

Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains

Manor Askenazi¹, Edward M. Driggers¹, Douglas A. Holtzman¹, Thea C. Norman¹, Sara Iverson¹, Daniel P. Zimmer¹, Mary-Ellen Boers¹, Paul R. Blomquist¹, Eduardo J. Martinez¹, Alex W. Monreal¹, Toby P. Feibelman¹, Maria E. Mayorga¹, Mary E. Maxon², Kristie Sykes¹, Jenny Vittum Tobin¹, Etchell Cordero¹, Sofie R. Salama³, Joshua Trueheart¹, John C. Royer¹, and Kevin T. Madden^{*1}

Published online 21 January 2003; doi:10.1038/nbt781

We describe a method to decipher the complex inter-relationships between metabolite production trends and gene expression events, and show how information gleaned from such studies can be applied to yield improved production strains. Genomic fragment microarrays were constructed for the *Aspergillus terreus* genome, and transcriptional profiles were generated from strains engineered to produce varying amounts of the medically significant natural product lovastatin. Metabolite detection methods were employed to quantify the polyketide-derived secondary metabolites lovastatin and (+)-geodin in broths from fermentations of the same strains. Association analysis of the resulting transcriptional and metabolic data sets provides mechanistic insight into the genetic and physiological control of lovastatin and (+)-geodin biosynthesis, and identifies novel components involved in the production of (+)-geodin, as well as other secondary metabolites. Furthermore, this analysis identifies specific tools, including promoters for reporter-based selection systems, that we employed to improve lovastatin production by *A. terreus*.

Microbes represent a valuable source of metabolites and enzymes that are essential for the production of many biomanufactured goods, including numerous commercial therapeutics and therapeutic precursors¹. In addition, biosynthetic enzymes produced by microbes represent a rich source of useful, novel, and highly stereospecific catalytic activities. To realize the full promise of metabolic engineering for industrial strain development, facile methods must be developed to understand the genetic control of metabolite production and to identify productive routes for engineering²⁻⁴.

The development of methods to comprehensively assess gene expression provides the opportunity to correlate patterns of global gene expression with the production of specific metabolites. We describe a method, referred to here as association analysis, that serves to reduce the complexity of profiling data sets to identify those genes whose expression is most tightly linked to metabolite production. Importantly, association analysis is applicable to all biological systems, including industrially useful organisms for which genome sequence information is often limited.

Association analysis was used to determine gene expression patterns that correlate with the yield of lovastatin and (+)-geodin (Fig. 1A), two secondary metabolites produced by the filamentous fungus *Aspergillus terreus*. Lovastatin is a potent hydroxymethylglutaryl coenzyme A (HMG-CoA) reductase inhibitor⁵⁻⁸ that is used clinically to reduce serum cholesterol levels^{9,10}. (+)-Geodin is derived from the anthraquinone emodin¹¹, an intermediate in the biosynthesis of many natural products¹²⁻¹⁴. Given the importance of *A. terreus* as a source of

lovastatin and other bioactive natural products, engineered *A. terreus* strains constitute a commercially relevant context for assessing the utility of association analysis to direct rational strain improvement efforts.

Results and discussion

Metabolite and gene expression data sets. In order to perform association analysis, we required data sets in which the levels of metabolite(s) and global gene expression patterns vary. To generate diversity, a collection of *A. terreus* strains was engineered to produce enhanced or reduced lovastatin titers (Table 1). These strains were engineered to express genes that were previously implicated in lovastatin production (for example, *lovE*)^{15,16}, or encode proteins expected to broadly modulate secondary metabolism (for example, *creA* and genes encoding G α proteins (*fadA*, *ganB*, *gna3*, *gpa1*, and *gna1*))¹⁷⁻²⁰, or were first identified through reporter-based genetic selections in *Saccharomyces cerevisiae* as filamentous fungal genes that promote expression from the yeast *FLO11* promoter (for example, *rfeC* and *rfeH*). Furthermore, enhanced variants of several of the proteins encoded by these genes were generated either through specific amino acid substitutions or by fusion of genes for transcription factors to the coding sequence of a transcriptional activation domain. A more detailed description of the isolation and modification of genes used to engineer metabolic and transcriptional diversity can be found as supplementary information online.

Secondary metabolite levels were analyzed by high-pressure liquid chromatography (HPLC) and electrospray mass spectrometry (MS).

¹Microbia, Inc., 320 Bent Street, Cambridge, MA 02141. ²Cytokinetics, Inc., 280 East Grand Ave., Suite 2, S. San Francisco, CA 94080. ³University of California, Santa Cruz, Jack Baskin Engineering Bldg., 1156 High St., Santa Cruz, CA 95064. *Corresponding author (kmadden@microbia.com).

[†]These authors contributed equally to this work.

Table 1. List of experimental strains and metabolite yields in engineered strains relative to reference strains

Number of transcript profiles	Engineered strain	Reference strain	Lovastatin relative concentration ^a	Geodin relative concentration ^a
4	MF22 + <i>lovE</i> (MF99)	MF22	9.28	0.42
3	MF22 + <i>VP16-rfeC</i>	MF22	4.89	0.12
2	MF22 + <i>rfeH</i>	MF22	3.00	NA
2	MF22 + <i>creA</i>	MF22+control vector	4.16	2.66
1	MF22 + <i>ganB</i>	MF22+control vector	4.44	1.95
1	MF22 + <i>gna3</i> ^{G44R}	MF22+control vector	4.07	1.62
1	MF22 + <i>ganB</i> ^{G45R}	MF22+control vector	3.18	1.43
1	MF22 + <i>gpa1</i> ^{Q204L}	MF22	<0.13	<0.02
1	MF22 + <i>gna1</i> ^{G42R}	MF22+control vector	<0.16	<0.06
3	MF22 + <i>fadA</i> ^{G42R}	MF22	<0.13	<0.02
2	MF99 + <i>fadA</i> ^{G42R}	MF99	<0.07	<0.05

^a Mean molar concentrations of metabolites were determined by HPLC from a population of reference and engineered strains. Relative concentrations represent ratio of mean concentration in engineered strains relative to mean concentration in the appropriate reference strain.

In addition to lovastatin and related monacolins, secondary metabolite profiling identified a variety of (+)-geodin related compounds, with (+)-geodin itself being the most abundant secondary metabolite in broths from control strains. The identity of this compound was confirmed by NMR spectroscopy. Quantitative lovastatin and (+)-geodin yields from engineered strains, relative to levels from appropriate reference strains, are listed in Table 1.

To identify gene expression patterns that correlate with the production of lovastatin, (+)-geodin, or both secondary metabolites, representative transformants from each set of manipulated strains and appropriate reference strains were used to generate transcriptional profiles for samples from 72 h fermentations. Since limited sequence information is available for the *A. terreus* genome, we developed a genomic fragment microarray methodology to monitor genome-wide expression patterns. Specifically, random genomic fragments with an average size of 2 kb were used to generate a microarray of approximately 21,000 elements. In addition, previously described *A. terreus* genes, such as genes encoding the lovastatin biosynthetic components, were included on the microarray. These arrays were used to generate profiles for 21 strains producing increased or decreased lovastatin levels and 19 strains with altered (+)-geodin levels (Table 1; complete transcriptional profiling data sets are available on the authors' website (see URL in Experimental protocol)).

Figure 1 provides an overview of the metabolic and transcriptional profiles obtained using the strains described above. Overexpression of different genes resulted in distinct patterns of lovastatin and (+)-geodin production (Fig. 1B), suggesting that these genes elicit metabolic responses via distinct mechanisms of action. Hierarchical clustering (Fig. 1C) of the transcriptional profiling data sets shows that strains displaying similar metabolite profiles tend to have relatively similar transcriptional profiles. For example, strains that produce high lev-

els of lovastatin and decreased levels of (+)-geodin cluster together, and cluster separately from strains that produce decreased levels of both metabolites. Principal component analysis (PCA) (Fig. 1D; see Supplementary Table 1 online) helps to determine the few linear combinations of genes that explain most of the variance present in the data sets, and provides further insight into the underlying variables that distinguish the various strains. The first two components (PC1 and PC2) account for 61% of the variance in the overall data sets. The variance accounted for by PC1 is best explained by genes whose expression was highly influenced by overexpression of a constitutively active form of FadA. Many of the affected genes are also involved in secondary metabolism. The loadings in PC2 demon-

strate that lovastatin biosynthetic genes account for a significant portion of the variance in these data sets. Our initial study of the global transcriptional response (by clustering and PCA) demonstrated that biologically meaningful variation exists in the transcriptional data. However, it also suggested that a large component of this variation might not be directly relevant to those secondary metabolites that we are most interested in studying (for example, lovastatin, (+)-geodin). We therefore required a more directed analytical approach.

Association analysis. Association analysis was performed using the combined metabolic and transcriptional data sets. Secondary metabolite and gene expression values were expressed as ratios that reflect a value from an engineered strain relative to that from a reference strain. Two statistical approaches were subsequently employed

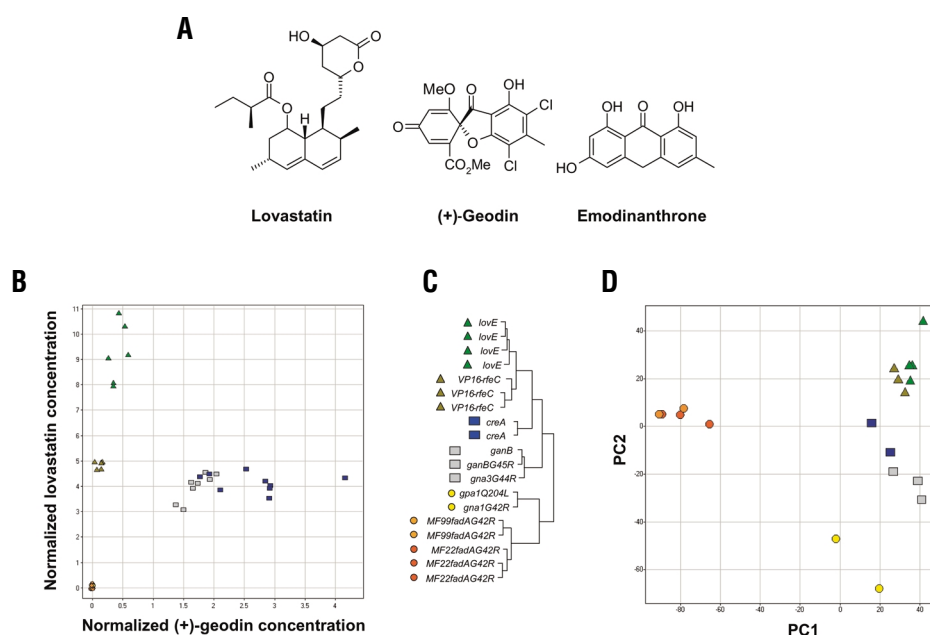


Figure 1. Characterization of metabolic and transcriptional diversity. (A) Chemical structures for lovastatin, (+)-geodin, and emodinanthrone. (B) A scatter plot of lovastatin and (+)-geodin titers, normalized relative to mean titers from relevant reference strains (see Table 1 for reference strains). (C) Hierarchical clustering (average linkage with Pearson correlation coefficients) of the transcriptional profiling data sets. (D) Scatter plot of the scores for the first and second components generated by the application of principal component analysis (PCA) to the same data sets.

Table 2. Proteins encoded on elements that positively associate with lovastatin and/or (+)-geodin production^a

Category	Both lovastatin and (+)-geodin	Lovastatin	(+)-Geodin
Lovastatin biosynthetic cluster proteins	LovA (cytochrome P450 monooxygenase), LovC (enoyl reductase)	LovB (nonaketide PKS), LovD (trans-esterase), LovF (diketide PKS), LvrA (HMG CoA reductase), ORF2, ORF5, ORF10 (ABC transporter), ORF17 (cytochrome P450 monooxygenase)	
Demonstrated and predicted (+)-geodin biosynthetic proteins	Dihydrogeodin oxidase, emodinanthrone PKS, flavin-binding monooxygenase, halogenase, O-methyltransferase B		
Additional secondary metabolite biosynthetic proteins	<i>A. oryzae</i> dimethylallyl-cycloacetyl-L-tryptophan synthase, fungal pigment/mycotoxin PKS, homolog of fungal pigment/Lov/bacterial PKS, <i>N. haematococca</i> pisatin demethylase (cytochrome P450 monooxygenase), phenol hydroxylase, salicylate hydroxylase	Fungal non-ribosomal peptide synthetase	<i>B. fuckeliana</i> protein related to DOPA 4,5-dioxygenase
Fatty acid metabolism	Acyl-CoA oxidase, C-14 sterol reductase, lysophospholipase	Delta-12 fatty acid desaturase	Acyl-CoA dehydrogenase
Sulfur assimilation and methionine biosynthesis	Adenohomocysteinase, homoserine O-acetyltransferase, O-acetylhomoserine (thiol)-lyase		
Other primary metabolic proteins	Formate dehydrogenase, serine hydroxymethyltransferase		Transaldolase
Carbohydrate utilization	Chitin deacetylase	Alpha-glucosidase (maltase), glycogen phosphorylase	
ABC transporters	<i>N. crassa</i> Yor1, <i>V. inaequalis</i> Abc1		Amino acid permease
Proteases	Alkaline protease (elastase)	Carboxypeptidase Y	
Sporulation		<i>N. crassa</i> conidiation-specific protein, <i>N. crassa</i> Pro1	<i>S. cerevisiae</i> YNL223w homolog (required for sporulation)
Translation	rDNA, multiple ribosomal proteins		
Contiguous genes encoding several enzymatic activities (possible secondary metabolite biosynthetic cluster)	AMP-binding protein (NRPS/CoA synthetase), fungal isotrichodermin C-15 hydroxylase, short-chain alcohol dehydrogenase, <i>S. cerevisiae</i> YOL119c monocarboxylate permease		
Other	Acid phosphatase, <i>B. graminis</i> Qde2, cytochrome b5 reductase, fungal Cot-1 Ser/Thr kinase, glutathione-S-transferase, GMC oxidoreductase, histone deacetylase, homology to AfIR, lectin, <i>N. crassa</i> hypothetical protein, NADPH:quinone oxidoreductase, potassium transporter, probable biotin-protein ligase	Aldehyde dehydrogenase, <i>S. cerevisiae</i> Gpr/Fun 34 family protein, ketoreductase, oxalate decarboxylase, peptidyl prolyl <i>cis-trans</i> isomerase with RNA binding region, potassium transporter, Rel-associated pp40, <i>S. pombe</i> hypothetical protein, UbcM4-interacting protein 83, ubiquitin-like protein, vanadate resistance protein (possible mannose transporter)	<i>A. oryzae</i> EST, aromatic amino acid aminotransferase, cyclophilin-like peptidyl prolyl <i>cis-trans</i> isomerase, Rho1 GTPase

^aAll proteins (or homologs of these proteins) listed are encoded by genes present on lovastatin- and/or (+)-geodin-associated elements ($P < 0.05$ for either gamma or correlation coefficient calculations). Complete information can be found in the Supplementary Note and Supplementary Tables 2–5 and Supplementary Table 8 online.

to define the relationships between gene(s) present on hybridizing elements and secondary metabolite levels. In the first approach, Pearson product-moment correlation coefficients were calculated from transcriptional profiling ratio values and metabolite ratios. However, integration of different data types presents inherent challenges, including technology-driven issues such as varying sensitivity and reliability between detection methods, as well as statistical concerns such as the impact of outlying data points on resulting correlation values. To avoid these potential pitfalls in our analysis, an alternative association method was also pursued. In this method, the

two data types were transformed into simplified ordinal representations using transformation techniques specific to each technology, and then standard categorical data analysis methods were applied to detect significant associations²¹.

For the ordinal approach, gene expression ratios were categorized as increased, unchanged, or decreased. The significance of ratio values was defined by the likelihood of a similar ratio being present in “self” experiments (see Experimental protocol). Since the variance of transcriptional profiling ratios increased considerably at the low end of the signal intensity range, we defined a significantly changed gene

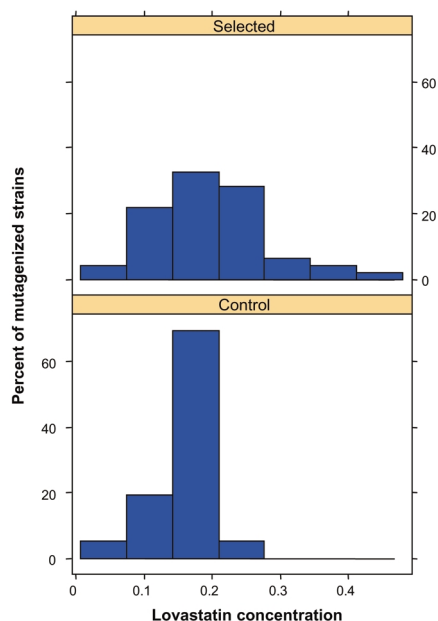


Figure 2. Selection systems identify improved strains. Histograms of lovastatin concentration distributions from fermentation broths of *lovF-ble* containing strains that were mutagenized and either plated on medium containing 50 µg/ml phleomycin (46 samples) (Selected) or control medium lacking phleomycin (36 samples) (Control).

expression ratio value as a function of its signal intensity. For metabolite measurements, HPLC-based methods were sufficiently robust and sensitive to reproducibly detect small changes between samples; all ratio values in this data set were defined as either increased or decreased. Next, Goodman and Kruskal's gamma was calculated for every pairing of array element and metabolite²². The vast majority of elements determined to be significantly associated with metabolite production by gamma are also present in the larger element sets generated by Pearson correlation. Thus, for these data sets, measures of association that use either ordinal or continuous data representations converge on a common set of elements.

Sequence information was obtained for many microarray elements showing expression patterns that were significantly associated with lovastatin or (+)-geodin production, or both. Following sequencing, homology searching, and contig analyses were performed; in many instances, multiple clones with similar expression patterns were found to contain overlapping sequences. Tables 2 and 3 summarize the results of association analysis by displaying proteins with homology to known sequences that are encoded upon elements with expression patterns that are positively or negatively associated with lovastatin and/or (+)-geodin production ($P < 0.05$). A complete listing of metabolite-associated elements can be found in Supplementary Tables 2–5 online.

Association analysis reveals biosynthetic clusters. Association analysis enables the rapid identification of genes required for the biosynthesis of secondary metabolites. The *A. terreus* lovastatin biosynthetic cluster is a 64 kb genomic region predicted to encode 18 proteins¹⁵, some of which are known to be required for lovastatin production¹⁶. Array elements containing *lovA*, *lovB*, *lovC*, *lovD*, *lovF*, *lvrA*, and open reading frames (ORFs) 2, 5, 10, and 17 were positively associated with lovastatin production (Table 2). It is likely that LovE, a transcription factor encoded within the cluster, at least partially regulates the coordinate expression of these genes¹⁵. The independent discovery of the regulated lovastatin biosynthetic genes by association analysis validates the method.

In addition, association analysis sheds light upon (+)-geodin biosynthetic mechanisms: it identifies the previously unknown polyketide synthase (PKS) required for (+)-geodin production (the emodinanthrone PKS); it demonstrates that expression of a known (+)-geodin biosynthetic gene, encoding the dihydrogeodin oxidase, correlates with (+)-geodin production; and it predicts several novel (+)-geodin biosynthetic genes^{11,23–25} (Table 2; and see Supplementary information and Supplementary Fig. 1 online). For the identification of the PKS required for (+)-geodin production, the combination of observed association scores, protein sequence homology to a known PKS class, and chemical similarities to the relevant polyketide metabolites led us to predict that several contiguous (+)-geodin-associated array elements encode the emodinanthrone PKS. These elements show significant homology to filamentous fungal enzymes required for pigment biosynthesis^{26–29}; the pigmented natural products are non-reduced fungal polyketides^{30,31}. The chemical structure of the (+)-geodin precursor, emodinanthrone, clearly defines it as a member of the non-reduced fungal polyketide class¹¹ (Fig. 1A). To verify that the identified PKS is required for (+)-geodin biosynthesis, strains were generated that contain a disruption in the putative emodinanthrone PKS gene. Strains that contain the genetic disruption did not produce detectable levels of (+)-geodin, whereas control strains produced robust (+)-geodin titers.

Association analysis identified many genes that encode proteins either predicted or known to play a role in the production of secondary metabolites other than lovastatin and (+)-geodin (Tables 2 and 3). These secondary metabolite biosynthetic enzymes include novel and known polyketide synthases (for example, PksM^{32,33}), a non-ribosomal peptide synthetase, and a dimethylallyl-cycloacetyl-L-tryptophan synthase homologous to enzymes required for the production of several fungal secondary metabolites (for example, cyclopiazonic acid³⁴).

Association analysis enhances current metabolic engineering approaches by facilitating both the discovery of novel biosynthetic genes and the development of rational tools for uncoupling the coordinate control of specific biosynthetic pathways. Newly discovered genes could be used to develop production systems; for example, the emodinanthrone PKS gene could facilitate development of strains for microbial production of useful anthraquinone-derived natural products. Furthermore, the elimination of these same genes (as shown above) can decrease the metabolic complexity of the cell and fermentation broth, both by promoting the flow of precursors toward desired metabolites and by decreasing the concentration of impurities that may hinder the purification of specific metabolites.

Association analysis reveals key metabolic trends. Analysis of gene expression patterns that correlate with lovastatin or (+)-geodin production, or both, provides insight into the physiological states that promote the biosynthesis of secondary metabolites. For example, a collection of genes expected to be expressed during growth phase or involved in the generation of energy (for example, enzymes of the glycolytic and tricarboxylic acid pathways, proteins involved in oxidative phosphorylation) are present on elements that negatively correlate with secondary metabolite production (Table 3). It is possible, though, that these expression patterns reflect a more specific cellular effect such as the level of cellular reserves of reductive equivalents or oxygen availability. Consistent with previous observations that many secondary metabolites are produced late during fermentation, several genes that are positively associated with secondary metabolite production encode enzymes that catabolize possible alternative sources of nutrients (for example, glycogen phosphorylase, lysophospholipase, alkaline protease) (Table 2).

In addition to the lovastatin- and (+)-geodin-associated genes that promote a general metabolic state that is conducive to secondary metabolite production, some sets of secondary metabolite-

Table 3. Proteins encoded on elements that negatively associate with lovastatin and/or (+)-geodin production^a

Category	Both lovastatin and (+)-geodin	Lovastatin	(+)-Geodin
Lovastatin biosynthetic cluster proteins			<i>LovC, LovE</i>
Additional secondary metabolite biosynthetic proteins	6-methylsalicylic acid PKS (PksM), bacterial PKS		<i>A. parasiticus</i> versicolorin B synthase
Central carbon metabolism	Aconitase, carnitine acetyl transferase FacC, citrate synthase, dihydroliipoamide succinyltransferase, L-lactate ferricytochrome c oxidoreductase, malate synthase, phosphoglyceromutase, pyruvate decarboxylase	2-phosphoglycerate dehydratase, glucose-6-phosphate isomerase, glyceraldehyde-3-phosphate dehydrogenase, methylenetetrahydrofolate phosphofructokinase, Yfi030w (alanine: glyoxylate aminotransferase)	Mannitol-1-phosphate 5-dehydrogenase, phosphoenolpyruvate carboxykinase
Fatty acid metabolism	Acyl-CoA oxidase, fatty acid desaturase, fatty acid synthase alpha (primary metabolic FAS), fatty acid synthase beta (primary metabolic FAS)		
Sterol metabolism	Coproporphyrinogen III oxidase	S-adenosyl-methionine-sterol-C-methyltransferase, sterol C5-desaturase	C-4 methyl sterol oxidase, C-22 sterol desaturase
Carbohydrate utilization	1,3-beta-glucan synthase, <i>Bacillus</i> sp. alpha-amylase, beta-glucosidase, <i>P. brasiliensis</i> chitin synthase, UTP-glucose-1-phosphate uridylyltransferase	<i>A. fumigatus</i> Gel1 protein (beta(1-3)glucanosyltransferase), <i>A. oryzae</i> alpha-amylase, beta(1-3)-glucanosyltransferase/Phr1, putative glucanase	Exo-alpha (1,4)-polygalacturonase
ABC transporters		<i>M. graminicola</i> Atr4	<i>A. fumigatus</i> Mdr1, <i>C. dubliniensis</i> Mdr1, homolog of secondary metabolite pumps, <i>N. crassa</i> cycloheximide resistance protein
Heat shock response	Heat shock protein 104	<i>A. nidulans</i> heat shock protein (Hsp12 homolog)	Possible heat shock protein (<i>S. cerevisiae</i> YOR285w homolog)
G protein signaling	Canine retinitis pigmentosa GTPase regulator		FadA GTP-binding protein, GTPase activating protein, opsin-1
Oxidative phosphorylation	Cytochrome b, cytochrome oxidase assembly factor (Cox15)		<i>S. cerevisiae</i> Ytp1
Sulfur assimilation and methionine biosynthesis		<i>A. nidulans</i> MetR, PAPS reductase	<i>A. nidulans</i> sulfur metabolite repression (SconB)
Vacuolar-targeting proteins		<i>S. cerevisiae</i> Vac8, <i>S. cerevisiae</i> Vps13, <i>S. cerevisiae</i> Vps55	
Translation	Mitochondrial rDNA, multiple ribosomal proteins		
Chromatin/ chromosome structure	<i>S. cerevisiae</i> Bdf1 homolog (required for sporulation)		Histone acetyltransferase
Other	<i>A. parasiticus</i> Hxt1 hexose transporter, ADP, ATP translocase, aldehyde dehydrogenase, asparagine synthetase, Cu-binding metallothionein, cullin, histidine kinase, lignostilbene-alpha, beta dioxygenase, <i>N. crassa</i> Tol protein, mitochondrial RNA polymerase, NADP-dependent alcohol dehydrogenase oxidoreductase, probable tricarboxylate transport protein, <i>S. cerevisiae</i> Gcn1, ser/arg rich protein, urate oxidase, <i>S. cerevisiae</i> YCR061w, zinc-containing dehydrogenase	1-Cys peroxiredoxin, <i>A. nidulans</i> NsdD, beta actin, <i>N. crassa</i> hypothetical protein, NADPH-cytochrome P450 oxidoreductase, outer mitochondrial membrane protein, peripheral-type benzodiazepine receptor, phospholipid transporting ATPase, SR protein kinase, uracil permease, vanadate resistance protein (possible mannose transporter), <i>S. cerevisiae</i> YOL119c monocarboxylate permease	<i>A. nidulans</i> QutA-like zinc binuclear cluster protein, aspartyl protease, aureobasidin-resistance protein (inositol phosphorylceramide (IPC) synthase), F-Box/WD repeat containing protein, filamentous fungal Ace1 protein, helicase-like protein, hypothetical <i>S. coelicolor</i> protein, kinesin-like protein Kif1C, <i>N. crassa</i> calcium-related spray protein, peptidylprolyl isomerase, probable helicase, putative peroxisomal membrane protein, putative quinone oxidoreductase, <i>S. cerevisiae</i> Cdc54, <i>S. cerevisiae</i> Ste20, <i>S. pombe</i> protein involved in Mn ²⁺ homeostasis, <i>S. pombe</i> hypothetical protein, septin, TATA-box binding protein

^aAll proteins (or homologs of these proteins) listed are encoded by genes present on lovastatin- and/or (+)-geodin-associated elements ($P < 0.05$ for either gamma or correlation coefficient calculations). Complete information can be found in the Supplementary Note and Supplementary Tables 2–5 and Supplementary Table 8 online.

associated genes may be more directly relevant to the production of these polyketide-derived metabolites. For instance, fatty acid metabolism genes that positively associate with lovastatin and (+)-geodin production tend to encode catabolic enzymes that are predicted to promote formation of the polyketide precursors acetyl- and malonyl-CoA (Table 2), whereas fatty acid metabolism genes that negatively associate with secondary metabolite production tend to encode anabolic enzymes (Table 3). Furthermore, S-adenosyl-L-methionine is required for the methylation of lovastatin and (+)-geodin intermediates, and several enzymes involved in sulfur assimilation and methionine biosynthesis positively associate with lovastatin and (+)-geodin production (Table 2).

Successful metabolic engineering strategies from association analysis. Metabolite-associated genes can be useful metabolic engineering tools. For example, promoter sequences from lovastatin-associated genes can be used to configure reporter-based selections in *A. terreus* to rapidly identify improved lovastatin-producing strains. To this end, a reporter-based selection was configured in *A. terreus* by constructing a strain that contains the promoter from the lovastatin biosynthetic gene *lovF* fused to the *ble* gene, which encodes a protein that confers increased resistance to phleomycin³⁵. The *lovF* promoter was of particular use because *lovF* expression correlated with lovastatin production, the intensity of *lovF* expression facilitated configuration of the selection system, and the transcriptional profiling data sets suggested that the *lovF* promoter integrates signaling from multiple pathways.

We determined that randomly generated mutants expressing *lovF* at increased levels also produced lovastatin at increased levels. The strain containing *lovF-ble* was mutagenized and plated on medium with (3×10^7 spores) or without (10^3 spores) phleomycin. Colonies from both sets of plates were picked and tested for lovastatin production. The population of strains from phleomycin plates displayed a significant increase in mean lovastatin yield ($P < 0.001$ in a bootstrap confidence interval for difference of means³⁶) (Fig. 2). These selection systems provide an enormous improvement over traditional screening methods. In this example, a single selection plate was used to identify ten strains that produce more lovastatin than any control strain (up to about twofold more). When expression from the promoter of the lovastatin-associated gene accurately reflects metabolite production, a single selection plate constitutes the functional equivalent of extremely large numbers of fermentations and assays for metabolite titer (e.g., 3×10^7 described above).

Integration of diverse experimental data types provides a unique vantage point on complex cellular responses that cannot be acquired from any single data type in isolation³⁷. Our experiments show that association analysis of transcriptional and metabolite profiling data has the power to elucidate key genetic components and physiological traits that impact the production of lovastatin and (+)-geodin. In addition, association analysis has identified genetic tools that can be utilized to drive the strain improvement process. Furthermore, this approach can be extended both to elucidate interrelationships between other physiological traits and to define multivariate relationships between traits. Within the context of metabolic engineering, association analysis promises to be a powerful tool to illuminate unexplored avenues to the rational development of industrial strains.

Experimental protocol

***A. terreus* strains.** All strains used were either MF22 (American Type Culture Collection (Manassas, VA) # 20542 (*A. terreus* Thom, anamorph)) or derivatives thereof.

Transformation and shake flask growth of *A. terreus*. Plasmids or linear fragments used to transform MF22 and transformation methods are as described in Supplementary information online and in Royer *et al.*³⁸, respectively. Transformants were colony purified prior to shake flask experiments. For

shake flask experiments, strains were grown in 25 ml of modified RPM medium³⁹ (see Supplementary Note online).

Genomic fragment microarrays. Approximately 21,000 microarray elements were generated from PCR products amplified from bacterial strains containing plasmids from an *A. terreus* genomic library prepared by partial digestion with *Sau3A* and ligation into the vector pZErO-2 (Invitrogen, Carlsbad, CA) (see Supplementary Note online). In addition, unique primer pairs were designed to amplify coding sequence for previously described genes from *A. terreus*, including the 18 predicted genes from the lovastatin biosynthetic cluster¹⁵. Purified amplification products were printed upon SuperAmine slides (TeleChem International, Sunnyvale, CA) using the OmniGrid microarrayer (GeneMachines, San Carlos, CA) and Stealth Micro Spotting Pins (TeleChem International). The Supplementary online files contain information demonstrating that the *A. terreus* genomic fragment microarrays can be used to reliably detect differential expression.

Identification and quantification of secondary metabolites. Lovastatin and (+)-geodin were resolved and identified from broths of *A. terreus* fermentations using HPLC-electrospray MS (see Supplementary Note online). Isocratic reverse-phase chromatography was used for initial HPLC-MS identification of secondary metabolites. Eluted compounds were detected by quadrupole-time of flight MS (Micromass Q-TOF, Milford, MA) in positive ion mode. High resolution mass values for eluted peaks were compared to a database of values for known secondary metabolites, modified to report monoisotopic mass⁴⁰. Following identification of secondary metabolites, a rapid HPLC assay was developed for monitoring lovastatin and (+)-geodin titers. Authentic lovastatin (Sigma, St. Louis, MO) and (+)-geodin (purified from *A. terreus* culture broth) samples were used to generate standard integration curves for absolute quantification. The identity of the isolated (+)-geodin was confirmed by ¹H-NMR (¹H-NMR (CDCl₃, 300 MHz): δ 2.58 (s, 3H), 3.71 (s, 3H), 3.74 (s, 3H), 5.82 (s, 1H), 7.15 (s, 1H), 7.43 (br s, 1H)) and high resolution MS (calculated for C₁₇H₁₃O₇Cl₂ (MH⁺): 399.0039; found 399.0034).

RNA preparation. Standard methods were used to prepare total RNA from biomass harvested from 72 h fermentations (see Supplementary Note online).

Transcriptional profiling. Standard methods were employed to generate and purify Cy3- and Cy5-labeled first-strand cDNA from *A. terreus* total RNA (see Supplementary Note online). Cy3- and Cy5-labeled cDNA were combined, lyophilized, resuspended in hybridization buffer, and then used for hybridization. Slides were analyzed using an Affymetrix 418 Array Scanner (Santa Clara, CA) to measure fluorescence of the Cy- and Cy5- labeled cDNAs (532 and 635 nm, respectively) bound to the DNA microarrays. GenePix Pro 3.0 (Axon Instruments, Union City, CA) was used to quantify the signal intensity of each element on the array.

Statistical analysis. Complete transcriptional profiling data sets are available at the Microbia, Inc. website (see URL). The Supplementary Note and Supplementary Tables 6 and 7 provide evidence that the *A. terreus* genomic fragment microarrays are valid tools for identifying genes with expression patterns that correlate with secondary metabolite production. Local background subtraction was performed prior to calculation of ratio values for each array element. An intensity threshold filter was applied to remove low intensity elements from the analysis (see Supplementary information online). The log₂ ratio values were then rescaled to a median of zero.

For calculating gamma, the rescaled ratio data was transformed into three ordinal categories: increased, unchanged, and decreased. A pooled set of "self experiments" (that is, experiments using the same sample labeled with both Cy3 and Cy5) was used to generate an empirical null distribution of signal ratios from which we defined critical values ($\alpha = 0.05$). Since increased variance was detected at low signal intensities, two separate critical values were used (critical values = ± 2.06 , when the log₂ of the geometric mean of background-subtracted signals ≤ 9.5 , and ± 1.07 at all other intensity levels).

When calculating correlation coefficients and gamma for a given gene-metabolite pair, a simple permutation test was employed to assess the significance of the result. Four hundred random permutations of the gene-metabolite value pairs were generated, and the resulting correlation coefficients and gamma values were calculated. The distribution of values obtained (under the assumption of independence between the gene and the metabolite) was used to estimate the *p* value for a two-sided test.

When applying hierarchical clustering and PCA to the transcriptional profiling data, the analyses were restricted to experiments for which both

lovastatin and (+)-geodin levels were available and to those array elements that passed the intensity threshold filter in at least 30% of the experiments. When applying PCA, missing data points were replaced with zeros.

Sequence analysis and annotation. First-strand sequence information was obtained for metabolite-associated elements after reamplification of DNA from source samples used for microarray printing. The resulting sequence was subjected to basic local alignment search tool (BLAST) analysis⁴¹. A short annotation tag was given to an element when significant homology to a known gene or genes was detected (see Supplementary Note online). Batch contig analysis was performed on all sequence data in order to determine whether subsets of clones contain overlapping sequence (Sequencher, Gene Codes Corporation, Ann Arbor MI).

Disruption of (+)-geodin biosynthesis. A gene disruption plasmid was constructed by replacing a significant portion of the putative emodinanthrone PKS coding sequence with a *gpd-ble* cassette (see Supplementary Note online). The resulting construct was used to transform wild-type *A. terreus* to phleomycin resistance. Phleomycin resistant strains were screened for (+)-geodin production. Disruption of the putative emodinanthrone PKS was confirmed by performing diagnostic PCR using genomic DNA prepared from transformants as template⁴².

lovF-ble reporter-based selection. Plasmid p3437 was generated by inserting a *lovF-ble* cassette into p3SR2 (ref. 43) (see Supplementary Note online). Wild-

type *A. terreus* was transformed with p3437 and transformants were selected. Approximately 3×10^7 spores of transformant 3437-4 were spread on a 150 mm minimal plate containing 50 µg/ml phleomycin, and 10^3 spores were spread on a minimal plate containing no drug. Both plates were UV irradiated to approximately 90% killing, and plates were incubated at 30 °C in the light for 5 d. The resulting strains (46 strains from selection plates, 36 strains from non-selected plates) were grown and lovastatin levels were quantified.

URL. For complete transcriptional profiling data sets, see <http://www.microbia.com/news/SupplDataTable1.exe>.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgments

We are grateful to Brian Cali, Gerry Fink, and Todd Milne for critical comments on this manuscript. In addition, the authors would like to thank all members of the Precision Engineering program at Microbia, Inc. for their support and helpful discussions.

Competing interests statement

The authors declare competing financial interests: see the Nature Biotechnology website (<http://www.nature.com/naturebiotechnology>) for details.

Received 1 July 2002; accepted 13 November 2002

- Masurekar, P.S. Therapeutic metabolites. *Biotechnology* **21**, 241–301 (1992).
- Parekh, S., Vinci, V.A. & Strobel, R.J. Improvement of microbial strains and fermentation processes. *Appl. Microbiol. Biotechnol.* **54**, 287–301 (2000).
- Nielsen, J. The role of metabolic engineering in the production of secondary metabolites. *Curr. Opin. Microbiol.* **1**, 330–336 (1998).
- Nielsen, J. Metabolic engineering. *Appl. Microbiol. Biotechnol.* **55**, 263–283 (2001).
- Endo, A. Monacolin K. A new hypocholesterolemic agent produced by a *Monascus* species. *J. Antibiot.* **32**, 852–854 (1979).
- Endo, A., Kuroda, M. & Tsujita, Y. ML-236A, ML-236B, and ML-236C, new inhibitors of cholesterologenesis produced by *Penicillium citrinum*. *J. Antibiot.* **29**, 1346–1348 (1976).
- Endo, A., Kuroda, M. & Tanazawa, K. Competitive inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase by ML236A and ML-236B fungal metabolites having hypocholesterolemic activity. *FEBS Lett.* **72**, 323–326 (1976).
- Tanazawa, K. & Endo, A. Kinetic analysis of the reaction catalyzed by rat-liver 3-hydroxy-3-methylglutaryl-coenzyme-A reductase using two specific inhibitors. *Eur. J. Biochem.* **98**, 195–201 (1979).
- Alberts, A.W. Discovery, biochemistry and biology of lovastatin. *Am. J. Cardiol.* **62**, 10J–15J (1988).
- Endo, A. Hypocholesterolemic agents. *Biotechnology* **26**, 301–320 (1994).
- Fujimoto, H., Flash, H. & Franck, B. Biosynthese der seco-anthraquinone geodin und dihydrogeodin aus emodin. *Chem. Ber.* **108**, 1224–1228 (1975).
- Sankawa, U., Ebizuka, Y. & Shibata, S. Biosynthetic incorporation of emodin and emodinanthrone into the anthraquinoids of *Penicillium brunneum* and *P. islandicum*. *Tetrahedron Lett.* 2125–2128 (1973).
- Franck, B., Huper, F., D., G. & Erge, D. Seco-Anthraquinone nach einem neugefundenen Biosyntheseprozess. *Angew. Chem.* **78**, 752–753 (1976).
- Birch, A.J., Baldas, J., Hlubucek, F.B., Simpson, T.J. & Westerman, P.W. Biosynthesis of the fungal xanthone ravenelin. *J. Chem. Soc. Perkin Trans. I* 898–904 (1976).
- Kennedy, J. *et al.* Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. *Science* **284**, 1368–1372 (1999).
- Hutchinson, C.R. *et al.* Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases. *Antonie Van Leeuwenhoek* **78**, 287–295 (2000).
- Bailey, C. & Arst, H.N., Jr. Carbon catabolite repression in *Aspergillus nidulans*. *Eur. J. Biochem.* **51**, 573–577 (1975).
- Dowzer, C.E. & Kelly, J.M. Cloning of the *creA* gene from *Aspergillus nidulans*: a gene involved in carbon catabolite repression. *Curr. Genet.* **15**, 457–459 (1989).
- Hicks, J.K., Yu, J.H., Keller, N.P. & Adams, T.H. *Aspergillus* sporulation and mycotoxin production both require inactivation of the FadA G alpha protein-dependent signaling pathway. *EMBO J.* **16**, 4916–4923 (1997).
- Tag, A. *et al.* G-protein signaling mediates differential production of toxic secondary metabolites. *Mol. Microbiol.* **38**, 658–665 (2000).
- Agresti, A. *Categorical Data Analysis*. (John Wiley & Sons, New York; 1990).
- Goodman, L.A. & Kruskal, W.H. Measures of association for cross classifications. *J. Am. Stat. Assoc.* **49**, 732–764 (1954).
- Curtis, R.F., Hassall, C.H. & Parry, D.R. The biosynthesis of phenols. XXIV. The conversion of the anthraquinone question into the benzophenone, sulochrin, in cultures of *Aspergillus terreus*. *J. Chem. Soc. Perkin Trans. I* **2**, 240–244 (1972).
- Fujii, I., Iijima, H., Tsukita, S., Ebizuka, Y. & Sankawa, U. Purification and properties of dihydrogeodin oxidase from *Aspergillus terreus*. *J. Biochem. (Tokyo)* **101**, 11–18 (1987).
- Gatenbeck, S. & Malmstrom, L. On the biosynthesis of sulochrin. *Acta. Chem. Scand.* **23**, 3493–3497 (1969).
- Mayorga, M.E. & Timberlake, W.E. The developmentally regulated *Aspergillus nidulans* wa gene encodes a polypeptide homologous to polyketide and fatty acid synthases. *Mol. Gen. Genet.* **235**, 205–212 (1992).
- Takano, Y. *et al.* Structural analysis of PKS1, a polyketide synthase gene involved in melanin biosynthesis in *Colletotrichum lagenarium*. *Mol. Gen. Genet.* **249**, 162–167 (1995).
- Tsai, H.F., Wheeler, M.H., Chang, Y.C. & Kwon-Chung, K.J. A developmentally regulated gene cluster involved in conidial pigment biosynthesis in *Aspergillus fumigatus*. *J. Bacteriol.* **181**, 6469–6477 (1999).
- Fulton, T.R., Ibrahim, N., Losada, M.C., Grzegorski, D. & Tkacz, J.S. A melanin polyketide synthase (PKS) gene from *Nodulisporium* sp. that shows homology to the *pkS1* gene of *Colletotrichum lagenarium*. *Mol. Gen. Genet.* **262**, 714–720 (1999).
- Bingle, L.E.H., Simpson, T.J. & Lazarus, C.M. Ketosynthase domain probes identify two subclasses of fungal polyketide synthase genes. *Fungal Genet. Biol.* **26**, 209–223 (1999).
- Nicholson, T.P. *et al.* Design and utility of oligonucleotide gene probes for fungal polyketide synthases. *Chem. Biol.* **8**, 157–178 (2001).
- Fujii, I. *et al.* Cloning of the polyketide synthase gene *atX* from *Aspergillus terreus* and its identification as the 6-methylsalicylic acid synthase gene by heterologous expression. *Mol. Gen. Genet.* **253**, 1–10 (1996).
- Pazoutova, S., Linka, M., Storkova, S. & Schwab, H. Polyketide synthase gene *pkS1* from *Aspergillus terreus* expressed during growth phase. *Folia Microbiol.* **42**, 419–430 (1997).
- Christensen, B.E. & Kaasgaard, S. Methods for producing polypeptides in *Aspergillus* mutant cells. World Patent Application WO 00/39322 (2000).
- Drocourt, D., Calmels, T., Reynes, J.P., Baron, M. & Tiraby, G. Cassettes of the *Streptoalloteichus hindustanus ble* gene for transformation of lower and higher eukaryotes to phleomycin resistance. *Nucleic Acids Res.* **18**, 4009 (1990).
- Good, P. Permutation tests: A practical guide to resampling methods for testing hypotheses (Springer-Verlag, New York, 2000).
- Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
- Royer, J.C. *et al.* *Fusarium graminearum* A 3/5 as a novel host for heterologous protein production. *Biotechnology* **13**, 1479–1483 (1995).
- Szakacs, G., Morovjan, G. & Tengerdy, R.P. Production of lovastatin by a wild strain of *Aspergillus terreus*. *Biotechnol. Lett.* **20**, 411–415 (1998).
- Buckingham, J. *Dictionary of Natural Products on CD-ROM*, vol. 10:1 (Chapman & Hall / CRC Press, Boca Raton FL, 2001).
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Timberlake, W.E. Isolation of stage- and cell-specific genes from fungi. in *Biology and Molecular Biology of Plant-Pathogen Interactions* NATO ASI series H1 (ed. Bailey, J.) 343–357 (Springer-Verlag, Berlin Heidelberg, 1986).
- Kelly, J.M. & Hynes, M.J. Transformation of *Aspergillus niger* by the *amdS* gene of *Aspergillus nidulans*. *EMBO J.* **4**, 475–479 (1985).