Classification of the plant-associated phenotype of *Pseudomonas* strains using genome properties and machine learning

Wasin Poncheewin[1], Anne D. van Diepeningen[2], Theo AJ van der Lee[2], Maria Suarez-Diez[1], Peter J. Schaap[1,3*]

[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, The Netherlands

[2]BU Biointeractions and Plant Health, Wageningen Plant Research, Wageningen University & Research, The Netherlands

[3]UNLOCK Large Scale Infrastructure for Microbial Communities, Wageningen University and Research, Wageningen, The Netherlands

* Corresponding author

Email: peter.schaap@wur.nl

Abstract

The rhizosphere, the region of soil surrounding roots of plants, is colonized by a unique population of Plant Growth Promoting Rhizobacteria (PGPR). By enhancing nutrient uptake from the soil and through modulation of plant phytohormone status and metabolism, PGPR can increase the stress tolerance, growth and yield of crop plants. Many important PGPR as well as plant pathogens belong to the genus *Pseudomonas*. There is, however, uncertainty on the divide between phytobeneficial and phytopathogenic strains as previously thought to be signifying genomic features have limited power to separate these strains. Here the Genome properties (GP) common biological pathways annotation system was applied to establish the relationship between the genome wide GP composition and the plant-associated phenotype of 91 *Pseudomonas* strains representing both phenotypes. GP enrichment analysis, Random Forest model fitting and feature selection revealed 28 discriminating features. A validation dataset of 67 new strains confirmed the importance of the selected features for classification. A number of unexpected discriminating features were found, suggesting involvement of novel molecular mechanisms. The results suggest that GP annotations provide a promising computational tool to better classify the plant-associated phenotype.

Author summary

With a growing population the need to double the agricultural food production is specified. Simultaneously, there is an urgent need to implement sustainable and climate change resilient agricultural practices that preserve natural ecosystems. Cooperative microbiomes play important positive roles in plant growth development and fitness. Properly tuned, these microbiomes can significantly reduce the need for synthetic fertilizers and can replace chemicals in crop pest control. To select beneficial candidates, their traits need to be described and likewise, potential detrimental traits should be avoided. Here we applied GP-based comparative functional genomics, enrichment analysis and Random Forest model fitting to compare known phytobeneficial and phytopathogenic

Pseudomonas strains. A number of unexpected discriminating features were found suggesting the involvement of novel molecular mechanisms.

Introduction

Among the targets set by the UN to achieve the zero-hunger goal, the need to double the agricultural food production is specified [1]. Earlier attempts to improve plant performance and production focused on plant breeding, pest control by chemical means and the implementation of synthetic fertilizers tapping into finite global reserves. While these strategies were successful in enhancing production, the increasing adverse effects on the environment challenges us to find sustainable alternatives [2–4].

A multitude of studies has demonstrated that cooperative microbiomes can play important positive roles in plant growth, development and fitness. One particular hotspot is the rhizosphere, the region of soil surrounding plant roots, colonized by Plant Growth Promoting Rhizobacteria (PGPR)[5]. A stable PGPR population can increase the stress tolerance, growth and yield of crop plants by enhancing nutrient uptake from the soil and through modulation of plant phytohormone status and metabolism [6–13]. As a result, a large catalogue of plant beneficial bacterial strains has been identified. The most studied are *Pseudomonas* spp., a functionally diverse group representing both plant beneficial and pathogenic strains [14–16].

A diverse spectrum of plant-host interaction pathways determines the plant-associated phenotype of a *Pseudomonas* strain. Correlational approaches have identified a number of marker genes contributing to the phenotype [17–19]. These genes are however, to a certain degree, shared between beneficial and pathogenic strains [20] and consequently, with each new genome addition the uncertainty on the divide between beneficial and the pathogenic strains increases. Until now, a generic description of presence and completeness of biological pathways contributing to the plant-associated phenotype of a *Pseudomonas* strain is lacking. Such knowledge would bring fundamental

insights into their potential to enhance plant performance and resilience. When genes are placed in context of biological pathways comparative functional genomics is possible. Genome Properties (GP) is an annotation system whereby functional attributes can be assigned to a genome [21]. The resource represents a collection of 1286 common biological pathways, and each GP is evidenced by a distinct set of protein domains.

Here we applied GP-based comparative functional genomics to compare known phytobeneficial and phytopathogenic *Pseudomonas* strains using both traditional statistical analysis and machine learning methods. This allowed us to accurately classify *Pseudomonas* strains, and to identify discriminating features for both the phytobeneficial and phytopathogenic lifestyle. In the discussion section these discriminating features are placed into biological context.

Results

Based on literature review, the complete genomes of 84 *Pseudomonas* strains were retrieved from the Pseudomonas Genome DB (version 17.2) [22] and categorized as encoding either a 'phytobeneficial' strain (51 strains) or a 'phytopathogenic' strain (33 strains). This selection was supplemented with the complete genomes of seven new or re-sequenced phytobeneficial strains; *P. putida* P9, *P. corrugata* IDV1, *P. fluorescens* R1 and WCS374, *P. protegens* Pf-5, *P. chlororaphis* Phz24 and *P. jessenii* RU47. To avoid gene and protein domain annotation inequality, the genome sequences of all 91 strains were *de novo* annotated. Subsequently, the two groups were compared using nucleotide sequence similarity, by protein domain presence and by presence and completeness of domain-based GPs (**Fig 1**). Domain content was subjected to enrichment analysis and the GP content of both groups was used to train a Random Forest (RF) model for classification and feature selection [23]. The performance of the classification methods was further validated using a set of 67 newly sequenced soil derived *Pseudomonas* genomes obtained from a newer version (V20.2) of the Pseudomonas Genome DB. Based on literature data. Using literature data, 17 strains of this validation set could be classified as phytobeneficial strains while 34 strains were involved in bioremediation. For

16 strains the classification was unclear however, a number of these strains were *P. chlororaphis* strains known to be phytobeneficial.

Sequence similarity

We first examined the genomic relatedness between the phytobeneficial and phytopathogenic group, by calculating the Average Nucleotide Identity (ANI) scores between all possible pairs (**Fig 2**). The ANI scores showed that corresponding with their phenotypic classification the genome sequences could be divided into two groups with *Pseudomonas sp.* M30-35 being less similar to the rest of the phytobeneficial group. The average sequence similarity within the phytopathogen and the phytobeneficial group was 90.01 ± 5.53 and 79.57 ± 4.27 respectively. The ANI-score measures genomic similarity between the coding regions of two genomes at nucleotide-level taking into account hits that have 70% or more identity and at least 70% coverage of the shorter gene. The ANI score does not take into account the fraction of coding sequences that actually contribute to this score and thus provides no insight in the degree of strain-specific functional adaptations. To study which strain-specific functional adaptations impact the phenotype, the protein domain content of each strain was considered.

Protein domain content

The 91 complete Pseudomonas genomes contained, on average, 5640 ± 643 protein encoding genes. For each genome, 9342 ± 709 domains were identified with an average domain copy number of 2.35 ± 0.12 (**S1 Table**). Using domain presence as input, a group-wise enrichment analysis was done and a total of 410 and 329 protein domains were found to be significantly enriched in respectively phytobeneficial and phytopathogenic strains (**S2 Table**).

Phytobeneficial strains were enriched with five domains linked to Type II secretion systems (T2SS), ten domains linked to the term “cytochrome”, eight domains linked to, “quinohemoprotein” and six domains linked to “biofilm” (Poly-beta-1,6-N-acetyl-D-glucosamine type) biosynthesis. Interestingly,

119 domains related to “quinohemoprotein” and “biofilm” were not only enriched but also exclusively
120 found in phytobeneficial strains.

121 Phytopathogenic strains were enriched with domains involved in various types of secretion
122 systems. Moreover, some of these domains were not present in any of the phytobeneficial strains.
123 Eighteen of those pathogen enriched domains are reported to be involved in the Type III secretion
124 system and five in the Type IV secretion system. In addition, the phytopathogen list showed
125 enrichment of nine different domain involved in phosphonate metabolism. Functional clustering of
126 enriched domains was further explored using genome properties.

127

128 **Genome properties**

129 Genome properties (GP) represent a collection of currently 1286 common biological pathways. Each
130 GP is constructed from a precomputed cluster of core protein domains which are used as essential
131 evidence for the presence of the biological pathway [21]. Genome derived protein domains were
132 used to construct for each strain a list of GPs with two possible evidence values: ‘YES’ indicating that
133 the complete set of precomputed evidences had been detected and ‘PARTIAL’, indicating that the GP
134 is likely present due to the presence of an incomplete set of evidences above a per GP specified
135 minimal threshold. In addition, we took into account that the bacterial genes encoding domains that
136 function in the same biological pathway are often arranged in operonic structures corresponding to
137 syntenic blocks. For each strain therefore GPs were reconstructed not only based on protein domain
138 presence/absence (GP-PA) but also on protein domain colocalization (GP-SND; synteny-non-
139 directional) and on domain colocalization and being encoded on the same strand (GP-SD; synteny-
140 directional). For domain colocalization a nearest neighbor approach was applied using a sliding
141 window of 20 protein domains**.**

142

143 **Table 1** summarizes the results obtained for the three approaches.

144

Table 1: Number of strain specific GP classes per approach

Approach	Complete	Partial	Not detected
GP-PA	440 ± 22	256 ± 14	590 ± 14
GP-SND	161 ± 11	362 ± 6	763 ± 12
GP-SD	158 ± 10	365 ± 7	763 ± 13

A total of 438 GPs were not present in any the investigated Pseudomonas strains. The majority of these GPs represented functions and processes typically found in eukaryotic species (**S3 Table**). Conversely, using the GP-PA method, a functional GP core of 154 complete GPs present in all strains could be obtained. When domain colocalization was used as an additional constraint a functional core of 37 complete, likely operonic, GPs was found with both domain colocalization methods. Note that overall, the GP-SND and GP-SD generated very similar output underpinning a strong linkage between operonic structures and functional genome properties in bacterial species (

Table 1).

Next, a principal component analysis (PCA) was applied to the GP data. With all three methods a clear separation between the pathogen and the biocontrol group were obtained (**S4 Fig**)**. Fig 3** shows the results obtained with the GP-SND approach.

To further understand the contribution of each GP to the separation, we performed an enrichment analysis on the results obtained with the GP-PA, GP-SD and GP-SND approach (**S3 Table**). The enrichment analysis was performed on the binary data of presence and absence of the properties by considering "PARTIAL" as presence or absence separately, creating two enriched sets per approach. Subsequently, the two enriched sets were intersected to create the enriched set for that particular approach. Lastly, an overall enriched set was constructed by considering only the GPs that were enriched in the GP-SD and GP-SND approaches (

Table 2).

To extend our analysis utilizing the full information of the classes and to capture feature importance, a Random Forest (RF) classifier was built. For 99% of the strains, the RF classifier correctly

predicted the phenotype. The only exception was *Pseudomonas cichorii* JBC1, which had been reported to be pathogenic but was classified by RF-classifier as phytobeneficial. To study the discriminating variables further, variable selection from RF was implemented (

Table 3 and **S3 Table**). These variables were integrated with the list of enriched GPs to generate a comprehensive list of key genomic features contributing to the plant-associated phenotype (**Fig 4**).

Table 2: Genome Properties related to the plant-associated phenotype: enrichment analysis

Genome Property	Description	Adjusted P-value
GPs enriched in phytobeneficial strains		
GenProp0238*	2-aminoethylphosphonate catabolism to acetaldehyde	$< 10^{-6}$
GenProp0721*	2-aminoethylphosphonate (AEP) ABC transporter, type II	$< 10^{-6}$
GenProp0613*	Cytochrome c reductase	$< 10^{-6}$
GenProp0907	Poly-beta-1,6 N-acetyl-D-glucosamine system, PgaABCD type	$< 10^{-6}$
GenProp0271	Trehalose utilization	$< 10^{-6}$
GenProp1745	GA12 biosynthesis	$< 10^{-6}$
GenProp1189	MqsRA toxin-antitoxin complex	$< 10^{-6}$
GenProp1645	Zeaxanthin biosynthesis	$< 10^{-6}$
GenProp0659	Tryptophan degradation to anthranilate	7.96×10^{-5}
GenProp0895	Alcohol ABC transporter, PedABC-type	7.01×10^{-4}
GenProp0902	Quinohemoprotein amine dehydrogenase	1.40×10^{-3}
GenProp1516	Phosphatidylcholine biosynthesis V	5.37×10^{-3}
GPs enriched in phytopathogenic strains		
GenProp0908*	2,3-diaminopropionic acid biosynthesis	$< 10^{-6}$
GenProp0813*	Pyrimidine utilization	$< 10^{-6}$
GenProp1165*	PhnGHIJKL complex	$< 10^{-6}$
GenProp1381	Methylphosphonate degradation I	$< 10^{-6}$
GenProp0236	Phosphonates ABC transport	2.62×10^{-3}
GenProp0710	Generic phosphonates utilization	2.62×10^{-3}
GenProp1193	RelBE toxin-antitoxin complex	3.19×10^{-2}
GenProp1566	D-galactonate degradation	3.64×10^{-2}

*These Genome Properties are also important random forest features (Table 3).

Table 3: Genome Properties related to the plant-associated phenotype: Random Forest features importance

Genome Property	Description	Predictive power**
GenProp0813*	Pyrimidine utilization	500
GenProp0908*	2,3-diaminopropionic acid biosynthesis	500
GenProp0721*	2-aminoethylphosphonate (AEP) ABC transporter, type II	329
GenProp0238*	2-aminoethylphosphonate catabolism to acetaldehyde	328
GenProp0615	Cytochrome c based oxygen reduction and quinone re-oxidation	251
GenProp0613*	Cytochrome c reductase	243
GenProp1629	Propanoyl-CoA degradation I	215
GenProp1572	L-carnitine degradation I	145
GenProp1562	Fatty acid salvage	53
GenProp1717	Fatty acid beta-oxidation I (GenProp1308, GenProp1510 and GenProp1544)	53
GenProp1165*	PhnGHIJKL complex	2
GenProp1251	L-tyrosine biosynthesis I	2
GenProp1281	Hydrogen sulfide biosynthesis I	1
GenProp1681	L-cysteine degradation III	1

*GP also found in the enrichment analysis. **Numbers were obtained using recursive feature elimination (500 iterations)

Prediction validation

A set of 67 newly retrieved *Pseudomonas* genome sequences were analyzed for the presence of GPs using the GP-SND approach and used in RF performance evaluation (**S1 Table**). Confirming the capability of GP content to predict the plant-associated phenotype, a PCA of the full dataset (training and validation) indicated that the separation between the phytobeneficial and the phytopathogenic strains was retained. Additionally, a clustering of bioremediation strains with phytobeneficial strains was observed (**Fig 5Error! Reference source not found.)**. Unclassified strain *Pseudomonas* sp. KBS0707 was positioned within the pathogen group. As all *P. syringae* are considered to be phytopathogenic, the unclassified *P. syringae* isolate inb918 was of interest as it appeared to be a phytobeneficial strain. The ANI score however suggested that strain inb918 might have been taxonomically misclassified as among the *P. syringae* strains the pair-wise score between this strain

and the others remained below 79% (**Fig 5**). Lastly, the RF classifier was applied to the validation set and yielded the same predictions as the PCA.

Discussion

Plants live in symbiotic interactions with microbial communities, which are complex networks composed of interacting microbiotic nodes. The sum of these interactions can be beneficial for plant growth and development, detrimental or neutral. Many important PGPR as well as plant pathogens belong to the genus *Pseudomonas*. The genomic diversity observed at species [22] and strain level suggests that Pseudomonas spp. have a broad potential for evolutionary adaptation to different environments. Consequently, the plant-associated lifestyle of a *Pseudomonas* strain is likely to be the result of a combinatorial accumulation and emergence of a diverse set of contributing traits. A selected isolated genome encoded feature therefore will have limited power to confidently predict the plant-associated phenotype.

Differences between phytopathogenic and phytobeneficial strains emerge at all levels of analysis. At genome sequence similarity level, a separation between the two groups was prominent. As most of the described phytopathogenic genomes in the scientific literature are *P. syringae* strains, a higher degree of sequence similarity was observed for the phytopathogenic group. The ANI score, however, does not take into account the most variable genomic regions that are likely to harbor genes that function in the biological relevant differences and would provide further insight in the functional diversity within the two phytotypes. The Genome properties (GP) annotation system was applied in this study to specifically address functional differences encoded in the genomes.

GPs represent not only metabolic pathways but also various other classes of functional attributes and provide, compared to KEGG and SEED, a better functional annotation coverage [21]. The GP annotation system is organized as a doubly linked rooted DAG. Leave nodes use domains as evidences, parent nodes, representing super-pathways, use leaf node GPs as evidences. For a functional genome comparisons at a larger scale, protein domains are better scalable and less

sensitive to sequence variation compared to techniques based on sequence similarity [24]. By focusing on the reconstruction of domain-based GPs only, feature independence is promoted, and the complexity of the RF-model is reduced. In total 848 domain-based GPs were annotated to be (likely) present in one or more of the here studied *Pseudomonas* strain. Underpinning the genomic diversity of the 91 *Pseudomonas* strains used in this study, in contrast a functional core of maximal 154 complete and persistently present GPs was obtained. While for obvious reasons by far most of the typical eukaryotic GPs were not detected, a limited number of the *Pseudomonas* GPs may have some domain overlap with GPs of similar function typically found eukaryotic species. An example is the domain overlap between GenProp1717 and the "peroxisomal" GPs GenProp1308, GenProp1510 and GenProp1544 all involved in fatty acid beta-oxidation which we treated as one.

Three different approaches were used to determine the domain-based GP content of each strain. Implementation of the domain colocalization constraint mirrors the operonic structure common in bacterial genomes [25]. For domain colocalization a sliding window of 20 domains was chosen as it covers 1255 of the 1286 GPs (98%) with the most abundance group of GPs being GPs with two evidences (396 GPs) (**S5 Fig**). As the average domain copy number is 2.3, indicating that the same domain could be assigned to multiple functions across the genome, inclusion of protein domain colocalization in GP reconstruction also increases the prediction certainty of those GPs and further promotes the selection accessory traits, some of which may be acquired by lateral transfer, as RF-variables. Very similar results were obtained with GP-SND and the strain specific GP-SD method, suggesting that domain clustering most likely yields operonic structures.

The validation data was used to explore the performance of the RF classifier. For most validation data the RF firmly supports the discrimination between the beneficial and the pathogenic strains. *P. cichorii* JBC1 was classified as non-pathogenic. However, that does not directly translate into it being beneficial. **Fig 4** shows that *P. cichorii* JBC1 still contains three GPs associated with pathogenicity: '2,3-diaminopropionic acid biosynthesis' (GenProp0908), 'RelBE toxin-antitoxin complex' (GenProp1193) and 'D-galactonate degradation' (GenProp1566). *P. cichorii* JBC1 has already

248 been reported to be quite different to other pathogenic *Pseudomonas* at the genome level [26] and
249 our results confirm this finding suggesting that there may be different mechanisms for pathogenicity
250 associated with this strain.

251 RF recursive feature elimination and GP enrichment analysis was used to select a minimal set
252 of GP-variables needed for a good prediction of the phenotype [27]. GenProp0238 and GenProp0721
253 are two of those important RF-variables and are shown to be enriched in phytobeneficial strains. The
254 two GPs are related to mechanisms of phosphonate utilization, which have been shown to occur in
255 *Pseudomonas* and also in other microorganisms [28]. Phosphonate is a form of phosphorus, which is
256 essential for many biological processes [29]. However, both groups show differences in the usable
257 form of phosphonate. Most phytobeneficial strains appear to be able to utilize only 2-
258 aminoethylphosphonate (AEP) via the genome properties: '2-aminoethylphosphonate catabolism to
259 acetaldehyde' (GenProp0238) and '2-aminoethylphosphonate (AEP) ABC transporter, type II'
260 (GenProp0721), whereas the phytopathogens are able to access broader forms of phosphonates, as
261 also shown by the enriched protein domain, via 'phosphonates ABC transport' (GenProp0236),
262 'generic phosphonates utilization' (GenProp0710), 'PhnGHIJKL complex' (GenProp1165) and
263 'methylphosphonate degradation I' (GenProp1381) [30]. AEP is the most abundant C-P compound in
264 nature while other phosphonates and their derivatives are substances used in agriculture (herbicides,
265 fungicides and insecticides) and pharmacy (antibiotics) [31]. It has been reported that the virulence of
266 pathogenic species was enhanced under conditions of orthophosphate limitation [32]. Thus, we
267 hypothesize this could be due to the presence of genome traits that enable them to access a wider
268 set of phosphate sources.

269 GenProp0908 is another important RF-variable. This GP was found to be enriched in
270 phytopathogenic strains and is involved in 2,3-diaminopropionic acid biosynthesis (DAP). DAP is a
271 precursor of several secondary metabolites, such as siderophores, neurotoxins and antibiotics [33].
272 Pyoverdins, the principal siderophores, have been reported to be produced exclusively by the
273 pathogens, such as *P. syringae* and *P. cichorii* [34]. Siderophores are important metabolites involved

in iron acquisition [35]. Iron is crucial to many metabolic processes and is therefore required to maintain cells in a healthy state [36]. The stronger ability to scavenge for iron, and the phosphonate previously mentioned, will increase the fitness of the pathogens.

Two GPs strongly enriched among the phytobeneficial strains are GenProp0907, and GenProp0902. GenProp0907 represents a cluster of four genes involved in the synthesis, modification and export of the biofilm adhesin poly-beta-1,6-N-acetyl-D-glucosamine and the four domain evidences represent the four genes required. The GP is not present in the phytopathogen group and found to be complete as likely operonic structures in 39 phytobeneficial strains. Biofilms of the PgaABCD type have been studied in *Escherichia coli* [37] but not in *Pseudomonas* species. GenProp0902 represents quinohemoprotein amine dehydrogenase (QHNDH). QHNDH is a three-subunit enzyme located in the periplasmic space of *P. putida* and part of the amine oxidation respiratory chain. QHNDH catalyzes the oxidative deamination of primary amines when used as a sole carbon and energy source [38]. The GP consists of four evidences, three domains representing the alpha-, beta- and gamma-subunit of the enzyme and one representing the QHNDH maturation protein. This likely operonic GP was found to be complete in 24 biocontrol strains and not present in the pathogen group. As these GPs are only present in subset of the phytobeneficial strains, they did not emerge as important RF-variables in recursive feature elimination.

Protein domains associated with Type II secretion system (T2SS) were found to be enriched among the phytobeneficial strains while domains involved in the type III secretion system (T3SS) were found to be enriched among the phytopathogenic strains. T2SS is described by GenProp0053 and consists of 10 non-optional evidences and 3 optional domains. GP results however, indicated for both phytobeneficial and phytopathogenic strains a “PARTIAL” status for this GP. Similarly, the type III secretion system, represented by GenProp0052 is considered to be a key virulence factor and has been considered as evidence for pathogenicity in many genome studies [17,39,40]. GenProp0052 is a complex GP consisting of 14 evidences and 28 optional domains. Due to the set zero threshold for “PARTIAL” for this specific GP, a single evidence domain will already result in a “PARTIAL” status.

Eighteen protein domains enriched in phytopathogens are described to be involved in Type III secretion systems. Eleven of those enriched domains are used as evidences for GenProp0052. One other, TIGR02551, did also occur in the pathogen set but was considered not to be enriched after the Bonferroni adjustment. In contrast, the two missing evidences, TIGR02105 and TIGR02546 are only present in five phytobeneficial genomes. Thus, amongst the 91 Pseudomonas strains all 14 evidences are present, but none of the strains used in this study have the complete set of 14 evidences.

Due to the 'Partial' status of GenProp0052 (T2SS) and GenProp0053 (T3SS) for both phytotypes these GPs were not enriched, nor were they selected as discriminating variables in RF classification. We further examined the distribution of the GenProp0053 and of GenProp0052 evidences over all strains (**S6 Fig**)**.** The distribution showed that protein domains linked to GenProp0052 more consistently occurred in the pathogen group with more variation in the phytobeneficial group. The result suggests that the abundance of T3SS related domain content could be sufficient for an indication of the pathogenicity. However, there is no guarantee that the feature is functional due to the missing evidences.

Specifically, for the phytobeneficial group a number of enriched GPs suggested a role for pathways involved in the degradation and utilization of trehalose (GenProp0271), tryptophan (GenProp0659) (Table 2), tyrosine (GenProp1251) and carnitine (GenProp1572) (Table 3). On the other hand, phytopathogenic strains appears to be more specialized in the degradation of galactonate (GenProp1566) and cysteine (GenProp1681). Carbon sources that were predicted to be degradable by preferably the phytobeneficial group could contribute to the agricultural industry. These substrates could be used as fertilizers, growth promotors, or as additives to alternate the microbial composition [41]. Similar to elicitors, which directly enhance plant defense and resistance, this indirect approach could be applied to the existing microbial community to select for the beneficial strains and potentially increase the productivity of the crop. [42]. On the other hand, carbon sources that might prolong saprobic growth and survival of pathogens should be avoided.

Other GPs found in the phytobeneficial group are linked to four 'human hormones', which are 'mineralocorticoid biosynthesis' (GenProp1644), 'estradiol biosynthesis II' (GenProp1417), 'glucocorticoid biosynthesis' (GenProp1666) and 'pregnenolone biosynthesis' (GenProp1740). The evidence shared by these hormones, domain PF00067 (cytochrome P450), is the same as for 'GA12 biosynthesis' (GenProp1745). Hence, only GA will be further discussed. Gibberellin 12 (GA_{12}), is the common precursor of all gibberellins (GA) [43]. GA phytohormones play important roles in influencing the growth and development of the host plants [44] and GA from *Pseudomonas* could increase seed germination [45].

Not all known traits are represented by a GP. Many of those are found in phytopathogenic strains such as, coronatine, cytokinin and auxin [46]. We examined the presence of the protein domains associated to these traits in our dataset (**S7 Fig**). The results showed that the associated protein domains are generally present in both groups. Among these domains, only PF08659 and PF16197 were enriched in the phytopathogenic group. This suggests that the occurrence of these, known to be, phytopathogenic traits may not be sufficient as a genetic marker to identify the pathogenicity of a strain.

In conclusion, domain-based Genome Properties appear to be robust computational features to differentiate between phytobeneficial and phytopathogenic *Pseudomonas* strains and our analysis shows that incorporation of domain colocation further increases their relevance. By combining traditional statistical analysis (enrichment analysis) and machine learning methods (random forest) we were able to identify new discriminating genome properties that can be used to identify species that promote plant growth. These could be applied in strategies to develop synthetic PGPR communities and to formulate soil additives to improve plant health and performance.

Materials and Methods

Genome retrieval and annotation: *Pseudomonas* genomes with were downloaded from Pseudomonas Genome DB version 17.2. The validation set was obtained from database version 20.2 (https://www.pseudomonas.com) [22]. Genomes were manually categorized according their phytotype using literature data. Additionally, 7 genome sequences were (re)sequenced from phytobeneficial strains *P. putida* P9 (accession ERS6670306), *P. Corrugata* IDV1 (accession ERS6652532), *P. fluorescens* R1 (accession ERS6670181), *P. protegens* Pf-5 (accession ERS6652530), *P. chlororaphis* Phz24 (accession ERS6670416), *P. jessenii* RU47 (accession ERS6670307) and *P. fluorescens* WCS374 (accession ERS6652531). DNA was extracted using the Epicenter Masterpure kit (Epicentre Technologies, USA) according to the manufacturer's protocol, quantified. For with the Infinite® 200 PRO (Tecan, Männedorf, Switzerland) using the Quant-iT™ PicoGreen™ dsDNA Assay Kit (ThermoFisher, Waltham, USA) according to the manufacturer's protocol. The strains were sequenced on the PacBio Platform (Pacific BioSciences, Menlo Park, USA). A total of 4 ug DNA was sheared to 7 Kb and two SMRT bell libraries were prepared using the kit Barcoded Adapters for Multiplex SMRT sequencing in combination with the Sequel Binding Kit V2.0 and the Sequel Polymerase 2.0 Kit. Per library, a pool with sheared DNA of all strains was used as input according to the manufacturer's protocol. Sequencing was done on a Sequel system operated at the services of Business Unit Bioscience, Wageningen Plant Research (Wageningen, The Netherlands). Subsequently, de-multiplexing was performed by aligning the barcodes to the sub-reads with pyPaSWAS version 3.0 [47]. Canu version 1.6 [48] was used to assemble the PacBio reads

The SAPP semantic annotation framework [49] was used to systematically (re)annotated the genomes. Briefly, protein encoding genes were de novo predicted using Prodigal 2.6.3 [50] and protein domains were characterized with InterProScan 5.36-75.0 using the Pfam and TIGRFAMs databases [51–53]. Annotation data and meta-data was stored in a semantic database using the GBOL ontology [54,55]. SPARQL queries were used to extract protein domain identifiers, and the location and direction of the corresponding gene.

Data processing: OrthoANI version 1.40 was used to calculate the Average Nucleotide Identity (ANI) score for all genomes [56]. PygenProp, was used to infer from each genome domain-based GPs, [55]. Three criteria were applied; "PA", considering only domain presence as evidence, "SND", synteny-non-directional, requiring the genome location of the corresponding domains to be in close proximity and "SD" that in addition to gene location also considers strandness. For SND and SD a nearest neighbor approach and a sliding window of 20 protein domains was applied. Each GP was classified as either 'YES', or 'PARTIAL' according to the completeness of the set of evidences.

Statistical analysis: The natural grouping of the data was visualized using principal component analysis (prcomp package). Then, with R packages; fisher.test and p.adjust, Fisher Exact Test with Bonferroni correction was applied to protein domains and the genome properties to test for enrichment. This analysis identified the over- and under-represented features. GP data was reassessed twice by considering 'PARTIAL' as either 'YES' or 'NO'. The enriched list was created by intersecting the two cases of 'PARTIAL'. Enrichments were considered significant if the adjusted p-value after Bonferroni correction of the GP is below 0.05.

The Random Forest classifier was created using R package randomForest v4.6-14 [58]. Labelled data were divided into training and test sets. The unbiased training set was created with equal numbers per group determined by using 75% of the smaller group, the pathogen group, resulting in 25 strains per group. Therefore, the test set remains with 33 phytobeneficials and 8 phytopathogens. The Variable Selection from Random Forests v 0.7-8 (varSelRF) package in R was used to determine variable importance. We used 5000 trees for the first forest and 2000 trees for all additional forests during the iteration. Vars.drop.frac, the portion of the variable that is excluded on each iteration, was set to 0.2.

Acknowledgements

WP is financially supported by a Royal Thai Government Scholarship, Thailand. TL acknowledges the support by the Dutch Ministry of Economic Affairs in the Topsector Program "Horticulture and Starting Materials" under the theme "Plant Health" (project number: TU 16022) and its partners (NAK, Naktuinbouw and BKD). PS and MSD acknowledge the Dutch national funding agency NWO, and Wageningen University and Research for their financial contribution to the Unlock initiative (NWO: 184.035.007).

References

1. Nations U. United Nations | Peace, dignity and equality on a healthy planet. In: United Nations [Internet]. United Nations; [cited 10 Feb 2021]. Available: https://www.un.org/en/

2. Arif I, Batool M, Schenk PM. Plant Microbiome Engineering: Expected Benefits for Improved Crop Growth and Resilience. Trends in Biotechnology. 2020;38: 1385–1396. doi:10.1016/j.tibtech.2020.04.015

3. Timmusk S, Behers L, Muthoni J, Muraya A, Aronsson A-C. Perspectives and Challenges of Microbial Application for Crop Improvement. Front Plant Sci. 2017;8: 49–49. doi:10.3389/fpls.2017.00049

4. Vejan P, Abdullah R, Khadiran T, Ismail S, Nasrulhaq Boyce A. Role of Plant Growth Promoting Rhizobacteria in Agricultural Sustainability-A Review. Molecules. 2016;21: 573. doi:10.3390/molecules21050573

5. Bakker PAHM, Berendsen RL, Doornbos RF, Wintermans PCA, Pieterse CMJ. The rhizosphere revisited: root microbiomics. Frontiers in plant science. 2013;4: 165. doi:10.3389/fpls.2013.00165

6. Backer R, Rokem JS, Ilangumaran G, Lamont J, Praslickova D, Ricci E, et al. Plant Growth-Promoting Rhizobacteria: Context, Mechanisms of Action, and Roadmap to Commercialization of Biostimulants for Sustainable Agriculture. Frontiers in Plant Science. 2018;9: 1473. doi:10.3389/fpls.2018.01473

7. Finkel OM, Castrillo G, Herrera Paredes S, Salas González I, Dangl JL. Understanding and exploiting plant beneficial microbes. Curr Opin Plant Biol. 2017/06/13 ed. 2017;38: 155–163. doi:10.1016/j.pbi.2017.04.018

8. Gupta G, Parihar SS, Ahirwar NK, Snehi SK, Singh V. Plant growth promoting rhizobacteria (PGPR): current and future prospects for development of sustainable agriculture. J Microb Biochem Technol. 2015;7: 096–102.

9. Ilangumaran G, Smith DL. Plant Growth Promoting Rhizobacteria in Amelioration of Salinity Stress: A Systems Biology Perspective. Frontiers in Plant Science. 2017;8: 1768. doi:10.3389/fpls.2017.01768

10. Köhl L, Oehl F, van der Heijden MGA. Agricultural practices indirectly influence plant productivity and ecosystem services through effects on soil biota. Ecological Applications. 2014;24: 1842–1853. doi:10.1890/13-1821.1

11. Kumar A, Patel JS, Meena VS, Srivastava R. Recent advances of PGPR based approaches for stress tolerance in plants for sustainable agriculture. Biocatalysis and Agricultural Biotechnology. 2019;20: 101271. doi:10.1016/j.bcab.2019.101271

12. Lugtenberg BJJ, Malfanova N, Kamilova F, Berg G. Microbial Control of Plant Root Diseases. Molecular Microbial Ecology of the Rhizosphere. John Wiley & Sons, Ltd; 2013. pp. 575–586. doi:10.1002/9781118297674.ch54

13. Vacheron J, Desbrosses G, Bouffaud M-L, Touraine B, Moënne-Loccoz Y, Muller D, et al. Plant growth-promoting rhizobacteria and root system functioning. Frontiers in Plant Science. 2013;4: 356. doi:10.3389/fpls.2013.00356

14. Qessaoui R, Bouharroud R, Furze JN, El Aalaoui M, Akroud H, Amarraque A, et al. Applications of New Rhizobacteria Pseudomonas Isolates in Agroecology via Fundamental Processes Complementing Plant Growth. Scientific Reports. 2019;9: 12832. doi:10.1038/s41598-019-49216-8

15. Shaikh S, Yadav N, Markande AR. Interactive potential of Pseudomonas species with plants. Journal of Applied Biology & Biotechnology Vol. 2020;8: 101–111.

16. Sitaraman R. Pseudomonas spp. as models for plant-microbe interactions. Front Plant Sci. 2015;6: 787–787. doi:10.3389/fpls.2015.00787

17. Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, Cherkis K, et al. Dynamic Evolution of Pathogenicity Revealed by Sequencing and Comparative Genomics of 19 Pseudomonas syringae Isolates. PLOS Pathogens. 2011;7: e1002132. doi:10.1371/journal.ppat.1002132

18. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 2019;47: D687–D692. doi:10.1093/nar/gky1080

19. Loper JE, Hassan KA, Mavrodi DV, Davis EW II, Lim CK, Shaffer BT, et al. Comparative Genomics of Plant-Associated Pseudomonas spp.: Insights into Diversity and Inheritance of Traits Involved in Multitrophic Interactions. PLOS Genetics. 2012;8: e1002784. doi:10.1371/journal.pgen.1002784

20. Passera A, Compant S, Casati P, Maturo MG, Battelli G, Quaglino F, et al. Not Just a Pathogen? Description of a Plant-Beneficial Pseudomonas syringae Strain. Front Microbiol. 2019;10: 1409–1409. doi:10.3389/fmicb.2019.01409

21. Richardson LJ, Rawlings ND, Salazar GA, Almeida A, Haft DR, Ducq G, et al. Genome properties in 2019: a new companion database to InterPro for the inference of complete functional attributes. Nucleic acids research. 2018;47: D564–D572.

22. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FS. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. Nucleic acids research. 2016;44: D646–D653.

23. Breiman L. Random Forests. Machine Learning. 2001;45: 5–32. doi:10.1023/A:1010933404324

24. Koehorst JJ, van Dam JCJ, van Heck RGA, Saccenti E, dos Santos VAPM, Suarez-Diez M, et al. Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data. Scientific Reports. 2016;6: 38699. doi:10.1038/srep38699

25. Bergman NH, Passalacqua KD, Hanna PC, Qin ZS. Operon Prediction for Sequenced Bacterial Genomes without Experimental Information. Appl Environ Microbiol. 2007;73: 846. doi:10.1128/AEM.01686-06

26. Ramkumar G, Lee SW, Weon H-Y, Kim B-Y, Lee YH. First report on the whole genome sequence of Pseudomonas cichorii strain JBC1 and comparison with other Pseudomonas species. Plant Pathology. 2015;64: 63–70. doi:10.1111/ppa.12259

27. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7: 3. doi:10.1186/1471-2105-7-3

28. Villarreal-Chiu JF, Quinn JP, McGrath JW. The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. Front Microbiol. 2012;3: 19–19. doi:10.3389/fmicb.2012.00019

29. Yu X, Doroghazi JR, Janga SC, Zhang JK, Circello B, Griffin BM, et al. Diversity and abundance of phosphonate biosynthetic genes in nature. Proc Natl Acad Sci U S A. 2013/12/02 ed. 2013;110: 20759–20764. doi:10.1073/pnas.1315107110

30. White AK, Metcalf WW. Microbial Metabolism of Reduced Phosphorus Compounds. Annu Rev Microbiol. 2007;61: 379–400. doi:10.1146/annurev.micro.61.080706.093357

31. Shiraishi T, Kuzuyama T. Biosynthetic pathways and enzymes involved in the production of phosphonic acid natural products. Bioscience, Biotechnology, and Biochemistry. 2021;85: 42–52. doi:10.1093/bbb/zbaa052

32. Lamarche MG, Wanner BL, Crépin S, Harel J. The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis. FEMS Microbiology Reviews. 2008;32: 461–473. doi:10.1111/j.1574-6976.2008.00101.x

33. Ernst DC, Anderson ME, Downs DM. L-2,3-diaminopropionate generates diverse metabolic stresses in Salmonella enterica. Mol Microbiol. 2016/05/06 ed. 2016;101: 210–223. doi:10.1111/mmi.13384

34. Bultreys A, Gheysen I. Siderophore Uses in Pseudomonas syringae Identification. In: Fatmi M, Collmer A, Iacobellis NS, Mansfield JW, Murillo J, Schaad NW, et al., editors. Pseudomonas syringae Pathovars and Related Pathogens – Identification, Epidemiology and Genomics. Dordrecht: Springer Netherlands; 2008. pp. 21–35. doi:10.1007/978-1-4020-6901-7_2

512 35. Kobylarz MJ, Grigg JC, Shin-ichi JT, Rai DK, Heinrichs DE, Murphy ME. Synthesis of
513 L-2, 3-diaminopropionic acid, a siderophore and antibiotic precursor. Chemistry &
514 biology. 2014;21: 379–388.

515 36. Aznar A, Dellagi A. New insights into the role of siderophores as triggers of plant
516 immunity: what can we learn from animals? Journal of experimental botany. 2015;66:
517 3001–3010.

518 37. Wang X, Preston JF 3rd, Romeo T. The pgaABCD locus of Escherichia coli promotes the
519 synthesis of a polysaccharide adhesin required for biofilm formation. J Bacteriol.
520 2004;186: 2724–2734. doi:10.1128/jb.186.9.2724-2734.2004

521 38. ADACHI O, KUBOTA T, HACISALIHOGLU A, TOYAMA H, SHINAGAWA E,
522 DUINE JA, et al. Characterization of Quinohemoprotein Amine Dehydrogenase from
523 Pseudomonas putida. Bioscience, Biotechnology, and Biochemistry. 1998;62: 469–478.
524 doi:10.1271/bbb.62.469

525 39. Büttner D. Protein Export According to Schedule: Architecture, Assembly, and
526 Regulation of Type III Secretion Systems from Plant- and Animal-Pathogenic Bacteria.
527 Microbiol Mol Biol Rev. 2012;76: 262. doi:10.1128/MMBR.05017-11

528 40. Lombardi C, Tolchard J, Bouillot S, Signor L, Gebus C, Liebl D, et al. Structural and
529 Functional Characterization of the Type Three Secretion System (T3SS) Needle of
530 Pseudomonas aeruginosa. Frontiers in Microbiology. 2019;10: 573.
531 doi:10.3389/fmicb.2019.00573

532 41. Wawrik B, Kerkhof L, Kukor J, Zylstra G. Effect of different carbon sources on
533 community composition of bacterial enrichments from soil. Appl Environ Microbiol.
534 2005;71: 6776–6783. doi:10.1128/AEM.71.11.6776-6783.2005

535 42. Thakur M, Sohal BS. Role of Elicitors in Inducing Resistance in Plants against Pathogen
536 Infection: A Review. ISRN Biochem. 2013;2013: 762412–762412.
537 doi:10.1155/2013/762412

538 43. Regnault T, Davière J-M, Wild M, Sakvarelidze-Achard L, Heintz D, Carrera Bergua E,
539 et al. The gibberellin precursor GA12 acts as a long-distance growth signal in
540 Arabidopsis. Nature Plants. 2015;1: 15073. doi:10.1038/nplants.2015.73

541 44. Morrone D, Chambers J, Lowry L, Kim G, Anterola A, Bender K, et al. Gibberellin
542 biosynthesis in bacteria: Separate ent-copalyl diphosphate and ent-kaurene synthases in
543 Bradyrhizobium japonicum. FEBS Letters. 2009;583: 475–480.
544 doi:10.1016/j.febslet.2008.12.052

545 45. Bharathi R, Vivekananthan R, Harish S, Ramanathan A, Samiyappan R. Rhizobacteria-
546 based bio-formulations for the management of fruit rot infection in chillies. Crop
547 Protection. 2004;23: 835–843. doi:10.1016/j.cropro.2004.01.007

548 46. Ruinelli M, Blom J, Smits THM, Pothier JF. Comparative genomics and pathogenicity
549 potential of members of the Pseudomonas syringae species complex on Prunus spp.
550 BMC Genomics. 2019;20: 172. doi:10.1186/s12864-019-5555-y

551 47. Warris S, Timal NRN, Kempenaar M, Poortinga AM, van de Geest H, Varbanescu AL, et
552 al. pyPaSWAS: Python-based multi-core CPU and GPU sequence alignment. PLOS
553 ONE. 2018;13: e0190279. doi:10.1371/journal.pone.0190279

554 48. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable
555 and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
556 Genome Res. 2017/03/15 ed. 2017;27: 722–736. doi:10.1101/gr.215087.116

557 49. Koehorst JJ, van Dam JCJ, Saccenti E, Martins dos Santos VAP, Suarez-Diez M, Schaap
558 PJ. SAPP: functional genome annotation and analysis through a semantic framework
559 using FAIR principles. Bioinformatics. 2017;34: 1401–1403.

560 50. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal:
561 prokaryotic gene recognition and translation initiation site identification. BMC
562 bioinformatics. 2010;11: 119.

563 51. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam
564 protein families database in 2019. Nucleic acids research. 2019;47: D427–D432.

565 52. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, et al. TIGRFAMs: a
566 protein family resource for the functional identification of proteins. Nucleic Acids
567 Research. 2001;29: 41–43. doi:10.1093/nar/29.1.41

568 53. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5:
569 genome-scale protein function classification. Bioinformatics. 2014;30: 1236–1240.

570 54. van Dam JCJ, Koehorst JJJ, Vik JO, Schaap PJ, Suarez-Diez M. Interoperable genome
571 annotation with GBOL, an extendable infrastructure for functional data mining.
572 bioRxiv. 2017; 184747.

573 55. van Dam JCJ, Koehorst JJ, Vik JO, Martins dos Santos VAP, Schaap PJ, Suarez-Diez M.
574 The Empusa code generator and its application to GBOL, an extendable ontology for
575 genome annotation. Scientific Data. 2019;6: 254. doi:10.1038/s41597-019-0263-7

576 56. Lee I, Kim YO, Park S-C, Chun J. OrthoANI: an improved algorithm and software for
577 calculating average nucleotide identity. International journal of systematic and
578 evolutionary microbiology. 2016;66: 1100–1103.

579 57. Bergstrand LH, Neufeld JD, Doxey AC. Pygenprop: a Python library for programmatic
580 exploration and comparison of organism Genome Properties. Bioinformatics. 2019.

581 58. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2: 18–
582 22.

583
584
585
586

Supporting information

S1 Table. List of strains.

List of strains used in this study. The dataset used for the initial analysis and the validation data are in different tabs. The list provides strain's name, their classification along with their corresponding annotation information.

(XLSX)

S2 Table. Enriched protein domains.

Enriched protein domains on phytopathogenic and phytobeneficial strains with the p-value and number of occurrences.

(XLSX)

S3 Table. Genome Properties analysis.

Genome Properties analysis results divided into 9 sheets. First three sheets are according to the analysis approaches: GP-PA (presence-absence), GP-SD (synteny-directional) and GP-SND (synteny-nondirectional). Sheets 4 and 5 are the enriched GP of the phytopathogen and beneficial respectively. Sheets 6 to 8 are the variable selection using the Random Forest using 3 analysis approaches. The final sheet are the GPs that are not presented according to any approaches.

(XLSX)

S4 Fig. PCA using 3 approaches.

(PDF)

S5 Fig. Distribution of number of evidences of the Genome Properties.

(PDF)

S6 Fig. Distribution of non-optional evidences of GenProp0053 and GenProp0052.

(PDF)

S7 Fig. Presence and absence of protein domains associated to genes related to selected pathogenic traits found in *P. syringae*.

(PDF)

Figure Captions

Fig 1: Workflow for GPs based functional genomics and classification. Genome sequences are analyzed using sequence similarity and protein domain content. (Colocalized) protein domain content is used to infer Genome Properties. Enrichment analysis and Random Forest feature selection was used obtain genomic features. Classification performance was evaluated using a validation dataset of 67 newly available genomes.

Fig 2: Pairwise Average Nucleotide Identity (ANI) scores between coding regions. Scores were calculated from alignments that have 70% or more identity and at least 70% coverage of the shorter gene**.**

Fig 3: PCA based on GP-SND content as variables. The fraction of the variance is given in parentheses. *P. cichorii* JBC1 and two strains of *P. cerasi* are outside 95% confidence ellipse of the phytopathogenic group.

Fig 4: Representative list of discriminating Genome Properties obtained with the GP-SND approach. Left panel: enrichment analysis, right panel: Random Forest feature selection. Red lines indicate the phytobeneficial strains (vertical) and enriched traits (horizontal). Blue lines indicate the phytopathogenic strains (vertical) and enriched traits (horizontal). Newly sequenced strains are in red. Enriched GPs that were also highlighted in the RF feature importance analysis are indicated in red.

Fig 5: Analysis of the validation set. (a) PCA of the complete set of SND-GP data: variance is indicated in brackets. Previously analyzed *Pseudomonas* strains and previous obtained 95% confidence ellipses are in gray. The validation set is composed of 3 classes: phytobeneficial strains (red squares),

638 bioremediation strains (green squares) and unclassified strains (purple squares). The arrow points at

639 *P. syringae* isolate inb918. (b) Average Nucleotide Identity (ANI) score among *P. syringae* strains**.**

640 *Pseudomonas syringae* isolate inb918 is at the top left.

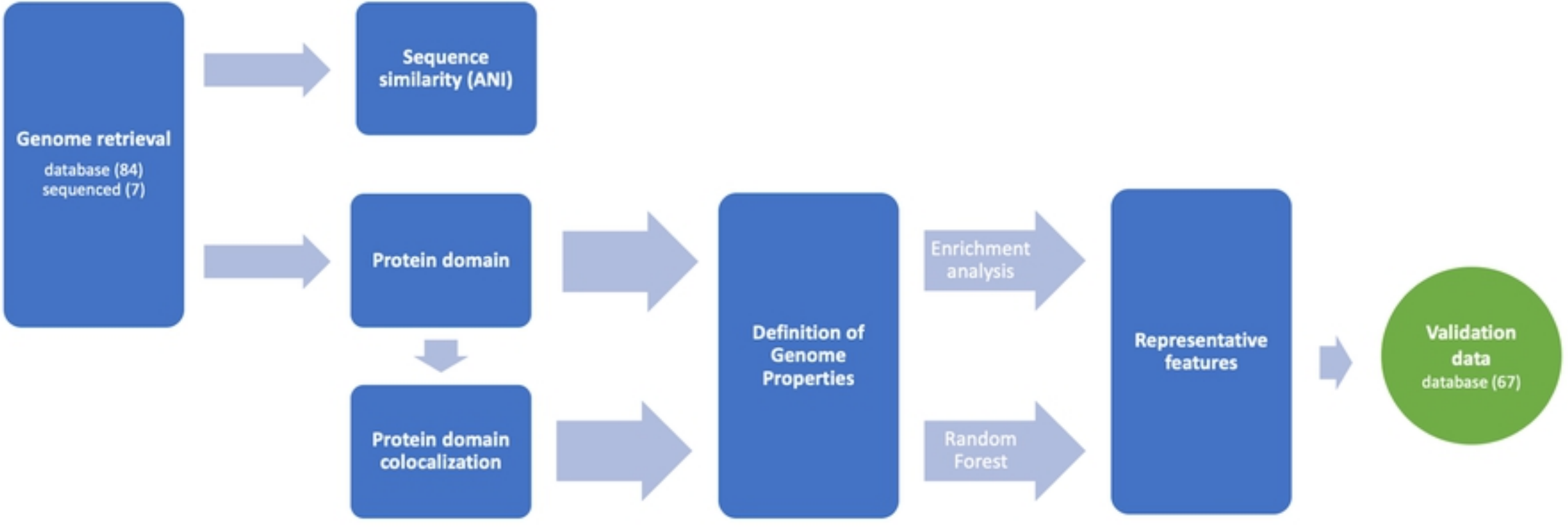
Genome retrieval
database (84)
sequenced (7)
Sequence similarity (ANI)
Protein domain
Protein domain colocalization
Definition of Genome Properties
Enrichment analysis
Random Forest
Representative features
Validation data
database (67)

Figure1

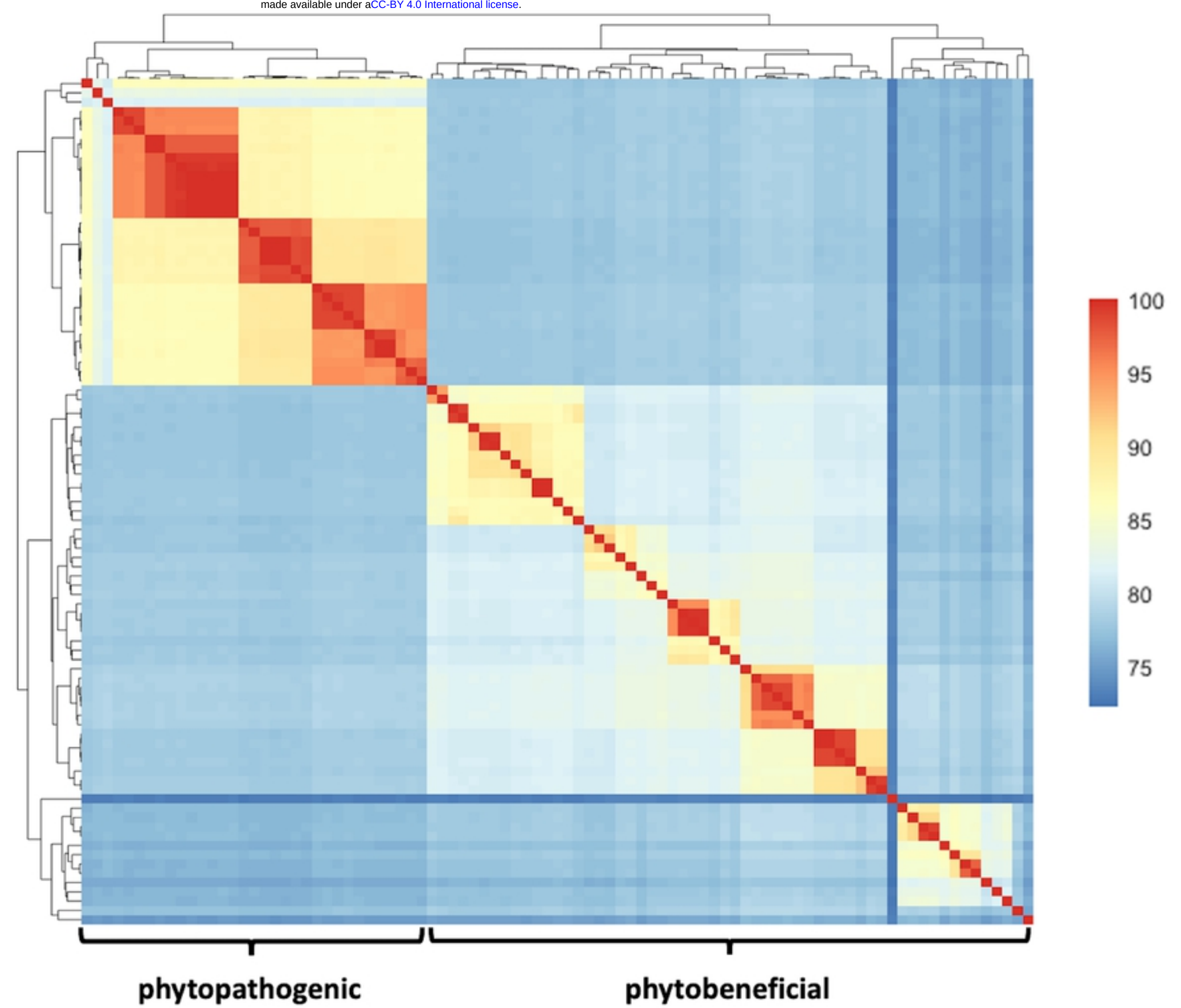

Figure2

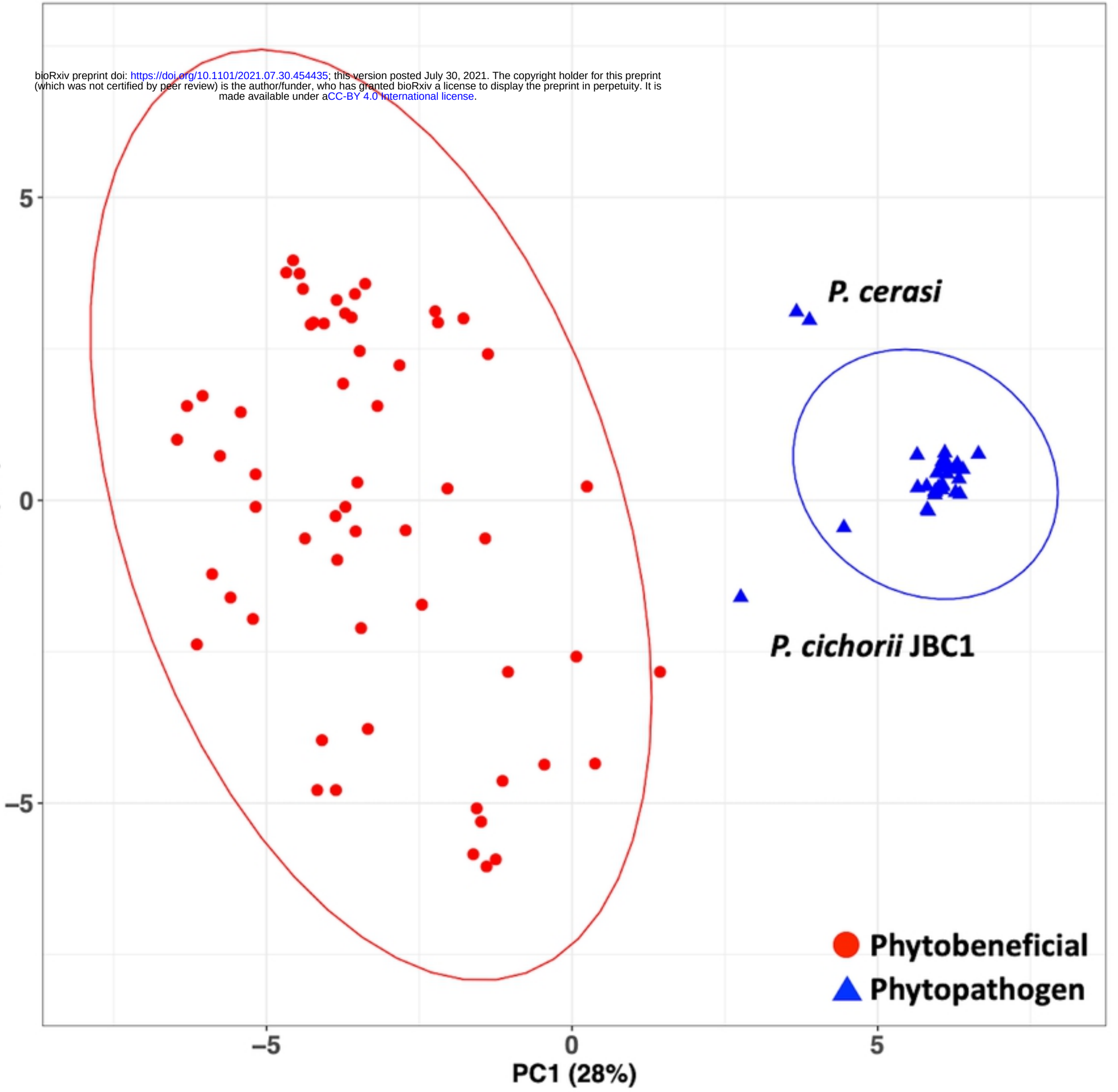

Figure3

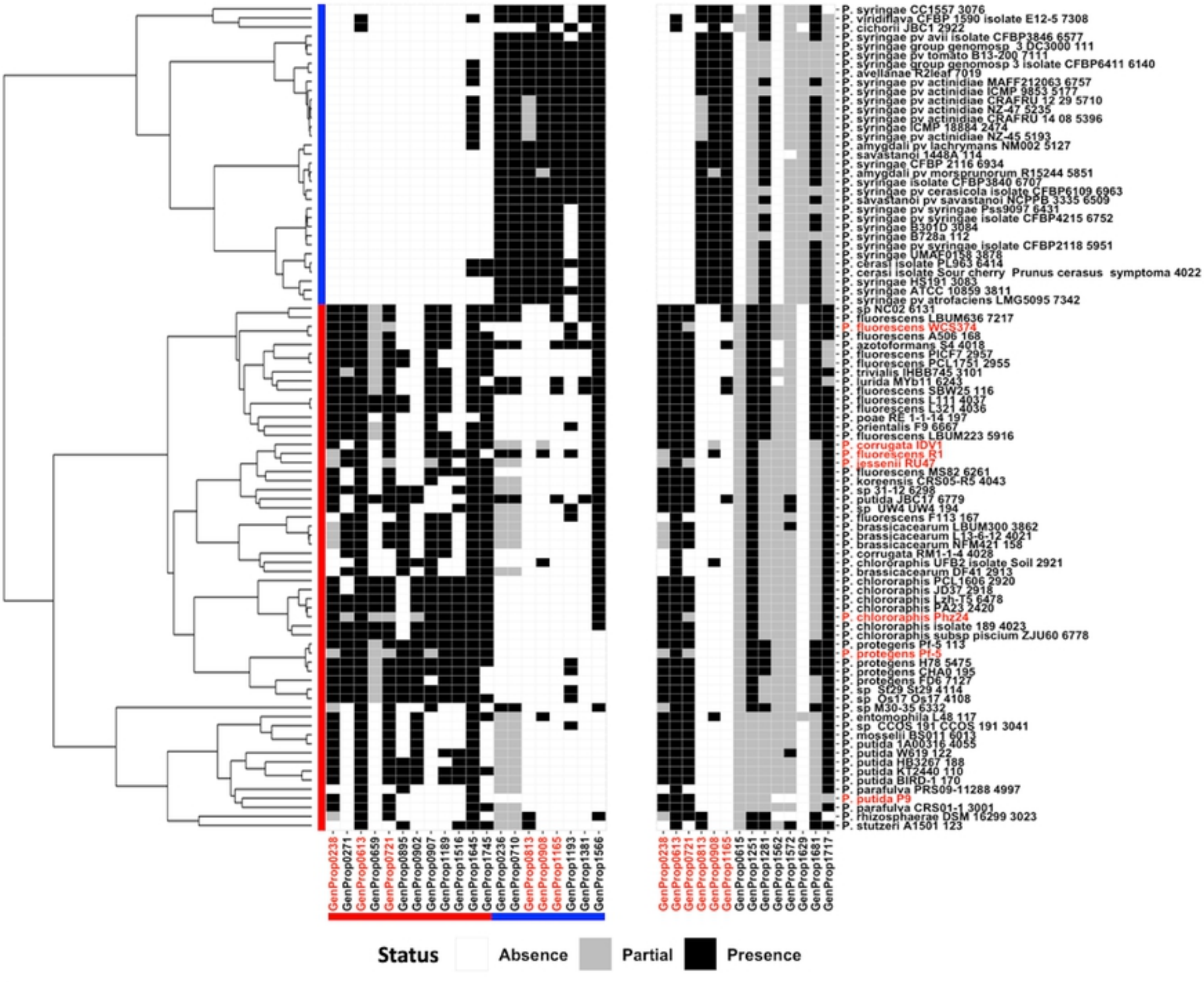
Enrichment analysis
Random Forest
P. syringae CC1557 3076
P. viridiflava CFBP 1590 isolate E12-5 7308
P. cichorii JBC1 2922
P. syringae pv avii isolate CFBP3846 6577
P. syringae group genomosp 3 DC3000 111
P. syringae pv tomato B13-200 7111
P. syringae group genomosp 3 isolate CFBP6411 6140
P. avellanae R2leaf 7019
P. syringae pv actinidiae MAFF212063 6757
P. syringae pv actinidiae ICMP 9853 5177
P. syringae pv actinidiae CRAFRU 12 29 5710
P. syringae pv actinidiae NZ-47 5235
P. syringae pv actinidiae CRAFRU 14 08 5396
P. syringae ICMP 18884 2474
P. syringae pv actinidiae NZ-45 5193
P. amygdali pv lachrymans NM002 5127
P. savastanoi 1448A 114
P. syringae CFBP 2116 6934
P. amygdali pv morsprunorum R15244 5851
P. syringae isolate CFBP3840 6707
P. syringae pv cerasicola isolate CFBP6109 6963
P. savastanoi pv savastanoi NCPPB 3335 6509
P. syringae pv syringae Pss9097 6431
P. syringae pv syringae isolate CFBP4215 6752
P. syringae B301D 3084
P. syringae B728a 112
P. syringae pv syringae isolate CFBP2118 5951
P. syringae UMAF0158 3878
P. cerasi isolate PL963 6414
P. cerasi isolate Sour cherry Prunus cerasus symptoma 4022
P. syringae HS191 3083
P. syringae ATCC 10859 3811
P. syringae pv atrofaciens LMG5095 7342
P. sp NC02 6131
P. fluorescens LBUM636 7217
P. fluorescens WCS374
P. fluorescens A506 168
P. azotoformans S4 4018
P. fluorescens PICF7 2957
P. fluorescens PCL1751 2955
P. trivialis IHBB745 3101
P. lurida MYb11 6243
P. fluorescens SBW25 116
P. fluorescens L111 4037
P. fluorescens L321 4036
P. poae RE 1-1-14 197
P. orientalis F9 6667
P. fluorescens LBUM223 5916
P. corrugata IDV1
P. fluorescens R1
P. jessenii RU47
P. fluorescens MS82 6261
P. koreensis CRS05-R5 4043
P. sp 31-12 6298
P. putida JBC17 6779
P. sp UW4 UW4 194
P. fluorescens F113 167
P. brassicacearum LBUM300 3862
P. brassicacearum L13-6-12 4021
P. brassicacearum NFM421 158
P. corrugata RM1-1-4 4028
P. chlororaphis UFB2 isolate Soil 2921
P. brassicacearum DF41 2913
P. chlororaphis PCL1606 2920
P. chlororaphis JD37 2918
P. chlororaphis Lzh-T5 6478
P. chlororaphis PA23 2420
P. chlororaphis Phz24
P. chlororaphis isolate 189 4023
P. chlororaphis subsp piscium ZJU60 6778
P. protegens Pf-5 113
P. protegens Pf-5
P. protegens H78 5475
P. protegens CHA0 195
P. protegens FD6 7127
P. sp St29 St29 4114
P. sp Os17 Os17 4108
P. sp M30-35 6332
P. entomophila L48 117
P. sp CCOS 191 CCOS 191 3041
P. mosselii BS011 6013
P. putida 1A00316 4055
P. putida W619 122
P. putida HB3267 188
P. putida KT2440 110
P. putida BIRD-1 170
P. parafulva PRS09-11288 4997
P. putida P9
P. parafulva CRS01-1 3001
P. rhizosphaerae DSM 16299 3023
P. stutzeri A1501 123
GenProp0238
GenProp0271
GenProp0613
GenProp0659
GenProp0721
GenProp0895
GenProp0902
GenProp0907
GenProp1189
GenProp1516
GenProp1645
GenProp1745
GenProp0236
GenProp0710
GenProp0813
GenProp0908
GenProp1165
GenProp1193
GenProp1381
GenProp1566
GenProp0238
GenProp0613
GenProp0721
GenProp0813
GenProp0908
GenProp1165
GenProp0615
GenProp1251
GenProp1281
GenProp1562
GenProp1572
GenProp1629
GenProp1681
GenProp1717
Status
Absence
Partial
Presence

Figure4

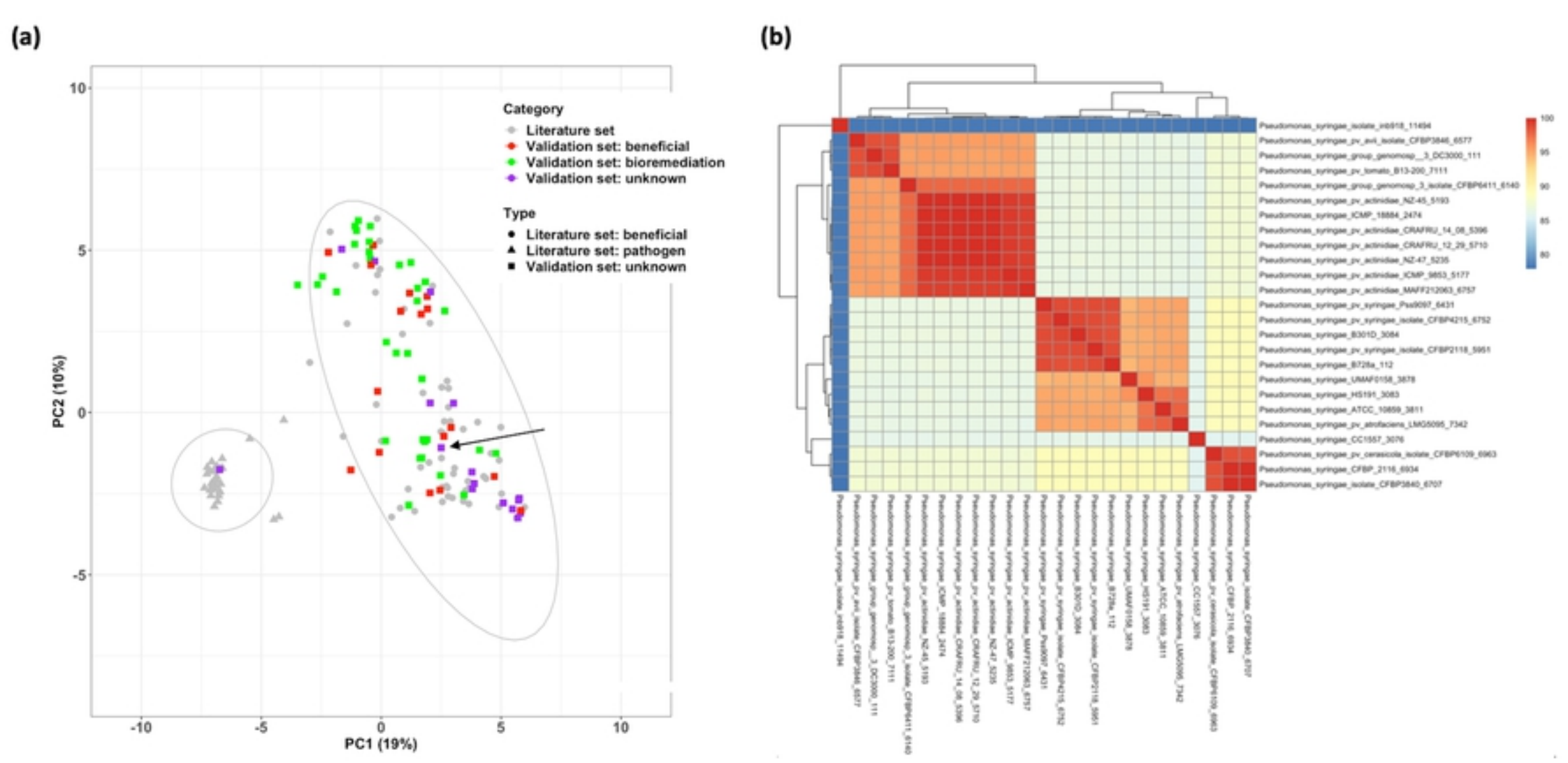
(a)
Category
Literature set
Validation set: beneficial
Validation set: bioremediation
Validation set: unknown
Type
Literature set: beneficial
Literature set: pathogen
Validation set: unknown
PC2 (10%)
PC1 (19%)
10
5
0
-5
-10
-5
0
5
10
(b)
Pseudomonas_syringae_isolate_mb918_11494
Pseudomonas_syringae_pv_avii_isolate_CFBP3846_6577
Pseudomonas_syringae_group_genomosp__3_DC3000_111
Pseudomonas_syringae_pv_tomato_B13-200_7111
Pseudomonas_syringae_group_genomosp_3_isolate_CFBP6411_6140
Pseudomonas_syringae_pv_actinidiae_NZ-45_5193
Pseudomonas_syringae_ICMP_18884_2474
Pseudomonas_syringae_pv_actinidiae_CRAFRU_14_08_5396
Pseudomonas_syringae_pv_actinidiae_CRAFRU_12_29_5710
Pseudomonas_syringae_pv_actinidiae_NZ-47_5235
Pseudomonas_syringae_pv_actinidiae_ICMP_9853_5177
Pseudomonas_syringae_pv_actinidiae_MAFF212063_6757
Pseudomonas_syringae_pv_syringae_Psa9097_6431
Pseudomonas_syringae_pv_syringae_isolate_CFBP4215_6752
Pseudomonas_syringae_B301D_3084
Pseudomonas_syringae_pv_syringae_isolate_CFBP2118_5951
Pseudomonas_syringae_B728a_112
Pseudomonas_syringae_UMAF0158_3878
Pseudomonas_syringae_HS191_3083
Pseudomonas_syringae_ATCC_10859_3811
Pseudomonas_syringae_pv_atrofaciens_LMG5095_7342
Pseudomonas_syringae_CC1557_3076
Pseudomonas_syringae_pv_cerasicola_isolate_CFBP6109_6963
Pseudomonas_syringae_CFBP_2116_6934
Pseudomonas_syringae_isolate_CFBP3840_6707
100
95
90
85
80

Figure5