**Introduction**

The modern world relies heavily on bioactive small molecules by finding use in medicine, agriculture, and industry. Although chemists often employ vast imaginative capacities to develop new molecules, nature continues to act as a wellspring of chemical inspiration and novelty. The vast majority of FDA approved drugs are natural products in their own right, or some derivative thereof. Anticancer and antitumor natural products have come under renewed interest as pharmacologists devise targeted drug delivery systems (Harvey et al, 2015). Several important chemotherapy drugs have been derived from botanical sources including paclitaxel from Pacific yew and vincristine from Madagascar periwinkle. Given supply constraints of Pacific yew bark, researchers (Stierle et al, 1993) were able to determine that biosynthesis of paclitaxel was accomplished by an endophytic fungus living within the tree. Conveniently, heterologous expression of biosynthetic enzymes is simpler for bacterial and fungal endophytes, and so supply was able to be met by industrial cell culture instead of botanical harvest (Williams et al, 2000).

Endophytes, like all microorganisms, have clusters of genes responsible for the production of individual secondary metabolites. Biosynthetic gene clusters (BGC) are generally organized in the genome with a core set of genes (synthases) that produce the backbone of the metabolite and are surrounded by modifying enzymes responsible for altering the backbone into a mature product. Enough BGCs have now been studied to allow prediction of the type of BGC responsible simply by classifying features of the molecule at hand (Greule et al, 2017).

Linking specific natural products from plants to their potential endophytic origin is hampered by several characteristics innate to these unique symbiotic relationships. Adaptation to symbioses within the host frequently renders endophytes unculturable (Liaquat and Eltem, 2016) or only able to produce natural products under particular environmental conditions (Stevens et al, 2013), which are difficult to predict. Historically, this has meant that such screening was inherently dependent on trial and error. Improvements in metagenomic sequencing technologies and competent software to identify BGCs have allowed for the rapid curation of vast BGC libraries, which would be otherwise inaccessible by culture-dependent methods. However, such an approach does little to link a specific BGC to a particular natural product: to do so efficiently requires a targeted approach.

A class of natural products known as acetogenins has been isolated exclusively from the custard apple family (Annonaceae). Annonaceous acetogenins (ACGs) have proved to have potent bioactivity including antitumor, pesticidal, and antioxidant properties. To date, over 500 ACGs have been isolated from various plants within Annonaceae. The ACGs are relatively simple in structure and yet diverse through combinatorial pairing of functional groups and chiral centers. ACGs are chemically characterized by a long, linear, sparsely functionalized C32-34 carbon chain containing 0 to 3 tetrahydrofuran (THF) rings and a terminal butyrolactone (TBL) ring (Figure 1). The biosynthetic origins have never been directly studied, though they are widely believed to be synthesized from alpha-linolenic acid and 2-propanol (Mclaughlin et al, 1993) catalyzed by uncharacterized plant enzymes.

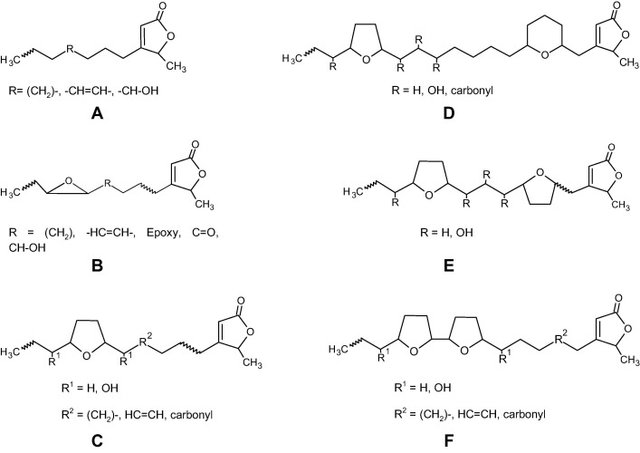


Figure 1: Acetogenins representing the 6 sub-types isolated from *Annona* muricata (Coria-Téllez et al, 2018) displaying the structural simplicity and diversity of acetogenins. ACGs are generally categorized based on the internal ring structures. C and F represent the two most common types of ACGs with 1 and 2 THF rings respectively.

In 2016, Taechowisan et al reported isolating ACGs from a non-annonaceous source for the first time. They isolated two previously described ACGs from the culture broth of a bacterial endophyte (Figure 2). The new bacterial species, given the temporary name *Streptomyces sp. VE2*, was isolated from the stem tissue of *Vernonia cinerea* of the sunflower family (Asteraceae). Although the authors conceded that this discovery was quite unexpected, they stopped short of suggesting a biosynthetic explanation. The likelihood that *S. VE2* and Annonaceae would, through convergent evolution, produce the exact same secondary metabolites is exceedingly small. The parsimonious explanation is that ACGs are produced by endophytic *Streptomyces* *spp.* in Annonaceae.

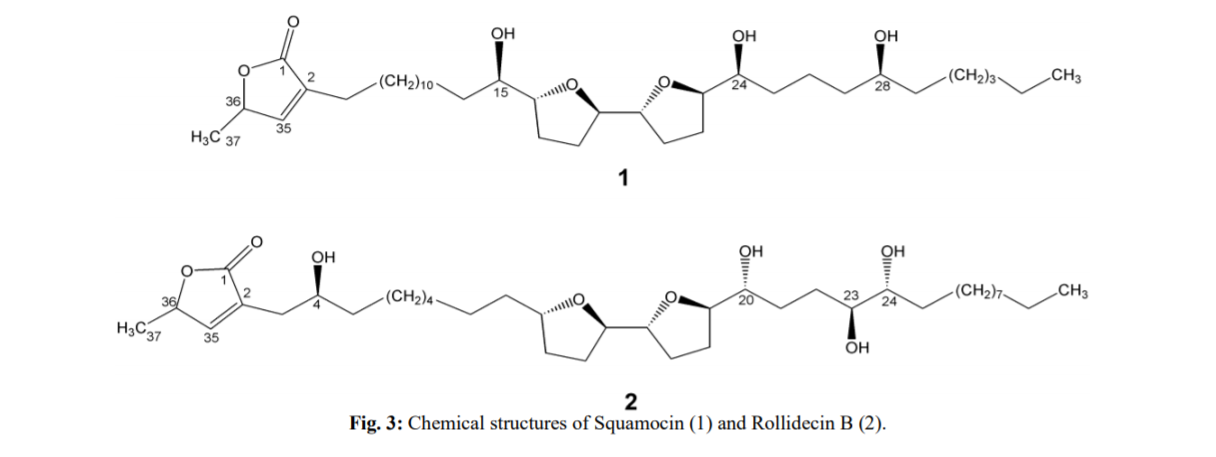


Figure 2: These two ACGs are similar but distinct. Both are of the bis-THF type but have different placement and chirality of hydroxyl groups.

*Streptomyces* is a “multicellular” bacterial genus typified by its fungal-like growth habit, exhibiting both thread-like mycelia and spores. The genus is known to produce an incredible array of natural products both in number and diversity. Many biosynthetic gene clusters in the genus are fully characterized and the biosynthetic pathways fully elucidated (DoBISCUIT: http://www.bio.nite.go.jp/pks/). By structural comparison alone, it can be hypothesized that acetogenins almost certainly come from the polyketide synthase (PKS) pathway.

PKSs are biochemically and genetically related to fatty acid synthases (FAS) (Fig. 3). Both PKS and FAS use modular, assembly-line like logic to synthesize products. The natural product is shuffled from one module to the next, adding a C2 malonyl unit for each module. Three protein domains are required for minimum module functionality: the ketosynthase (KS), acyl-transferase (AT), and acyl-carrier protein (ACP). Three optional domains reduce the lactone of the malonyl extender unit first to hydroxyl, then enoyl (double C=C bond), and finally the fully reduced alkyl (single C-C bond). FASs practically always employ all three reducing domains to yield a non-functionalized carbon chain. Believed to have evolved from FASs, PKSs have lost some domains in select modules, producing a vast diversity of carbon skeletons with different combinations of functional groups (lactone, hydroxyl, enoyl, and alkyl) (Figure 3: B, C). Generally, a single PKS gene cluster yields one or a few very closely related compounds. Bacterial genomes, in particular *Streptomyces*, may have dozens of polyketide BGCs taking up a significant portion of the genome (Komaki et al, 2018). Type I modular PKSs are responsible for long unbranching polyketides. Type II iterative PKSs are responsible for polycyclic and aromatic metabolites, and type III PKSs are responsible for small molecules with diverse architectures. Acetogenins most closely resemble known *Streptomyces* compounds synthesized by Type I modular PKSs. Although the differences in Type I, II, and III PKSs are biochemically, evolutionarily, and genetically nuanced, such an overview will be left to other publications to more sufficiently address this complex issue (Hertweck, 2009). For the purposes of this proposal, it is only necessary to understand that by looking at the final natural product, it is possible to assume with some certainty the PKS type from which it came (Figure 3: A).

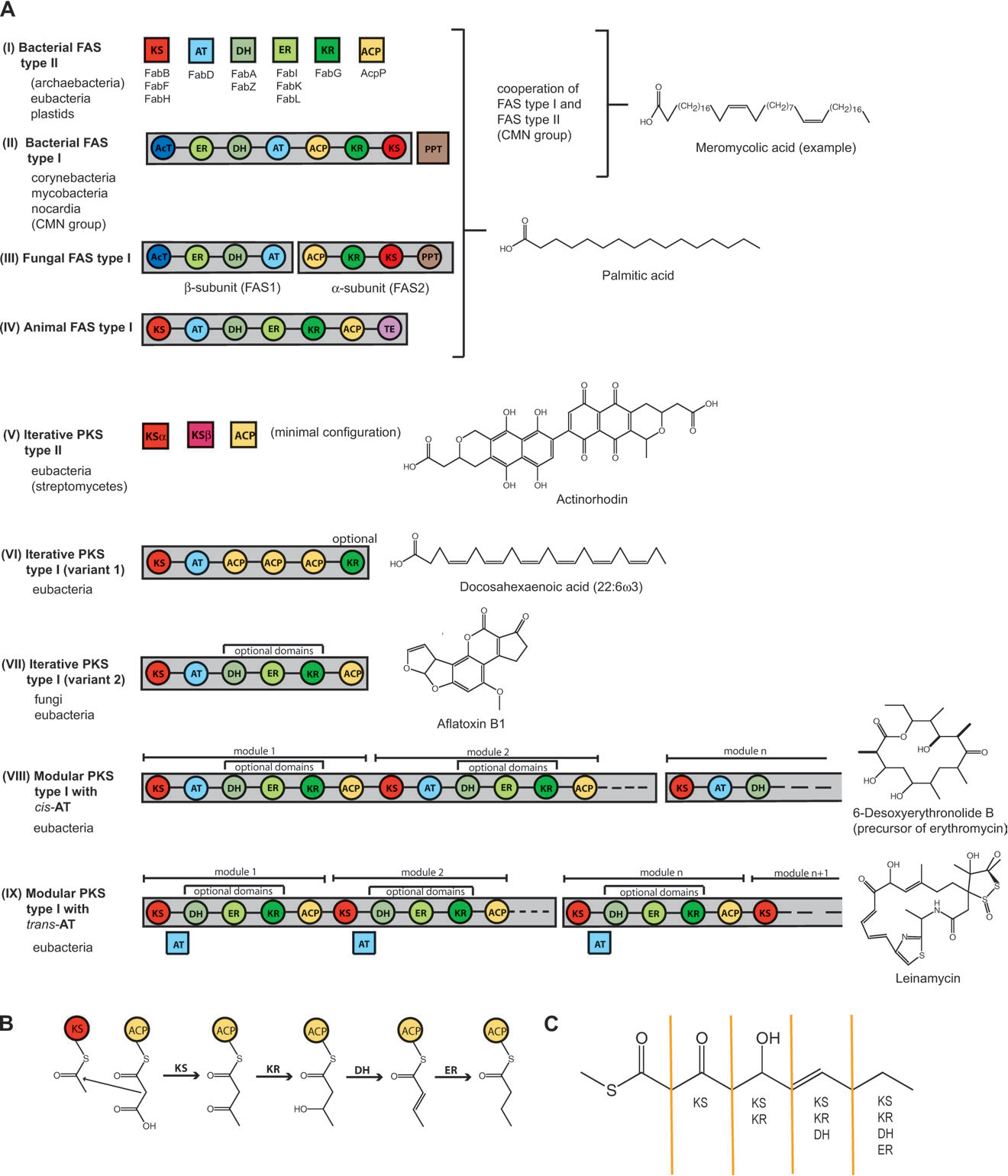


Fig. 3: A) examples of products produced by certain PKS/FAS types. B) Assembly-line like logic of PKS/FASs. C) explanation of how presence/absence of reducing domains in a module alters product. (Jenke-Kodama et al 2005)

Bacterial Type I modular PKSs, similar to Animal Type I FASs, are encoded by one or a handful of very long open reading frames (ORFs), with each module responsible for the addition of every two carbons found in the backbone of the product. Because ACGs have 32 or 34 but never 33 carbons in the backbone, this is often cited as evidence that a single C-2 monomer separates the two, a feature prominent among PKSs.

*Streptomyces* PKSs are known to produce THF rings similar to those found in ACGs. In monensin biosynthesis, THF ring formation is accomplished by epoxidation of adjacent C=C double bonds, followed by domino ring opening (Gallimore, 2009). A careful stereochemical analysis performed by Andrew Gallimore showed that ACGs are formed by this monensin-type THF ring formation. However, the enzymes that are responsible for these post-PKS modifications are ubiquitous in *Streptomyces* and do not represent a good target for elucidating ACG biosynthesis.

The other chemical motif of ACGs, the terminal gamma-lactone, is similar to that found in several *Streptomyces* natural product classes. Chief among these are the ‘bacterial hormones’: A-factor-like autoregulators and the tetronates. The latter is biosynthesized by four genes in a conserved “glycerate utilization operon” (Vieweg et al, 2014) which act as post-PKS modification enzymes to generate the TBL. Tetronate biosynthesis was not pursued as a target in this proposal, due to difficulty in curating the operon sequences across multiple species of *Streptomyces* (see Discussion).

A-factor-like compounds act as regulators of secondary metabolite synthesis, pigment synthesis, and spore formation (Du et al, 2011). The core gene in A-factor synthesis, AfsA, is solely responsible for the lactone ring formation (Nishida et al, 2007). SRB1 and SRB2 (Figure 3) are two autoregulators isolated from *S. rochei,* which most closely resemble the TBL of ACGs (Arakawa et al, 2012). Based on 16sRNA phylogenies, *S. rochei* and *S. maritimus* are the two most closely related species to *S. VE2* (Taechowisan, 2016). Additionally, AfsA homologues have been found to be very near or within *Streptomyces* PKS BGCs (AntiSMASH data not shown), multi-copy genes, and present in some plasmids (Mochizuki, 2003). AfsA homolog enzymes catalyze the reaction between an unincorporated ACP or CoA bound diketone fatty acid, and a (DHAP; dihydroxy acetone phosphate) releasing the product from ACP or CoA (###). Both tetronate and AfsA biosynthesis share similar reactants through convergent evolution.

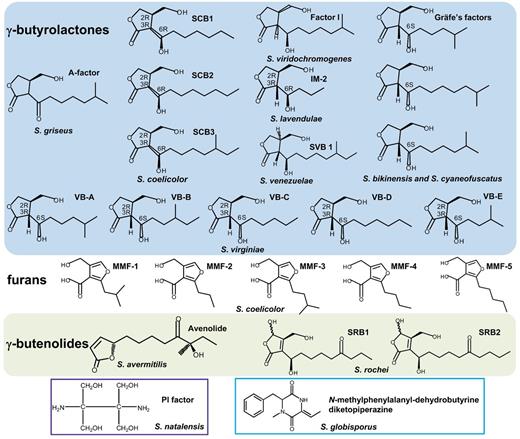


Figure 4: Diversity of *Streptomyces* hormones (Niu et al, 2016). Notice the similarity in terminal lactone between ACGs and SRB1/2. *S. rochei* is closely related to the ACG producing *S. sp. VE2*. The butyrolactones, furans, and SRB1/2 all come from AfsA homologues

Ultimately AfsA homologues were chosen as a target for potential ACG biosynthesis. AfsA homologues alone cannot account for ACG biosynthesis. Instead, incorporation of AfsA into a PKS would predict both ring formation and release from the PKS megasynthase. Normally, a thioesterase domain found at the C-terminal of the megasynthase would release the immature polyketide by reaction with water or by distal portion of the polyketide forming a macrolide, as found in many antibiotics. Searching for examples of PKS/AfsA hybrid proteins, both pfam and InterPro report examples of AfsA genes fused with Type I and III PKSs from gamma proteobacteria and cyanobacteria. These proteins are uncharacterized so it is unknown what the products of these hybrid synthases would be. It is not clear what the evolutionary relationships between these proteins are, whether common ancient gene fusion, HGT of recent gene fusion, or convergent evolution of many independent gene fusion events (see Discussion). It is also not clear what role AfsA genes, hybrid or not, play in phyla outside of actinobacteria, as they have their own distinct hormone systems (Shaefer et al, 2013). Furthermore, it is unclear what relationship said hybrid proteins would have with any potential *Streptomyces* hybrid proteins, but it is sufficient to suggest that incorporation of AfsA into a PKS is 1) biochemically compatible and 2) does not alter AfsA to the degree that it is unable to be aligned to pfam and Interpro HMMs. A full analysis of these hybrids including the gene phylogenies of both components would be informative as to the number of times such a fusion event has occurred and to what extent the hybrid and non-hybrid components share sequence similarities. Such an analysis, though feasible, was outside of the scope of the current proposal.

Hypothetical PKS/AfsA hybrid(s) were chosen as a molecular target for this proposal. Detection of a hybrid protein would be a strong putative lead for ACG biosynthesis. Indeed, none of the evidence alone so far points to a hybrid as a likely hypothesis, but taken together, it makes such an explanation plausible. Perhaps *Streptomyces* endophytes in Annonaceae produce acetogenins by a fused PKS-AfsA gene. The testable question remains: can such a hybrid sequence be detected from an Annonaceae sample directly?

**Experiment**

The experimental logic of the current proposal is as follows: if a *Streptomyces* PKS-AfsA hybrid gene can be detected from an Annonaceous tissue sample, this gene (and the corresponding BGC) is a strong candidate for ACG biosynthesis.

Detection was chosen to be performed by polymerase chain reaction (PCR) because of the method’s strength, simplicity, flexibility, and sensitivity. Normally, PCR primer sequences are chosen so that they are specific only to the target of interest. This approach presupposes knowledge of the sequence of interest. When PCR is used as a discovery tool, primers are designed so that they are conserved and specific to sequences similar to those wished to be discovered.

In the case of detecting unknown genes based on their similarities to known gene sequences, degenerate primers are employed. Degenerate primers are designed from conserved protein motifs present in the protein family of interest but absent in all other protein families (or the combination of forward and reverse motifs is unique). Because of the degeneracy of the genetic code, these primers are actually a population of primers covering every possible DNA sequence that could produce a given protein motif. While others have detected fusion events by designing a primer to span the junction sequence (ArcherDX), such an approach is presently inappropriate because the hypothetical junction would be very poorly conserved. Instead, designing the forward primer to target the PKS sequence and the reverse primer to target the AfsA sequence is more appropriate.

Separation of prokaryotic and eukaryotic DNA ranges from difficult to impossible, making amplification of DNA unsatisfactory. Additionally, the relative concentration of PKS/AfsA DNA sequences is predicted to be low in comparison to RNA during the active production of ACGs. In the genome, as little as one copy is predicted to exist, whereas a multitude of RNA transcripts should be present during ACG production. Unlike DNA, separation of bacterial and prokaryotic mRNA can be accomplished by poly-A depletion (Szabo and Salzman, 2017), effectively increasing the concentration of bacterial mRNA in the sample. Use of RNA is also predicted to drastically reduce non-target amplification, due to decoupling of non-target probe sites and reduction in size from the genome to the *in situ* transcriptome. For these reasons, reverse transcription PCR (RT-PCR) was chosen.

PKS genes have three minimal domains: KS, AT, and ACP. Many PKSs have trans-AT domains, meaning that they are encoded elsewhere in the genome. ACP domains are small and highly divergent in primary, secondary, and tertiary organization, making them unsuitable for motif discovery. KS domains are the least divergent of all PKS domains and are also present in all types of PKSs, making them the most suitable for motif discovery.

Existing degenerate primer design software has three major shortcomings towards the current proposal. Firstly, such software attempts to find two motifs at least N residues apart, whereas only one motif for KS and one for AfsA is necessary. Secondly, it requires a protein global alignment input. Even if *Streptomyces* PKS/AfsA hybrids were available for input at the time of this writing, AfsA produces poor global alignments in its own right, making such an approach not feasible (data not shown). Thirdly, when one motif is insufficient, such software does not allow for multiple motifs from hidden subtypes of proteins.

In *Streptomyces* AfsA (pfam), no single 6 amino acid motif exists for greater than ~60% of the sequences. Multiple motifs must be found for the reverse primer of AfsA used in individual PCR reactions. Arguably, the most powerful software for determining protein motifs is the MEME suite. Although this software does identify motifs in AfsA, it is designed to discover functional motifs. As such, it neither provides tools to determine the percentage of sequences covered by each additional motif, nor a way to handle overlapping motifs. For example, if one motif is reported in 50% of sequences and a second is reported in 50% of sequences, it is not clear whether the co-occurrence is 100%, 0% or some value in between. In the case of 100% co-occurrence, only one of the motifs is necessary to represent that half of the sequences while other motifs will need to be discovered to represent the remaining half. In the 0% case, the two motifs are adequate to represent all of the sequences. A custom program was used to surmount these issues.

The basic concept behind the custom program was to find motifs unique to known sequences homologous to the unknown sequence of interest. The uniqueness of a motif was determined by its absence from distantly homologous, non-target sequences for which motif conservation is a distinct possibility. The program was implemented for forward primer development of KS and reverse primer development of AfsA. Included sequences consisted of AfsA (pfam) protein sequences only within *Streptomyces*. Excluded sequences consisted of AfsA found in other phyla (1 group). Distantly-related, non-target homologues consisted of protein families in the same pfam ‘clan’ as AfsA (‘hotdog fold containing domain’; cl0050; 13 families) that make up the remaining portion of excluded sequences (12 groups). A greedy algorithm (see Discussion) was employed so that the most common motif (MCM) was considered at the start of each iteration. If the putative MCM was present in more than 1% of any excluded group (ex. AfsA from non-Actinobacteria, Polyketide Dehydratase from *Streptomyces*, etc.), it was deemed non-unique and discarded, and the next MCM was considered. All sequences containing the unique MCM were removed and a MCM was found from the remaining sequences. Iteration was continued until 10% of divergent sequences remained, which were not considered for primer development. This process yielded two homologous and three heterologous motifs (Table 1) from which degenerate reverse primers were designed.

Being less divergent, motif discovery in KS was not as problematic as AfsA. Pfam does not distinguish between KS domains from FAS and Type I, II, and III PKSs. To have more specificity in the excluded sequences a small, hand-curated list of KS domain sequences was taken from fully characterized Type I modular PKSs in *Streptomyces* (DoBISCUIT database; 17 sequences). Excluded sequences (9 groups) consisted of protein families in the pfam thiolase clan (cl0046). These sequences were given the same treatment as AfsA but the kmer length was extended to eight amino acids, yielding two homologous motifs representing all [100%] of the 17 sequences (see Table 1). A single nucleic acid substitution accounts for the difference between the two motifs. The two were combined giving a single degenerate forward primer for *Streptomyces* KS domain.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1** | | | |
| primer direction | homology | motif | primer |
| f | KS | VE[A|G]HGTGT | GTNGARGSNCAYGGNACNGGNAC |
| r | AfsA | DHVPGM | CATNCCNGGNACRTGRTC |
| DHVPGL | NARNCCNGGNACRTGRTC |
| EVFVTD | RTCNGTNACRAANACYTC |
| RYVELD | RTCNARYTCNACRTANCK |
| AETIRQ | YTGNCKDATNGTYTCNGC |

Primer specificity was tested by *in silico* PCR of both *S. coelicolor A3(2)* and *E. coli* transcriptomes with BIOBIKE function ‘Run-PCR’ -*degenerate*. Neither transcriptome gave an amplicon. After introduction of an artificial PKS/AfsA hybrid transcript (tautomycin-PKS/srrX), both transcriptomes gave five amplicons, consistent with the number of modules present in the PKS portion from tautomycin (each containing one KS domain). The tautomycin PKS gene portion represents a Type I modular PKS. KS domains from tautomycin were not among the 17 sequences used above for forward primer design. The srrX gene portion is responsible for synthesis of the A-factor homologues SRB1 and SRB2 from Figure 4 in *S. rochei*.

Relatively few publications exist concerning amplification of endophytic transcripts directly from host tissue. Terakado-Tonooka et al (2008) showed this was possible when they amplified diazotrophic (nitrogen-fixing) endophyte transcripts from sweet potato. The authors noticed that protein amounts in sweet potatoes were considerably higher than what could be explained by fertilizer and soil nitrogen alone. They hypothesized that nitrogen fixing bacteria were responsible for the discrepancy between nitrogen and protein observed. To overcome the difficulties with studying endophyte gene expression *in vitro* (see Introduction), they opted to measure transcript levels directly from host tissue. They chose to target the gene nifH, a bacterial nitrogenase reductase, unique and specific to nitrogen fixing bacteria. The workflow proceeded as follows. First sweet potatoes were harvested, surface sterilized (to ensure only true endophytes were observed). Sweet potato was frozen under liquid nitrogen to preserve the relatively unstable bacterial mRNA. Tissue was then ground to a fine powder with mortar and pestle under liquid nitrogen. RNA was extracted using the SDS method followed by further purification and finally DNase I treatment. First strand cDNA synthesis was prepared by incubation with random hexamers. PCR was performed first with four degenerate primers targeting a wide variety of bacterial nifH genes (explicit design steps were not discussed). Nested PCR was carried out by two degenerate primer sequences conserved in all nifH genes. High concentrations of nifH transcripts (360bp) in the sweet potato samples implicated bacterial endophytes for the excess protein observed in the host.

The sweet potato/diazotroph example is useful to show the logical parallels between it and the current proposal. In both cases, an anomalous trait (protein/ACG) in a plant host (sweet potato/Annonaceae) is linked to a hypothetical endophyte (diazotroph/*Streptomyces*) by direct amplification of bacteria-specific gene transcript (nifH/PKS-AfsA). While the current proposal will largely adhere to the methodology of Terakado-Tonooka et al, it will diverge in important ways due to a difference in predicted amplicon size and quantity. For RNA extraction, Holmes et al (2014) reports an optimized method for RNA extraction from a mixed bacterial/host sample by a combination of bead beating and chemical treatment (LiCl) to recover larger amounts of bacterial RNA than that used by Terakado-Tonooka et al. This is important since the concentration of PKS-AfsA transcripts could be very low. During cDNA synthesis instead of random primers, the AfsA reverse primers will be employed since the former yields many truncated transcripts. PCR will be performed by regular PCR using the previously discussed KS and AfsA degenerate primers. The extension step for each cycle will be lengthened to account for the long expected amplicon. Per Holmes et al (2014), amplification of a bacterial house-keeping gene indicates RNA extraction and PCR have not had major issues. This has the added benefit of showing whether or not the expected bacteria is present and metabolically active in the host tissue. Laskaris et al (2012) shows that the atpD (an ATP synthase) house-keeping gene is highly conserved and *Streptomyces-*specific primers could be designed for atpD. cDNA preparation and PCR of atpD will be performed per the original article. Finally, gel electrophoresis will be used to visualize the experimental PKS-AfsA and atpD amplicons.

**Discussion**

**Result Interpretations**

I. Control Positive

The most likely explanation is the presence of metabolically active *Streptomyces*. Laskaris et al report that their atpD *Streptomyces* primers occasionally non-specifically amplify [*Nocardioides*](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0035756) atpD. This genus is relatively rare and should not present significant issues.

II. Control Negative

If no atpD band is present, this could be explained by several things. Firstly and most likely, PCR could have gone awry for a multitude of reasons. The problem should be identified, corrected, and the experiment attempted again. Secondly and less likely, the transcript is below the detection limit of PCR. More extreme efforts to concentrate bacterial transcripts need to be taken, such as those known to deplete poly-A mRNA and rRNA (Szabo and Salzman, 2016). Lastly and least likely, there is a true absence of metabolically active *Streptomyces* within the sample. Several authors have described endophytic *Streptomyces* from Annonaceae (Wu et al, 2009) (Zin et al, 2007). One possible explanation is that endophytes are often both specific to certain plant organs and vary with the season or host growth phase.

III. Experimental Positive (with Control Positive)

Best possible results would be the visualization of a single PKS-AfsA band in at least one of the experimental AfsA motif lanes, and a single atpD band in the control lane. Experimental amplicon(s) are expected in the range of ~2-6 kb depending on the number of optional reducing PKS domains present (see Introduction), while atpD is expected around 466 bp. To determine whether this result is a true or false positive, the bands would be purified and sequenced to determine their architecture and homology to known sequences. A single ORF encompassing both PKS and AfsA would indicate a true positive, while separate ORFs and presence of a terminal thioesterase domain would indicate a polycistronic mRNA that is unlikely to be involved in ACG biosynthesis. Such a false positive, while unlikely, is not particularly unexpected, given that AfsA-homolgues are occasionally found within *Streptomyces* PKS gene clusters (anti-SMASH data not shown) where AfsA’s relative position varies considerably (high recombination). If the sequencing of any particular band reveals that one or either gene part is missing, this indicates a false positive by non-specific amplification. If all experimental band(s) are false positives, further steps would follow the Experimental False Negative below.

IV. Experimental False Negative (with Control Positive)

The most likely explanation of this result would be issues revolving around cDNA synthesis of long transcripts and primer/PCR issues. Bead beating was not shown to shear long RNA (4kb) sequences significantly (Zecca, 2016), and RNase contamination would remove the control band as well; this, however, is not predicted to be a significant cause of false negatives. In the case of an experimental negative, the best way to correct potentially false negatives is by cDNA/PCR optimization. Because the primers need to be optimized experimentally, it becomes necessary to synthesize an artificial PKS-AfsA hybrid RNA for this purpose. The hybrid could then also be used as an internal standard to be run in parallel with each experimental reaction mixture. A whole host of issues could be the cause of primer failure, including the unknown/low concentration of target transcript, unknown secondary structure in transcript, and primer issues such as Tm, primer-dimers, etc. The preparation of an artificial chimeric transcript is outside of this project’s scope. However, Vasl et al (2004) describes a rapid three step PCR preparation of a hybrid gene without need for cloning. After *in vitro* transcription, a known quantum of hybrid could be prepared and the sensitivity of different PCR reactions determined.

If it is assumed that the AfsA domain in the hybrid is equally likely to be similar to any non-hybrid AfsA sequence known, then there is a 10% probability that this domain will not be amplified by the employed reverse primers (see Methods). There is even more cause for concern if incorporation into PKS significantly altered the AfsA domain sequence so that it cannot be predicted. Careful motif analysis of incorporated AfsA and non-incorporated AfsA homologues in proteobacteria and cynaobacteria may give clues on what to expect in the way of motif divergence after a fusion event. An improvement to the algorithm itself may actually help with these issues. The algorithm is deemed ‘greedy’ because it does not always find the minimum number of motifs. For example, assuming complete independence of motifs under a greedy algorithm, a hypothetical motif A may be present in 60% of the sequences. Motif B is present in 30% (total) and motif C 10% of sequences. Under a maximally efficient algorithm, motif D may represent 51% and motif E 49%. Even though the greedy algorithm identified the most common motif (A), it did not identify the fewest possible motifs (D and E) necessary to represent all sequences. The strength of the greedy algorithm resides in its ability to simultaneously identify both very common motifs and motifs representative of rare subtypes. The latter is considerably more complicated, making it an area for future improvement.

The final scenario leading to a false negative (that can be presently predicted) involves a trans-acting AfsA gene. This AfsA protein would be not be incorporated into a PKS and interact non-covalently with the PKS protein to perform the functions predicted of a true PKS-AfsA hybrid. A genome mining approach would be the only way to overcome this obstacle (see Introduction). It is not possible to predict how likely this scenario is.

**Alternative hypotheses (Experimental true negative)**

As discussed in the Introduction, the most plausible alternative biosynthesis pathway is the tetronate biosynthesis pathway typified by the antibiotic RK-682 in *Streptomyces*. The problem with testing this approach is two-fold. First, one would have to determine the uniqueness of the four biosynthesis genes to tetronate biosynthesis. For example, this arrangement could have resulted from a neofunctionalization of a common operon, in which case amplification would be universal instead of specific. Additionally, one must consider the number of transcripts amplified from non-ACG tetronate biosynthesis. If this number is sufficiently large, or contains many uncharacterized pathways, parsing the ACG pathway would be tedious. Contrastingly, any true positive PKS-AfsA hybrid will be surprising and relatively easy to associate with ACG biosynthesis.

The remaining hypotheses operate under the assumption that Taechowisan et al were mistaken in their isolation of ACGs from *Streptomyces sp. VE2*. Dr. Feras Qasem Alali, an expert in the field of ACG identification, had more than a few concerns about their methods (personal communication). The first hypothesis considers that the biosynthesis of ACGs is performed by enzymes native to the Annonaceae. In one hypothesis by Zeng et al (1993), a C32-34 fatty acid reacts with 2-propanol to form a proto-ACG. Wong and Brown (2002) noted a possible biosynthesis of an unusual ACG, artapetalin, in which alpha-lineolic acid is reacted with pyruvate followed by an unknown reaction mechanism to remove the beta-carbon –O-CH3 group. Neither of these papers offer any putative enzymes that could accomplish such reactions. Attempting to accumulate a list of biochemically capable plant enzymes is likely possible but has not been attempted here.

A final alternative considers acaterin, an inhibitor of acyl-CoA: cholesterol acyltransferase (ACAT), isolated from the soil-dwelling *Pseudomonas sp. A92*. Acaterin is the only known biomolecule to have the identical terminal lactone as ACGs to have direct experimental evidence for its biosythesis. Feeding experiments suggest synthesis by C-5 unit with octanoate in an unknown biosynthesis pathway (###). Because *Pseudomonas sp. A92* is culturable and acaterin is unique to it, genome mining of *Pseudomonas sp. A92* may be the key to unlock ACG biosynthesis. This approach would only be useful if similarities in chemical structure between ACGs and acaterin are not coincidental (convergent evolution). It is possible that both *Streptomyces VE2* and *Pseudomonas sp. A92* produce identical terminal lactones through genetically related biosynthesis, since these genera are known to conjugate and exchange genetic information (Mazodier et al, 1989).

Next steps from best possible results (for future studies):

1. Sequencing of experimental band(s)
2. Characterize sequence using antiSMASH and/or pfam
3. If 1 and 2, predict single hybrid protein product: design specific primers (non-degenerate)
4. Use primers for PCR guided isolation of strain
5. Attempt extraction/detection of acetogenins from strain
6. Reverse PCR/genome walking to characterize full BGC sequence
7. Knockout/feeding experiments; monitor acetogenin production
8. Characterize BGC on antiSMASH; compare to known BGCs
9. Attempt axenic growth of Annonaceae, inoculate with acetogenin producing *Streptomyces* for comparative analysis; determine symbiotic biosynthesis effects
10. Determine plant organs that harbor *Streptomyces*
11. Isolate acetogenin producing *Streptomyces* from other Annonaceae sp. for comparative analysis
12. Co-evolutionary studies between *Streptomyces* and Annonaceae species

**References**

Arakawa K, Tsuda N, Taniguchi A, Kinashi H. The Butenolide Signaling Molecules SRB1 and SRB2 Induce Lankacidin and Lankamycin Production in Streptomyces rochei. ChemBioChem. 2012;13:1447–57.

ArcherDX. Why is it better to use RNA instead of DNA for fusion detection? [Internet]. Brand. [cited 2018Nov25]. Available from: https://archerdx.com/support/faqs/archer-fusionplex-variantplex-faqs/general-faqs/better-to-use-rna-instead-of-dna-fusion-detection

Coria-Téllez AV, Montalvo-Gónzalez E, Yahia EM, Obledo-Vázquez EN. Annona muricata: A comprehensive review on its traditional medicinal uses, phytochemicals, pharmacological activities, mechanisms of action and toxicity. Arabian Journal of Chemistry. 2018;11:662–91.

Du Y-L, Shen X-L, Yu P, Bai L-Q, Li Y-Q. Gamma-Butyrolactone Regulatory System of Streptomyces chattanoogensis Links Nutrient Utilization, Metabolism, and Development. Applied and Environmental Microbiology. 2011;77:8415–26.

Gallimore AR. ChemInform Abstract: The Biosynthesis of Polyketide-Derived Polycyclic Ethers. ChemInform. 2009;40.

Greule A, Zhang S, Paululat T, Bechthold A. From a Natural Product to Its Biosynthetic Gene Cluster: A Demonstration Using Polyketomycin from Streptomyces diastatochromogenesTü6028. Journal of Visualized Experiments. 2017;

Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nature Reviews Drug Discovery [Internet]. 2015;14:111–29. Available from: https://www.nature.com/articles/nrd4510

Hertweck C. ChemInform Abstract: The Biosynthetic Logic of Polyketide Diversity. ChemInform. 2009;40.

Holmes A, Birse L, Jackson RW, Holden NJ. An optimized method for the extraction of bacterial mRNA from plant roots infected with Escherichia coli O157:H7. Frontiers in Microbiology. 2014;5.

Jenke-Kodama H, Sandmann A, Müller Rolf, Dittmann E. Evolutionary Implications of Bacterial Polyketide Synthases. Molecular Biology and Evolution. 2005;22:2027–39.

Komaki H, Sakurai K, Hosoyama A, Kimura A, Igarashi Y, Tamura T. Diversity of nonribosomal peptide synthetase and polyketide synthase gene clusters among taxonomically close Streptomyces strains. Scientific Reports. 2018;8.

Laskaris P, Sekine T, Wellington EMH. Diversity Analysis of Streptomycetes and Associated Phosphotranspherase Genes in Soil. PLoS ONE. 2012;7.

Liaqat F, Eltem R. Identification and characterization of endophytic bacteria isolated from in vitro cultures of peach and pear rootstocks. 3 Biotech. 2016;6.

Mazodier P, Petter R, Thompson C. Intergeneric conjugation between Escherichia coli and Streptomyces species. Journal of Bacteriology. 1989;171:3583–5.

Mclaughlin JL, Gu Z-M, Fang X-P, Zeng L, Wood KV. Bullacin: A New Cytotoxic Annonaceous Acetogenin from Annona bullata. Heterocycles. 1993;36:2221.

Mochizuki S, Hiratsu K, Suwa M, Ishii T, Sugino F, Yamada K, et al. The large linear plasmid pSLA2-L of Streptomyces rochei has an unusually condensed gene organization for secondary metabolism. Molecular Microbiology. 2003;48:1501–10.

Nishida H, Ohnishi Y, Beppu T, Horinouchi S. Evolution of ?-butyrolactone synthases and receptors in Streptomyces. Environmental Microbiology. 2007;9:1986–94.

Niu G, Chater KF, Tian Y, Zhang J, Tan H. Specialised metabolites regulating antibiotic biosynthesis in Streptomyces spp. FEMS Microbiology Reviews. 2016;40:554–73.

Schaefer AL, Lappala CR, Morlen RP, Pelletier DA, Lu T-YS, Lankford PK, et al. LuxR- and LuxI-Type Quorum-Sensing Circuits Are Prevalent in Members of the Populus deltoides Microbiome. Applied and Environmental Microbiology. 2013;79:5745–52.

Stevens DC, Conway KR, Pearce N, Villegas-Peñaranda LR, Garza AG, Boddy CN. Alternative Sigma Factor Over-Expression Enables Heterologous Expression of a Type II Polyketide Biosynthetic Pathway in Escherichia coli. PLoS ONE. 2013;8.

Stierle A, Strobel G, Stierle D. Taxol and taxane production by Taxomyces andreanae, an endophytic fungus of Pacific yew. Science [Internet]. 1993;260:214–6. Available from: http://science.sciencemag.org/content/260/5105/214.long

Szabo L, Salzman J. Detecting circular RNAs: bioinformatic and experimental challenges. Nature Reviews Genetics. 2016;17:679–92.

Taechowisan T, Singtotong C, Phutdhawong W. Antibacterial and Antioxidant Activities of Acetogenins from Streptomyces sp. VE2; An Endophyte in Vernonia cinerea (L.) Less. Journal of Applied Pharmaceutical Science. 2016;:067–72.

Terakado-Tonooka J, Ohwaki Y, Yamakawa H, Tanaka F, Yoneyama T, Fujihara S. Expressed nifH Genes of Endophytic Bacteria Detected in Field-Grown Sweet Potatoes (Ipomoea batatas L.). Microbes and Environments. 2008;23:89–93.

Vasl J, Panter G, Bencina M, Jerala R. Preparation of chimeric genes without subcloning. BioTechniques. 2004;37:726–30.

Vieweg L, Reichau S, Schobert R, Leadlay PF, Süssmuth RD. Recent advances in the field of bioactive tetronates. Nat. Prod. Rep. 2014;31:1554–84.

Williams DC, Wildung MR, Jin AQ, Dalal D, Oliver JS, Coates RM, et al. Heterologous Expression and Characterization of a “Pseudomature” Form of Taxadiene Synthase Involved in Paclitaxel (Taxol) Biosynthesis and Evaluation of a Potential Intermediate and Inhibitors of the Multistep Diterpene Cyclization Reaction. Archives of Biochemistry and Biophysics. 2000;379:137–46.

Wong H-F, Brown GD. β-Methoxy-γ-methylene-α,β-unsaturated-γ-butyrolactones from Artabotrys hexapetalus. Phytochemistry. 2002;59:99–104.

Wu Y, Lu C, Qian X, Huang Y, Shen Y. Diversities Within Genotypes, Bioactivity and Biosynthetic Genes of Endophytic Actinomycetes Isolated from Three Pharmaceutical Plants. Current Microbiology. 2009;59:475–82.

Zecca J. The Effect of Bead Beating on RNA Integrity [Internet]. opsdiagnostics.com. OPS Diagnostics; Available from: https://opsdiagnostics.com/notes/rnaintegritypaper.html

Zin NM, Sarmin NIM, Ghadin N, Basri DF, Sidik NM, Hess WM, et al. Bioactive endophytic streptomycetes from the Malay Peninsula. FEMS Microbiology Letters. 2007;274:83–8.