unit is shown as a thick black bond. The ValB-derived amino acid residues are shown in pink. The structure of valactamide A was determined by mass spectrometry and NMR spectroscopy; all other structures shown are hypothetical. KS, ketosynthase domain; AT, acyltransferase domain; DH, dehydratase domain; ER, enoylreductase domain; KR, ketoreductase domain.

On the basis of molecular genetic evidence from deletions of *valB* and *valA* and high-resolution tandem mass spectrometry data for precursor and fragmented ions, we assert that AtFAC9J20 produces a family of valactamide metabolites, termed valactamides A–H (**Supplementary Fig. 20**). We were able to distinguish these family members by their incorporation of different  $\alpha$  numbers of ketide units (5–8) and the presence or absence of the  $\alpha,\beta$ -alkene. All shared the macrolactone architecture of valactamide A, with ring sizes ranging from 16 to 22 atoms. Deletion of either *valA* or *valB* eliminated all these metabolites from the metabolome of AtFAC9J20, further supporting the idea that they were produced by the valactamide gene cluster (**Supplementary Fig. 21**). Each metabolite also shared the same diagnostic tandem mass spectrometry fragment ions with valactamide A that correspond to the presence of Val and Ile residues, and several additional fragments in the 50–200  $m/z$  region, which confirmed their structural relationship (**Supplementary Fig. 22**).

The proposed biosynthesis of valactamide A is consistent both with the structure of valactamide A and with the domain architectures of *valA* and *valB* (**Fig. 4**). An acetyl starter unit is incorporated by the ketosynthase domain of the iterative HRPKS enzyme ValA, and subsequently six malonyl-CoA-derived ketide units are incorporated and fully reduced to their respective alkane forms by the action of the ketoreductase, dehydratase, and enoylreductase domains (except for the penultimate unit, which is reduced only

to the alkene). The final five ketide units are each proposed to be  $\alpha$ -methylated by the methyltransferase domain before ketone reduction by the ketoreductase domain. The C<sub>1</sub> domain of ValB then catalyzes amide bond formation between the heptaketide chain and L-Val attached to the T<sub>1</sub> domain (Fig. 4), in agreement with prior biochemical studies on hybrid HRPKS–NRPS systems<sup>34,35</sup>. The C<sub>2</sub> domain incorporating L-Ile then carries out chain elongation, which is followed by macrolactonization by the C<sub>T</sub> domain to release the final product. Product release and ring formation by the C<sub>T</sub> domain align well with reports that terminal C domains typically carry out cyclizations in fungal assembly lines<sup>32,36</sup>.

The relative LC–mass spectrometry abundances for the eight metabolites in the valactamide family showed that valactamide A contributed ~95% of the ions attributable to the cluster. The hexaketide products, valactamides B and C, contributed about 3% and 1%, respectively (Supplementary Table 6). Products with either five or eight ketides (valactamides E–H) seemed to be much less abundant, accounting for between 0.1% and 0.004% of the total valactamide signal. Interestingly, the  $\alpha,\beta$ -saturated form of valactamide A, named valactamide D, accounted for only 1% of the signal from the family. This suggests that reduction of the  $\alpha,\beta$ -alkene is strongly disfavored at this site for the heptaketide intermediate, either because it hinders downstream steps of the biosynthesis (and thus not much of the final macrolactone is produced), or because the enoylreductase domain of ValA is selective against the heptaketide intermediate. It was shown previously that  $\alpha$ - and  $\gamma$ -carbon methylation is important for intermediate transfer between iterative HRPKS and NRPS assembly lines in some systems<sup>32,34</sup>. Our result may suggest that an  $\alpha,\beta$ -alkene can also be recognized by the ‘programming’ of iterative HRPKS enzymes to control intermediate length and release.

## DISCUSSION

Untargeted metabolomics, empowered by the increasing availability of Fourier transform mass spectrometers, allows natural product discovery by accurate mass determination<sup>37,38</sup>. However, a major challenge in domesticating microbial sources of natural products lies in differentiating LC–mass spectrometry signals for secondary metabolites from those for other molecules produced by the host or organism under study. For example, although the fungus *A. nidulans* harbors gene clusters for ~50 secondary metabolite scaffolds and is likely to express only a fraction of these at any given time<sup>39</sup>, untargeted LC–mass spectrometry analysis of its metabolome leads to the detection of ~5,000 compounds. Even after advanced data processing with programs such as CAMERA<sup>40</sup>, which help filter out noise and isotope peaks, ~1,000 compounds typically remain, the majority of which probably do not come from secondary metabolites.

To address this challenge, we used FAC-Scores combined with data from large-scale LC–mass spectrometry screens of 56 heterologously expressed BGCs to simplify the process of discovery from three different species. Importantly, because FACs are fungal–*E. coli* shuttle vectors, quick cycles of genetic deletions and targeted testing enabled the direct validation of metabolite–gene cluster identifications (Table 1, Supplementary Fig. 2). Genetic validation was an important part of this study because it reduced the likelihood that predicted metabolites were artifacts of the host metabolome. Genetic disruption also facilitated our investigation of biosynthesis, as demonstrated for both benzomalvin A and the newly identified macrolactone valactamide A. It is notable that we were able to solve the structure of valactamide A without extensive FAC engineering (such as by the introduction of an inducible promoter). However, for some metabolites discovered by FAC-MS, overexpression may be required to enable structure elucidation.

Intriguingly, of the 15 new metabolites reported in this study, 11 were not detected in their parent fungi under the same growth conditions (Supplementary Fig. 2). This might indicate that the FAC-MS platform can force the expression of clusters that are silent

under similar growth conditions in the parent fungus (Table 1). Two out of seven *A. terreus* metabolites were silent in the parent strain, whereas seven out of seven *A. wentii* metabolites were silent. One possible explanation for this difference might be phylogenetic relatedness between the native hosts and the heterologous host. *A. terreus* may retain more prorepressive regulatory mechanisms, as it is more closely related to the heterologous host than *A. wentii*<sup>20</sup>. Future work could reveal the determinants of successful FAC-based metabolite production, and whether there is a relationship between the rate of silent metabolite activation and the biological origin of FAC-encoded BGCs. We also plan to develop methods to ‘turn on’ FAC-encoded clusters for those metabolites that cannot be detected confidently by the current FAC-MS implementation.

Here we have demonstrated that FAC-MS can be used to systematically express, discover, and characterize novel secondary metabolites from the phylogenetically diverse aspergilli *A. terreus*, *A. aculeatus*, and *A. wentii*. As the FAC-MS platform is applied to increasingly exotic fungi, the high-throughput discovery of new molecules from hundreds and even thousands of fungal species will become feasible. The greatly improved access to new compounds will enable both the systematic mapping of fungal secondary metabolism and the regularized introduction of natural products into pharmaceutical and agricultural screening campaigns.

Received 14 December 2016; accepted 13 March 2017;  
published online 12 June 2017

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

## References

- Blackwell, M. The fungi: 1, 2, 3 . . . 5.1 million species? *Am. J. Bot.* **98**, 426–438 (2011).
- Khalidi, N. *et al.* SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47**, 736–741 (2010).
- Inglis, D.O. *et al.* Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of *Aspergillus nidulans*, *A. fumigatus*, *A. niger* and *A. oryzae*. *BMC Microbiol.* **13**, 91 (2013).
- Han, X., Chakraborti, A., Zhu, J., Liang, Z.X. & Li, J. Sequencing and functional annotation of the whole genome of the filamentous fungus *Aspergillus westerdijkiae*. *BMC Genomics* **17**, 633 (2016).
- Andersen, M.R. *et al.* Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc. Natl. Acad. Sci. USA* **110**, E99–E107 (2013).
- Brown, D.W. & Proctor, R.H. Insights into natural products biosynthesis from analysis of 490 polyketide synthases from *Fusarium*. *Fungal Genet. Biol.* **89**, 37–51 (2016).
- Medema, M.H. *et al.* Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
- Anyagou, D.C. & Mortensen, U.H. Heterologous production of fungal secondary metabolites in *Aspergillus*. *Front. Microbiol.* **6**, 77 (2015).
- Bailey, A.M. *et al.* Characterisation of 3-methylorcinolaldehyde synthase (MOS) in *Acremonium strictum*: first observation of a reductive release mechanism during polyketide biosynthesis. *Chem. Commun. (Camb.)* **2007**, 4053–4055 (2007).
- Holm, D.K. *et al.* Molecular and chemical characterization of the biosynthesis of the 6-MSA-derived meroterpenoid yanuthone D in *Aspergillus niger*. *Chem. Biol.* **21**, 519–529 (2014).
- Richter, L. *et al.* Engineering of *Aspergillus niger* for the production of secondary metabolites. *Fungal Biol. Biotechnol.* <http://dx.doi.org/10.1186/s40694-014-0004-9> (2014).
- Nielsen, M.T. *et al.* Heterologous reconstitution of the intact geodin gene cluster in *Aspergillus nidulans* through a simple and versatile PCR based approach. *PLoS One* **8**, e72871 (2013).
- Heneghan, M.N. *et al.* First heterologous reconstruction of a complete functional fungal biosynthetic multigene cluster. *ChemBioChem* **11**, 1508–1512 (2010).
- Chiang, Y.M. *et al.* An efficient system for heterologous expression of secondary metabolite genes in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **135**, 7720–7731 (2013).
- Smith, D.J., Burnham, M.K., Edwards, J., Earl, A.J. & Turner, G. Cloning and heterologous expression of the penicillin biosynthetic gene cluster from *Penicillium chrysogenum*. *Bio/Technology* **8**, 39–41 (1990).

16. Gressler, M., Hortschansky, P., Geib, E. & Brock, M. A new high-performance heterologous fungal expression system based on regulatory elements from the *Aspergillus terreus* terrein gene cluster. *Front. Microbiol.* **6**, 184 (2015).
17. Lazarus, C.M., Williams, K. & Bailey, A.M. Reconstructing fungal natural product biosynthetic pathways. *Nat. Prod. Rep.* **31**, 1339–1347 (2014).
18. Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, W237–W243 (2015).
19. Bok, J.W. *et al.* Fungal artificial chromosomes for mining of the fungal secondary metabolome. *BMC Genomics* **16**, 343 (2015).
20. Samson, R.A. *et al.* Phylogeny, identification and nomenclature of the genus *Aspergillus*. *Stud. Mycol.* **78**, 141–173 (2014).
21. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787 (2006).
22. Ridder, L. *et al.* Automatic chemical structure annotation of an LC-MS(n) based metabolic profile from green tea. *Anal. Chem.* **85**, 6033–6040 (2013).
23. Sun, H.H., Barrow, C.J. & Cooper, R. Benzomalvin D, a new 1,4-benzodiazepine atropisomer. *J. Nat. Prod.* **58**, 1575–1580 (1995).
24. Jang, J.P. *et al.* Benzomalvin E, an indoleamine 2,3-dioxygenase inhibitor isolated from *Penicillium* sp. FN070315. *J. Antibiot. (Tokyo)* **65**, 215–217 (2012).
25. Sun, H.H., Barrow, C.J., Sedlock, D.M., Gillum, A.M. & Cooper, R. Benzomalvins, new substance P inhibitors from a *Penicillium* sp. *J. Antibiot. (Tokyo)* **47**, 515–522 (1994).
26. Wei, H. *et al.* Cytotoxic sesterterpenes, 6-epi-ophiobolin G and 6-epi-ophiobolin N, from marine derived fungus *Emericella varicolor* GF10. *Tetrahedron* **60**, 6015–6019 (2004).
27. Yoganathan, K. *et al.* Inhibition of the human chemokine receptor CCR5 by varicocolin and varicocol and isolation of four new varicocolin analogues, emericolins A–D, from *Emericella aurantiobrunnea*. *J. Nat. Prod.* **67**, 1681–1684 (2004).
28. Liu, H.-B. *et al.* Ophiobolin sesterterpenoids and pyrrolidine alkaloids from the sponge-derived fungus *Aspergillus ustus*. *Helv. Chim. Acta* **94**, 623–631 (2011).
29. Molander, G.A., Quirnbach, M.S., Silva, L.F. Jr., Spencer, K.C. & Balsells, J. Toward the total synthesis of varicocolin. *Org. Lett.* **3**, 2257–2260 (2001).
30. Stachelhaus, T., Mootz, H.D. & Marahiel, M.A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).
31. Chiang, Y.M. *et al.* Molecular genetic mining of the *Aspergillus* secondary metabolome: discovery of the emericellamide biosynthetic pathway. *Chem. Biol.* **15**, 527–532 (2008).
32. Cacho, R.A. *et al.* Understanding programming of fungal iterative polyketide synthases: the biochemical basis for regioselectivity by the methyltransferase domain in the lovastatin megasynthase. *J. Am. Chem. Soc.* **137**, 15688–15691 (2015).
33. Meehan, M.J. *et al.* FT-ICR-MS characterization of intermediates in the biosynthesis of the  $\alpha$ -methylbutyrate side chain of lovastatin by the 277 kDa polyketide synthase LovF. *Biochemistry* **50**, 287–299 (2011).
34. Zou, Y. *et al.* Methylation-dependent acyl transfer between polyketide synthase and nonribosomal peptide synthetase modules in fungal natural product biosynthesis. *Org. Lett.* **16**, 6390–6393 (2014).
35. Gatto, G.J. Jr., McLoughlin, S.M., Kelleher, N.L. & Walsh, C.T. Elucidating the substrate specificity and condensation domain activity of FkbP, the FK520 pipecolate-incorporating enzyme. *Biochemistry* **44**, 5993–6002 (2005).
36. Gao, X. *et al.* Cyclization of fungal nonribosomal peptides by a terminal condensation-like domain. *Nat. Chem. Biol.* **8**, 823–830 (2012).
37. Henke, M.T. & Kelleher, N.L. Modern mass spectrometry for synthetic biology and structure-based discovery of natural products. *Nat. Prod. Rep.* **33**, 942–950 (2016).
38. Bouslimani, A., Sanchez, L.M., Garg, N. & Dorrestein, P.C. Mass spectrometry of natural products: current, emerging and future technologies. *Nat. Prod. Rep.* **31**, 718–729 (2014).
39. Albright, J.C. *et al.* Large-scale metabolomics reveals a complex response of *Aspergillus nidulans* to epigenetic perturbation. *ACS Chem. Biol.* **10**, 1535–1541 (2015).
40. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T.R. & Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **84**, 283–289 (2012).

## Acknowledgments

This work was supported in part by the US National Institutes of Health (National Institute of Allergy and Infectious Diseases SBIR award in the form of grant R44AI094885 to C.C.W., J.W.B., and N.L.K.; grant R01AT009143 to N.L.K.; grant R01-AI065728 to N.P.K.; grant 5T32GM105538-04 to G.P.M.). NMR instrumentation and assistance was provided by the Integrated Molecular Structure Education and Research Center (IMSERC) at Northwestern University.

## Author contributions

R.Y., C.C.W., M.L., and M.N.I. worked on the FAC library assembly, FAC end sequencing, FAC DNA preparation, and FAC engineering. C.C.W., J.M.P., C.C., and M.L. carried out fungal secondary metabolite gene cluster prediction, FAC bioinformatic analyses, and FAC next-generation sequencing assembly and annotation. J.W.B., T.V., and K.H.Y. performed the *A. nidulans* transformation with FACs and prepared samples for metabolite identification and structure determination. K.D.C. and P.G. carried out LC–mass spectrometry analyses. K.D.C. conducted analysis of LC–mass spectrometry data, including development of the analysis pipeline, invention of the FAC-Score, discovery of valactamide A, and bioinformatic analyses of the benzomalvin and valactamide gene clusters, under the supervision of P.M.T. and N.L.K. P.M.T. identified benzomalvin A. G.P.M. carried out total synthesis of benzomalvin A/D. M.H.V., G.P.M., and K.D.C. carried out purification and structural characterization of valactamide A. The paper was written by K.D.C. under the supervision of N.L.K. The Online Methods and supplementary material were prepared by K.D.C., N.P.K., C.C.W., J.W.B., R.Y., G.P.M., M.H.V., and M.T.R. All authors read and approved the final draft of the manuscript. C.C.W., N.P.K., and N.L.K. conceived of and supervised the project.

## Competing financial interests

The authors declare competing financial interests: details are available in the [online version of the paper](#).

## Additional information

Any supplementary information, chemical compound information, and source data are available in the [online version of the paper](#). Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to C.C.W., N.P.K., or N.L.K.

## ONLINE METHODS

**Generation of FACs from aspergilli.** We generated unbiased random-shear FAC libraries of fungi from high-molecular-weight genomic DNA of the *Aspergillus* species *A. aculeatus* and *A. wentii*. BGC-containing FACs were identified by FAC end sequences aligned with their reference genomes and confirmed through PCR. FAC BCGs from the *A. aculeatus*, *A. terreus*, and *A. wentii* genomes were predicted by the software program Secondary Metabolite Unique Regions Finder (SMURF; [http://jcv.org/smurf/run\\_smurf.php](http://jcv.org/smurf/run_smurf.php)). We were able to capture the entire sets of 48 BGCs of *A. aculeatus* and 47 BGCs from *A. wentii* each on individual FAC clones, and we added these to our previous library of 61 BGC-containing FACs from *A. terreus* (for a total library size of 156 FACs). Of the 61 *A. terreus* FACs, 56 were previously reported by our team, and 5 were drawn from the same FAC clone library as previously reported but were added to the original 56 for the current study only after the identification of new putative BGCs. Alignment with the appropriate reference genomes led to the confirmation of 156 FACs, each encoding at least one predicted BGC of *A. wentii*, *A. terreus*, or *A. aculeatus*. The backbone of the FAC vector allows self-propagation in both *A. nidulans* and *E. coli* through inclusion of the autonomous fungal replicating element AMA1 (ref. 19), which makes amplification, manipulation, and heterologous expression of secondary metabolite (SM) gene clusters facile at all stages (Fig. 1). After excluding previously characterized BGCs by multigene sequence alignment, we randomly selected 56 FACs for further study (drawn from all three species; **Supplementary Table 7**). We prepared high-quality DNA from 200 mL each of FAC LB cultures with a FAC/BAC DNA preparation kit (Intact Genomics), and dissolved FAC DNA in 200  $\mu$ L of 10 mM Tris-HCl, pH 8.0, for subsequent transformation and secondary metabolite screening.

**Transformation of FACs into the host strain.** To improve the transformation yield, we modified a published method<sup>19</sup>. We mixed 2  $\mu$ g of FAC DNA with 10 mM Tris-HCl in a 1.5-mL microcentrifuge tube to a final volume of 50  $\mu$ L and incubated the mixture on ice for 1 h to improve FAC DNA solubility. Then we mixed that volume with 50  $\mu$ L of sorbitol-Tris-HCl-CaCl<sub>2</sub> (STC buffer; 1.2 M sorbitol, 10 mM Tris-HCl, 50 mM CaCl<sub>2</sub>, pH 7.5) just before transformation. We gently mixed 100  $\mu$ L of protoplasts (10<sup>7</sup> protoplasts/mL) with the preincubated 100- $\mu$ L sample containing 2  $\mu$ g of FAC DNA. The tube with the protoplast-DNA mixture was incubated for 30 min on ice, 200  $\mu$ L of 30% PEG-4,000 with 50 mM CaCl<sub>2</sub> was laid under the sample, and the mixture was centrifuged for 5 min at 250g. The sample was gently mixed with an autopipette with a wide tip before gentle tapping. This mixture was incubated for an additional 15 min at room temperature, and then 1 mL of STC buffer was added. After transferring the mixture into a 13-mL tube, we added an additional 5 mL of STC buffer and gently mixed the solution. We distributed 1 mL of this final solution onto regeneration media (glucose minimal medium (GMM) with 1.2 M sorbitol and pyridoxine (1 mL of a 0.1% stock solution) as a supplement) and incubated it for 3 d at 37 °C in an incubator to obtain transformants.

**Fungal culture and extraction of secondary metabolites.** Each *A. nidulans* FAC transformant was inoculated on four solid GMM plates supplemented with pyridoxine and incubated for 6 d at 37 °C according to standard methods<sup>41</sup>. Also, each parental wild-type strain, including *A. terreus* NIH 2624, *A. terreus* ATCC 20542, *A. wentii*, and *A. aculeatus*, was inoculated on four GMM plates and incubated for 10 d at 30 °C. Subsequently, the entire contents of the plates were collected and lyophilized for 48 h. Samples were then pulverized with a mortar and pestle, and methanol was added. Air-dried methanol extracts were prepared with a SpeedVac system (Savant SpeedVac Concentrator, SC250EXP-115) and then further extracted with organic solvent (chloroform:methanol:ethylacetate, 8:1:1). Organic extracts were evaporated to dryness in a SpeedVac and stored at -20 °C until analysis.

**Untargeted metabolomic screening of FAC extracts.** Dried SM extracts from four complete biological replicates of each *A. nidulans* FAC strain, or extracts from *A. terreus* NIH 2624, *A. terreus* ATCC 20542, *A. wentii* and *A. aculeatus*, were resuspended to a concentration of 2 mg/mL by the addition of 50% acetonitrile, and then subjected to bath sonication for 10 min. Insoluble material was removed by centrifugation at 21,000g for 10 min at room temperature.

The supernatant of each reconstituted extract was then transferred to an LC-mass spectrometry autosampler vial and stored at 4 °C until analysis on the same day. Four biological replicates were used according to standard practices for untargeted metabolomic analyses of lab-cultured microbial samples. After analysis, samples were stored at -80 °C, and it was our experience that samples could be stored this way indefinitely.

Samples were analyzed by high-resolution high-performance liquid chromatography (HPLC)-tandem mass spectrometry with a Thermo Q-Exactive in line with an electrospray source and an Agilent 1200 series HPLC stack consisting of a binary pump, degasser, and autosampler, outfitted with a Phenomenex Luna C-18 column with dimensions of 2 mm  $\times$  150 mm, 3  $\mu$ m *d<sub>p</sub>*. A binary linear gradient of water and acetonitrile balanced with 0.1% formic acid was used (buffer A, H<sub>2</sub>O; buffer B, acetonitrile), and 50  $\mu$ g of extract (25  $\mu$ L) was injected. The gradient was as follows: at 0 min, 2% B; 35 min, 70% B; and 54 min, 98% B, with a flow rate of 200  $\mu$ L/min. A 1:4 split was used to transfer sample from the column to the electrospray source, so the flow rate into the electrospray ionization source was 40  $\mu$ L/min. The capillary of the ESI source was set to 275 °C, with sheath gas set to 5 arbitrary units and spray voltage set to 3.5 kV. Precursor ion data were collected at 70,000 resolution from 150 to 2,000 *m/z*. Product ion fragmentation scans for the five most abundant ions in each scan from the first stage of the tandem mass spectrometry analysis were collected at a resolution of 35,000, with a 60-s dynamic exclusion list. Fragmentation was achieved with the higher-energy collisional dissociation cell set to a normalized collisional energy value of 30. To minimize the effect of instrumental drift over time, we ran samples in a randomized order. All FACs from a given species were analyzed together (biological quadruplicates) in a single sequence, along with a set of biological quadruplicates from the *A. nidulans* host harboring an empty FAC vector with no insert. Extracts from *A. terreus* NIH 2624, *A. terreus* ATCC 20542, *A. wentii*, and *A. aculeatus* were run in the same manner, but separate from FAC strains to prevent carryover. Untargeted LC-mass spectrometry data were uploaded to the GNPS database and assigned the MassIVE ID **MSV000081097** (ref. 42).

**Feature detection of untargeted metabolomic data.** Chromatographic and *m/z* features were extracted and grouped with the open source metabolomics software XCMS, running in the R environment<sup>21,40</sup>. Importantly, however, any untargeted metabolomic peak-picking software could be used that allows output to a Microsoft Excel spreadsheet, such as the user-friendly cloud-based tool XCMS Online (<https://xcmsonline.scripps.edu>). Thermo .RAW files were converted to mzML format with the *mconvert* algorithm from Proteowizard<sup>43</sup>. FACs from the same species, run in the same batch, were analyzed together. First, all mzML files for a species and set of runs (i.e., *A. terreus*, *A. wentii*, or *A. aculeatus*) were placed in a parent directory along with the empty FAC controls with which they were run (in the same block of samples). This ranged from 48 to 104 files depending on the species. Then a subdirectory was created for each treatment (i.e., sets of biological quadruplicates corresponding to a specific FAC or empty FAC negative control). XCMS was run in R with the centwave algorithm for feature detection with the following parameters: *p.p.m.* = 3, *peakwidth* = *c*(20,100), *snthresh* = 10, *prefilter* = *c*(5,10000), *mzCenterFun* = "wMean", *integrate* = 1, *mzdiff* = 0.001, *fitgauss* = FALSE, *noise* = 1000. Grouping was then carried out with the *group* command and the parameters *bw* = 30, *minfrac* = 0.5, *mzwid* = 0.01. After grouping, retention-time correction was carried out with the algorithm *retcor* with the following parameters: *family* = "s", *plottype* = "m". Typical retention-time deviations ranged from 0 to 100 s, with the vast majority of features having deviations of less than 10 s. After retention-time correction, features were regrouped and then subjected to peak filling with the algorithm *pkfill*. Results from XCMS were output into a .csv file. Generally, about 5,000 features were detected for each FAC strain, and the abundance of each feature in each extract and treatment was recorded. The commands used for XCMS and CAMERA are given in **Supplementary Figure 23**. We found that the use of CAMERA could be omitted without any substantial effect on the downstream FAC-Score analysis (described below).

**Generation of FAC-Scores to filter untargeted metabolomic data.** We generated a FAC-Score for each feature in each FAC to allow the ranking of hits within each FAC and the identification of FAC strains with high expression

of unique (relative to all other FACs and the negative control for a given species) heterologous products. Typical untargeted metabolomic analyses search for statistically significant differences in relative feature abundances between treatment groups; however, we were interested in features that were qualitatively present in a single FAC strain at a meaningfully high abundance, but absent from all others—that is, products that were expressed by only one FAC strain and thus were likely to be heterologous expression products. Therefore, in order to rapidly triage data and highlight heterologously expressed FAC metabolites, we developed a scoring system to rank detected compounds on the basis of their average abundance across biological replicates of a particular FAC-transformed strain and their uniqueness to each FAC-transformed strain. Scoring was carried out in Microsoft Excel using equations (1) and (2) and the data generated by XCMS. As stated above, any ‘peak-picking’ program that outputs data into a spreadsheet can be used for FAC-scoring.

In the FAC-scoring method, uniqueness was represented by a ratio, namely, the average abundance of an ion across all replicates of a FAC strain divided by the average abundance of that same ion from every other FAC strain (equation (1)). Theoretically, a feature present in one FAC strain but absent in all others will have a ratio of infinity. In reality, because of inevitable (trace) carryover between samples, as well as the tendency of untargeted metabolomic peak-picking algorithms to occasionally assign nonzero values to noise within a retention-time and  $m/z$  window, features that are unique to a specific FAC strain are often assigned high ratio values ranging from  $5 \times 10^1$  to  $1 \times 10^9$  (rather than infinity). In cases where this did not occur and features were indeed assigned an abundance of zero, leading to division by zero and a ratio of infinity, the ratio was assigned an arbitrary value of  $9.99 \times 10^4$ , which is roughly the logarithmic median of the dynamic range of observed ion abundances within the entire study, to facilitate downstream analysis.

$$\text{Ratio of feature } x \text{ in FAC } y = \frac{\text{Average abundance of feature } x \text{ in FAC } y}{\text{Average abundance of feature } x \text{ in all other FACs}} \quad (1)$$

To allow the comparison of compounds from different FACs and highlight likely heterologous expression products, we combined the ratio and raw relative abundance of each feature in each FAC to create a FAC-Score (equation (2)). The  $\log_{10}$  of abundance and ratio values were used to account for the large dynamic range of signals for each parameter (which spanned several orders of magnitude). Then the  $\log_{10}$  abundance values were standardized to a normal distribution, which allowed comparison between FAC strains that globally expressed SMs at very different levels and/or were measured on different days with differing instrument performance. The standardized  $\log_{10}$  abundance was then multiplied by the  $\log_{10}$  of the ratio for each feature, after features with negative scores for either  $\log_{10}$  ratios or standardized  $\log_{10}$  abundances had been filtered out (as multiplying two negative values could conceivably lead to a high FAC-Score, i.e., a false positive). Thus, the final equation used was as follows:

$$\text{FAC-Score} = (\text{Standardized log(abundance)} \times \log(\text{ratio})) - \text{MaxScore}_{\text{negative control}} \quad (2)$$

The result of these calculations was a score for each feature. The highest scores for compounds from empty, negative control FAC strains ranged from 3.2 to 3.8, with an average of 3.4. The average maximum negative control score of 3.4 was subtracted from the score of each compound for each FAC to give the final FAC-Score, shown in **Supplementary Figure 1**. Thus, FACs with compounds with positive scores (29/56) were deemed likely to express heterologous BGC products. Manual analysis of these hits showed that in 16 cases the top feature was clearly unique to the relevant FAC and was not a known metabolite of the heterologous host strain (*A. nidulans*), based on dereplication using the Dictionary of Natural Products<sup>44</sup> (DNP) and the Antibase<sup>45</sup> database of natural products. In 13 cases the top-scoring compound was not determined to be a unique FAC product; in one of these (AwFAC3B4), the top-scoring feature was identified as an *A. nidulans* SM. In the other 12 cases, the features of interest were detected at trace levels in extracts from other FAC strains in manual analysis of LC–mass spectrometry data. It is likely that some of this resulted from carryover between LC–mass spectrometry runs or cross-contamination of samples, which suggests that these cases could also be hits.

However, we chose a conservative process for hit designation in this study and did not consider these FAC strains for further consideration.

**Gene cluster editing via FAC and production of FAC deletants.** Red/ET tools have been developed for efficient large DNA or BAC-based recombineering and transgenic animal models<sup>46,47</sup>. Here we modified the methodology described for those models for FAC recombineering to create a one-step targeting protocol using kanamycin antibiotic selection with the neomycin-resistance gene from the plasmid PL451 (<https://ncifrederick.cancer.gov/research/brb/productDataSheets/recombineering/plasmid.aspx>). For gene deletion, we created primers that contained DNA sequences allowing fusion of the kanamycin-resistance gene (5′ and 3′ ends) to the homologous DNA sequences of the flanking regions of the FAC locus to be deleted. Each primer contained a 50-bp homologous sequence flanking either side of the deletion, followed by 21 bp of forward or reverse primer for the kanamycin-resistance gene, to give a total length of 71 bp (**Supplementary Table 8**).

All PCR products therefore resulted in DNA fragments that contained a full kanamycin-resistance gene flanked by 50 bp of fungal DNA on both sides to allow for replacement of the FAC locus of interest with the kanamycin-resistance gene. To engineer a FAC, we first transformed FAC DNA into the Red/ET-inducible *E. coli* strain SW102. A gel-purified kanamycin-resistance gene construct was transformed into the FAC-containing *E. coli* strain and plated on YT medium amended with kanamycin. Kanamycin-resistant *E. coli* colonies were grown, and the FACs were extracted from each colony. Replacement of the FAC DNA of interest with the kanamycin-resistance gene was assessed by PCR, restriction digestion, pulse field gel electrophoresis, and sequencing (if needed). The correct recombineered FAC DNA was then extracted from *E. coli* and prepared with a FAC/BAC DNA preparation kit (Intact Genomics) as described above for transformation into *A. nidulans*.

**Validation of hits with FAC engineering.** Strains with targeted genetic deletions of the predicted backbone gene in each FAC identified as a positive ‘hit’ were prepared, transformed, and screened by LC–mass spectrometry (see ‘Gene cluster editing via FAC and production of FAC deletants,’ above) to confirm the FAC-Score analysis (**Table 1**). Of the 16 FAC deletants tested, in 15 the previously observed putative FAC-associated ion was eliminated by the deletion without substantial additional perturbations to the observed metabolome. In 1 case out of 16, the predicted product was still observed despite the deletion. In summary, 15 FAC–SM associations were validated by backbone-gene deletions, and 1 was refuted. Thus, we report 15 hits in this study (**Supplementary Fig. 1, Table 1**).

**Analysis of AtFAC9J20 biosynthetic products and intermediates.** To analyze the levels of AtFAC9J20 products and intermediates in AtFAC9J20 deletants (including products and intermediates of benzomalvin, valactamide, and ses-terterpenoid biosynthesis), we wrote a program in-house to determine the integrated total ion current (TIC) for each compound, as well as the integrated peak area for each compound of interest based on a 4-min retention-time window and a 5-p.p.m.  $m/z$  tolerance. We used the resulting values to calculate the fraction of TIC accounted for by each feature ( $\text{frac}_{\text{TIC}}$ ) in a given .RAW file. Values from biological quadruplicates were averaged, and error bars were generated as the s.d. of the four values. We determined AtFAC9J20 normalized abundances for features of interest by dividing the  $\text{frac}_{\text{TIC}}$  of a feature from a given deletant by the  $\text{frac}_{\text{TIC}}$  for AtFAC9J20. Error bars for AtFAC9J20 normalized values were generated through propagation of error, using the s.d. for each value and the appropriate equation for propagation of error through division.

**Identification of benzomalvin A/D.** Because it expresses an unusually high number of abundant and unique metabolites, we selected AtFAC9J20 for intensive follow-up in this study. Manual analysis showed that all of the top-scoring features were adducts of the ion with  $m/z$  382.1547 and  $R_f$  values of 38 and 41 min. This  $m/z$  did not match any known fungal metabolites in either DNP or Antibase. To broaden the dereplication search, we used the online tool MAGMA to search simulated tandem mass spectrometry data for every molecule with the same mass in PubChem against actual tandem mass spectrometry data obtained for AtFAC9J20 (ref. 22). The top three hits were synthetic

compounds, but the fourth was benzomalvin A/D. Fragment ions were then assigned with the program Mass Frontier 7.0 (Thermo, San Jose, CA, USA) and a starting hypothesis of benzomalvin A/D as the parent compound, which resulted in assignment of all major fragments (Supplementary Fig. 5).

**Sequence analysis of AtFAC9J20 (Chr. 4: 2,151,734–telomeric end).** A bar-coded Illumina next-generation sequencing library of AtFAC9J20 was made with the true-seq kit and pooled for sequencing with other samples at the University of Wisconsin–Madison Biotechnology Center DNA Sequencing Facility. We achieved 1,333× coverage of the AtFAC9J20 DNA sequence. A single contig (102.722 kb, completed and finished FAC clone sequence) was obtained by the DNASTar next-generation sequencing assembling program. The entire FAC sequence of 102.722 kb was confirmed by FAC end sequences, *PmeI*, *NotI* digestion, and contour-clamped homogeneous-electric-field gel electrophoresis. From this sequencing, we identified 35,769 kb of extra sequence toward the telomere end that were not present in the reference genome. We also discovered 1,164 single-nucleotide polymorphisms, one inversion, four duplications, 59 insertions, and 77 deletions, including an 8,773-kb insertion and a 1,851-kb deletion in AtFAC9J20 compared with the NIH 2624 strain. In total there were 1,305 genomic variations in the 66,953-kb genomic region compared with the *A. terreus* NIH 2624 reference genome sequence. Interestingly, the majority of the genomic variations were within the first 25 kb of the genomic region. Sixteen predicted open reading frames (ORFs), eight of which are not present in the reference genome data of *A. terreus* NIH 2624 (Supplementary Table 2), were detected in total. The eight unique ORFs included two predicted NRPS enzymes, *benY* and *benZ*, as well as the predicted methyltransferase *benX* (Supplementary Table 2, Supplementary Fig. 3). The final sequence revealed that the benzomalvin, sesterterpenoid, and valactamide BGCs were captured on this single FAC clone (Supplementary Table 2). The DNA sequence of AtFAC9J20 was deposited at GenBank with the accession code [KX449366](https://www.ncbi.nlm.nih.gov/nuclseq/KX449366).

The extracted 10-amino-acid A domain signatures for both the A domain of *benY* and the A<sub>1</sub> domain of *benZ* matched the distinct signature expected for anthranilate-encoding A domains<sup>30,48</sup>. The A<sub>2</sub> domain of *benZ* is predicted to encode a Phe, NmPhe, or Tyr residue (Supplementary Table 4). In addition to *benY* and *benZ*, we discovered the gene *benX*, which encodes a predicted SAM-binding domain and is the only gene likely to encode a methyltransferase enzyme in AtFAC9J20. Additionally, ORFs encoding a predicted PKS, a third NRPS enzyme, and an isoprenoid synthase enzyme were found on the FAC and annotated (Supplementary Table 2, Supplementary Fig. 6).

**Total synthesis and LC–tandem mass spectrometry analysis of benzomalvin A/D.** We synthesized benzomalvin A/D as previously described to confirm the identity of the AtFAC9J20 ion with *m/z* 382.1547 as follows. We synthesized N-desmethylbenzomalvin A/D essentially as previously reported<sup>49</sup>, but replacing Pd/C, H<sub>2</sub> nitro reduction with Zn/AcOH, at room temperature for 12 h. N-desmethylbenzomalvin A/D was methylated with lithium bis(trimethylsilyl) amide (1.1 eq.) and methyl iodide (1.2 eq.) in tetrahydrofuran to produce crude benzomalvin A/D. An analytical sample for LC–tandem mass spectrometry was prepared by preparative HPLC with a Higgins Analytical 250 × 10 mm TARGA C18 5-μm column eluted with buffers A (water, 0.1% TFA) and B (acetonitrile, 0.1% TFA) at 7.5 ml/min. The gradient was ramped from 95% A to 95% B over 12 min and then re-equilibrated. <sup>1</sup>H NMR spectroscopy (500 MHz, acetone-*d*<sub>6</sub>): δ 8.23 (dd, *J* = 7.9, 1.5 Hz, 1H), 7.90 (ddd, *J* = 8.5, 7.0, 1.5 Hz, 1H), 7.85 (dd, *J* = 8.1, 1.3 Hz, 1H), 7.83 (dd, *J* = 7.7, 1.7 Hz, 1H), 7.70–7.67 (m, 1H), 7.67–7.62 (m, 1H), 7.59 (dddd, *J* = 7.2, 5.5, 4.2, 1.4 Hz, 2H), 7.37 (d, *J* = 6.8 Hz, 1H), 7.32–7.20 (m, 3H), 7.19–7.13 (m, 1H), 5.08 (t, *J* = 7.4 Hz, 1H), 3.77 (dd, *J* = 14.6, 7.4 Hz, 1H), 3.55 (dd, *J* = 14.6, 7.5 Hz, 1H), 3.03 (s, 3H). <sup>13</sup>C NMR spectroscopy (126 MHz, acetone-*d*<sub>6</sub>): δ 161.85, 153.72, 138.42, 135.60, 134.21, 132.97, 131.30, 130.26, 130.03, 129.93, 129.71, 129.46, 129.35, 129.04, 128.53, 128.24, 127.86, 127.47, 122.82, 59.29, 33.74, 27.91.

Synthetic benzomalvin A/D was then analyzed by LC–tandem mass spectrometry as described above for FAC extracts, and the selected ion chromatogram for benzomalvin A/D was compared to that for *m/z* 382.1547 in extracts from *A. terreus* and AtFAC9J20, which confirmed the identification of the unknown compound as benzomalvin A/D. To confirm that the compound

produced by AtFAC9J20 was identical to the synthetic standard, we also mixed the two together and coinjected them (Supplementary Fig. 3d).

**Purification and structural analysis of valactamide A.** One hundred plates of AtFAC9J20-Δ*benY* were grown and extracted as described above to yield roughly 130 mg of dry extract. This was re-extracted five times with 10 mL of hexanes, evaporated *in vacuo*, and resuspended in 2 mL of methylene chloride. The resuspended extract was loaded on a 12-g spherical silica gel column and eluted with a linear gradient of 100:0 to 60:40 dichloromethane:ethylacetate (DCM:EA) over 30 min. Fractions were collected and screened by LC–mass spectrometry. Fractions containing the target compound were pooled and evaporated *in vacuo* to yield 800 μg of dry material, which was further refined by preparative thin-layer chromatography eluting with 80:20 DCM:EA, yielding 200 μg of purified compound. For NMR spectroscopy, the dried extract was resuspended in 300 μL of anhydrous CDCl<sub>3</sub> and loaded into a Shigemi tube.

The extract was analyzed by NMR spectroscopy, including by <sup>1</sup>H, <sup>13</sup>C, COSY, HSQC, and HMBC methods. <sup>1</sup>H, HSQC, HMBC, and COSY NMR spectra were acquired with an Agilent DD2 600-MHz spectrometer, and the <sup>13</sup>C spectrum was acquired in a Bruker Avance III 500-MHz spectrometer equipped with a DCH CryoProbe.

<sup>1</sup>H NMR (600 MHz, CDCl<sub>3</sub>): δ 0.80 (d, *J* = 6.3 Hz, 3H), 0.81 (d, *J* = 6.5 Hz, 3H), 0.82 (m, 1H), 0.86 (d, *J* = 6.6 Hz, 3H), 0.87 (d, *J* = 6.9 Hz, 3H), 0.93 (d, *J* = 6.8 Hz, 3H), 0.93 (dd, *J* = 7.4, 7.3 Hz, 3H), 0.94 (m, 1H), 0.96 (d, *J* = 6.7 Hz, 3H), 0.99 (m, 1H), 1.00 (ddd, *J* = 13.4, 7.6, 5.1 Hz, 1H), 1.02 (d, *J* = 6.9 Hz, 3H), 1.12 (ddd, *J* = 13.0, 10.6, 3.4 Hz, 1H), 1.20 (m, 1H), 1.23 (d, *J* = 6.4 Hz, 3H), 1.26 (ddd, *J* = 13.6, 9.4, 3.3 Hz, 1H), 1.28 (m, 1H), 1.37 (ddd, *J* = 13.0, 10.9, 2.6 Hz, 1H), 1.43 (m, 1H), 1.44 (m, 1H), 1.48 (m, 1H), 1.49 (m, 1H), 1.51 (m, 1H), 1.62 (m, 1H), 1.90 (m, 1H), 1.93 (d, *J* = 0.9 Hz, 3H), 2.33 (ddd, *J* = 6.9, 6.8, 5.0 Hz, 1H), 2.57 (dddd, *J* = 10.6, 10.2, 6.7, 2.6 Hz, 1H), 4.28 (dd, *J* = 6.5, 6.2 Hz, 1H), 4.33 (dd, *J* = 7.6, 5.0 Hz, 1H), 4.93 (ddd, *J* = 7.0, 6.4, 4.4 Hz, 1H), 6.18 (d, *J* = 7.5 Hz, 1H), 6.34 (dq, *J* = 10.2, 0.9 Hz, 1H), 6.56 (d, *J* = 6.2 Hz, 1H).

<sup>13</sup>C NMR (126 MHz, CDCl<sub>3</sub>): δ 11.5, 12.6, 15.1, 17.5, 19.5, 19.6, 20.3, 20.8, 20.9, 21.2, 25.9, 28.5, 29.0, 29.9, 30.2, 30.5, 30.8, 32.6, 37.4, 44.9, 46.1, 48.0, 57.3, 58.8, 72.5, 128.2, 144.6, 169.4, 171.0, 171.2.

The <sup>1</sup>H NMR spectrum easily distinguished ten methyl groups (δ 0.80–1.94), two CH protons (δ 2.33 and 2.57), one O-CH proton (δ 4.93), two NH protons (δ 6.18 and 6.56) and one olefinic proton (δ 6.34). The aliphatic region of the spectrum was complex and required elucidation via 2D NMR techniques.

The <sup>13</sup>C NMR spectrum has 30 signals, which is in accordance with the molecular formula. Three of the signals are carbonyls (δ 171.2–169.4); two are olefinic carbons, one CH at δ 144.6 and the other quaternary at δ 128.2. We observed one O-CH carbon (δ 72.5) that showed single-range correlation with the proton at δ 4.93, as well as two N-CH carbons at δ 57.3 (isoleucine) and δ 58.8 (valine).

The <sup>1</sup>H–<sup>13</sup>C HSQC spectrum successfully deconvoluted the complex aliphatic region in the 1D <sup>1</sup>H, matching 26 <sup>1</sup>H–<sup>13</sup>C pairs, including the identification of six carbons bearing diastereomeric CH<sub>2</sub> protons. <sup>1</sup>H–<sup>13</sup>C HMBC and COSY 2D experiments were able to characterize the structure. Beginning at C1, a clear path can be traced through the isoleucine, across the amide bond to valine and into the branched aliphatic chain.

With H-2 (δ 4.28) as the entry point, a COSY sequential coupling path can be traced through H-3 (δ 1.90) and on to H-4 (δ 0.87) and H-5 (δ 1.20). From H-5, a coupling to H-6 is apparent (δ 0.93). H-2 (δ 4.28) also experiences coupling from HN-Ile (δ 6.56), as shown by COSY. HMBC confirmed the structural configuration provided by COSY and added a connection to carbonyl C-1 (δ 171.2) and, through HN-Ile (δ 6.56), a connection to a second carbonyl C-7 (δ 171.0). This evidence, in combination with <sup>1</sup>H integrations and HSQC polarity, unambiguously identified this region as an isoleucine residue.

Working from α-carbon C-8 (δ 58.8)/H-8 (δ 4.33), a COSY spin system is seen beginning at H-9 (δ 2.33) and proceeding to methyls C-10 (δ 0.93) and C-11 (δ 1.02). H-9 (δ 2.33) is also coupled to HN-Val (δ 6.18). HMBC confirmed this arrangement and additionally showed coupling between H-9 (δ 2.33) and C-7 (δ 171.0), connecting this spin system with the isoleucine fragment identified previously. This evidence, in combination with <sup>1</sup>H integrations and HSQC polarity, unambiguously identified this region as a valine residue.

From HN-Val ( $\delta$  6.18), an HMBC coupling is apparent to carbonyl C-12 ( $\delta$  169.4). Tracing the HMBC path from this point yields protons and carbons with characteristic olefin shifts, C-13 ( $\delta$  128.2) and H/C-15 ( $\delta$  6.34, 144.6). The presence and relative location of the double bond between C-13 ( $\delta$  128.2) and C-15 ( $\delta$  144.6) is confirmed by the chemical shifts; as it is close to a carbonyl group (C-12), C-15 should be more deshielded owing to electron  $\pi$  delocalization. The remainder of the branched aliphatic chain was well characterized by HMBC and COSY correlations. C-29 ( $\delta$  72.5) was chosen as the site of cyclization, as it must be bonded to an oxygen, and no OH proton was observed. Further, if carbonyl C-1 ( $\delta$  171.2) were a ketone, the chemical shift of it on the isoleucine unit would be around  $\delta$  175, but it is shielded owing to the resonance effect of the oxygen atom. These assignments, all chemical shifts, coupling constants, and correlations observed on the spectra were in accordance with the structure proposed for valactamide A (**Supplementary Fig. 18**).

**Marfey's analysis of valactamide A.** Free amino acids (L-Val, D-Val, L-Ile, and D-Ile; 5 mM, 100  $\mu$ L) were individually added to 1-mL tubes. To each tube was added 200  $\mu$ L of FDLA (1-fluoro-2-4-dinitrophenyl-5-L-leucine amide; 0.4 mg/mL) and 40  $\mu$ L of 1.0 M sodium bicarbonate solution. After heating at 40 °C for 1 h, the reactions were quenched by the addition of 20  $\mu$ L of 2.0 M HCl. Samples were prepared for LC–mass spectrometry analysis by 1:1 dilution in 50% water, 50% acetonitrile.

200  $\mu$ g of valactamide A was added to a 1-mL vacuum hydrolysis tube with 600  $\mu$ L of degassed, constantly boiling 6 N HCl. The mixture was subjected to two freeze–pump–thaw cycles and then was placed into a bath heated to 100 °C for 16 h. The resulting hydrolysate was concentrated *in vacuo* and resuspended in 100  $\mu$ L of water. Fifty microliters of resuspended hydrolysate was combined with 100  $\mu$ L of FDLA (0.4 mg/mL) and 50  $\mu$ L of 1.0 M sodium bicarbonate solution. After heating at 40 °C for 1 h, the derivatization reactions were quenched by the addition of 25  $\mu$ L of 2.0 M HCl. Samples were prepared for LC–mass spectrometry analysis by 1:1 dilution in 50% water, 50% acetonitrile.

LC–mass spectrometry analysis was conducted with a Phenomenex-Luna-C18(2) column at 200  $\mu$ L/min, coupled to an electrospray source, with the mass spectrometer and source settings described above. Five microliters of each sample was injected on the column, and the gradient ran linearly from 5% to 75% acetonitrile in water over 33 min, with 0.1% formic acid (vol/vol) held constant throughout the run. For coinjection experiments, we prepared

samples by mixing derivatized valactamide A hydrolysate and derivatives of either L- or D-Val and L- or D-Ile 1:1:1. Separate analyses of valactamide A and L-Val revealed a 2-min discrepancy in retention time, potentially suggesting a substantial matrix effect. Coinjection of valactamide A and L-Val eliminated this matrix effect, yielding only one peak at the retention time observed for valactamide A alone, which indicated that valactamide A contains L-Val (**Supplementary Fig. 19a**). By contrast, coinjection of valactamide A and D-Val gave two distinct peaks, one of which was at the same retention time as for coinjection with L-Val (**Supplementary Fig. 19a**), and the other of which came off 2.3 min earlier and matched the retention time of D-Val. We observed similar results for isoleucine, indicating that valactamide A contained L-Ile (**Supplementary Fig. 19b**). The presence of L-amino acids and the lack of D-amino acids in valactamide A were consistent with the lack of an epimerase domain in the valactamide gene cluster.

**Data availability.** The DNA sequence of AtFAC9J20 was deposited at GenBank under accession [KX449366](https://www.ncbi.nlm.nih.gov/nuccore/KX449366) and is freely available for download. The LC–mass spectrometry data generated in this study are available from the corresponding author upon reasonable request.

41. Bok, J.W. & Keller, N.P. LaeA, a regulator of secondary metabolism in *Aspergillus* spp. *Eukaryot. Cell* **3**, 527–535 (2004).
42. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
43. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).
44. Running, W. Computer software reviews. Chapman and Hall Dictionary of Natural Products on CD-ROM. *J. Chem. Inf. Comput. Sci.* **33**, 934–935 (1993).
45. Laatsch, H. *Antibase 2011* (Wiley VCH, 2011).
46. Copeland, N.G., Jenkins, N.A. & Court, D.L. Recombineering: a powerful new tool for mouse functional genomics. *Nat. Rev. Genet.* **2**, 769–779 (2001).
47. Muyrers, J.P., Zhang, Y. & Stewart, A.F. Techniques: recombinogenic engineering—new options for cloning and manipulating DNA. *Trends Biochem. Sci.* **26**, 325–331 (2001).
48. Ames, B.D. & Walsh, C.T. Anthranilate-activating modules from fungal nonribosomal peptide assembly lines. *Biochemistry* **49**, 3351–3365 (2010).
49. Al-Said, N.H. Effective formal synthesis of benzomalvin A. *Monatsh. Chem.* **141**, 1249–1251 (2010).