Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at *Molecules* **2019**, *24*, 1411; doi:10.3390/molecules24071411

Article

The distribution of tryptophan-dependent indole-3-acetic acid synthesis pathways in bacteria unraveled by large-scale genomic analysis

Pengfan Zhang[1,2], Tao Jin[2], Sunil Kumar Sahu[2,4], Jin Xu[3], Qiong Shi[1,2], Huan Liu[2,4], Yayu Wang[2,4,5,*]

1. BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China.

2. BGI-Shenzhen, Shenzhen 518083, China.

3. Citrus Research and Education Center, Department of Microbiology and Cell Science, IFAS, University of Florida, Lake Alfred 33885 FL, USA.

4. State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China.

5. Department of Biotechnology and Biomedicine, Technical University of Denmark, Søltofts Plads 2800 Kgs. Lyngby, Denmark.

*Corresponding author: Yayu Wang, Email: wangyayu@genomics.cn

Abstract: Bacterial indole-3-acetic acid (IAA), an effector molecule in microbial physiology, plays an important role in plant growth-promotion. Here, we comprehensively analyzed about 7282 prokaryotic genomes representing diverse bacterial phyla, combined with root-associated metagenomic data to unravel the distribution of tryptophan-dependent IAA synthesis pathways and to quantify the IAA synthesis-related genes in the plant root environments. We found that 82.2% of the analyzed bacterial genomes were potentially capable of synthesizing IAA from tryptophan (Trp) or intermediates. Interestingly, several phylogenetically diverse bacteria showed a preferential tendency to utilize different pathways and tryptamine and indole-3-pyruvate pathways are most prevalent in bacteria. About 45.3% of the studied genomes displayed multiple coexisting pathways, constituting complex IAA synthesis systems. Furthermore, root-associated metagenomic analyses revealed that rhizobacteria mainly synthesize IAA via indole-3-acetamide (IAM) and tryptamine (TMP) pathways and might possess stronger IAA synthesis abilities than bacteria colonizing other environments. The obtained results refurbished our understanding of bacterial IAA synthesis pathways and provided a faster and less labor-intensive alternative to physiological screening based on genome collections. The better understanding of IAA synthesis among bacterial communities could maximize the utilization of bacterial IAA to augment the crop growth and physiological function.

Keywords: Indole-3-acetic acid; Bacteria; Pathways; Genomes; Metagenomes

1. Introduction

Indole-3-acetic acid (IAA) is a crucial phytohormone in plant development, controlling many important physiological processes including cell enlargement and division, tissue differentiation, and responses to light and gravity [1]. The ability to synthesize IAA is a well-characterized trait in many plant growth-promoting bacteria [2]. Several experiments have proven that bacterial IAA enhances the plant growth [3-6]. Paradoxically, bacterial IAA is also known as a notorious virulence factor which could induce gall tumor [7, 8]. In addition to its involvement in bacteria-plant interactions, IAA also plays an important role in microbial physiology. For instance, IAA acts as a signaling molecule. The addition of IAA represses the *vir* regulon and *chv* genes of *Agrobacterium tumefaciens* [9]. Overproduction of IAA in *Sinorhizobium meliloti RD64* exhibits resistance to various stresses, such as UV and salinity; and also activates the tricarboxylic acid (TCA) cycle [10].

© 2019 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

IAA is generally synthesized in two ways, the Trp-dependent and the Trp-independent. However, the Trp-independent pathway is still elusive in bacteria [2]. There are 4 major Trp-dependent IAA synthesis pathways in bacteria (Figure 1): indole-3-pyruvate (IPA), tryptamine (TPM), indole-3-acetonitrile (IAN) and indole-3-acetamide (IAM). The IAM pathway is the best characterized pathway in both phytopathogens and symbiotic bacteria [11]. It converts tryptophan into IAA in two steps. Tryptophan is first converted to IAM by the enzyme tryptophan-2-monooxygenase (*iaaM*). In the second step, IAM is converted to IAA by an IAM hydrolase (*iaaH*). In IPA pathway, IAA is initially transformed into IPA by aminotransferases or L-amino-acid oxidases, and then IPA can be converted into IAA by a two-step ipdC pathway in which IPA is decarboxylated to indole-3-acetaldehyde (IAAld) and then IAAld is converted to IAA by dehydrogenases [2], but in plants IAA can be synthesized from IPA one-step YUCCA pathway (flavin monooxygenase-like enzyme) [12]. For the TPM pathway, Trp is primarily converted into TPM by a decarboxylase and TPM is directly converted to IAAld by an amine oxidase [13]. IAAld is eventually transformed to IAA by dehydrogenases. Trp can be directly converted into IAAld by a Trp side chain monooxygenase enzyme, which is known as the Trp side chain oxidase (TSO) pathway, however, this pathway is only observed in strains from *Pseudomonas fluorescens* [14-16], so the TSO pathway is not described here. The IAN pathway is not well studied in bacteria. The first step is to convert Trp into indole-3-acetaldoxime (IAOX) by an oxidoreductase, however, no orthologues have been identified in bacteria thus far. In the second step, IAOX is converted into indole-3-acetonitrile (IAN) by an indoleacetaldoxime dehydratase. Once IAN is formed, it can be converted to IAA from two pathways, one by nitrilases and the other by NHase/amidase system. NHase links the IAN and IAM pathways by catalyzing IAN into IAM [17]. It has also been confirmed that multiple synthesis pathways can be present in a single bacterium [18]. But how bacteria regulate these pathways to avoid excessive IAA accumulations and whether these pathways are functionally redundant is still enigmatic. It is reported that nitrilase-based and NHase/amidase system-based synthesis pathways have very different regulatory mechanisms [19]. What's more, different pathways function at different status. *Pantoea agglomerans* is a gall-inducing bacterium and possesses IPA and IAM pathways. It's reported that the IPA pathway is related to plant colonization and the IAM pathway is associated with pathogenicity [20].

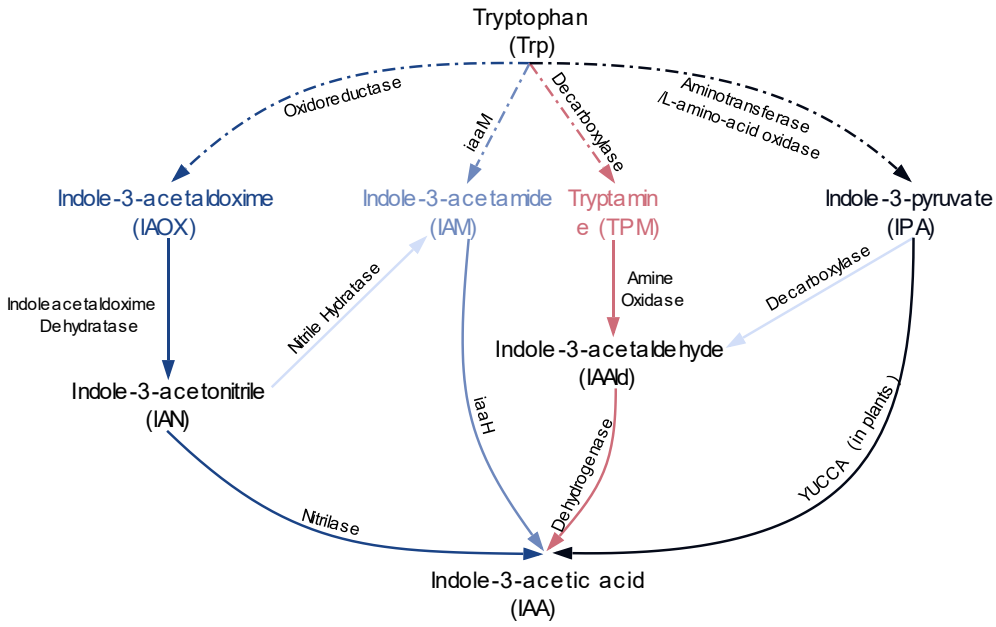


Figure 1. The major microbial Trp-dependent IAA synthesis pathways. Different pathways are depicted with different colors. In our study, we decompose each pathway into three parts: the first part represents the first step of IAA synthesis or the Trp-catalyzing step (Dashed lines in the figure); the second part represents synthesizing IAA from Trp or the de novo IAA synthesis; the third part represents synthesizing IAA from intermediates (IPA, IAM, TPM, IAM).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

The bacterial IAA biosynthetic capacity was studied and confirmed primarily by experiments and *in silico* analysis of individual genome mainly affiliated to Proteobacteria or Actinobacteria [2]. There's a dearth of information regarding the IAA biosynthetic potential in bacteria across diverse phyla and/or in complex communities. The large-scale genomic analysis could possibly provide insights into whether certain species are potentially capable of IAA synthesis, and how they diversify across different phylogenetic clades. Therefore, in this study, we used thousands of bacterial genomes covering different phyla along with plant's root metagenomic quantification data and bins to draw an integrated atlas of Trp-dependent IAA synthesis pathways in bacteria. Our findings greatly advanced our understanding of the complex IAA synthesis systems in bacteria and emphasized that IAA is not only a plant growth-promoting trait, but also plays a potentially important role in microbial physiology.

2. Results

2.1. IAA biosynthesis is widely distributed among different bacterial phyla

We retrieved 7282 complete and draft genomes of microbes across 5 phyla from NCBI, comprising 114 Acidobacteria, 923 Actinobacteria, 281 Bacteroidetes, 1705 Firmicutes and 4259 Proteobacteria, mostly prevailing in soil and root environment [21]. We surveyed the four major Trp-dependent IAA synthesis pathways in these genomes. In the initial step of each IAA synthesis pathway, Trp is first converted into different intermediates followed by distinct pathways to synthesize IAA. We found that different phyla tended to catalyze Trp into disparate intermediates (P value<0.05, Chi-square Test, Figure 2a). In Acidobacteria, Bacteroidetes and Firmicutes, 65.8% and 48.2%, or 17.4% and 13.5%, or 7.8% and 3.8% of genomes could catalyze Trp into TPM and IAM, respectively. In Actinobacteria, 29.7% and 14.2% could catalyze Trp into IAM and TPM, respectively; In Proteobacteria, 68.5% and 11.9% could catalyze Trp into IPA and TPM. For the IPA pathway, Trp is firstly converted into IPA by aromatic amino acid or Trp-specific aminotransferases [22] or L-amino acid oxidases [23] *in vitro*. However, no orthologues of Trp-specific aminotransferases were found in any genomes and only L-amino acid oxidases were observed in Bacteroidetes. In accordance with the previous reports, no genomes exhibited the presence of oxidoreductase orthologues that are capable of transforming Trp into IAOX in our data. Overall, these results showed that Trp was likely to be mainly transformed into TPM via Acidobacteria, Bacteroidetes and Firmicutes, but mainly into IAM by Actinobacteria and into IPA by Proteobacteria.

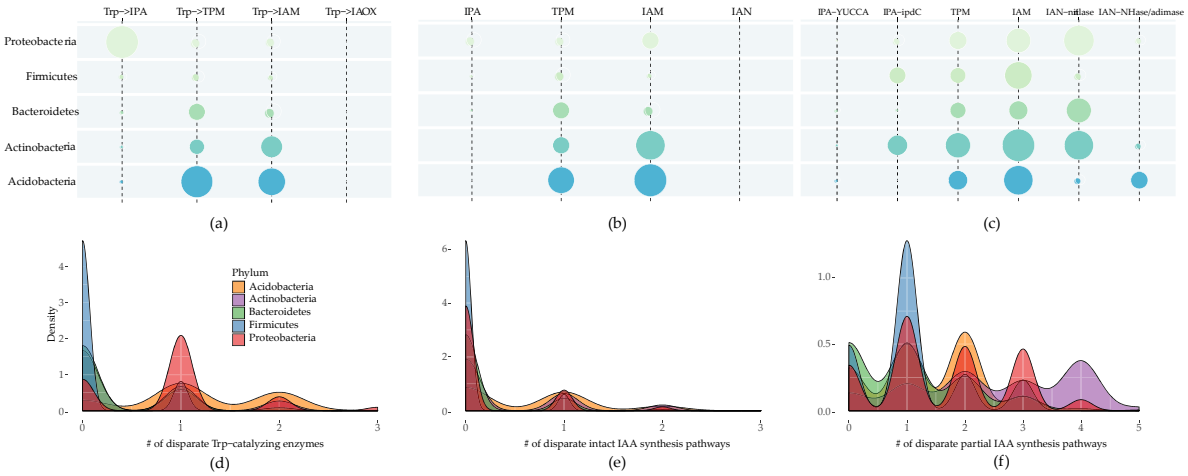


Figure 2. The distribution of IAA synthesis pathways across distinct phylum. The size of bubbles in each bubble plot represents the percentages of detected genomes in each phylum. a) The number of genomes possessing the Trp-catalyzing enzymes. b) The number of genomes potentially capable of *de novo* synthesizing IAA from Trp. c) The number of genomes potentially competent to synthesize IAA from intermediates. The density plots (from d to f) show the distribution of the number of genomes possessing coexisting pathways. The x axes represent the number of coexisting disparate synthesis pathways in an individual genome and y axes represent the density of the number of genomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

After examining the *de novo* IAA synthesis capability of a single bacterial genome from Trp, we found that 50% Acidobacteria, 30.8% Actinobacteria, 15.3% Bacteroidetes, 8.8% Firmicutes and 19.0% Proteobacteria could potentially *de novo* synthesize IAA from Trp. Different preferences for biosynthetic pathways were found across phyla (P value < 0.05, Chi-square Test). In detail, 36.0% and 23.7%, 29.0% and 9.4%, 7.1% and 9.3%, 1.8% and 6.5% or 9.7% and 5.9% of Acidobacteria, Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria could synthesize IAA by IAM and TPM pathways, respectively (Figure 2b, Table S1). Though 68.5% of Proteobacteria could catalyze Trp into IPA, only 6.6% of those bacteria possessed the intact IPA pathway. Taken together, IAM and TPM were the dominant *de novo* IAA synthesis pathways in bacteria, but varied in phyla.

Although many genomes couldn't assimilate Trp, some experimental assays verified that some bacteria could synthesize IAA when being supplied with exogenous intermediate metabolites [19], so we scanned how many genomes could potentially synthesize IAA in the presence of key intermediate metabolites (IPA, TPM, IAM and IAN). Surprisingly, the majority of genomes owned this ability (78.9% of Acidobacteria, 89.4% of Actinobacteria, 63.3% of Bacteroidetes, 78.4% of Firmicutes and 83.6% of Proteobacteria). IPA could be transformed into IAA by the one-step YUCCA pathway in plants, in our data, we found a small proportion of bacteria contained the homologues of YUCCA, however, the vast majority of bacteria (14.3% of total genomes) used the two-step ipdC pathway (Figure 2c, Supplementary Table 1). IAN could also be transformed into IAA by two distinct pathways, first by one-step nitrilase and the second by two-step NHase/amidase system. Opposite to the IPA pathway, 56.8% of bacterial genomes possess the one-step nitrilase pathway. However, only a minority of bacterial nitrilases identified by in silico analysis were experimentally reported to be capable of converting IAN into IAA [24]. A strong preference for the IAM and TPM pathways were observed in Acidobacteria, and IAM and IPA pathways were dominant in Firmicutes; and one-step nitrilase and IAM pathways were primary in Actinobacteria, Bacteroidetes, and Proteobacteria. As aforementioned, only several genomes in Actinobacteria (11 genomes) and Firmicutes (53 genomes) were capable of converting Trp to IPA, but 298 and 379 genomes from these two phyla could potentially synthesize IAA by the two-step ipdC pathway from IPA. These results suggested that majority of genomes potentially participated in synthesizing IAA from intermediates despite lacking the genes responsible for utilizing Trp, and bacteria from different phyla tended to use incongruent pathways (P value < 0.05, Chi-square Test).

2.2. Multiple distinct biosynthetic pathways coexist in bacteria

It is already known that multiple IAA synthesis pathways are present in a single bacterium, however, the quantified description of how many IAA synthesis pathways could coexist in a genome is still unexplored. Here, we calculated the number of coexisting disparate IAA synthesis pathways in each genome. The result showed that up to 3 disparate Trp-catalyzing enzymes coexisted in 119 genomes (1 from Acidobacteria, 1 from Actinobacteria and 117 from Proteobacteria, Figure 2d, Table S2). For the majority of genomes, only 1 Trp-catalyzing enzyme existed. 32.5% of Acidobacteria contained 2 disparate Trp-catalyzing enzymes, followed by 10.6% of Actinobacteria. As for the coexistence of various intact IAA synthesis pathways in the individual genome, we observed that 5 genomes from Proteobacteria possessed 3 different intact IAA synthesis pathways, both of which simultaneously possess IPA, IAM and TPM pathways (Figure 2e), but a plethora of genomes exclusively owned one complete IAM or TPM synthesis pathway. In particular, all the genomes from Firmicutes only contained a single complete pathway. IAM and TPM were the most frequently coexisting intact pathways (Table S3). Further detailed investigation of these genomes revealed that 45.3% of genomes possessed unexpectedly complex yet incomplete IAA synthesis pathways, which could potentially produce IAA from intermediate metabolites (Figure 2f). Moreover, for the genomes which were potentially capable of synthesizing IAA from intermediates, 57.7% of Acidobacteria, 82.6% of Actinobacteria, 43.2% of Bacteroidetes, 28.4% of Firmicutes and 59.4% of Proteobacteria genomes possessed multiple IAA synthesis pathways (≥ 2 diverse pathways). A total of 11, 13, 14, 19 and 26 various combinations of coexisting complete or incomplete pathways were found in Acidobacteria, Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria, respectively (Table S4). Rarefaction analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

showed that the number of distinct combinations in Actinobacteria, Firmicutes and Proteobacteria had come to saturations due to the great number of genome repertoires (Figure S1). These results suggest that Firmicutes genomes have the simplest Trp-dependent IAA synthetic systems and a plethora of bacteria could potentially synthesize IAA via multiple incomplete pathways.

2.3. Quantification of IAA synthetic genes in root environments

IAA production is a common feature for plant growth-promoting bacteria, however little is known about how IAA is produced by the complex microbial communities in plant-associated environment. Here we quantified the abundance of different IAA synthesis-related genes in root environments and determined which pathways were dominant in such an ecosystem. We conducted metagenomic sequencing of 27 rhizosphere (RS, the soil microbiota loosely attached to roots) and 19 rhizoplane (RP, the soil microbiota closely attached to roots) samples of foxtail millet (*Setaria italica*) and obtained a non-redundant gene catalogue comprising 67 Mb genes. Even in our deep sequencing data, no orthologues of oxidoreductases were identified that could convert Trp into IAXO. The results also showed that there was no obvious difference in the compositions of IAA synthesis-related genes between rhizosphere and rhizoplane (Figure 3a and 3b). Trp was found to be mainly converted into IAM (The abundance relative to Trp-catalyzing enzymes: 54.2% in RP, 54.7% in RS) or TPM (35.9% in RP, 41.4% in RS) and IAA was generally produced by IAM (The abundance relative to IAA-producing enzymes: 37.4% in RP, 43.7% in RS) or IAAld (53.5% in RP, 46.5% in RS) in both rhizosphere and rhizoplane compartments.

Many Trp-catalyzing genes were affiliated to unknown taxa, followed by Proteobacteria and Actinobacteria in both of the two sampling compartments (Figure 3c, Figure S2a). However, the taxonomic compositions of enzymes catalyzing IAA synthesis through intermediate metabolites exhibited distinct results in the two compartments. In rhizosphere, still a large proportion of uncharacterized taxa prevailed, whereas, discrepant to rhizosphere, Proteobacteria and Actinobacteria flourished in rhizoplane (Figure 3d, Figure S2). Thus, this result suggested that several unreported taxa potentially participated in IAA synthesis and Proteobacteria and Actinobacteria were important agents for IAA synthesis in root environments, indicating that the IAA-producing trait was widely distributed in root-colonizing bacteria. Though there was negligibly slight difference in IAA-synthesis related functional abundance in the two compartments, the taxonomic composition differed a lot. This phenomenon could be attributed to the highly functional redundancy of the IAA-producing trait in root-colonizing bacteria.

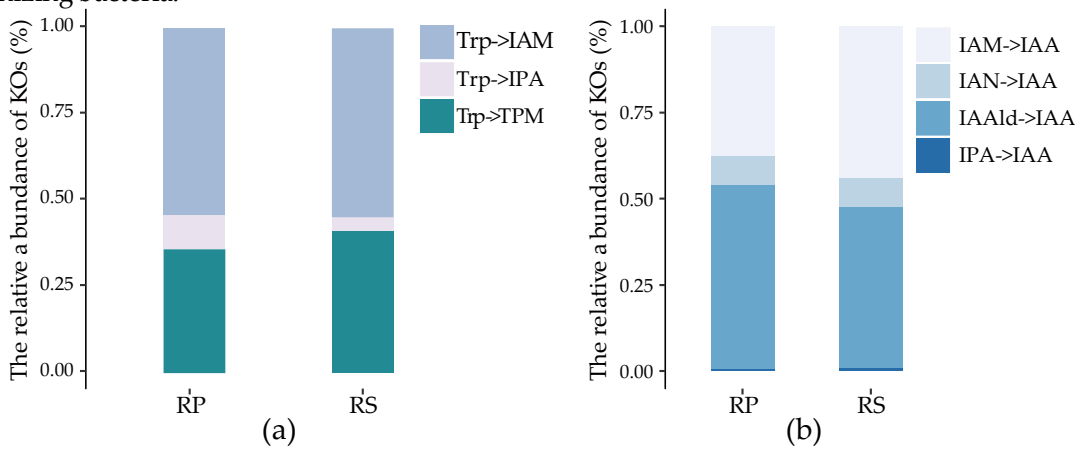
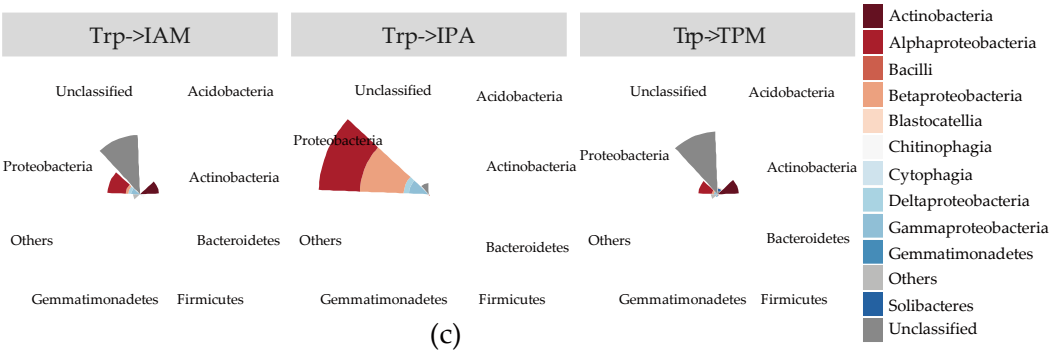


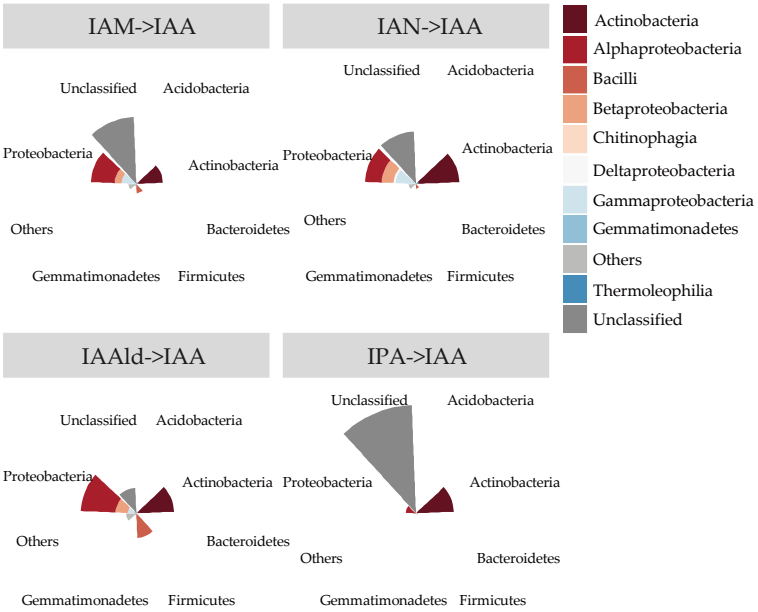
Figure 3. *Cont.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411



(c)



(d)

Figure 3. Quantification of the IAA synthesis related genes in root environments. The bar plots show the relative abundance of a) Trp-catalyzing enzymes and b) enzymes in last step of IAA synthesis in rhizosphere (RS) and rhizoplane (RP), respectively. c and d represent the taxonomic structures of Trp-catalyzing enzymes and IAA synthesis enzymes respectively.

2.4. The stronger capacity of IAA synthesis for rhizobacteria

To compare the IAA synthesis abilities between rhizobacteria and retrieved genomes from diverse environments, we binned 563 high-quality metagenomic assembled genomes (≥ 70% completeness, ≤10% contamination, here we termed these bins as "rhizobacteria") from our metagenomic data. Phylogenetic annotations showed that 117 bins were affiliated to Proteobacteria and 60 were from Acidobacteria, 49 from Bacteroidetes, 29 from Firmicutes and 24 from Actinobacteria (for a more detailed annotations of all the bins, please refer to Table S5). We found that 75.0% (422) of these rhizobacteria could convert Trp into intermediate metabolites in IAA synthesis pathway. While 60.0% and 92.2% of rhizobacteria could synthesize IAA from Trp or intermediate metabolites, respectively. However, the corresponding values in the retrieved genomes were 53.8%, 18.5% and 82.2% respectively. Additionally, bins tended to possess multiple distinct IAA synthesis pathways in each genome (the percent of bins and retrieved genomes possessing multiple complete pathways: 38.3% vs. 3.0%; multiple Trp-catalyzing enzymes: 45.6% vs. 10.2%, Figure S3). These results indicated that rhizobacteria have very high potential to synthesize IAA. Moreover, the ability to synthesize IAA from intermediates IAM and TPM is widely distributed in almost all the rhizobacteria (Figure 4), which was consistent with the abovementioned results from metagenomic profiling. Intriguingly, 59 out of 60 rhizobacteria from Acidobacteria could transform Trp into IPA but only 2 out of 114 retrieved Acidobacteria owned this capacity. The *iaaM* genes were lost in all the rhizobacteria from Acidobacteria

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

and were only observed in Proteobacteria and Planctomycetes (Figure 4). Rhizobacteria from Firmicutes were absent of all the genes participating in Trp transformations, but 14.7% of retrieved Firmicutes possessed these genes. One-step IPA-YUCCA pathway was exclusively observed in 1 Bacteroidetes, 3 Firmicutes and 8 Gemmatimonadetes rhizobacterial genomes. These results highlighted that the patterns of IAA synthesis pathways in rhizobacteria were different from bacteria within the same phyla but isolated from other environments, which might be selected by the specific secretions from plant roots.

IAA synthesis-related genes are supposed to locate either on chromosomes or plasmids. Those genes which are located in plasmids usually represent higher levels of transcription, as plasmids frequently occur with high copy numbers in bacterial genomes and have the opportunity to transfer and insert into the plant chromosomes [25]. To identify the location of related genes in each rhizobacterial genomes, we predicted the plasmid scaffolds by Plasflow [26] and mapped those genes in the predicted plasmids. We found that the majority of genes were located on chromosomes, however one clade (leaves from bins_290 to bins_328 in the Figure 4) from Proteobacteria showed the highest proportions of genes located on plasmids. Rhizobacteria affiliated to Proteobacteria and Acidobacteria tended to locate IAA synthesis-related genes on plasmids compared with rhizobacteria from other phyla. The two phyla showed a close phylogenetic relationship, so this tendency might arise with the bacterial evolution and is possibly maintained by the vertical inheritance.

Figure 4. The phylogeny and distribution of IAA synthesis pathways in rhizobacteria. The tips of the tree are masked with different colors to indicate corresponding phylum. The histograms represent the percent of genes in each pathway located on plasmids. The 3 inner rings and 6 outside rings show the distribution of different Trp-catalyzing enzymes and capacities for synthesizing IAA from intermediates by different pathways, respectively. The blank space along the rings represent no related genes or pathways are found in the rhizobacterium.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

3. Discussion

The access to tens of thousands of genome data facilitates the understanding of the genetic foundations of numerous phenotypes. In this study, we conducted a large-scale genomic analysis to uncover the Trp-dependent IAA synthesis pathways in bacteria and quantified related genes in the root environments by metagenomics. Bacterial IAA is reported to mediate microbial physiology, such as overcoming stress, behaving as a signaling molecule for different cellular processes and serving as a carbon and nitrogen source [27-29]. We found that 18.5% or 82.2% of bacteria from diverse environments could potentially synthesize IAA from Trp or intermediates, respectively, so this phenomenon emphasized that IAA probably has a profound impact on shaping microbial physiology. We speculate that either the required intermediates are rich in their habitat or the transfer of intermediates among individuals in communities might lead to the plethora of bacterial genomes which can synthesize IAA from the intermediates. Similarly, a recent study on cobamide biosynthesis in bacteria also revealed that many bacteria only contained incomplete pathways and the authors experimentally verified that bacteria could truly synthesize cobamide from intermediates [30]. Intriguingly, oxidoreductase orthologues that catalyze Trp into IAOX does not showed any hit, neither in the individual genomes nor in the deeply sequenced metagenomes, implying that the sequences are too divergent or the strategy for producing IAOX in microbes is quite different from plants. IPA pathway is the sole pathway used to study the regulation of IAA synthetic pathways, as IPA is deemed to be the most common pathway in bacteria [2, 17]. Different from previous knowledge in our data, we found bacteria from different phyla favored different pathways, but IAM and TPM showed the dominance. In the previous studies, the identification of the IPA pathway in bacteria were mainly confirmed by the detection of aminotransferases or ipdC, and most of those scanned bacteria were from Proteobacteria [2]. Consistent with our results, most bacteria from Proteobacteria did contain aminotransferases, but many of them lacked the enzymes required in the subsequent steps of IPA pathway. Hence, mere identification of the genes essential for the IPA pathway doesn't imply that bacteria could synthesize IAA via this pathway, as multiple pathways could also coexist in a single genome. Therefore, we also characterized the coexisting pathways in each genome. Intriguingly, there were up to 3 distinct yet intact pathways or 4 various incomplete pathways were found to coexist in several genomes. Different pathways were linked to a difference in regulation modes [20], which could help to maintain continuous IAA synthesis under diverse environments.

IAA plays a profound role in plant-microbe interactions. Both beneficial and pathogenic bacteria are able to produce IAA. IAA synthesized by plant growth-promoting bacteria isolated from root environments could participate in nodule formation and increase host root surfaces by promoting lateral roots' growth [31, 32]. To the best of our knowledge, this is the first time to quantify the IAA biosynthesis related genes in root environments. Though, both IAAld and IAM could be produced by two pathways respectively, the high abundance of decarboxylases and iaaM in metagenomic data might suggest that TPM and IAM were the major IAA synthesis pathways in root environments, besides, Proteobacteria and Actinobacteria were essential performers for the IAA synthesis. We binned 563 high-quality rhizobacterial genomes from our metagenomic data. Pioneer works screened the auxin productivity in 1151 plant-associated bacterial strains and verified that 229 strains produced auxin from Trp [33]. However, much higher potential was found in our metagenomic bins, of which 60.0% could potentially *de novo* synthesize IAA from Trp. This deviation might be caused by erroneous clustering in bins or the inappropriate conditions for detecting IAA synthesis in culture media. When compared with the retrieved genomes from diverse environments, we found that rhizobacteria likely showed more effective IAA synthetic abilities. Having the powerful ability to synthesize IAA must give rhizobacteria a selective superiority in root environments. It has speculated that microbial IAA could promote plants' susceptibility to both pathogen and plant growth-promoting bacteria [34]. Microbial IAA is assumed to alter auxin homeostasis in plants and the outcomes of this interaction are determined by the endogenous IAA level in plants [35]. Root exudates are rich in carbohydrates and IAA could facilitate microbes to uptake carbohydrates by increasing the expression of ATP-binding cassette transporters and tripartite ATP-independent periplasmic transporters [28]. Microbes also can obtain Trp from root exudates [36] and this might explain why rhizobacteria

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

have the stronger capacity to *de novo* synthesize IAA from Trp. In addition, IAA could facilitate the expressions of the type VI secretion systems and effectors of type III secretion system which participate in plant-microbe and microbe-microbe interactions in root environments [28, 37]. Furthermore, IAA is engaged in inducing antibiotic production which could be equipped as a weapon against other microbes competing for the same resources in root environments or prevent host plants from pathogens [38]. Taken together, microbial IAA could contribute to the successful colonization of microbes in root environments.

4. Materials and Methods

4.1. Genome retrieval and analysis

We searched and downloaded the complete and draft genomes across five phyla from NCBI, including Acidobacteria, Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria. Totally, 7282 genomes were retrieved. The downloading ftp for the analyzed genomes could be found in the supplementary information (Table S6). We re-predicted the genes in those genomes by Prodigal [39] and then mapped the genes against the KEGG database (Ver. 81) [40] by Diamond [41] to search for the Trp-dependent IAA synthesis-related genes (map00380, the KOs for each pathway are listed in Table S7). The alignment results were filtered by using E-value < 1e-5 and alignment score ≥ 60. Only the best hits were retained after filtration. To evaluate the accuracy of functional annotation, all the genomes from Acidobacteria and Bacteroidetes were aligned against COG database by Blastp [42] and only the KOs with COG tags were considered. The results showed 80.6% of the KEGG-identified genes had the same annotations with the COG-identified genes. Though COG database identified more IAA synthesis-related genes than KEGG database did, to decrease the ambiguous alignments, only the annotations from KEGG database were retained.

4.2. Sample collection for metagenomic sequencing

Foxtail millet seeds were sowed in the field in Yangling (Shanxi, China). For each cultivar, the root samples from three individuals were harvested during the ripening stage and separated into rhizosphere and rhizoplane samples using an established protocol as our previous study [21]. All the samples were placed in 15ml clean microcentrifuge tubes and transferred back to the lab under -20°C. Rhizosphere samples were collected by scraping the soil loosely attached to the root surface and rhizoplane samples were collected from the tightly adhering soil that was washed off by vortexing for 30 minutes using the mixture of 1×PBS and adsorbent Silwet-77. DNA extraction procedure was done according to the protocols by the MoBio PowerSoil toolkit. Unfortunately, several samples failed to yield a good quality DNA from rhizosphere or rhizoplanes. However, we managed to obtain 27 DNA samples from rhizosphere and 19 DNA samples from rhizoplanes.

4.3. Metagenomic sequencing and processing

Four samples were pooled together in equal amounts and libraries were constructed to be sequenced on BGISEQ500 platform. The raw sequencing reads have been deposited in NCBI Bioproject database under the accession number PRJNA450295 and is also accessible at the CNGBdb (China National GeneBank DataBase) under the project CNP0000240. Raw reads were filtered by SOAPnuke (v1.5.3) [43] with parameters setting as "filterMeta -Q 2 -S -L 15 -N 3 -P 0.5 -q 20 -l 60 -R 0.5 -5 0". Then the clean reads were aligned against the foxtail millet genome (v2.3, http://foxtailmillet.genomics.org.cn/page/species/download.jsp) [44] to remove host contaminations using bowtie2 [45]. Assembled contigs with length less than 500 bp were excluded from the subsequent analyses. Genes were predicted over the contigs assembled from each test sample by Prodigal [46] under the "Meta" mode and only genes with a length greater than or equal to 150bp were retained. All of the nucleic acid sequences of genes were merged and input into CD-HIT [47] with 0.9 coverage and 0.95 identity to generate the non-redundant gene catalog.

Functional annotations were implemented by mapping non-redundant genes against KEGG (version.81) [48] by Diamond [49], and filtering criteria were the same as the genome annotation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

step. Taxonomic annotations were executed by mapping genes to the NCBI "nr" database (release date: 20160219) with Diamond, and alignments were filtered with coverage ≥0.8 and identity ≥0.65. Taxonomic information was obtained using an in-house lowest common ancestor (LCA) pipeline [50].

The clean reads from each sample sequenced on the BGISEQ500 platform were aligned against the gene catalog by Bowtie2 with parameters "--sensitive --dpad 0 --gbar 99,999,999 --mp 1,1 --np 1 -I 100 -X 500 --score-min L,0,-0.1 -k 50--no-discordant--no-unal --no-sq --no-head". The abundance of each gene in each sample was calculated as the number of reads mapped to the gene divided by the length of the gene. KO profile was generated by summing up the abundance of genes that affiliated to the same KO. Taxonomy profile at the phylum or class level was generated by summing up the abundance of genes affiliated to the same phylum or class.

4.4. Metagenomic binning

Metagenomic assembled genomes were generated by the following steps: 1) All of the 46 clean metagenomic data were assembled by Megahit [51]. A total of 8 samples were co-assembled and while the remaining 38 samples were assembled separately. Finally, 39 assemblies were generated in total. 2) Only the scaffolds with length ≥ 2000 bp were retained for metagenomic binning. The clean metagenomic reads from all the 46 samples were mapped to each assembly by bowtie2 [45], resulting in 46*39 bam files. 3) Scaffolds from each assembly and the corresponding bam files were used as input files of Metabat2 [52]. 4) The bins from each assembly were evaluated for completeness and contaminations by CheckM [53]. Only those bins with completeness ≥ 70% and contamination ≤ 10% were retained for the following analysis. 5) To remove the redundancy of bins from different assemblies, we calculated the gANI by ANIcalculator [54]. Bins were clustered based on the gANI ≥ 99.9% and alignment fraction (AF) ≥ 90%. However, no bins could be clustered due to the low AF. Each bin was as input into Plasflow [26] to predict the plasmid sequences.

4.5. Phylogenetic annotations

Retained bins were searched for the 16 single-copy marker genes as previously described [55] by HMMsearch [56]. Bins with ≥ 8 marker genes and all single-copy genes were kept for the phylogenetic analysis. Also, prokaryotic genomes marked as "Reference" or "Representative" in Refseq [57] were downloaded and searched for these marker genes. The protein sequences of each marker gene were multiple aligned by MUSCLE [58] and the columns containing more than 70% of gaps were removed. The concatenated alignments were used to generate the phylogenetic trees by Fastree [59] and visualized using iTOL [60]. The taxonomic annotations of bins were assigned by Tax2tree [61].

5. Conclusions

In summary, our work provided a comprehensive atlas of Trp-dependent IAA synthesis pathways across different bacterial phyla. However, experimental assays are needed to verify these conclusions in future. A full understanding of how microbes synthesize IAA may enable the maximization of the beneficial effects of plant growth-promoting bacteria and reduction of the virulence of IAA-producing phytopathogens. Finally, the identification of high percentage of microbes that could potentially synthesize IAA further calls for an in-depth research to explore the role of IAA in mediating microbial physiology.

Supplementary Materials: The following are available online, Figure S1: Rarefaction curves displaying the number of combinations of IAA synthesis routes in each phylum, Figure S2: The bacterial taxonomic compositions of IAA synthesis-related genes in rhizosphere, Figure S3: The comparison of IAA synthesis capacity between metagenomic bins and retrieved genomes, Table S1: The number of genomes possess IAA-synthesis related enzymes in each pathway, Table S2: The number of genomes in each phylum own multiple coexisting pathways, Table S3: The number of genomes contain complete synthesis pathways, Table S4: The different combinations of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at Molecules 2019, 24, 1411; doi:10.3390/molecules24071411

synthesis pathways in each phylum, Table S5: The taxonomic assignments of metagenomic bins, Table S6: The ftp downloading paths of retrieved genomes, Table S7: The KOs scanned in each synthesis pathway.

Author Contributions: P. Z. and Y. W. conceived the project. P. Z. collected the root samples, bacterial genomes and analyzed the data. P. Z. wrote the draft manuscript. S. S., Y. W., T. J., J. X., H. L. and Q. S. revised the manuscript.

Acknowledgments: This project was financially supported by the Shenzhen Municipal Government of China (grant numbers JCYJ20150831201643396), the State Key Laboratory of Agricultural Genomics (grant number 2011DQ782025), and the Guangdong Provincial Key Laboratory of Genome Read and Write (grant number 2017B030301011).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhao, Y., Auxin biosynthesis and its role in plant development. *Annual review of plant biology* **2010,** 61, 49-64.
2. Duca, D.; Lorv, J.; Patten, C. L.; Rose, D.; Glick, B. R., Indole-3-acetic acid in plant–microbe interactions. *Antonie Van Leeuwenhoek* **2014,** 106, (1), 85-125.
3. Fierro-Coronado, R. A.; Quiroz-Figueroa, F. R.; García-Pérez, L. M.; Ramírez-Chávez, E.; Molina-Torres, J.; Maldonado-Mendoza, I. E., IAA-producing rhizobacteria from chickpea (Cicer arietinum L.) induce changes in root architecture and increase root biomass. *Canadian journal of microbiology* **2014,** 60, (10), 639-648.
4. Gamalero, E.; Glick, B. R., Mechanisms used by plant growth-promoting bacteria. In *Bacteria in agrobiology: Plant nutrient management*, Springer: 2011; pp 17-46.
5. Patten, C. L.; Glick, B. R., Role of Pseudomonas putida indoleacetic acid in development of the host plant root system. *Applied and environmental microbiology* **2002,** 68, (8), 3795-3801.
6. Bianco, C.; Defez, R., Improvement of phosphate solubilization and Medicago plant yield by an indole-3-acetic acid-overproducing strain of Sinorhizobium meliloti. *Applied and environmental microbiology* **2010,** 76, (14), 4626-4632.
7. Hooykaas, P.; Beijersbergen, A. G., The virulence system of Agrobacterium tumefaciens. *Annual review of phytopathology* **1994,** 32, (1), 157-181.
8. Chalupowicz, L.; Barash, I.; Schwartz, M.; Aloni, R.; Manulis, S., Comparative anatomy of gall development on Gypsophila paniculata induced by bacteria with different mechanisms of pathogenicity. *Planta* **2006,** 224, (2), 429-437.
9. Yuan, Z.-C.; Liu, P.; Saenkham, P.; Kerr, K.; Nester, E. W., Transcriptome profiling and functional analysis of Agrobacterium tumefaciens reveals a general conserved response to acidic conditions (pH 5.5) and a complex acid-mediated signaling involved in Agrobacterium-plant interactions. *Journal of Bacteriology* **2008,** 190, (2), 494-507.
10. Imperlini, E.; Bianco, C.; Lonardo, E.; Camerini, S.; Cermola, M.; Moschetti, G.; Defez, R., Effects of indole-3-acetic acid on Sinorhizobium meliloti survival and on symbiotic nitrogen fixation and stem dry weight production. *Applied microbiology and biotechnology* **2009,** 83, (4), 727.
11. Spaepen, S.; Vanderleyden, J.; Remans, R., Indole-3-acetic acid in microbial and microorganism-plant signaling. *FEMS microbiology reviews* **2007,** 31, (4), 425-448.
12. Mano, Y.; Nemoto, K., The pathway of auxin biosynthesis in plants. *Journal of experimental Botany* **2012,** 63, (8), 2853-2872.
13. Hartmann, A.; Singh, M.; Klingmüller, W., Isolation and characterization of Azospirillum mutants excreting high amounts of indoleacetic acid. *Canadian Journal of Microbiology* **1983,** 29, (8), 916-923.
14. Suzuki, S.; He, Y.; Oyaizu, H., Indole-3-acetic acid production in Pseudomonas fluorescens HP72 and its association with suppression of creeping bentgrass brown patch. *Current Microbiology* **2003,** 47, (2), 0138-0143.
15. Oberhänsli, T.; Défago, G.; Haas, D., Indole-3-acetic acid (IAA) synthesis in the biocontrol strain CHA0 of Pseudomonas fluorescens: role of tryptophan side chain oxidase. *Microbiology* **1991,** 137, (10), 2273-2279.
16. Whistler, C. A.; Corbell, N. A.; Sarniguet, A.; Ream, W.; Loper, J. E., The Two-Component Regulators GacS and GacA Influence Accumulation of the Stationary-Phase Sigma Factor ς S and the Stress Response in Pseudomonas fluorescensPf-5. *Journal of bacteriology* **1998,** 180, (24), 6635-6641.
17. Spaepen, S.; Vanderleyden, J., Auxin and plant-microbe interactions. *Cold Spring Harbor perspectives in biology* **2010**, a001438.
18. Patten, C. L.; Blakney, A. J.; Coulson, T. J., Activity, distribution and function of indole-3-acetic acid biosynthetic pathways in bacteria. *Critical reviews in microbiology* **2013,** 39, (4), 395-415.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at *Molecules* **2019**, *24*, 1411; doi:10.3390/molecules24071411

19. Sun, S.-L.; Yang, W.-L.; Fang, W.-W.; Zhao, Y.-X.; Guo, L.; Dai, Y.-J., The plant growth-promoting rhizobacterium Variovorax boronicumulans CGMCC 4969 regulates the level of indole-3-acetic acid synthesized from indole-3-acetonitrile. *Applied and environmental microbiology* **2018**, AEM. 00298-18.
20. Manulis, S.; Haviv-Chesner, A.; Brandl, M. T.; Lindow, S. E.; Barash, I., Differential involvement of indole-3-acetic acid biosynthetic pathways in pathogenicity and epiphytic fitness of Erwinia herbicola pv. gypsophilae. *Molecular Plant-Microbe Interactions* **1998,** 11, (7), 634-642.
21. Jin, T.; Wang, Y.; Huang, Y.; Xu, J.; Zhang, P.; Wang, N.; Liu, X.; Chu, H.; Liu, G.; Jiang, H., Taxonomic structure and functional association of foxtail millet root microbiome. *GigaScience* **2017**.
22. Kittell, B. L.; Helinski, D. R.; Ditta, G. S., Aromatic aminotransferase activity and indoleacetic acid production in Rhizobium meliloti. *Journal of bacteriology* **1989,** 171, (10), 5458-5466.
23. Law, D. M., Gibberellin - enhanced indole - 3 - acetic acid biosynthesis: D - Tryptophan as the precursor of indole - 3 - acetic acid. *Physiologia plantarum* **1987,** 70, (4), 626-632.
24. Podar, M.; Eads, J. R.; Richardson, T. H., Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. *BMC evolutionary biology* **2005,** 5, (1), 42.
25. Follin, A.; Inzé, D.; Budar, F.; Genetello, C.; Van Montagu, M.; Schell, J., Genetic evidence that the tryptophan 2-mono-oxygenase gene of Pseudomonas savastonoi is functionally equivalent to one of the T-DNA genes involved in plant tumour formation by Agrobacterium tumefaciens. *Molecular and General Genetics MGG* **1985,** 201, (2), 178-185.
26. Krawczyk, P. S.; Lipinski, L.; Dziembowski, A., PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic acids research* **2018,** 46, (6), e35-e35.
27. Bianco, C.; Imperlini, E.; Calogero, R.; Senatore, B.; Amoresano, A.; Carpentieri, A.; Pucci, P.; Defez, R., Indole-3-acetic acid improves Escherichia coli's defences to stress. *Archives of Microbiology* **2006,** 185, (5), 373-382.
28. Van Puyvelde, S.; Cloots, L.; Engelen, K.; Das, F.; Marchal, K.; Vanderleyden, J.; Spaepen, S., Transcriptome analysis of the rhizosphere bacterium Azospirillum brasilense reveals an extensive auxin response. *Microbial ecology* **2011,** 61, (4), 723-728.
29. Tromas, A.; Perrot-Rechenmann, C., Recent progress in auxin biology. *Comptes rendus biologies* **2010,** 333, (4), 297-306.
30. Shelton, A. N.; Seth, E. C.; Mok, K. C.; Han, A. W.; Jackson, S. N.; Haft, D. R.; Taga, M. E., Uneven distribution of cobamide biosynthesis and dependence in bacteria predicted by comparative genomics. *bioRxiv* **2018**, 342006.
31. Mathesius, U., Goldacre paper: Auxin: at the root of nodule development? *Functional Plant Biology* **2008,** 35, (8), 651-668.
32. Xie, H.; Pasternak, J.; Glick, B. R., Isolation and characterization of mutants of the plant growth-promoting rhizobacterium Pseudomonas putida GR12-2 that overproduce indoleacetic acid. *Current Microbiology* **1996,** 32, (2), 67-71.
33. Finkel, O. M.; Castrillo, G.; Paredes, S. H.; González, I. S.; Dangl, J. L., Understanding and exploiting plant beneficial microbes. *Current opinion in plant biology* **2017,** 38, 155-163.
34. Fu, J.; Liu, H.; Li, Y.; Yu, H.; Li, X.; Xiao, J.; Wang, S., Manipulating broad-spectrum disease resistance by suppressing pathogen-induced auxin accumulation in rice. *Plant physiology* **2010**, pp. 110.163774.
35. Patten, C. L.; Glick, B. R., Bacterial biosynthesis of indole-3-acetic acid. *Canadian journal of microbiology* **1996,** 42, (3), 207-220.
36. Ryu, R. J.; Patten, C. L., Aromatic amino acid-dependent expression of indole-3-pyruvate decarboxylase is regulated by TyrR in Enterobacter cloacae UW5. *Journal of bacteriology* **2008,** 190, (21), 7200-7208.
37. Yang, S.; Zhang, Q.; Guo, J.; Charkowski, A. O.; Glick, B. R.; Ibekwe, A. M.; Cooksey, D. A.; Yang, C.-H., Global effect of indole-3-acetic acid biosynthesis on multiple virulence factors of Erwinia chrysanthemi 3937. *Applied and environmental microbiology* **2007,** 73, (4), 1079-1088.
38. Matsukawa, E.; Nakagawa, Y.; Iimura, Y.; Hayakawa, M., Stimulatory effect of indole-3-acetic acid on aerial mycelium formation and antibiotic production in Streptomyces spp. *Actinomycetologica* **2007,** 21, (1), 32-39.
39. Hyatt, D.; Chen, G.-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J., Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **2010,** 11, (1), 119.
40. Kanehisa, M.; Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **2000,** 28, (1), 27-30.
41. Buchfink, B.; Xie, C.; Huson, D. H., Fast and sensitive protein alignment using DIAMOND. *Nature methods* **2014,** 12, (1), 59.
42. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997,** 25, (17), 3389-3402.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2019 doi:10.20944/preprints201904.0115.v1

Peer-reviewed version available at *Molecules* **2019**, *24*, 1411; doi:10.3390/molecules24071411

43. Chen, Y.; Shi, C.; Huang, Z.; Zhang, Y.; Li, S.; Li, Y.; Ye, J.; Yu, C.; Li, Z., SOAPnuke: A MapReduce Acceleration supported Software for integrated Quality Control and Preprocessing of High-Throughput Sequencing Data. *Gigascience* **2017**.
44. Zhang, G.; Liu, X.; Quan, Z.; Cheng, S.; Xu, X.; Pan, S.; Xie, M.; Zeng, P.; Yue, Z.; Wang, W., Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nature biotechnology* **2012,** 30, (6), 549.
45. Langmead, B.; Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nature methods* **2012,** 9, (4), 357.
46. Hyatt, D.; Chen, G. L.; Locascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J., Prodigal prokaryotic dynamic programming genefinding algorithm. *Bmc Bioinformatics* **2010,** 11.
47. Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006,** 22, (13), 1658.
48. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000,** 27, (1), 29-34.
49. Buchfink, B.; Xie, C.; Huson, D. H., Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **2015,** 12, (1), 59.
50. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K. S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T., A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010,** 464, (7285), 59.
51. Li, D.; Luo, R.; Liu, C.-M.; Leung, C.-M.; Ting, H.-F.; Sadakane, K.; Yamashita, H.; Lam, T.-W., MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **2016,** 102, 3-11.
52. Kang, D. D.; Froula, J.; Egan, R.; Wang, Z., MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **2015,** 3, e1165.
53. Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W., CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **2015**, gr. 186072.114.
54. Varghese, N. J.; Mukherjee, S.; Ivanova, N.; Konstantinidis, K. T.; Mavrommatis, K.; Kyrpides, N. C.; Pati, A., Microbial species delineation using whole genome sequences. *Nucleic acids research* **2015,** 43, (14), 6761-6771.
55. Anantharaman, K.; Brown, C. T.; Hug, L. A.; Sharon, I.; Castelle, C. J.; Probst, A. J.; Thomas, B. C.; Singh, A.; Wilkins, M. J.; Karaoz, U., Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature communications* **2016,** 7, 13219.
56. Eddy, S. R., Profile hidden Markov models. *Bioinformatics (Oxford, England)* **1998,** 14, (9), 755-763.
57. Pruitt, K. D.; Tatusova, T.; Maglott, D. R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **2006,** 35, (suppl_1), D61-D65.
58. Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **2004,** 32, (5), 1792-1797.
59. Price, M. N.; Dehal, P. S.; Arkin, A. P., FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **2010,** 5, (3), e9490.
60. Letunic, I.; Bork, P., Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **2006,** 23, (1), 127-128.
61. McDonald, D.; Price, M. N.; Goodrich, J.; Nawrocki, E. P.; DeSantis, T. Z.; Probst, A.; Andersen, G. L.; Knight, R.; Hugenholtz, P., An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal* **2012,** 6, (3), 610.

.