nature chemical biology

ARTICLE

PUBLISHED ONLINE: 3 OCTOBER 2016 | DOI: 10.1038/NCHEMBIO.2188

Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching

Chris A Dejong[1–3], Gregory M Chen[1–3], Haoxin Li[1–3], Chad W Johnston[1,2], Mclean R Edwards[1,2], Philip N Rees[1,2], Michael A Skinnider[1,2], Andrew L H Webster[1,2] & Nathan A Magarvey[1,2]*

Polyketides (PKs) and nonribosomal peptides (NRPs) are profoundly important natural products, forming the foundations of many therapeutic regimes. Decades of research have revealed over 11,000 PK and NRP structures, and genome sequencing is uncovering new PK and NRP gene clusters at an unprecedented rate. However, only ~10% of PK and NRPs are currently associated with gene clusters, and it is unclear how many of these orphan gene clusters encode previously isolated molecules. Therefore, to efficiently guide the discovery of new molecules, we must first systematically de-orphan emergent gene clusters from genomes. Here we provide to our knowledge the first comprehensive retro-biosynthetic program, generalized retro-biosynthetic assembly prediction engine (GRAPE), for PK and NRP families and introduce a computational pipeline, global alignment for natural products cheminformatics (GARLIC), to uncover how observed biosynthetic gene clusters relate to known molecules, leading to the identification of gene clusters that encode new molecules.

Microorganisms craft a wide range of small molecules from modular assembly lines, such as PK synthases (PKSs) and NRP synthetases (NRPSs), which are intrinsically capable of creating unique molecular architectures. After the discovery of penicillin, bioactivity-guided fractionation and screening of microbial cultures has revealed >11,000 PK and NRP products (**Fig. 1**). Genome sequencing efforts are uncovering PKS-NRPS gene clusters at an unprecedented rate, but only a relatively small portion of these (~10%) have been associated with known products (**Fig. 1**). This disparity suggests that a number of new gene clusters encode known molecules, but many others likely produce valuable new natural products[1–3].

The initial discovery of the erythromycin gene cluster provided a prime example of how Nature uses modular biosynthetic logic to craft bioactive molecules[4]. Knowledge of the biosynthetic origins of NRP and PK molecules has been thoroughly confirmed by gene knockout studies, which have matched ~568 PK and NRP products to their respective gene clusters[5]. The localized nature of biosynthetic genes also expedited rigorous enzymology studies that defined the unifying principles and specialized reactions in NRP and PK systems[6–8]. Now, next-generation sequencing is exponentially accelerating the rate of gene-cluster discovery, revealing active biosynthetic loci as well as 'cryptic', seemingly silent clusters. Several of these cryptic clusters are not entirely silent, and have been shown to possess minimal or conditional activity, yielding low-abundance bioactive products[9–12]. With this in mind, we must now decide how best to focus future efforts to maximize the discovery of new compounds. By leveraging both classical natural-product chemistry and our knowledge of biosynthesis, we can now develop cutting-edge bioinformatic algorithms to determine which gene clusters produce the >11,000 known NRPs and PKs, and which will yield highly valuable new molecules.

Natural-product genome mining offers tools to identify clusters, but lacks a means to differentiate those encoding known versus new products[3]. Increased computational accuracy that can more closely emulate Nature has been illustrated in recent updates to AntiSMASH[13] and the highly accurate 'prediction informatics for secondary metabolomes' (PRISM) engine[14]. Still, predicting natural products from gene clusters remains a challenge because of frequent deviations in colinearity principles (order of genes and modular enzymes to products) and difficulty inferring reactions (e.g., regio-chemistry) from genes[15,16]. For other natural biomolecules (e.g., proteins), the genetic code translates effectively, and algorithms such as basic local alignment search tool (BLAST) readily define relationships and relatedness[17], enabling focused investigations. Applying these principles to small molecules would provide a unique tool to rapidly assess the novelty of downstream natural products, allowing efforts to be focused on new gene clusters with medical and industrial biosynthetic potential. Here we present a pipeline that links gene clusters to known natural products and defines clusters encoding new compounds based on a retro-biosynthesis tool (GRAPE) and an alignment algorithm (GARLIC) for PK and NRP small molecules (http://www.magarveylab.ca/garlic).

RESULTS

Predicting natural-product building blocks

Molecular studies have provided insight into the biosynthetic transformations and catalysts that promote PK and NRP biosynthesis. Advanced bioinformatic algorithms have likewise increased our capacity to identify biosynthetic clusters. Previously, we had defined a new bio- and chemo-informatic platform, PRISM (http://www.magarveylab.ca/prism), which uses a catalog of hidden Markov models (HMMs)[14,18] in an attempt to better catalog PK and NRP biosynthetic reactions. Here we provide the predictive capacity of PRISM for PK and NRP building blocks across 171 test clusters to define its accuracy for proteinogenic amino acids (93% accurate across 393 test cases), non-proteinogenic amino acids (94% accurate across 115 test cases), and PK acyltransferase (AT) domain substrates (74% based on 383 randomly selected AT domains; **Supplementary Results**, **Supplementary Table 1a**). PRISM also identifies 257 building blocks added to NRP and PK scaffolds[14], including sulfurs, hydroxyl functionalities, formyl and methyl groups, halogens (100% accuracy across 30 clusters), fatty acyl units (100% for 20 clusters

[1]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada. [2]Department of Chemistry and Chemical Biology, M. G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada. [3]These authors contributed equally to this work. *e-mail: magarv@mcmaster.ca

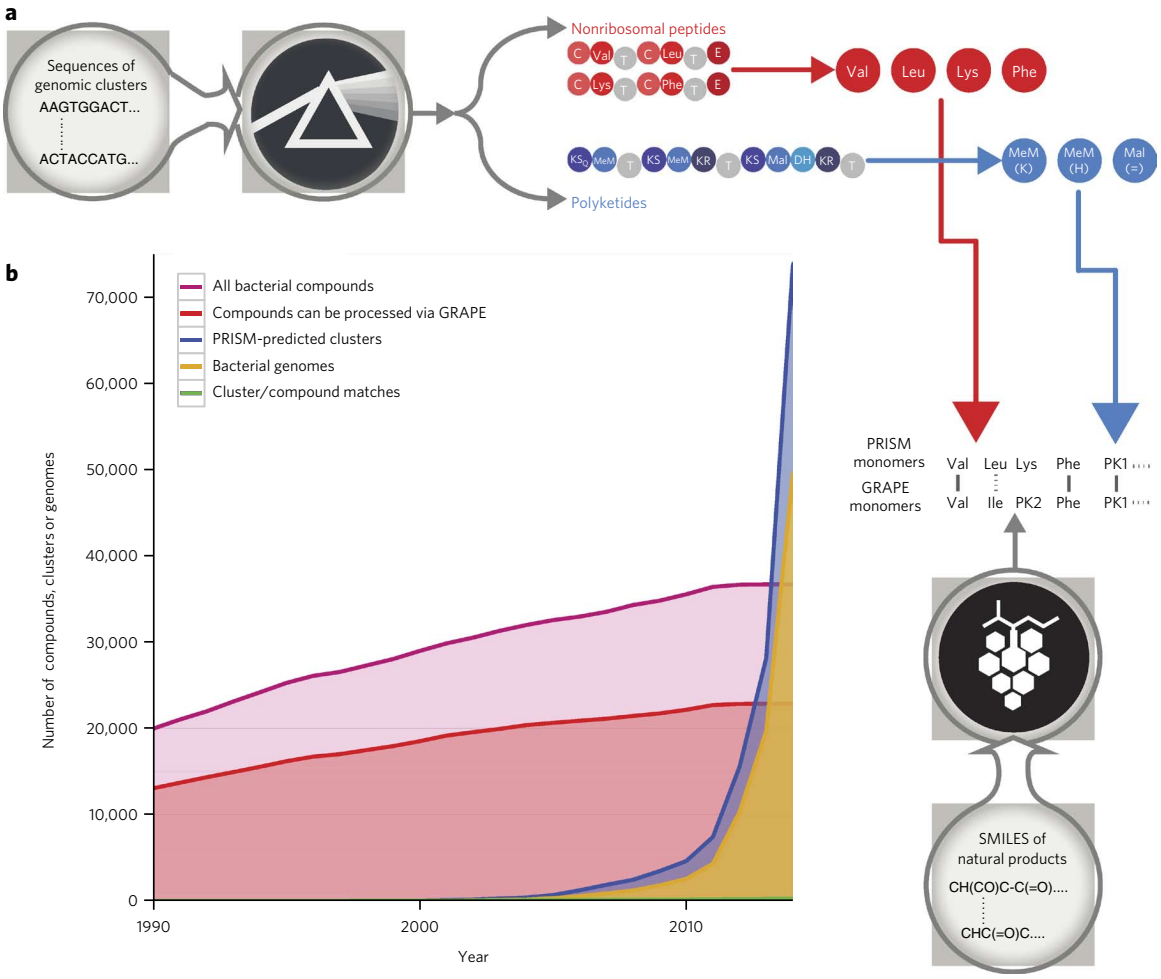

Figure 1 | Historical perspective of microbial polyketide and nonribosomal peptide natural-product discovery and associated genetic information. (**a**) A pipeline created to match unknown gene clusters with known natural products: PRISM takes in genetic information to infer assembly line monomers and tailoring enzymes, and GRAPE takes in small molecules to produce analogous information, which can be compared. (**b**) Comparison of natural-product discovery with sequencing rates of gene clusters and genomes from 1990 to 2015. The compounds in purple are bacterial natural products. The compounds in red are microbial natural products that can be processed in GRAPE. The genetic sequences in yellow are nucleotide sequences that are over 100 kb, from the US National Center for Biotechnology Information (NCBI) database. In blue are the number of clusters identified via PRISM using all nucleotide sequences that are over 100 kb, and the predicted products can be processed in GRAPE. Matches in green are all known natural products with known biosynthetic clusters.

tested) and sugar molecules (64% accuracy of the correct sugar over 30 clusters)[14]. In the cases of deoxysugars, PRISM infers the numbers of sugars placed on a given PK-NRP scaffold from the number of associated glycosyltransferases (GTs), and determines their identities using annotated sugar genes, facilitating monomer prediction. PRISM detects 5 dicarboxylic acid substrates, 132 fatty acids, 15 hydroxy acid substrates, 47 amino acids and several unique non-assembly line features, including 12 tailoring domains that are likewise detected via GRAPE (**Supplementary Data Set**). PRISM can also detect genes associated with minimal PKS, polyketide chain length and cyclization of type II polyketides and enediynes to predict their scaffolds[14]. For *trans*-acting acyltransferase domain-containing type I PKS, PRISM applies a recursive algorithm to insert acyltransferase into the open reading frame to form a PKS architecture similar to *cis*-acyltansferase PKS[14]. Detailed discussion regarding PRISM analysis of type II polyketides, enediynes and type I PKS with *trans*-acting acyltransferase has been published previously[14].

Retro-biosynthetic analysis of PK and NRP core tailorings

Systematic retro-biosynthesis of microbial PKs and NRPs requires a strategy to reverse the various ring patterns, heterocycles and other backbone elaborations (**Fig. 2** and **Supplementary Fig. 1**). To this end, we developed GRAPE (**Fig. 2**). GRAPE uses SMILES[19] structures as inputs, and uses protocols from the chemistry development toolkit (CDK)[20] to identify valences and bonds in order to perform theoretical deconstruction of PK and NRP structures. We designed the deconstruction processes of GRAPE with a series of retro-biosynthetic reactions to handle the exceptional complexity of PK and NRP molecules, including well-known structures, such as vancomycin, penicillin and erythromycin, as well as heavily tailored structures (kendomycin, anthramycin and avermectin) (**Fig. 2**, and **Supplementary Tables 3** and **4**). By identifying predictable moieties and functional groups, GRAPE can leverage our understanding of biosynthesis to reverse each reaction and reach the core components generated by an assembly-line enzyme.

In addition to backbone generation, amides of peptidic natural products are often modified during synthesis, including N-, O- and C-methylations, imines, thioesters, esters and heterocycles (oxazoles, thiazoles, thiazolines and thiazolidines) (**Supplementary Table 1b**). GRAPE theoretically reverses these modifications and annotates the subsequent fragment with the reversed tailoring (**Supplementary Table 4**). Other prospective tailorings occur after scaffold assembly, such as β-lactam ring formation, prenylation, halogenation, sulfation,

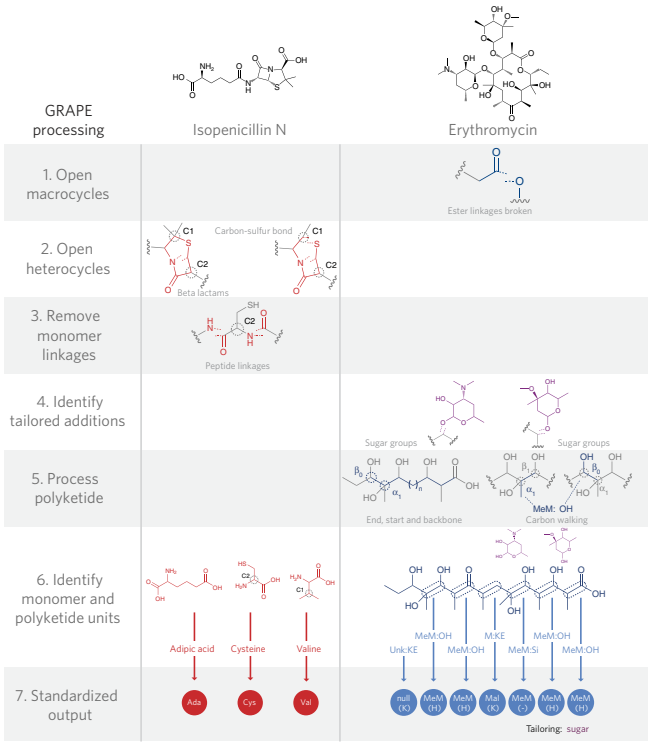

Figure 2 | Workflow of GRAPE. Two distinct natural products: isopenicillin, a β-lactam, and erythromycin, a macrolide, broken down via GRAPE. The overall GRAPE process in described under the chemical structures. Red and blue outputs are amino acid and polyketide monomer blocks, respectively. For detailed GRAPE overview, see **Supplementary Figure 1**. For retro-biosynthetic reactions used in GRAPE, see **Supplementary Tables 1** and **2**. For identification of amino acids as acyl keto-extension units and PK substrates, respectively, see **Supplementary Figures 2** and **3**. For full names of the abbreviations, see the **Supplementary Data Set**.

epoxidation, hydroxylation, and bi-aryl and disulfide formation, all of which GRAPE documents (**Supplementary Tables 3** and **4**). GRAPE uses maximum common substructure (MCS)[21] matching to identify liberated fragments and monomers by comparison with an annotated library of PK and NRP monomer and tailoring units, including modified and unmodified amino acids (297), singlet and doublet PK fragments (7), acyl adenylating units (67), fatty acids (132), uncategorized monomers (69) and sugars (71) (**Supplementary Figs. 1** and **2**). GRAPE also identifies oxygen, nitrogen, and carbon-linked hexose and deoxysugars, which are subsequently removed and logged for MCS searches. All remaining unknown fragments are then scanned with a fatty-acid-determining algorithm that has two filters. The first filter checks whether the fragment has a carboxylic acid and no other elements other than carbon or hydrogen for the remainder of the molecule, unless it is an oxygen on the γ-carbon. The second filter checks for the presence of a linear, nonbranching, saturated chain of at least four carbons. If the fragment passes both filters, it is deemed a fatty acid.

Retro-biosynthetic analysis of complex PK and PK-NRP hybrids

After MCS analysis is complete on all of the fragments, GRAPE identifies amino acids, acyl adenylating units, fatty acids, sugars and biosynthetically uncategorized monomers. Remaining are three general groups of fragments, all of which may or may not be polyketide-related fragments: ketide-extended hybrids, fatty acyl chains and standard polyketides. Remaining fragments that contain amine and carboxylic acid groups are potentially ketide unit extended hybrids. To analyze those, GRAPE identifies the longest carbon-only chain from the α-carbon of the amine to the carboxylate carbon of the furthest carboxylic acid (**Supplementary Fig. 3**). If the carbon chain has an odd number of carbons, the ketide-extended amino acid is identified as a β-amino acid. The bond between γ-carbon and δ-carbon is then theoretically broken and the γ-carbon is converted to a carboxylic acid to create the β-amino acid. If the carbon chain has an even number of carbons, the keto-extended amino acid is identified as an α-amino acid. The bond between the corresponding β-carbon and γ-carbon is then theoretically broken, and the β-carbon is converted to a carboxylic acid to create the α-amino acid. The amino acid fragment is reanalyzed by MCS to determine the exact amino acid, and the remaining polyketide fragment is then analyzed for its monomers.

To determine PK monomers, GRAPE reveals the longest carbon-only chain, starting from a carboxylate carbon, and predicts this to be the PK backbone. If the chain contains an even number of carbons, the second furthest carbon from the carboxylate carbon is selected as the biosynthetic starting β-carbon. In the case of an odd number of carbons in a chain, the furthest carbon from the carboxylate carbon is selected as the biosynthetic starting β-carbon. In an iterative analysis, two carbon atoms from the backbone, marked as β and α, are processed at a time until the entire backbone is analyzed. The α-carbon chemical environments are used by GRAPE to derive the biosynthetic dicarboxylic acid that would have been selected by the PKS AT domain in biosynthesis. For instance, if the α-carbon has a hydrogen, CH_3, OCH_3 or CH_2CH_3, this infers malonate (Mal), methylmalonate (MeMal), methoxymalonate or ethylmalonate, respectively (**Supplementary Fig. 2**). Similarly, β-carbon chemical environments define the oxidative status of β-ketone. Salient features to discern a PK chain from fatty acyl chain are also considered after the polyketide prediction is complete. If the majority of the β-carbons are fully saturated, the fragment is ambiguously labeled as a fatty acid or a polyketide, as its biosynthetic loading modules cannot be definitively determined based on structure.

Type I PKs can include complex post-assembly line cyclizations, which can occasionally remove the generalized predictability of a carboxylate (end carbon) or the start carbon, so predicting the PK scaffold is not possible without reverting these post-assembly modifications. Building on established biosynthetic paradigms, GRAPE includes a series of retro-biosynthetic operations to process these challenging structures (**Supplementary Table 3**). A number of other, more esoteric chemistries are also processed and recorded (**Supplementary Tables 3** and **4**). For instance, ether-containing rings, including polyethers such as monensin, are theoretically opened and attached to the appropriate atoms depending on which of the carbons is a β-carbon, in the case of even-numbered carbon rings. With an odd number of carbons, it is not possible to infer which carbon had the hydroxyl group and which had the ketone, so both are inserted as potential states at that site (**Supplementary Table 3**).

Having constructed breakage rules that cover a wide spectrum of chemistries found in the PK and NRP family, we next tested GRAPE on several sample molecules. Outputs of the molecules erythromycin, thiocoraline, ML-449, salinosporimide, cephalosporin, penicillin, nocardicin, chivosazole A, curacin, bleomycin, kendomycin, avermectin, piercidin, anthramycin, eponemycin, monensin, mupirocin, arthrofactin, mycoplanecin, SW163 C, vancomycin, A-47934, apoptolidin and yersiniabactin are shown in **Supplementary Tables 3** and **4**. In addition to the chemistries above, which are synthesized from multimodular assembly lines, GRAPE also includes a strategy for type II and enediyne polyketides. In these instances, GRAPE first removes tailorings and, because they are generated iteratively, it uses the entire core scaffold for matching. The GRAPE-derived skeletons are then compared with a repository of scaffolds for each type that has been compiled, and a substructure search is done on the query compound before it is further broken down by GRAPE. If the scaffold is found to be in the query compound, it is

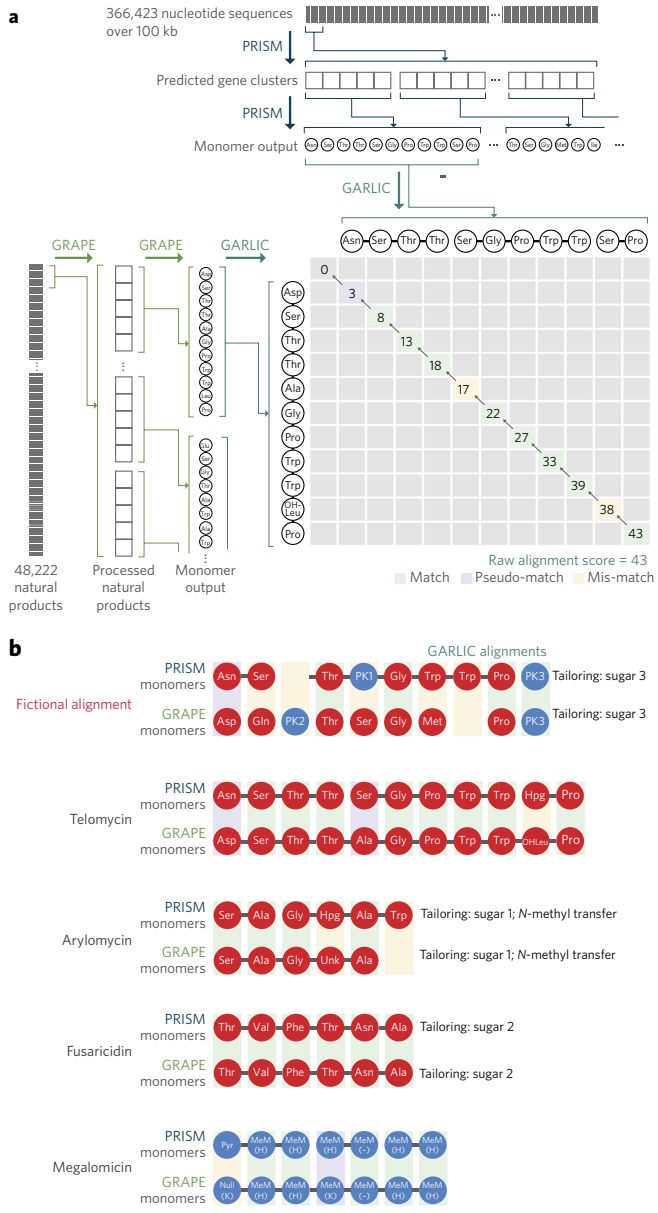

Figure 3 | Matching algorithm of GARLIC and examples of natural products matched. (**a**) Matching algorithm of GARLIC between PRISM and GRAPE outputs. The genetic information used in PRISM consists of nucleotide sequences that are over 100 kb from the NCBI database. Compounds used in GRAPE are microbial natural products from an in-house database. (**b**) Examples of natural products (telomycin, arylomycin, fusaricidin and megalomicin) matched between PRISM and GRAPE. For details of scoring configuration, see **Supplementary Table 5**. For comparison of different scoring methods, see **Supplementary Figure 4**. For full names of the abbreviations, see the **Supplementary Data Set**.

then labeled as that type of molecule. Analogous to the sugar example above, a collection of genes is ascribed to each type II aromatic and enediyne scaffold type, and these gene sets are outputted for comparison with PRISM (**Supplementary Table 6**).

Connecting natural products to orphaned gene clusters

To link biosynthetic clusters and their products, we created a program called GARLIC to align monomers from cluster predictions and small molecule breakdowns (**Fig. 3a**). Alignment algorithms are used to match other biomolecules, such as nucleic acids and proteins, often with query searches through a database of subjects, and generally are DNA on DNA or protein on protein. For natural PK and NRP biomolecules, this requires an added dimension of comparing gene clusters (DNA) to final products (small molecules), and with it are several unique challenges that hinder algorithm development. These range from degeneracy of the code (from protein to small molecule), inconsistent colinearity (from gene to small molecule) and the wider spectrum of comparable traits/monomer blocks (fatty acyl units, sugars, amino acids and carboxylic acids). In nucleic acid- and protein-based alignment, the numbers of monomers are relatively small, being 4 and 20 residues, respectively. According to our GRAPE analysis, within the PK and NRP realm are 20 PK monomers (five substrates, each with four possible oxidation states), 47 amino acids and a collection of tailoring substrates (e.g., sugars) and fatty acids, to name a few. Moreover, code degeneracy and colinearity for PK and NRP systems leads to many error possibilities (differential substrates incorporated and ordering of incorporation), which is different from the genetic code degeneracy (e.g., multiple tRNA species for the same amino acid) and defined colinearity from DNA to RNA to protein. PKS and NRPS small molecule machineries are well noted for their skipping, stuttering and code degeneracy, leading to errors and numerous nuances in colinearity[22,23]. To better communicate the complexity of PKS and NRPS colinearity and the considerations necessary to relate them to PK and NRP small molecules, the following scenario is provided. For a gene cluster whose final product contains blocks denoted 'ABCDE', there are a number of derivations. In one instance, ABC may be encoded by one gene, with the others separate, represented as 'ABC-D-E' (dashes indicate different genes). The order of ABC monomers may be fixed in this case as they are encoded in the same open reading frame (ORF), but the order of the others, D and E, cannot be assumed, as a series of combinations are plausible: D-E-ABC, E-D-ABC, E-ABC-D, etc. The number of permutations increase the more dissociated the modules are across ORFs, all of which would relate to the same compound containing A, B, C, D and E blocks. If there are eight monomers encoded by eight separate genes, the number of permutations is eight factorial or 40,320.

As it is computationally prohibitive to align every possible permutation, GARLIC takes a random sample of permutations using the Fisher-Yates shuffle[24]. Each permutation is scored and the top-scoring subset is retained; each remaining permutation is used as a seed to create new permutations, by swapping the positions of two ORFs picked randomly, per permutation. This process is repeated several times, approaching the optimal alignment without searching the entire permutation space. The final score is then determined by the highest-scoring alignment from the permutations. For this analysis, when there were six or fewer ORFs (720 permutations), all permutations were considered and aligned to GRAPE outputs; when there were more than six ORFs, the heuristic refinement method stated previously was used.

Multiple parameters of the alignment algorithm demand consideration, given the diversity of monomers and assembly line alterations, as well as the difficulty translating from genes to small molecule blocks. In total, we developed 26 different parameters that could be considered and integrated them into alignment schemes (**Supplementary Table 5**). Some of this included weighting based on the known distribution of monomers found in NRP and PK molecules. Surveying the known PK and NRP chemical space via GRAPE revealed that 6.7% of all amino acids are non-proteinogenic, 62.4% are proteinogenic but not aromatic, and 30.9% are aromatic. In PKSs, malonate is the most widely incorporated (70.8%), followed by methylmalonate (23.7%), leaving the more rare units at 5.5%. Other parameters for consideration include development of potential scoring based on the accuracy of PRISM predictions for the respective substrates and tailorings. Examples of these include hydroxylases, chlorinases, sulfotransferases and different sugar types. To match liberated sugar components from

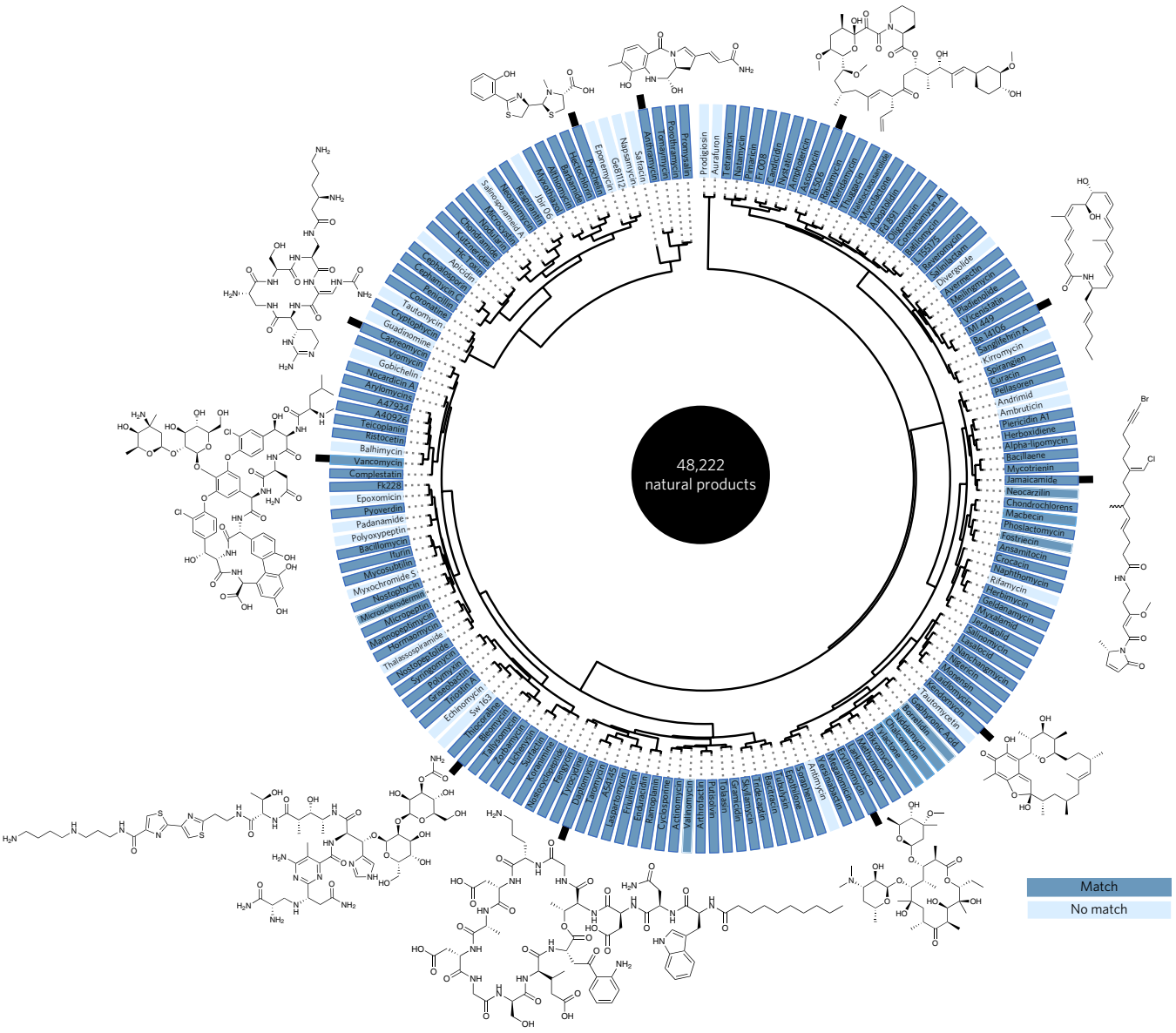

Figure 4 | Matching results of the test clusters (171 NRP, type 1 PK and hybrid PK-NRP) to in-house compound database (48,222 microbial natural products that can be processed via GRAPE). There were 144 direct matches (dark blue) and 27 'no matches' (light blue), which comprised 84% and 16% of the test clusters, respectively. Structures of 11 compounds (FK506, BE14106, jamaicamide, kendomycin, erythromycin, daptomycin, bleomycin, vancomycin, capreomycin, pyochelin and anthramycin) are provided as examples. For GARLIC matching results of type 2 PK and enediynes, see **Supplementary Figure 5**.

GRAPE, we constructed a sugar gene repository that contains a list of possible genes for each sugar from GRAPE to directly match to PRISM's predicted sugar genes (**Supplementary Table 2**). Following theoretical reversal of the tailoring reactions, results of the liberated monomer matches are recorded and output in a format consistent with PRISM gene cluster prediction outputs, facilitating cross-platform comparisons that correlate deconstructed molecules with their corresponding biosynthetic gene clusters (**Fig. 3b**, and **Supplementary Tables 1** and **2**).

Two types of alignment algorithms are commonly used, depending on the data domain and the purpose of the alignment: Smith-Waterman (local)[25] or Needleman-Wunsch (global)[26]. Local alignments may consider only a subset of the original sequences to produce an optimal score, whereas global alignments are often deployed when one wishes to define the likeness between two sequences as a whole. Given the challenges of converting data from genome to small molecules, we created a test set with 171 diverse PKs and NRPs having annotated gene clusters. To conduct an unbiased analysis, we considered these structures and their respective GRAPE breakdowns with all of the other known NRP and PK structures from our database of 48,222 compounds. To assist in establishing cluster matches, we developed a relative scoring metric, where the final score was derived as a fraction between a score of a given GARLIC alignment and the score of the cluster matched to itself.

For each of the 171 biosynthetic gene clusters, we performed GARLIC scoring under various algorithm configurations against each of the compounds, including both compounds made by the gene clusters and all others from the database. We generated a number of criteria and through continuous empirical testing, developed a refined algorithm that tuned each of the given parameters (**Supplementary Table 5**). We performed this analysis on seven algorithm configurations, including local and global alignment under basic scoring schemes, and global alignment based on a scoring scheme that factored in the above listed distribution of monomers in PK and NRP products, with heightened scores for

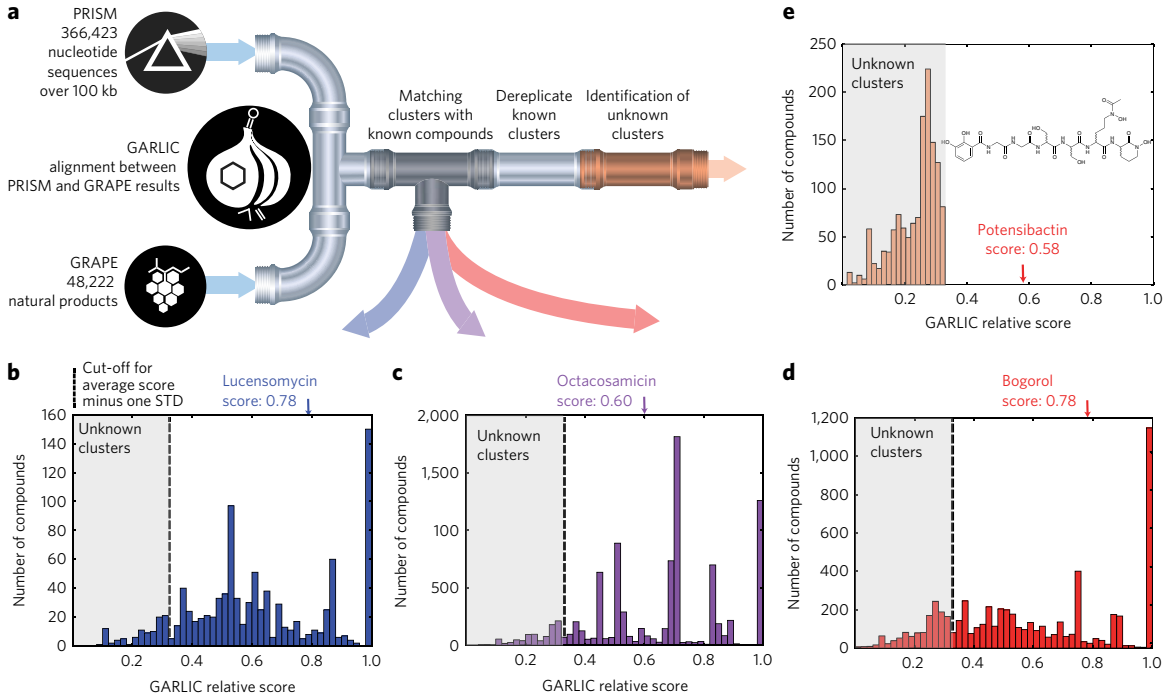

Figure 5 | Discovery of gene clusters for orphaned microbial natural products, and identification of new microbial natural products using the GARLIC pipeline. (**a**) The visualization of GARLIC workflow. A screenshot of the front end of GARLIC is in **Supplementary Figure 6**. (**b**–**d**) The relative scores of potential true matches of PRISM predicted clusters of PK, NRP and hybrid PK/NRP from over 300,000 sequences with a length of >100 kb (the unique top ten GARLIC-matched known compounds of all clusters with at least four modules) are plotted. Compounds with previously unknown biosynthetic clusters, lucensomycin, octacosamicin and bogorol, were identified in the PK, hybrid NRP-PK and NRP chemical space, using GARLIC cluster hits. The production of these compounds was confirmed via liquid chromatography-mass spectrometry (LC-MS). GARLIC and LC-MS data for the three de-orphaned compounds are in **Supplementary Figures 11**, **12** and **14**. (**e**) Unknown NRPs predicted through GARLIC (clusters identified as NRP with at least four modules and a GARLIC score lower than 0.33) are plotted. A new microbial NRP, potensibactin, was identified using GARLIC. GARLIC and LC-MS data for potensibactin are in **Supplementary Figure 16**.

rare or uncommon blocks. The algorithm and scoring configurations are given in **Supplementary Table 3**, and results are shown in **Supplementary Figure 4**. Though the global alignment outperformed the local alignment with basic scoring schemes, we noted improvements when the scoring was refined. Particularly important refinements were the sugar-gene matching score, the introduction of partial matching for amino acids, and bonus scores for rare amino acids and polyketide substrates. Our refinements to the scoring scheme based on biological knowledge gained from previously mentioned GRAPE monomer analysis and PRISM accuracy measurements led to marginal improvements. Additional score refinements of information external to the biosynthetic assembly line, such as sugars and chlorinations, further increased the number of correct matches.

To obtain the final scoring, we optimized the parameters using Powell's method[27], an automated method similar to the heuristic empirical tuning approach used previously. For a training set, we excluded 4 of the 171 clusters because their close relation to other clusters or errors in either GRAPE or PRISM would reduce the applicability of GARLIC outside the training set. We scored the effectiveness of each set of parameters according to how well GARLIC ranked the correct compound for each cluster. We used 167 clusters matched to their correct compound, and included 297 selected decoys to represent various families of compounds, using our empirically refined parameters as a starting point.

Overall, the final algorithm matched 144 of the 167 (84%) clusters in the top five alignments from the entire compound database. Those that did not match tended to have PRISM outputs that had predicted an incorrect substrate. The diversity of the test set's clusters, along with those that were successfully matched using the final scoring scheme are represented in **Figure 4**. Of the 144 matched clusters, 95% of the clusters had a final score that is higher than or equal to 0.33, which is 1 s.d. below the average final score (GF). We also determined the average final score for each compound class: 0.70 for PK, 0.71 for NRP and 0.55 for PK-NRP. Our program works on type II aromatic PK and enediynes as well, with the matches represented in **Supplementary Figure 5**.

To validate the algorithm and ensure the results were not due to overfitting to each cluster, we performed a leave-one-out analysis. We derived new parameter scoring systems from the Powell method by removing each cluster in turn from the training set, generating 167 scoring schemes in all. To eliminate potential bias, we started each optimization from the original basic scoring scheme, as opposed to the empirically derived one. Even though starting from the latter scoring yielded better results, we developed it using the 167 clusters and thus could not use it for this analysis. When we tested each cluster against the scoring scheme learned in its absence against the database of compounds, 110 of the 167 gene clusters (66%) correctly matched by highest score to its associated compound, and 125 (75%) matched in the top five. As the leave-one-out analysis performed similarly to our final algorithm, and as our training set was diverse (**Fig. 4**), we would expect GARLIC to exhibit comparable accuracy on any supplied biosynthetic gene cluster.

Matching clusters to known and new natural products

To build a comprehensive collection of bacterial NRP and PK biosynthetic gene clusters, we developed a script to extract and profile microbial genomes (both from NCBI and our internal library) using PRISM. We analyzed over 300,000 sequences with a length of >100 kilobases (kb), leading to the identification of 16,831 potentially

completely assembled PKS and NRPS clusters. We defined complete clusters as those possessing at least four modules and having 20 kb flanking both ends of the cluster (1,018 PK, 6,351 NRP and 9,462 hybrid PK-NRP gene clusters). To assess which gene clusters corresponded to known compounds, we ran PRISM results through GARLIC and obtained final matching scores for known products. Included in these results was a series of dereplicated gene clusters that were previously related to natural products, including accurate matches for telomycin[28] (final score: 0.72; second highest score: 0.49), acidobactin[9] (final score: 0.61; second highest score: 0.24), thanamycin[9,29] (final score: 0.68; second highest score: 0.56) and potensimicin[9] (final score: 0.77; second highest score: 0.73), further extending beyond the 144 listed above (**Supplementary Figs. 7–10**). Other candidate gene clusters to compound matches were revealed that had not previously been identified from this metagenomic data set (**Fig. 5**). Using the final score averages to the true positive matches as a guide (see above), we focused on compounds that were above these scores as representative candidates from each class: PK, PK-NRP and NRP. Moreover, we did not use strains previously documented in the literature to produce a given molecule. We cultivated the strains in the laboratory and processed them to generate crude extracts. For all cases we properly identified the prospective candidate, and assigned *Streptomyces achromogenes* NRRL 3125 as a producer of the PK lucensomycin (final score: 0.78; second highest score: 0.77)[30], *Amycolatopsis sp.* NAM 50 as a producer of the PK-NRP octacosamicin (final score: 0.60; second highest score: 0.49)[31], *Brevibacillus laterosporus* DSM 25 as a producer of the NRP bogorol (final score: 0.78; second highest score: 0.25)[32] and tauramamide[33] (**Supplementary Figs. 11–14**, respectively; **Supplementary Note**). Additional evidence for cluster matching is provided in **Supplementary Figure 15** and **Supplementary Table 7**.

As 95% of known compounds matched to known clusters in our test set with a final score of 0.33 or higher (see above), we can hypothesize that any full clusters with the top match score lower than 0.33 are likely to code for novel compounds. Among the 16,831 full clusters, 2,532 (15%) had scores < 0.33, suggesting they code for novel products. As an initial demonstration, we selected a cluster from the potensimicin producer *Nocardiopsis potens* DSM 45234 (ref. 9) that had a low match score (0.27) to all known compounds. Using metabolomic profiling through the genome to natural products platform (GNP)[9], we identified the orphan metabolite based in part on its prediction (**Supplementary Fig. 16**). We structurally characterized the isolated product, determined that it was a new natural product, and named it potensibactin (**Supplementary Fig. 16** and **Supplementary Note**). Loading the potensibactin structure into the identified natural-product database and re-running GARLIC validated this product as a match for the deorphaned cluster (final score of 0.58).

DISCUSSION

Since the dawn of the 'golden age of antibiotics', much effort has been taken to collect and solve the structures of microbial natural small molecules. Success in these efforts is illustrated by the vast caches of products isolated and the activities of numerous agents that have been determined and are currently used in medical and biotechnological applications. Genomics is a more recent addition to the natural-product workflow, and is assisting with how we define microorganisms with biosynthetic potential and the overall distribution of PK and NRP clusters in microbial genomes. Reconciling the rapidly expanding amount of genomic data with natural-product chemistry is now a requirement in order to define the next frontiers for pursuing natural small molecules. Further, the sequenced genomes that we currently have are biased toward the "classical" producers of bacterial natural products (actinomycetes) and pathogenic bacteria (*Escherichia coli* and *Pseudomonas* spp.), which is not the true representation of the genomic spaces. Therefore, the potential for novel cluster discovery would be much higher if we have more diversity in the sequenced genomes. Merging such tools with other developments in comparative metabolomics should now make it achievable to construct targeted libraries of strictly new microbial natural products, bypassing the challenges of dereplication in bioactivity-guided fractionation.

Here we present the unified tools of gene cluster prediction (PRISM), known natural-product retro-biosynthesis (GRAPE) and alignment processes (GARLIC) that work as a pipeline to define new clusters and those for known compounds. GRAPE has limitations with regard to a number of post-assembly line modifications, such as carbon deletions via Favorskii rearrangements, cyclization through Diels-Alder reactions and decarboxylation of the terminal carboxylic acid. There are also issues when a single fragment contains multiple carboxylic acids once polyketide prediction begins, as it is currently not possible to know for certain which carboxylic acid defines the end of the polyketide extension. As more information becomes available for the different reaction types, however, additional logic will be added to GRAPE's process. A central reasoning for developing the GARLIC platform and creating the retro-biosynthetic analysis was to define assembly lines that code for new PK or NRP molecules. The development of a scoring metric for the system enabled us to also define examples of assembly lines that have a high likelihood for encoding new molecules. The GARLIC algorithm is free for public use (accessible via http://www.magarveylab.ca/garlic), which allows for the comparison of clusters to small molecules through the upload of PRISM results or through manual input of scaffold and tailoring information obtained from other sources. The new clusters and their encoded products may be the next step for medicine to develop effective new agents, particularly in this current era of antibiotic resistance.

Received 9 November 2015; accepted 20 July 2016;
published online 3 October 2016

METHODS

Methods and any associated references are available in the online version of the paper.

References

1. Doroghazi, J.R. *et al.* A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
2. Jensen, P.R., Chavarria, K.L., Fenical, W., Moore, B.S. & Ziemert, N. Challenges and triumphs to genomics-based natural product discovery. *J. Ind. Microbiol. Biotechnol.* **41**, 203–209 (2014).
3. Rutledge, P.J. & Challis, G.L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.* **13**, 509–523 (2015).
4. Weber, J.M. *et al.* Organization of a cluster of erythromycin genes in *Saccharopolyspora erythraea*. *J. Bacteriol.* **172**, 2372–2383 (1990).
5. Medema, M.H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
6. Fischbach, M.A. & Walsh, C.T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468–3496 (2006).
7. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.* **48**, 4688–4716 (2009).
8. Walsh, C.T., O'Brien, R.V. & Khosla, C. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. *Angew. Chem. Int. Ed. Engl.* **52**, 7098–7124 (2013).
9. Johnston, C.W. *et al.* An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat. Commun.* **6**, 8421 (2015).
10. Vizcaino, M.I. & Crawford, J.M. The colibactin warhead crosslinks DNA. *Nat. Chem.* **7**, 411–417 (2015).
11. Lincke, T., Behnken, S., Ishida, K., Roth, M. & Hertweck, C. Closthioamide: an unprecedented polythioamide antibiotic from the strictly anaerobic bacterium *Clostridium cellulolyticum*. *Angew. Chem. Int. Ed. Engl.* **49**, 2011–2013 (2010).

12. Ishida, K., Lincke, T., Behnken, S. & Hertweck, C. Induced biosynthesis of cryptic polyketide metabolites in a *Burkholderia thailandensis* quorum sensing mutant. *J. Am. Chem. Soc.* **132**, 13966–13968 (2010).
13. 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).
14. Skinnider, M.A. *et al.* Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.* **43**, 9645–9662 (2015).
15. Bachmann, B.O. Biosynthesis: is it time to go retro? *Nat. Chem. Biol.* **6**, 390–393 (2010).
16. Bachmann, B.O. & Ravel, J. Complex enzymes in microbial natural product biosynthesis, part A: overview articles and peptides. in *Methods in Enzymology* **458**, 181–217 (Academic Press, 2009).
17. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
18. Khayatt, B.I., Overmars, L., Siezen, R.J. & Francke, C. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One* **8**, e62136 (2013).
19. Anderson, E., Veith, G.D. & Weininger, D. SMILES: a line notation and computerized interpreter for chemical structures. Report No. EPA/600/M-87/021 (US Environmental Protection Agency Environmental Research Laboratory-Duluth, 1987).
20. Steinbeck, C. *et al.* Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **12**, 2111–2120 (2006).
21. Rahman, S.A., Bashton, M., Holliday, G.L., Schrader, R. & Thornton, J.M. Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.* **1**, 12 (2009).
22. Callahan, B., Thattai, M. & Shraiman, B.I. Emergent gene order in a model of modular polyketide synthases. *Proc. Natl. Acad. Sci. USA* **106**, 19410–19415 (2009).
23. Challis, G.L. & Naismith, J.H. Structural aspects of non-ribosomal peptide biosynthesis. *Curr. Opin. Struct. Biol.* **14**, 748–756 (2004).
24. Fisher, R.A.Y. *Frank Statistical Tables for Biological, Agricultural and Medical Research* 3rd edn. (Oliver & Boyd, London, 1948).
25. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
26. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
27. Powell, M.J.D. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comput. J.* **7**, 155–162 (1964).
28. Sheehan, J.C., Mania, D., Nakamura, S., Stock, J.A. & Maeda, K. The structure of telomycin. *J. Am. Chem. Soc.* **90**, 462–470 (1968).
29. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. USA* **109**, E1743–E1752 (2012).
30. Gaudiano, G., Bravo, P. & Qiulico, A. The structure of lucensomycin. Part I. *Tetrahedr. Lett.* **7**, 3559–3565 (1966).
31. Dobashi, K., Naganawa, H., Takahashi, Y., Takita, T. & Takeuchi, T. Novel antifungal antibiotics octacosamicins A and B. II. The structure elucidation using various NMR spectroscopic methods. *J. Antibiot. (Tokyo)* **41**, 1533–1541 (1988).
32. Barsby, T., Kelly, M.T., Gagné, S.M. & Andersen, R.J. Bogorol A produced in culture by a marine *Bacillus sp.* reveals a novel template for cationic peptide antibiotics. *Org. Lett.* **3**, 437–440 (2001).
33. Desjardine, K. *et al.* Tauramamide, a lipopeptide antibiotic produced in culture by *Brevibacillus laterosporus* isolated from a marine habitat: structure elucidation and synthesis. *J. Nat. Prod.* **70**, 1850–1853 (2007).

Acknowledgments

This work was funded through an Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery grant (RGPIN 371576-2014) (N.A.M.) and a Joint Programme Initiative on Antimicrobial Resistance funded through the Canadian Institutes of Health Research (CIHR) (grant 138739) (N.A.M.). C.W.J. is funded through a CIHR Doctoral Research Award. N.A.M. is supported by the Canada Research Chairs Program (grant 950228183). We thank J. Cao for rendering trees, A. Luo for curating sugar genes and structures, and B. Furman for valuable communications.

Author contributions

C.A.D. and G.M.C. developed GRAPE and GARLIC, devised scoring strategies, contributed to study design, and wrote the manuscript. H.L. consulted on GRAPE and GARLIC's logic, devised scoring strategies, curated data sets, contributed to study design, and wrote the manuscript. C.W.J. isolated compounds and characterized structures. M.R.E. revised GARLIC, and designed and performed the optimization analysis of GARLIC scoring. M.A.S. developed PRISM, contributed to study design, and completed analysis for **Figure 1**. P.N.R. curated data sets, and contributed to study design. A.L.H.W. curated data sets. N.A.M. contributed to study design and wrote the manuscript.

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. Correspondence and requests for materials should be addressed to N.A.M.

ONLINE METHODS

PRISM analysis. We recently described PRISM, a Java web application designed to identify biosynthetic gene clusters in microbial genomes and predict the structures of genetically encoded secondary metabolites[14]. PRISM leverages a library of 498 HMMs to identify biosynthetic domains for modular, trans-acyltransferase, enediyne, and iterative type I polyketides, type II polyketides and nonribosomal peptides. Biosynthetic domains are grouped into gene clusters, and monomer units are predicted by analyzing identified adenylation, acyl-adenylating and acyltransferase domains with libraries of 66, 26 and 15 substrate-specific profile HMMs, respectively. To minimize false positives, a subset of rare monomers require the concurrent identification of one or more prerequisite domains in order to be conclusively identified, including methoxymalonate (3-hydroxyacyl-CoA dehydrogenase and acyl-CoA dehydrogenase), diaminopropionate (diaminopropionate synthase), capreomycidine (L-arginine hydroxylase and capreomycidine synthase), 3-hydroxypipecolic acid (cyclodeaminase and pipecolic acid hydroxylase), and 3-hydroxyanthranilic acid (tryptophan dioxygenase and aryl formamidase). A library of 54 tailoring reaction domains is implemented to account for the diverse set of enzymatic reactions that tailor the nascent natural product scaffold, including: C- and O-glycosylation, O-, N- and C-methylation, heterocyclization, macrocyclization, aromatization, oxidation/reduction, mono- and dioxygenation, Baeyer-Villiger rearrangement, halogenation, carbamoylation, sulfonation, amination and acyl group transfer. PRISM uses identified biosynthetic information to generate a combinatorial library of natural product structures that could putatively be produced by the biosynthetic gene cluster. This combinatorial scaffold library is not used to compare GRAPE and PRISM data via the GARLIC alignment algorithm, however; instead, monomer units and tailoring reactions identified by retro-biosynthesis or genomic prediction are aligned, while a series of substructure searches are performed to identify conserved biosynthetic scaffolds (e.g., in enediynes and type II polyketides; **Supplementary Table 6**), and subunits with variable sites of attachment (e.g., deoxysugars and acyl units; **Supplementary Table 2**).

GRAPE analysis. GRAPE was designed for the prediction of natural product bio-assembly based on chemical structure (**Fig. 2**). GRAPE was developed in the Java programming language using libraries from the Chemistry Development Kit[20], and the Maximum Common Subgraph (MCS) algorithm from the Small Molecule Subgraph Detector toolkit (SMSD)[21]. Given chemical structures as input in SMILES format[19], GRAPE performs matching against a database of scaffolds and then reverse biosynthetic chemical reactions. Each compound is first compared against scaffolds for nonmodular polyketides, enedyines and terpenes. This comparison checks whether any of these scaffolds are a substructure of the compound for initial classification. By rapidly annotating large chemical structural databases, with a focus on NRP and PK assembly units, GRAPE performs a collection of chemical reactions in reverse for each chemical structure, storing relevant biosynthetic information at each step (**Supplementary Tables 3** and **4**, and **Supplementary Fig. 1**). The output from GRAPE consists of two components: an ordered list of monomeric units that correspond to an NRPS or PKS assembly line, and a list of chemical features that correspond to non-assembly line biosynthetic enzymes (**Supplementary Table 1**).

The reverse biosynthetic chemical reactions fall under four major steps (**Fig. 2** and **Supplementary Fig. 1**). First, macrocycle-forming chemical bridges are reversed, such as disulfide bonds and ether linkages between aromatic rings. Second, reactions forming heterocyclic structures, such as thiazoles, oxazoles, penams and penems are performed in reverse. In the case of thiazoles and oxazoles, the atoms involved are tracked to output the predicted presence of a cyclization domain in the biosynthetic assembly line. In the case of penams and penems, the presence of these structures is stored for prediction of synthetic enzymes in the biosynthetic gene cluster. Rare chemistries, such as those found in kendomycin, avermectin and piericidin as well as di-cystine linkages, are scanned by substructure matching, and then their structures are reversed to pre-tailoring state. Third, core linking bonds, such as peptide bonds, thioesters and ester linkages, are reversed, and the connectivity and direction of peptide bonds are stored for the purpose of preserving order in the final output. If there is no ester bond in a cyclic molecule, then the starting monomer for biosynthetic alignments is considered unknown; otherwise the monomer with a N after the ester cleavage is considered the start. Fourth, additional added groups, such as sugars, sulfate groups, N and O methylations, and chlorines are detected, and their synthesis reactions are reversed.

After the reversals of these chemical reactions, the resulting monomeric chemical structures are identified as amino acids, fatty acids, sugars or polyketides. A list of known amino acids, hydroxy acids, fatty acids and sugars was curated, including substructures found in a PK and NRP products, and used to identify each monomeric structure. For structures with a hexose ring per the reverse biosynthetic analysis, these are identified as sugars, whereas remaining structures are compared to the curated list of sugar structures and identified if the maximum common substructure Tanimoto score from the MCS[21] is greater than a cutoff of 0.8. Structures still unclassified are compared to every known amino acid structure using the MCS and identified if the maximum common substructure Tanimoto score from the MCS is greater than a cutoff of 0.9. After this, any structures still unidentified are analyzed to determine if they are a potential polyketide or fatty acid. To this end, the remaining fragment is checked for a carboxylic acid, which is considered the biosynthetic end of a polyketide or fatty acid. Once the carboxylic acid carbon is determined, the longest carbon-only chain is determined. GRAPE then finds the shortest chain, with any atoms, between the start and end. The logic behind this is that all atoms that are not carbon (generally oxygen) or the bonds that create rings after the initial biosynthesis of the polyketide are made on this shortest path. By then checking along this path for any discrepancies between the carbon-only chain and the shortest chain where the ring is opened to be consistent with biosynthesis becomes clear. If there are multiple carboxylic acids, the one with the longest carbon-only chain is considered the biosynthetic end.

The program tracks from the newly found start C and checks that it is on a β carbon, or where the ketone is originally. This is done by ensuring that there is an odd number of carbons on the carbon-only chain; if the number is even, then the next closest carbon is chosen. GRAPE determines the state of the connected oxygen: ketone or hydroxyl. If no oxygen is present, it checks to see whether the carbon has a single or double bond. This information gives the putative domains for this single unit. If an epoxide is detected, as in mupirocin, it is removed and a double bond is left in its place, and this epoxide removal is recorded by the program (**Supplementary Table 3**). GRAPE also checks the chemical environment around each β carbon to determine whether there is an unusual cyclization event, such as those for avermectin, and reverses the chemistries (**Supplementary Table 3**). The α carbon is also checked for connected atoms to predict which initial substrate is being incorporated into the molecule. The program counts two carbons at a time on the backbone until it reaches the end. Specific chemistries are performed when the carbon in the backbone is determined to be in a ring structure. In monensin, for example, there are several rings with an oxygen that make up the entire molecule (**Supplementary Table 3**). As the program traverses the backbone, it opens these rings and attaches the appropriate atoms depending on which of the carbons is a beta carbon if there is an even number of carbons in the ring. If there is an odd number of a carbon in the ring, it is not possible to tell which carbon had the hydroxyl group and which had the ketone, so both are inserted as potential states at that site.

Owing to the chemical similarity of some fatty acids and polyketides, it is not always possible to infer from structure alone whether these structures are synthesized by a hybrid PK-NRP assembly line or through modifications to a fatty acid. To address this, structures are identified as fatty acids only if they contain clear identifying features, such as the presence of a saturated carbon chain and the carboxylic acid end of a saturated or 3-hydroxy fatty acid. Other structures from amino-acid-containing natural products are annotated as possible fatty acids and possible polyketides; this is assigned a separate score in the GARLIC scoring scheme (**Supplementary Table 5**). Structures with a carboxylic acid end and a carbon chain are processed as polyketides and identified unit by unit. Through the reverse biosynthesis steps, the lactone bonds in macrocyclic polyketides are reversed, resulting in linearized polyketide structures which are processed in this step.

Some compounds, such as lipomycin and jamaicamide, contain amino acids that have been decarboxylated and elongated through polyketide biosynthetic machinery. The reaction mechanism for these structures is reversed,

leaving identified amino acid and polyketide components (**Supplementary Fig. 3**). Remaining structures are run through chemical checks to identify whether they resemble polyketides (**Supplementary Fig. 2**); for these chemical fragments, GRAPE identifies the main polyketide carbon backbone by first predicting where the biosynthetic end is located (often a carboxylic group), then the start carbon, and then analyzes the structure in a stepwise manner to identify each oxidation state and substrate.

GARLIC scoring logic. To correlate the output of GRAPE with that of PRISM, we developed GARLIC, which is an algorithm that assigns a match score between GRAPE and PRISM output. PRISM and GRAPE each output two major components: an ordered sequence of monomeric units corresponding to assembly line modules and a list of enzymes or chemical modifications that are external to the biosynthetic assembly line. GARLIC computes a similarity score between GRAPE and PRISM output using a global alignment created with the Needleman and Wunsch algorithm[26]. Each site of the alignment is scored for matches, mismatches, tailoring events and gaps dictated by a customizable scoring logic. As biosynthetic gene clusters may consist of assembly line PKS and NRPS encoded on multiple open reading frames, GARLIC initially identifies all permutations and returns the score corresponding to the best-matching permutation. If there are too many permutations to search all space in a reasonable time, a random sample is taken, and the top scoring alignments are taken and reordered. This is repeated several times to narrow down the true alignment while searching a fraction of the permutable space. Biosynthetic features external to the assembly-line sequence (**Supplementary Table 1**) are matched between PRISM and GRAPE outputs, and added to the score. Tailorings, such as sulfonations and halogenations, recognized from GRAPE are matched to the tailoring enzymes predicted by PRISM. Sugar additions identified from GRPAE are matched by combinatoralizing the potential glycosylation enzymes predicted in PRISM as a different combination of enzymes will create a different sugar structure, or grouping the enzymes differently will yield a different set of sugar predictions (**Supplementary Table 2**). The identified molecule can then be matched to PRISM results, as PRISM also predicts the type of molecule based on known genes responsible for that scaffold. Adding scores in this manner are particularly important for type 2 polyketides and enedyines, whose scaffolds are not broken down by GRAPE to yield monomers.

Given PRISM sequence $P = (p_1, p_2, \ldots, p_n)$ and GRAPE sequence $G = (g_1, g_2, \ldots, g_m)$, an alignment of P and Q can be described as the aligned sequences

$$P^* = (p_1^*, p_2^*, \ldots, p_R^*)$$
$$G^* = (g_1^*, g_2^*, \ldots, g_R^*)$$

Where each of p_i^* and g_i^* may be a monomer unit or a gap, and P and G are subsequences of P^* and G^*, respectively.

Each aligned sequence pair (P^*, G^*) is assigned a score $S(P^*, G^*)$, where

$$S(P^*, G^*) = \sum_{i=1}^{R} s(p_i^*, g_i^*)$$

for a scoring function $s(p_i^*, g_i^*)$ that is defined by match scores, gap penalties, and substitution penalties (**Supplementary Table 5**).

For a given P and G, let A be the set of all possible aligned pairs (P^*, G^*). The optimal global alignment score between the sequence pair (P, G) is

$$rawScore(P,G) = \max_{(P^*,G^*)\in A} \{S(P^*, G^*)\}$$

We implemented the Needleman-Wunsch algorithm[26] (**Fig. 3**), which identifies the optimal global alignment using dynamic programming in time complexity $O(|P||G|)$. Scaling is performed by the length of the PRISM sequence in order to normalize the effect of large sequences. Information about the cluster and compound external to the alignments (for example, addition of sugars and tailored modifications) are then considered and scored. For external feature sets E_P and E_G from PRISM and GRAPE respectively, we added a bonus score to the concordant prediction of each feature, leading to the final score.

$$scaledScore(P,G) = \frac{rawScore(P,G)}{|P|} + s_E(|E_P \cap E_G|)$$

For score function s_E, which adds a score for each external feature (**Supplementary Table 5**), a normalization step to self-alignments is performed to produce the final score:

$$finalScore(P,G) = \frac{scaledScore(P,G)}{\max\{scaledScore(P,P), scaledScore(G,G)\}}$$

We assessed the impact of various substitution and gap penalties, as well as the Smith-Waterman algorithm[25] for local alignment.

GARLIC scoring and algorithm comparison. We sought to address the impact of local and global alignment, as well as the scoring scheme, on the ability of GARLIC to identify a matching compound out of a database of >40,000 small molecules on which GRAPE analysis was performed. Additionally, we compiled a list of 171 biosynthetic gene clusters with known products, and ran them through PRISM. For each of the 171 biosynthetic gene clusters, we performed GARLIC under various algorithm configurations against each of the >40,000 compounds, and ranked the compounds by score. We performed this analysis on several algorithm configurations, including local and global alignment under a basic scoring scheme, 'refined' scoring based on biological prior knowledge, and the final optimized score (GF). The algorithm and scoring configurations are given in **Supplementary Table 5**.

A 'leave-one-out' analysis was performed, using the Scientific Python software library[34], to validate GARLIC and the methods used to obtain the parameter scores. For each cluster to be left out, we used a different training set, which consists of the 166 other clusters matched to their compounds and 297 decoy compounds, including the compound associated with the left-out cluster. The cluster-associated compound was left in order to mimic the most likely scenario for GARLIC use: testing a new biosynthetic gene cluster against a large existing compound database. Performance was measured using a rank-based metric: the sum of the inverse of the rank of the true hit for each cluster as ranked by GARLIC score against the 463 compounds. We used Powell's method[27] to optimize parameters against this metric for each subset of 166 clusters. In each case, once local optimum was reached and a scoring scheme obtained, the scheme was used to test the left out gene cluster against the full set of GRAPE compounds. We used the basic scoring scheme to start the parameter optimization to avoid potential pre-fitting. We chose Powell's method because it was similar to the empirical method used to derive the refined scores and was able to obtain local optima some distance from the start.

Code availability. GRAPE and GARLIC code is available at https://github.com/magarveylab/grape-release and https://github.com/magarveylab/garlic-release.

NMR and mass spectrometry. 1D (^{1}H and ^{13}C) and 2D (^{1}H-^{13}C HMBC, HSQC, HSQC-TOCSY, and ^{1}H-^{1}H NOESY, TOCSY, and COSY) nuclear magnetic resonance (NMR) spectra for lucensomycin and potensibactin were recorded on a Bruker AVIII 700 MHz NMR spectrometer in d_6-DMSO (Sigma-Aldrich). High-resolution MS spectra were collected on a Thermo LTQ OrbiTrap XL mass spectrometer (ThermoFisher Scientific) with an electrospray ionization (ESI) source. For analytical and preparatory separations, LC-MS was used, employing a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using a Luna C18 column (150 mm × 4.6 mm, or 250 mm × 15 mm, Phenomenex), running acetonitrile with 0.1% formic acid and ddH_2O with 0.1% formic acid as the mobile phase.

Microbial strains. *Brevibacillus laterosporus* (DSM 25) and *Nocardiopsis potens* (DSM 45234) were obtained from the German Resource Centre for Biological Material (DSMZ) and maintained on LB agar and Bennet's agar, respectively. Environmental isolate NAM50 was recovered from soil samples collected from McMaster University in 2010 and maintained on Bennet's agar. *Streptomyces achromogenes* was obtained from the Northern Regional Research Lab (NRRL; no. 3125) and was maintained on Bennet's agar.

Production of natural products. All bacteria were initially grown for 3 d at 30 °C, then inoculated into fresh medium and grown for 3 d at 30 °C before

doi:10.1038/nchembio.2188

cultures were centrifuged to remove cells, extracting the supernatant with 2% HP-20 resin, and eluting the resin with excess methanol. To produce octacosamicin, NAM50 was initially cultured in KE medium, followed by Bennet's medium. To produce bogorol and tauramamide, *B. laterosporus* was cultured in LB medium.

To produce lucensomycin, *S. achromogenes* was initially cultured in KE medium, followed by aricidin production medium (AriP). For preparative isolation of lucensomycin, 12 L of AriP *S. achromogenes* culture was harvested by centrifugation at 7,000 r.p.m., followed by methanol extraction of the cell pellet, and Diaion HP-20 (20 g/L) extraction of the supernatant. Methanol eluent of the HP-20 resin was pooled with the methanol extract of the cell pellet and dried under rotary vacuum. This extract was dissolved and extracted with a 1:1 mixture of butanol and water. The butanol fraction was isolated, evaporated to dryness, resuspended in a minimal volume of methanol, and applied to an open gravity column of LH-20 size-exclusion resin (Sephadex) with methanol as a mobile phase. Fractions containing lucensomycin were pooled, evaporated to dryness and resuspended in methanol. Lucensomycin was isolated by preparative scale LC-MS using a Luna 5 µm C_{18} column (Phenomenex, 250 mm × 15 mm) with water (0.1% formic acid) and acetonitrile (0.1% formic acid) as the mobile phase, at a flow rate of 10 mL/min. After 4 min, acetonitrile was increased in a linear manner (curve 5) from 5% to 27% at 10 min, held until 16 min, and then increased to 41% by 40 min, followed by a wash of 100% acetonitrile. Lucensomycin eluted at 27 min.

To produce potensibactin, *N. potens* was initially cultured in KE medium, followed by Bennet's medium. For preparative isolation of potensibactin, 12 L of Bennet's medium *N. potens* culture was harvested by centrifugation at 7,000 r.p.m., followed by methanol extraction of the cell pellet, and Diaion HP-20 (20 g/L) extraction of the supernatant. Methanol eluent of the HP-20 resin was pooled with the methanol extract of the cell pellet and dried under rotary vacuum. The sample was resuspended in a minimal volume of methanol, and applied to an open gravity column of LH-20 size exclusion resin (Sephadex) with methanol as a mobile phase. Fractions containing potensibactin were pooled, evaporated to dryness, and resuspended in methanol. Potensibactin was isolated by preparative scale LC-MS using a Luna 5 µm C_{18} column (Phenomenex, 250 × 15 mm) with water (0.1% formic acid) and acetonitrile (0.1% formic acid) as the mobile phase, at a flow rate of 10 mL/min. After 3 min, acetonitrile was increased in a linear manner (curve 5) from 5% to 10% at 5 min, then increased to 27.5% by 22 min, followed by a wash of 100% acetonitrile. Potensibactin eluted at 14 min.

Genome sequencing. A single colony of *B. laterosporus* was grown in 3 mL LB overnight at 30 °C with shaking at 250 r.p.m. Genomic DNA was harvested using a GenElute Bacterial Genomic DNA Kit (Sigma). A single colony of *S. achromogenes* and NAM50 were used to inoculate 50 mL cultures of GYM media containing 0.5% glycine and grown for 96 h at 30 °C and 250 r.p.m. 500 µL of culture was centrifuged at 12,000*g* for 5 min and resuspended in 500 µL SET buffer (75 mM NaCl, 25 mM EDTA pH 8.0, 20 mM Tris HCl pH 7.5, 2 mg/mL lysozyme) to lyse for 2 h at 37 °C. Proteinase K and SDS were added after lysis to final concentrations of 0.5 mg/mL and 1%, respectively. The lysis mixture was incubated at 55 °C for 2 h before adjusting the concentration of NaCl to 1.25 M and extracting twice with phenol-chloroform. Isopropanol was added (equivalent to 60% the volume of the solution) to precipitate genomic DNA, followed by two washes with 70% ethanol and drying of the DNA, before resuspension in nuclease-free dH_2O. Genomic DNA was sequenced at the Farncombe Metagenomics Facility (McMaster University), using an Illumina HiSeq DNA sequencer. Contigs were assembled using the ABySS genome assembly program and Geneious bioinformatic software.

34. Jones, E., Oliphant, T. & Peterson, P. *SciPy: Open Source Scientific Tools for Python*. (2014).