 doi:10.20944/preprints202107.0400.v1

Article

Pan-genome of novel *Pantoea stewartii* subsp. *indologenes* reveal genes involved in onion pathogenicity and evidence of lateral gene transfer

Gaurav Agarwal*[1], Ronald D. Gitaitis[1] and Bhabesh Dutta*[1]

[1]Department of Plant Pathology, Coastal Plain Experiment Station, University of Georgia, Tifton, GA 31793
* Correspondence: authors: bhabesh@uga.edu (BD); gaurav.agarwal@uga.edu (GA)

Abstract: *Pantoea stewartii* subsp. *indologenes* (*Psi*) is a causative agent of leafspot of foxtail millet and pearl millet; however, novel strains were recently identified that are pathogenic on onion. Our recent host range evaluation study identified two pathovars; *P. stewartii* subsp. *indologenes* pv. *cepacicola* pv. nov. and *P. stewartii* subsp. *indologenes* pv. *setariae* pv. nov. that are pathogenic on onion and millets or on millets only, respectively. In the current study we developed a pan-genome using the whole genome sequencing of newly identified/classified *Psi* strains from both pathovars [pv. *cepacicola* (*n*= 4) and pv. *setariae* (*n*=13)]. The full spectrum of the pan-genome contained 7,030 genes. Among these, 3,546 (present in genomes of all 17 strains) were the core genes that were a subset of 3,682 soft-core genes (present in ≥16 strains). The accessory genome included 1,308 shell genes and 2,040 cloud genes (present in ≤ 2 strains). The pan-genome showed a clear liner progression with >6,000 genes, suggesting the pan-genome of *Psi* is open. Comparative phylogenetic analysis showed differences in phylogenetic clustering of *Pantoea* spp. using PAVs/wgMLST approach in comparison to core genome SNP-based phylogeny. Further, we conducted a horizontal gene transfer (HGT) study including four other *Pantoea* species namely, *P. stewartii* subsp. *stewartii* LMG 2715^T, *P. ananatis* LMG 2665^T, *P. agglomerans* LMG L15, and *P. allii* LMG 24248^T. A total of 317 HGT events among four *Pantoea* species were identified with most gene transfers observed between *Psi* pv. *cepacicola* and *Psi* pv. *setariae*. Pan-GWAS analysis predicted a total of 154 genes including seven cluster of genes associated with the pathogenicity phenotype on onion. One of the clusters contain 11 genes with known functions and are found to be chromosomally located.

Keywords: Pangenome; horizontal gene transfer (HGT); core genome; accessory genome

1. Introduction

Pantoea complex is constituted by four species namely, *P. ananatis*, *P. stewartii*, *P. allii* and *P. agglomerans* that causes center rot of onion [1-4]. Three out of the four species in *Pantoea* complex, *P. ananatis, P. agglomerans* and *P. stewartii* subsp. *indologenes* are responsible for more than 80% of the reported cases of disease in onions [5]. Earlier, Mergaert et al. [6] reclassified *Eriwinia stewartii* as *P. stewartii* and proposed two subspecies *P. stewartii* subsp. *stewartii* (*Pss*) and *P. stewartii* subsp. *indologenes* (*Psi*). Recently, we phenotypically and genotypically characterized seventeen *Psi* strains that are either pathogenic on both onions and millets or on millets only [7]. Based on the host-range evaluation we proposed two new pathovars of *Psi* namely *Psi* pv. *cepacicola* pv. *nov*. and *Psi* pv. *setariae* pv. nov [7]. The pathovar *Psi* pv. *cepacicola* causes symptoms on *Allium* species (leek, onion, chive and Japanese bunching onion) and also on foxtail millet, pearl millet and oat. However, *Psi* pv. *setariae* pv. nov can only infect the members of *Poaceae* (foxtail millet, pearl millet and oat) [7].

There has been a huge turnaround in terms of generating genomic resource due to simultaneous decrease in sequencing costs and generating next generating sequencing (NGS) based big data. As a result, several studies have been conducted to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

comprehensively explore features specific to each genome. Most widely explored genomic variants used in genome-wide studies are SNPs both in prokaryotes [8] and eukaryotes [9-14]. Another most widely explored variations in prokaryotic genomes are presence and absence variants (PAVs). These PAVs captures often evolving "accessory genome" of an organism (bacteria in current study). Another part of genome(s), which is conserved is regarded as a "core genome". Together the core and accessory genome constitute a pan-genome of a species or sometimes across species for a given genus (also regarded super pan-genome) [15]. Hence, the core genome refers to key genes that commonly exist in every member of a specific genome set, and the accessory genome represent dispensable genes, which only exist in some of the genomes [16]. In prokaryotes, a pan-genome can be open or closed depending on the similarity of gene content. Genomes with highly similar gene content make a closed pan-genome or conversely an open-pangenome [17]. A pan-genome of a species is dependent on the number of genomes involved in the dataset, the ability to integrate exogenous DNA into its genome via horizontal gene transfer (HGT) and environment of the species [18]. HGT and gene loss are key processes in bacterial evolution and often involved in gain or loss of function [19]. HGT can result in the replacement of genetic segments with donor homologues, often within species via homologous recombination, or via. acquisition of new genetic material.

In our earlier pan-genome study we used 81 strains of *P. ananatis* (included both pathogenic and non-pathogenic strains on onion) and performed pan-GWAS (pan-genome wide association study) to predict genes involved in onion pathogenicity (Agarwal et al. 2021). Our pan-GWAS study was able to predict genes and gene clusters potentially involved in onion pathogenicity and predicted several HGT events that occurred between onion-pathogenic vs. onion-non-pathogenic strains. In addition, phylogeny based on PAVs were also able to differentiate onion-pathogenic vs. non-pathogenic strains. In the current study, we utilized the available genome resource of *Psi* strains from both pathovars [pv. *cepacicola* (*n*= 4) and pv. *setariae* (*n*=13)] and developed a pan-genome with a conserved core and a flexible accessory genome. Further, we performed a pan-GWAS study to identify genes in *Psi* pv. *cepacicola* that are associated with onion pathogenicity and, also predicted several HGT events that occurred between *Psi* strains and among other *Pantoea* spp. (*P. stewartii* subsp. *stewartii* LMG 2715^T, *P. ananatis* LMG 2665^T, *P. agglomerans* LMG L15, *P. allii* LMG 24248^T). We further utilized SNPs and PAVs of the core-genome to assess phylogeny of both *Psi* pathovars.

2. Methods

2.1. Bacterial strains, identification, and culture preparation

Seventeen *Psi* strains from both pathovars [pv. *cepacicola* (*n*= 4) and pv. *setariae* (*n*=13)] (**Table 1**) were used in this study that were phenotypically characterized in our earlier study [7]. Out of the 17 strains, two strains were previously sequenced *Psi* pv. *setariae* LMG 2632^T (NZ_JPKO00000000.1) and *Psi* pv. *setariae* PNA 03-3 (GCA_003201175.1). Genome assemblies of the rest of the 15 *Psi* strains have been submitted to NCBI under Bioproject ID PRJNA670043 (genome submission: SUB8606059). Whole-genome sequences of already available four type strains of *Pantoea* spp. [*P. stewartii* subsp. *indologenes* pv. *setariae* LMG 2632^T (NZ_JPKO00000000.1); *P. stewartii* subsp. *stewartii* LMG 2715^T (GCA_008801695.1); *P. ananatis* LMG 2665^T (NZ_JMJJ00000000); *P. allii* LMG 24248^T (NZ_NTMH00000000)] and WGS based assembly of *P. agglomerans* L15 (NZ_CP034148) were used. Genomic features of all 21 strains used in this study are listed in **Table 1**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

Table 1. Genome architecture details of *Pantoea stewartii* subsp. *indologenes* (pv. *cepacicola* and pv. *setariae*) and other *Pantoea* species used in this study.

Strain name	*Pantoea* spp.	Biosample accession	Genome accession	Size (Mbp)	Contigs	CDSs	Genes	tRNAs
L15	*P. agglomerans*	SAMN07109613	GCA_003860325.1	4.85	4	4456	4538	81
LMG 24248[T$]	*P. allii*	SAMN07625522	NZ_NTMH00000000	5.24	57	4855	4925	69
LMG 2632[T$]	*Psi setariae*	SAMN02905159	NZ_JPKO00000000.1	4.68	35	4455	4521	65
LMG 2665[T$]	*P. ananatis*	SAMN02740635	NZ_JMJJ00000000	4.93	17	4560	4632	71
LMG 2715[T$]	*P. stewartii stewartii*	SAMN12697580	GCA_008801695.1	4.52	1	4603	4677	73
NCPPB 1562*	*Psi setariae*	SAMN16866628	JADWWO000000000	4.87	96	4524	4602	77
NCPPB 1877*	*Psi setariae*	SAMN16866626	JADWWM000000000	4.77	83	4410	4487	76
NCPPB 2275*	*Psi setariae*	SAMN16866625	JADWWL000000000	4.77	79	4406	4481	74
NCPPB 2281*	*Psi setariae*	SAMN16866629	JADWWP000000000	4.70	103	4323	4399	75
NCPPB 2282*	*Psi setariae*	SAMN16866627	JADWWN000000000	4.87	101	4529	4608	78
PANS_07_10*	*Psi setariae*	SAMN16866621	JADWWH000000000	4.95	102	4603	4678	74
PANS_07_12*	*Psi setariae*	SAMN16866622	JADWWI000000000	4.95	86	4602	4674	71
PANS_07_14*	*Psi setariae*	SAMN16866623	JADWWJ000000000	4.78	90	4429	4404	74
PANS_07_4*	*Psi setariae*	SAMN16866619	JADWWF000000000	5.05	125	4686	4760	73
PANS_07_6*	*Psi setariae*	SAMN16866620	JADWWG000000000	5.10	117	4744	4816	71
PANS_99_15*	*Psi setariae*	SAMN16866624	JADWWK000000000	4.81	97	4425	4498	72
PNA_15_2*	*Psi setariae*	SAMN16866618	JADWWE000000000	4.66	92	4266	4341	74
PNA_03_3*	*Psi cepacicola*	SAMN08776223	GCA_003201175.1	4.93	22	4571	4641	69
PNA_14_11*	*Psi cepacicola*	SAMN16866616	JADWWC000000000	4.68	77	4317	4390	72
PNA_14_12[T*]	*Psi cepacicola*	SAMN16866617	JADWWD000000000	4.68	73	4307	4380	72
PNA_14_9*	*Psi cepacicola*	SAMN16866615	JADWWB000000000	4.69	92	4325	4400	74

[T] Denotes type strains.
[$] Sequences were downloaded from the NCBI.
* Sequences utilized from Koirala et al., 2021 study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

2.2. Identification of presence and absence variations (PAVs)

The gbk (genebank format) files of the draft-assembled and annotated genomes of *Psi* strains [7] were used for pan-genome analyses using get_homologues [20]. These gbk files were used to get the syntenic sequence clusters by get_homologues.pl using OrthoMCL (OMCL) algorithm. The syntenic clusters generated were used to a develop a pan-genome matrix showing presence and absence variants (PAVs) using compare_clusters.pl. The matrix was also used to classify genes into core, soft-core, shell and cloud genes using parse_pangenome_matrix.pl (auxiliary script of get_homologues.pl). Core genes are defined as those present in all 17 *Psi* genomes whereas accessory genes are present in a subset of the 17 genomes. The accessory gene cluster was further divided into cloud and shell gene clusters. Soft-core genes occurred in 95% of the genomes. Cloud genes were present in ≤2 genomes and shell genes comprised of remaining genes [20]. Distribution of cluster sizes as a function of the number of genomes these clusters contained was displayed using R with parse_pangenome_matrix.pl. Gower's distance matrix was generated using the tab delimited pan-genome PAV file as input by executing shell script hcluster_pangenome_matrix.sh (auxiliary script of get_homologues) when used to call R function hclust.

2.3. Horizontal gene transfer (HGT) and phylogenetic analysis of genomes of Pantoea complex

Phylogenetic tree was built for all input genomes using the protein sequences of universal single copy genes (SCGs). To carry out HGT analysis we included genomes of four additional strains representing the four *Pantoea* species of *Pantoea* complex namely, *P. stewartii* subsp. *stewartii* LMG 2715^T, *P. ananatis* LMG 2665^T, *P. allii* LMG 24248^T and *P. agglomerans* L15. Predicted protein sequences of all 21 genomes (17 *Ps*i and four *Pantoea* spp. stated above; table 1) were searched for the HMM (PFAM and TIGRFAM) profiles of these SCG proteins using HMMER. Protein sequences for each HMM profile aligned using HMMER were concatenated into a single multiple sequence alignment using GTDB tool kit [21]. Further, we utilized multiple sequence alignment file to build a phylogenetic tree. The tree file was visualized in iTOL (https://github.com/songweizhi/BioSAK/tree/master/BioSAK_tutorial/Demo_tree_visualizationwith_iTol). Customized groupings (A, B, C and D) were made based on the tree (**Figure 4a**). These customized groups of genomes were used as input to study HGT events among the *Pantoea* spp. mentioned above. MetaCHIP [22] was used to identify horizontal gene transfer (HGT) among the customized assigned groups. MetaCHIP identified putative donor and recipient transfer events within the 21 *Pantoea* strains (17 *Ps*i and four *Pantoea* spp. stated above; **Table 1**) based on combined similarity and phylogenetic incongruency.

SNPs from core genome and PAVs of *Psi* strains (*n*=17) and other *Pantoea* spp. including *P. stewartii* subsp. *stewartii* LMG 2715^T, *P. ananatis* LMG 2665^T, *P. allii* LMG 24248^T and *P. agglomerans* L15 were identified, and a phylogenetic tree was constructed. Pan-seq pipeline [23] was used to identify the core SNPs, and PAVs from the accessory genomes. The output phylip files were used to construct phylogenetic trees based on SNPs and PAVs. The PHYLIP files were imported into RaxML software and PHYLIP trees were constructed by using the 'neighbor-joining' method, with a bootstrap setting of 1000. Further, whole genome multi locus sequence typing (wgMLST) tree was also constructed [24]. To carry out this analysis assembled contigs file of each of the 21 *Pantoea* spp. strains were uploaded to PGAdb builder and a pan-genome allele database was constructed with 1000 iterations.

2.4. Pan-genome-wide association analysis and annotations

We utilized phenotypic data from the Koirala et al. (2021) study [7]where both *Psi* pv. *cepacicola* and *Psi* pv. *setariae* strains were phenotyped based on their ability to cause symptoms on onion seedlings. Only four of the *Psi* pv. *cepacicola* strains (PNA 03-3, PNA 14-9, PNA 14-11 and PNA 14-12) strains were able to cause foliar symptoms. Scoary was used to calculate associations among genes in the pan-genome and the pathogenic

 doi:10.20944/preprints202107.0400.v1

phenotype on onion seedlings (a qualitative assessment; pathogenic vs. non-pathogenic association). The output of this program comprised of a list of genes sorted by strength of association with these traits. Genes with a naïve p-value <=0.005, and corrected p-value (Benjamini-Hochberg) of association <0.25 were considered significant. The core, soft-core, shell and cloud genes were retrieved from the *P. stewartii* pan-genome and subject to blastX against the NR database. The blast output files generated in .xml format were used as input to blast2GO. First GO mapping was done to retrieve the GO terms associated with blast to form a pool of GO terms. Then GO annotation was carried out where the GO terms from the pool of GO terms were assigned to query sequences. All sequences with GO annotations were also annotated for enzyme code. GO term associations were classified and plotted as biological process (BP), molecular function (MF) and cellular component (CC). FatiGO [25] package integrated into Blast2GO was used for statistical assessment of annotation differences as following: core vs accessory, soft-core vs core, shell vs core and cloud vs core genes. This package uses the Fisher's Exact Test and corrects for multiple testing. Adjusted p-values of each GO term was reported based on the corrected p-value by False Discovery Rate (FDR) control. Genes involved in horizontal gene transfer (HGT) were annotated using Blast2GO pipeline [26].

3. Results

3.1. The P. stewartii subsp. indologenes pan-genome architecture and phylogeny

Genome assembly is a pre-requisite to study pan-genome of a given species. Assembly sizes of the seventeen strains of *Psi* ranged from 4.6 (PNA 15-2) to 5.1 Mbp (PANS 07-6) with number of contigs ranging from 22 (PNA 03-3) to 125 (PANS 07-4) and number of genes ranging from 4341 (PNA 15-2) to 4816 (PANS 07-6) (**Table 1**). The full spectrum of the pan-genome of *Psi* contained 7,030 genes. Among these, 3,546 genes (present in all 17 strains) were part of the core genes and is a part of a subset of 3,682 soft-core genes (present in >=16 strains). The accessory genome included 1,308 shell genes and 2,040 cloud genes (present in <=2 strains) (**Figure 1a**). Genome of each strain contributed a conserved set of 3,546 core genes and a variable number of accessory genes. Overall, soft-core genes contributed by each genome ranged from 3,580 to 3,682 genes including a conserved set of 3,546 genes. Shell genes ranged from 394 to 703 genes and the cloud genes ranged from a minimum of five genes contributed by *Psi* pv. *cepacicola* PNA 14-12 to a maximum of 382 genes contributed by *Psi* pv. *setariae* LMG 2632^{T} (**Figure 1b**). Details of the number of core and accessory genes contributed by each strain are listed in (**Table S1)**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

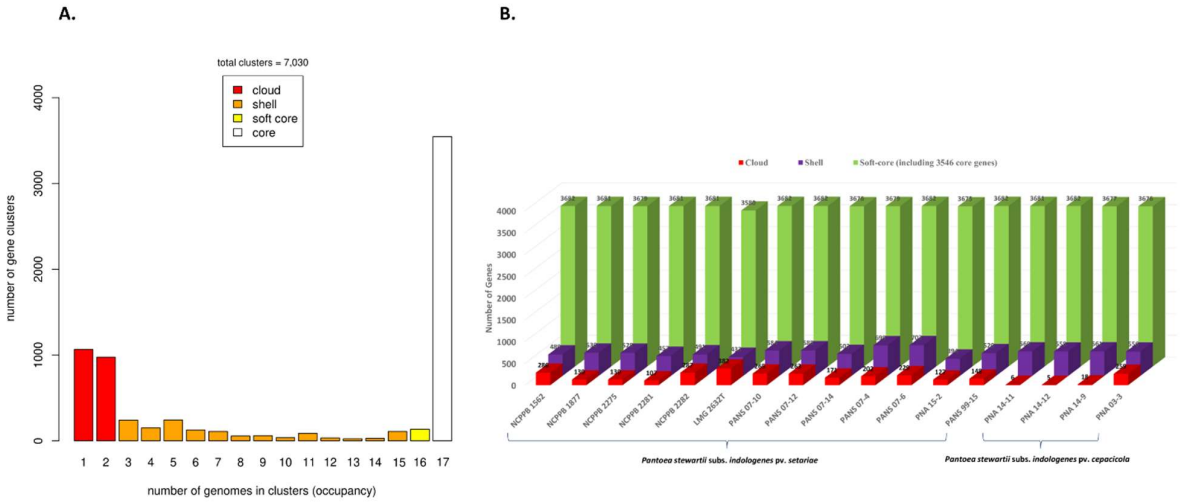

Figure 1. Pan-genome analysis of 17 *Pantoea stewartii* subsp. *indologenes* genomes. (A) Distribution of gene (cluster) sizes as a function of the number of genomes they contain showing the partition of OMCL pan-genomic matrix into shell, cloud, soft-core and core compartments. (B) Genes contributed to pan-genome by individual genomes.

Pan-genome architecture of the 17 *Psi* genomes including both pathovars revealed an open pan-genome (**Figure 2**). Exponential decay models [16, 27] that fitted the core gene clusters predicted a theoretical core genome of 3,598 and 3,524 genes (**Figure 2a**). Further, to confirm the openness/closeness of the pan-genome, theoretical estimation of pan-genome size was carried out using Tettelin's exponential model fitted to the OMCL accessory gene clusters. The pan-genome showed a clear linear progression with >6,000 genes, with ~43 new genes being added on an average to the pan-genome as each new *Psi* genome will be added (**Figure 2b**). This indicates an open *Psi* pan-genome.

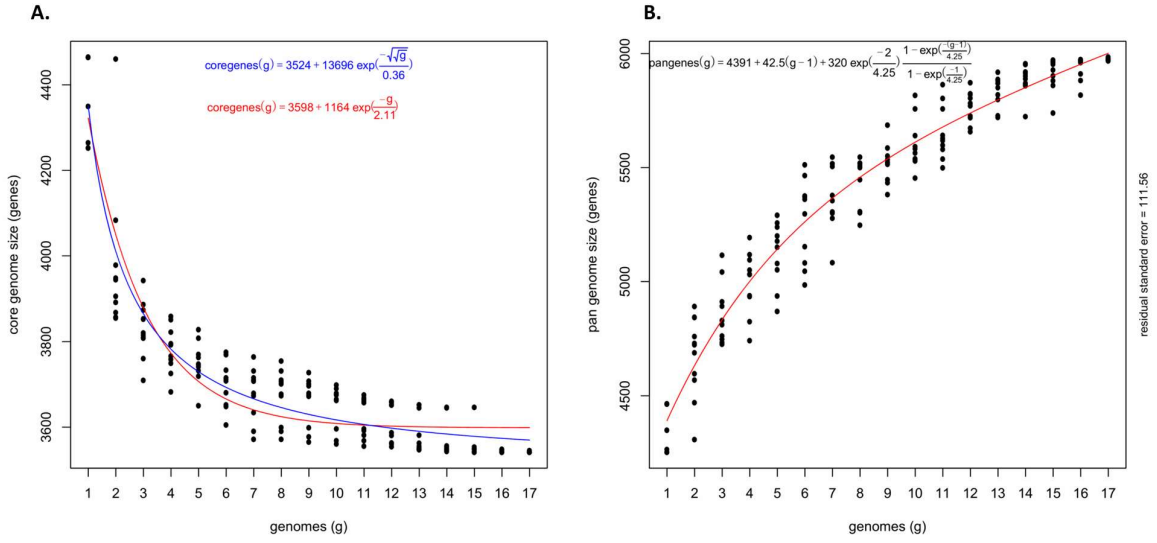

Figure 2. Theoretical estimation of the core and pan-genome sizes based on the exponential decay model. (A) Estimation of core genome size based on Willenbrock model fit to OMCL clusters. (B) Estimation of pan-genome size based on Tettelin model fit to OMCL clusters.

Dendrograms were plotted based on the core and accessory genes identified in the *Psi* pan-genome. The four components of the developed pan-genome of *Psi* namely, core, soft core, shell and cloud genes were used to assess the phylogeny of 17 *Psi* strains. All four components of pan-genome when used individually clustered the *Psi* pv. *cepacicola* strains separately from the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

Psi pv. *setariae* strains. However, only one *Psi* pv. *cepacicola* (PNA 03-3) was distantly clustered or with *Psi* pv. *setariae* strains when soft-core, shell and cloud genes were used. The *Psi* pv. *cepacicola* strain PNA 03-3 clustered close to other three *Psi* pv. *cepacicola* strains (PNA 14-12, PNA 14-11, PNA 14-9) only when core genes were used (**Figure 3a-d**).

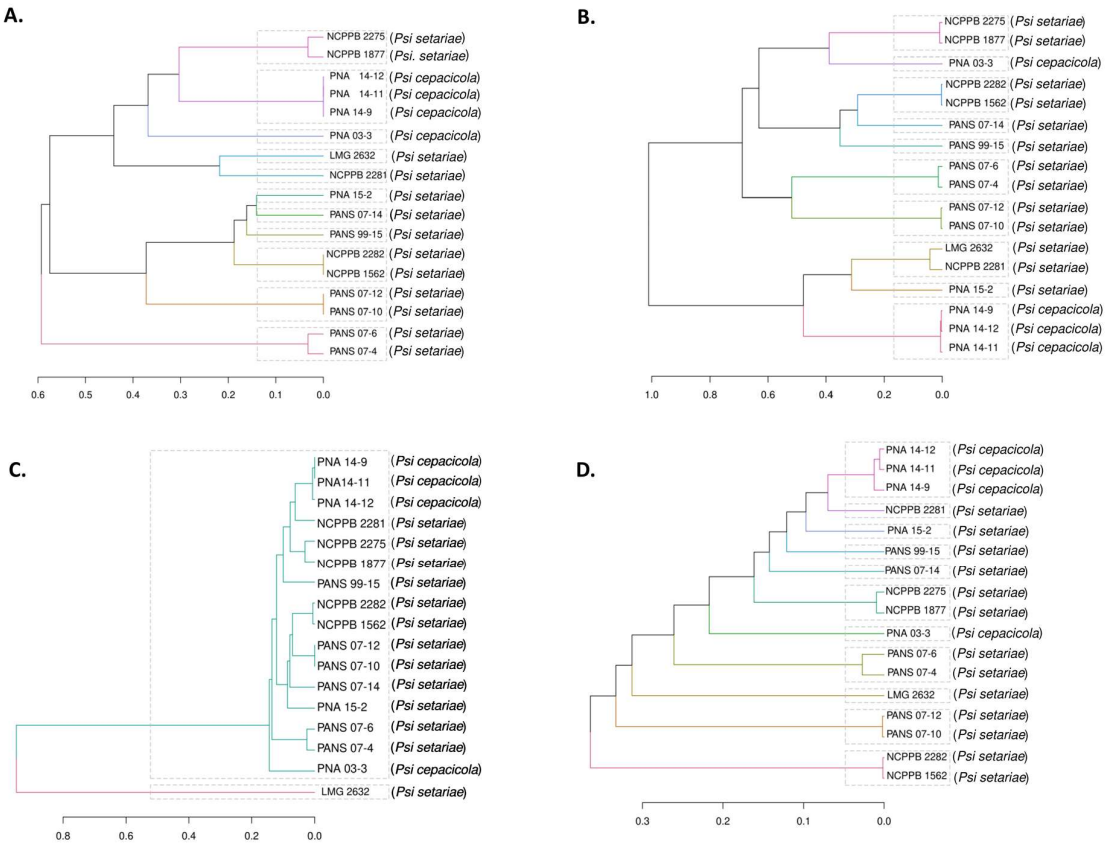

Figure 3. Dendrogram of 17 strains of *Pantoea stewartii* subsp. *indologenes* based on the core and accessory genes. (A) Based on core genes i.e., genes present in all 17 strains used in the study. (B) Based on soft-core genes i.e., genes present in at least 95% of the strains. (C) Based on shell genes. (D) Based on cloud genes i.e., the genes specific to each strain or shared by a maximum of two strains.

3.2. Core genome genes differentiated the pathogenic P. stewartii subsp. indologenes pv. cepacicola strains from the non-pathogenic P. stewartii subsp. indologenes pv. setariae strains in onion

Among the seventeen newly discovered *Psi* strains, four *Psi* pv. *cepacicola* strains were pathogenic on *Allium* spp. (onion, leek, chive and bunching onion) and *Poacea* (oat, rye, foxtail millet) species [7]. Rest of the strains belonged to *Psi* pv. *setariae*, which were pathogenic only on *Poacea* species (oat, rye, foxtail millet). Clustering based on the ANI matrix of pan-genome resulted in phylogenic trees based on core, soft-core, shell and cloud genes (**Figure 3**). Core genes ANI matrix resulted in eleven clusters with three of the four *Psi* pv. *cepacicola* strains clustered together and the fourth strain (PNA 03-3) clustered separately (**Figure 3a**). Soft-core ANI resulted in ten clusters (**Figure 3b**). Shell and cloud genes (accessory genome) ANI resulted in two and eleven clusters, respectively. Two shell gene clusters contained 16 out of the 17 *Psi* strains in one cluster and one strain (LMG 2632^T) in a separate clade (**Figure 3c**). Three out of the four *Psi* pv.*cepacicola* strains (PNA 14-9, PNA 14-11 and PNA 14-12) consistently clustered together with core or accessory genes-based ANI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

3.3. Horizontal gene transfer (HGT) and annotation of genes involved in HGT

Five *Pantoea* species including 17 *Psi* strains (four *Psi* pv. *cepacicola* and 13 *Psi* pv. *setariae* strains) and one strain each of *P. ananatis* (LMG 2665^{T}), *P. allii* (LMG 24248^{T}), *P. agglomerans* (L15) and *P. stewartii* subsp. *stewartii* (LMG 2715^{T}) were used for HGT analysis. Phylogenetic classification based on the conserved SCG resulted in four groups (a-d) (**Figure 4a**). Strains classified within these four groups were used to study the HGT as explained in methods section. A total of 317 HGT events including 314 donor and 299 recipient genes among the five *Pantoea* species/pathovars were identified (**Figure 4b**, **Table S2**). Most of the gene transfers (*n*=95) occurred from PNA 03-3 (*Psi* pv. *cepacicola*) to PANS 07-4 (*Psi* pv. *setariae*) followed by 76 HGTs from NCPPB 2275 (*Psi* pv. *setariae*) to PANS 07-4 (*Psi* pv. *setariae*), 32 from NCPPB2281 (*Psi* pv. *setariae*) to PANS 07-4 (*Psi* pv. *setariae*) and 27 from NCPPB 2275 to PNA 07-10 (*Psi* pv. *setariae*). The strain NCPPB 2275 (*Psi* pv. *setariae*) transferred (donated) a maximum number of 112 genes, followed by PNA 03-3 (*Psi* pv. *cepacicola*; *n*=100), NCPPB 2281 (*Psi* pv. *setariae*; *n*=35) and others donated less than 20 genes. Similarly, among the recipient strains, PANS 07-4 (*Psi* pv. *setariae*) received a maximum of 211 genes followed by PANS 07-10 (*Psi* pv. *setariae*; *n*=40) and rest of the strains received less than 20 genes (**Figure 4b**, **Table S2**). There were two compulsive donor strains (*Psi* pv. *setariae* NCPPB 2275 and *Psi* pv. *setariae* NCPPB 2281) that did not feature as recipients. Similarly, two recipient strains (*Psi* pv. *setariae* LMG 2632^{T} and *Psi* pv. *setariae* NCPPB 1562) that did not feature as donors (**Figure 4b, Table S2**).

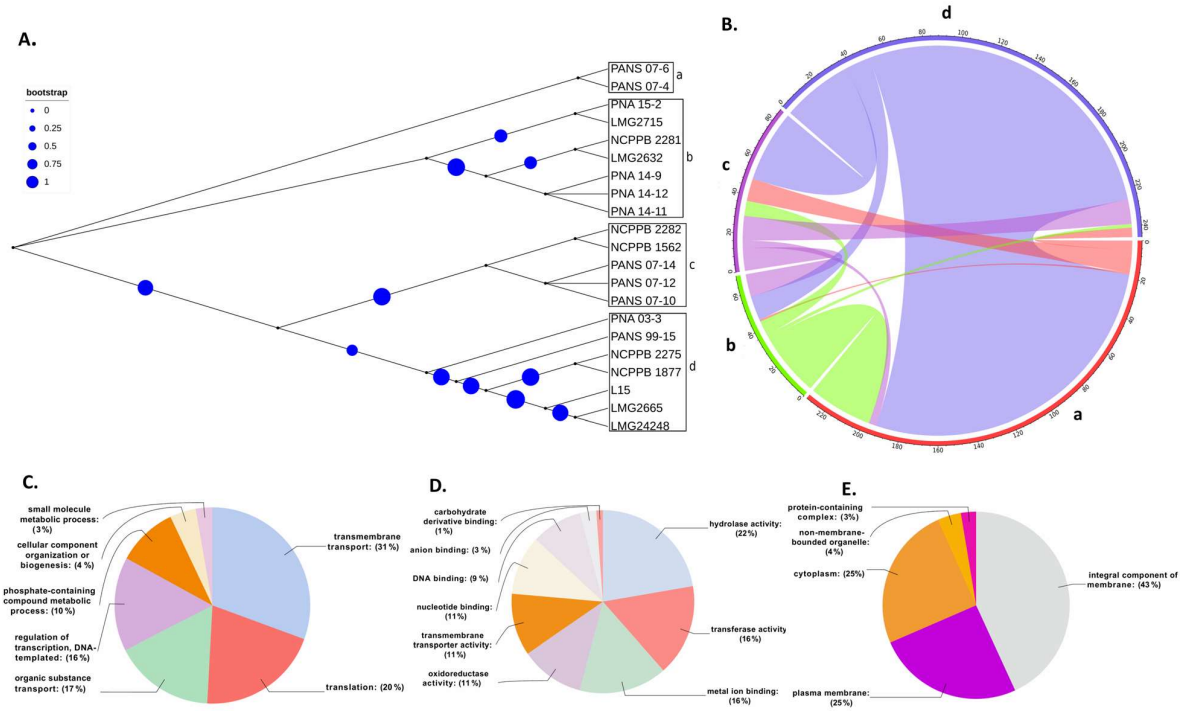

Figure 4. Phylogeny based horizontal gene transfer among 17 strains of of *Pantoea stewartii* subsp. *indologenes* strains and four other species of *Pantoea* complex namely, *P. ananatis*, *P. stewartii* subs. *stewartii*, *P. agglomerans* and *P. allii*. (A) Phylogenetic tree of *Pantoea* spp. (*n*=21) strains based on multiple sequence alignment. Phylogenetic tree resulted in four clusters (a-d). Size of circles represent the bootstrap values in that order; (B) Predicted gene flow within the four phylogenetic clusters of *Pantoea* spp. Bands connect donors and recipients, with the width of the band correlating to the number of HGTs and the color corresponding to the donors. Numbers on the circumference of circos plot represent the number of genes that undergo horizontal gene transfers. The four arcs (a-d) of circos represent the four phylogenetic clusters of *Pantoea* spp. strains; (C-E) Graphical annotations of sequences involved in horizontal genes transfer (HGT) as per the assigned GO terms: (C) Shows the function of genes assigned to biological process, (D) Shows the function of genes assigned to molecular function, and (E) Represent the function of genes assigned to cellular component.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

3.4. Annotations of genes involved in horizontal gene transfer (HGT)

Further, the donor and recipient proteins coded by genes involved in HGT were annotated. A non-redundant set of 607 genes coding for proteins involved in HGT showed blast hits. A total of 499 out of 607 genes that showed blast hits were mapped and annotated (**Table S3**). Each of the 499 genes involved in HGT were assigned GO IDs from a minimum of one to a maximum of eight. Eight genes were enriched by a maximum of eight GO IDs followed by 23 genes enriched by seven, 26 genes by six, 57 genes by five, 86 genes by four, 107 by three, 80 genes by two and the rest 112 genes by one GO ID (**Table S4**). Among all the HGTs, ABC transporter permease featured in a maximum of eight HGT events followed by GNAT family N-acetyltransferase in seven, cytochrome ubiquinol oxidase subunit I in six, outer membrane lipoprotein chaperone in five HGT events (**Table S3**). Based on the assigned GO IDs HGT genes were classified into biological process (BP), molecular function (MF) and cellular component (CC). Under BP the maximum number of genes involved in HGT were involved in transmembrane transport (31%) followed by genes involved in translation (20%) and only 3% were involved in small molecule metabolic process (**Figure 4c**). Under MF, 22% were categorized with hydrolase activity, followed by 16% each categorized with transferase activity and metal binding. The least number of HGTs were categorized as carbohydrate derivative binding (**Figure 4d**). In CC, most of the genes were categorized under integral component of membrane (43%) followed by 25% each under plasma membrane and cytoplasm and the minimum as protein-containing complex (3%) (**Figure 4e**). Further we investigated various pathways that these donor and recipient genes were involved. A total of 66 biochemical pathways were identified using KEGG database where HGT genes were involved. Purine fatty acid biosynthesis, pyrimidine, nicotinate and nicotinamide metabolic pathways featured most of the HGT genes. Purine metabolism showed 11 genes, fatty acid biosynthesis showed six, pyrimidine and nicotinamide metabolism showed five genes each. Rest of the metabolic pathways involved three, two or one gene(s) (**Figure 5, Table S5**).

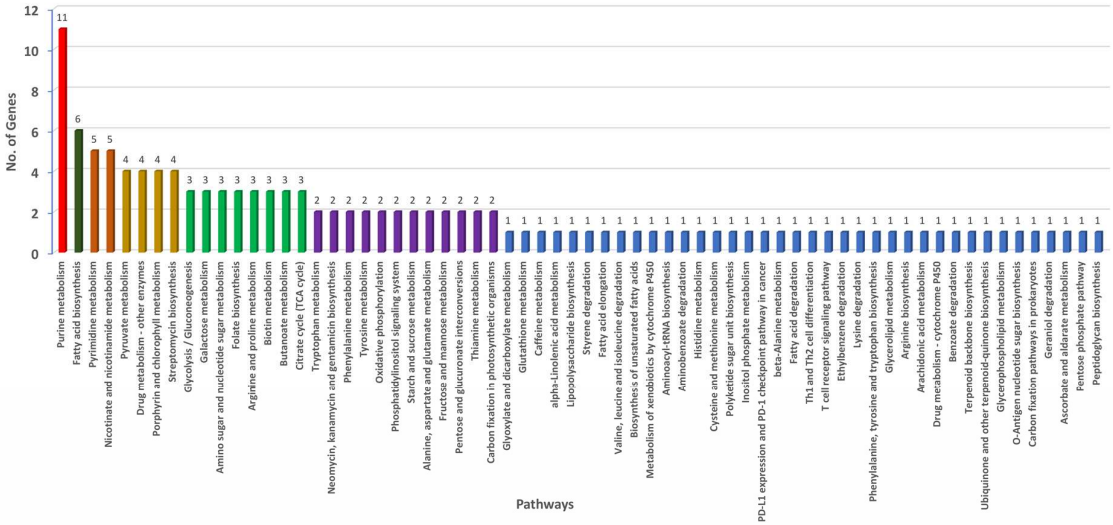

Figure 5. Annotated genes involved in horizontal gene transfer (HGT) with functions under biochemical, metabolic, and physiological pathways operational in *Pantoea stewartii* subs. *indologenes*.

 doi:10.20944/preprints202107.0400.v1

3.5. Comparative PAVs, core SNPs and whole genome multi locus sequence typing (wgMLST) based phylogeny

A comparative phylogenetic analysis was conducted using PAVs and core genome SNPs. Phylogenomic analysis using PAVs and SNPs showed both *P. ananatis* (LMG 2665^T) and *P. allii* (LMG 24248^T) were outliers. However, *P. agglomerans* (L15) grouped together with the same *Psi* pv. *setariae* strains using both PAVs and SNPs (**Figure 6**). The strain *P. stewartii* subsp. *stewartii* (LMG 2715^T) clustered together with PANS 07-4 (*Psi* pv. *setariae*) and PANS 07-6 (*Psi* pv. *setariae*) when plotted using PAVs. However, core SNPs were used it clustered with *Psi cepacicola;* although did not share the same node. Three of the four *Psi* pv. *cepacicola* strains (PNA 14-12, PNA 14-11 and PNA 14-9) clustered together when both with PAVs and core SNPs were used. The fourth *Psi* pv. *cepacicola* strain (PNA 03-3) clustered with *Psi* pv. *setariae* strains when PAVs were used. However, when core SNPs were used, PNA 03-3 formed a separate clade between LMG 2632^T (*Psi* pv. *setariae*) and LMG 2665^T (*P. ananatis*). Overall, with both PAVs and SNPs the *Psi* pv. *cepacicola* and *Psi* pv. *setariae* strains clustered in two separate groups except for PNA 03-3 (**Figure 6**). PAVs along with SNPs were identified using pan-seq to conduct comparative phylogenetic analysis. As seen with SNPs and PAVs, wgMLST clustered *Psi* pv. *cepacicola* strains together except for the strain PNA 03-3, which clustered with *Psi* pv. *setariae* strains (**Figure 7**). Strains of four other species of *Pantoea* used in this study (*P. agglomerans*, *P. ananatis*, *P. allii*, *P. stewartii* subs. *stewartii*) branched separately from *Psi* strains used. Like PAV based phylogenetic tree, wgMLST tree showed *P. stewartii* subsp. *stewartii* (LMG 2715^T) to be the closest to *Psi* pv. *setariae* out of the four Pantoea spp. compared (**Figure 6, 7**). However, wgMLST based phylogeny branched out four *Pantoea* species from *Psi* strains, which was not observed in PAV or SNP based phylogeny.

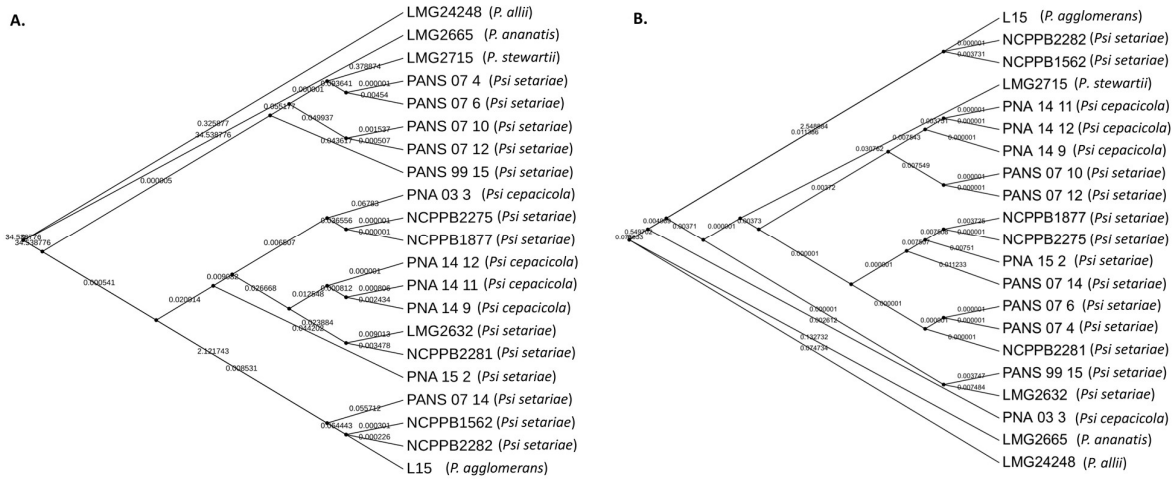

Figure 6. Comparative phylogeny of onion-pathogenic and onion-non-pathogenic strains of *Pantoea stewartii* subs. *indologenes* based on core genome SNPs and presence and absence variations (PAVs). (A) Phylogenetic tree constructed using PAVs using RAxML. (B) Phylogenetic tree constructed using core SNPs using RAxML. Numerical values in decimal represent the branch length. Longer branch length mean higher genetic divergence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

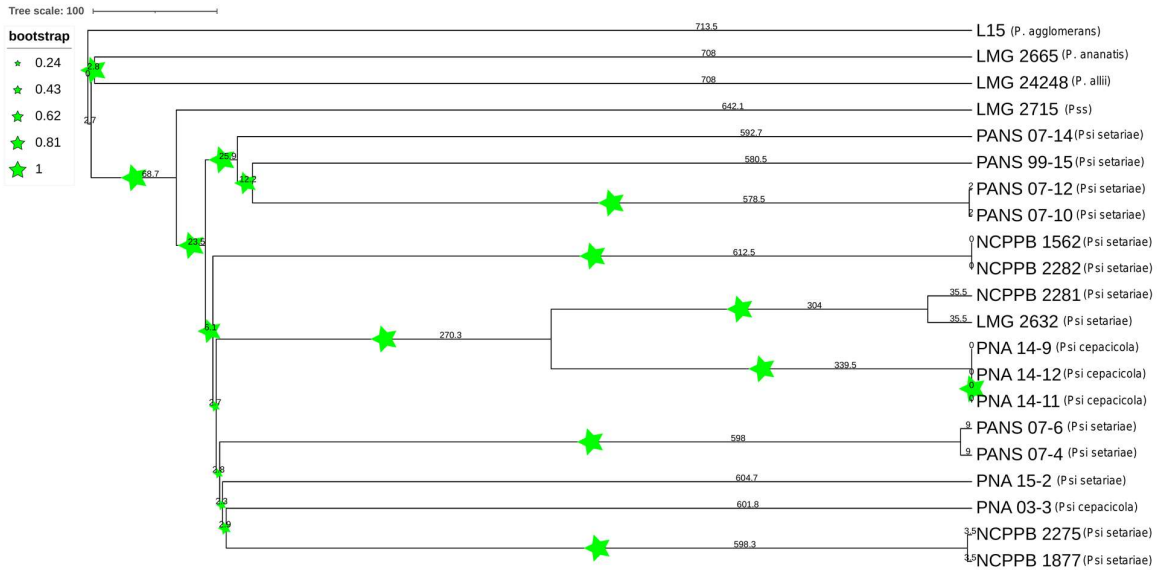

Figure 7. Phylogenetic tree based on whole genome multi-locus sequence typing (WgMLST). Dendrogram for 21 *Pantoea* spp. strains was constructed using assembled genome contigs. Size of stars represent the boot strap values and numbers represent the branch length. PSS is *Pantoea stewartii* subs. *stewartii.*

3.6. Pan-genome-wide association study

Association study was conducted using the qualitative phenotyping data (pathogenicity on onion seedling), and the core and accessory genes. The study aimed at identifying genes responsible for pathogenicity of *Psi* strains in onion. Scoary predicted a total of 154 genes associated with the phenotype (**Table S6**). Among the 154 genes, we found seven cluster of genes associated with the phenotype. There were three gene clusters with five, 12 and 11 genes in them that were highly significant based on p-values (p-value <= 0.005). Two out of the three clusters contained only non-annotated hypothetical protein coding genes. However, one cluster contained eleven protein coding genes (**Table 2**). These 11 genes coded for Pyridoxal 5'-phosphate synthase, AMP-binding protein, MFS transporter, phosphoglycerate kinase, FAD-NAD(P)-binding protein, phosphoenolpyruvate phosphomutase, NAD(P)-binding domain-containing protein, N-acetyl-gamma-glutamyl-phosphate reductase, alcohol dehydrogenase catalytic domain-containing protein, Iron containing alcohol dehydrogenase, LysE-family-translocator. Blast search against PNA 97-1R genome assembly (GCA_002952035.2) identified all genes except the last gene (LysE-family translocator) in cluster. Interestingly, first ten genes in the cluster are chromosomally localized but the last gene is localized in the large plasmid (**Table S7**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

Table 2. List of genes with annotated function in the cluster identified in *Pantoea stewartii* subsp. *indologenes* by Pan-GWAS analysis.

Gene ID	Function*	Sensitivity	Specificity	Naive_p
58220_pdxH_2	Pyridoxal 5'-phosphate synthase	75	100	0.0059
58221_dltA	AMP-binding protein	75	100	0.0059
58222_ydeE	MFS transporter	75	100	0.0059
58223_pgk-tpi	Phosphoglycerate kinase	75	100	0.0059
58225_spuC	FAD/NAD(P)-binding protein	75	100	0.0059
58226_pepM	Phosphoenolpyruvate mutase	100	100	0.0004
58227_Hydroxypyruvate_reductase	NAD(P)-binding domain-containing protein	75	100	0.0059
58228_argC_1	N-acetyl-gamma-glutamyl-phosphate reductase	75	100	0.0059
58229_lgoD_1	Alcohol dehydrogenase catalytic domain-containing protein	75	100	0.0059
58230_Iron_containing_alcohol dehydrogenase	Iron containing alcohol dehydrogenase	75	100	0.0059
58231_rhtC_1	LysE-family-translocator	75	100	0.0059

*Gene annotation was conducted using a blast search against NR database on NCBI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2021 doi:10.20944/preprints202107.0400.v1

3.7. Annotation of Pantoea stewartii subsp. indologenes pan-genome

The core and accessory genomes were annotated as genes involved in BP, MF and CC (**Figure 8**). Under BP: response to stimulus, metabolic process and cellular process were common in core, soft-core, shell and cloud genes. Regulation of biological process was specific to soft-core genes and shell genes lacked biological regulation and regulation of biological process functions (**Figure 8**). Under MF: catalytic and binding activities were common in all the core and accessory genes. However, genes coding for molecular function regulators were specific to soft-core component of the pan-genome and shell genes lacked transporter activity. The genes in CC category were enriched for cellular anatomical entity and intracellular component. Further, we performed statistical assessment of annotation differences between core and accessory genes. When core genome was compared with accessory genome, catalytic activity, cellular process, metabolic (cellular) process, binding and nitrogen compound metabolic process were the top six highly represented GO terms. Soft-core when compared with core, organic cyclic/heterocyclic compound binding, intrinsic/integral membrane component, small molecule binding and localization were observed to be highly represented GO terms. Shell genes when compared with core-genes showed nucleic acid binding, heterocyclic/organic cyclic compound binding, DNA binding, macromolecule (cellular) metabolic process as highly represented GO terms. Similarly, cloud genes against the core-genes showed the same highly represented GO terms. However, cloud vs. core was different from shell vs. cloud in terms of intracellular (membrane) bounded organelle, extracellular space and multicellular organismal process GO terms specific to cloud genome. Cytoskeleton organization and symbiotic process were specific to shell only (**Figure 8**).

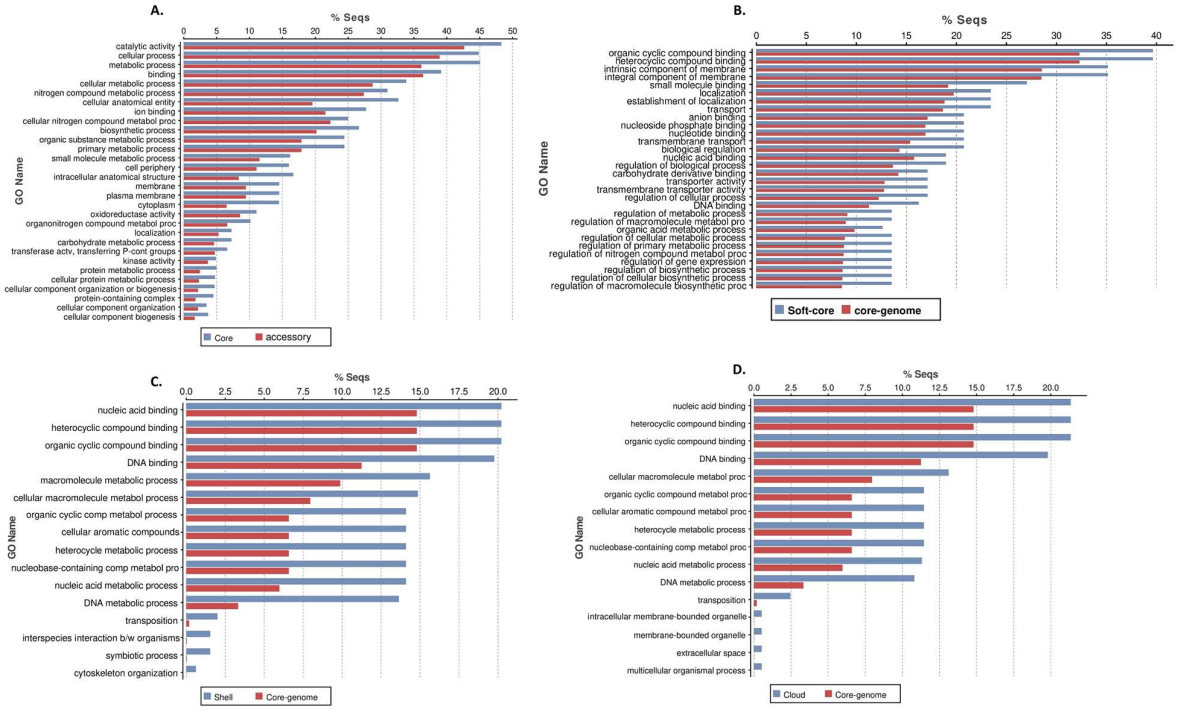

Figure 8. Statistical assessment of annotation differences in *Pantoea stewartii* subs. *indologenes* pan-genome. (A) core vs accessory. (B) soft-core vs core. (C) shell vs core. (D) cloud vs core genes.

 doi:10.20944/preprints202107.0400.v1

4. Discussion

Pantoea stewartii subsp. *indologenes* causes a leafspot of foxtail millet and pearl millet, a rot of pineapple and one strain has also been isolated from cluster bean (*Cyamopsis tetragonolobus*) [6]. Recently, *P. stewartii* subsp. *indologenes* strains (*Psi* pv. *cepacicola*) were identified that caused symptoms similar to center rot of onion [2]. These bacterial species/pathovar are not as prevalent as *P. ananatis* in Georgia or elsewhere but are known to cause center rot disease in onions. Genome analysis will provide insights about the genes involved in pathogenicity and probable virulence mechanism(s). Moreover, it is important to know how different or similar the mechanism(s) and pathogenicity and virulence genes are involved in causing center rot in onion by different *Pantoea spp*. We therefore conducted a pan-genome study of seventeen newly identified *Psi* strains from both pathovars [pv. *cepacicola* (*n*= 4) and pv. *setariae* (*n*=13)] and developed a core and a pan-genome followed by annotation of the core and accessory genes. Further, we carried out pan-GWAS study and identified gene(s) associated with pathogenicity in onion. A phylogenetic and HGT studies were also conducted to understand the role of SNPs, PAVs and HGTs on phylogeny of these strains.

4.1. Pantoea stewartii subsp. indologenes pan-genome and horizontal gene transfer

In the current study we identified 3,546 core genes, 5 to 382 cloud genes and 394 to 703 shell genes. A similar study using 81 *P. ananatis* strains identified 3,153 core genes and cloud genes ranging from 1000 to 6,808 [8]. We observed a stark difference in the number of cloud genes when compared to our earlier study on *P. ananatis*. Slightly higher number of core genes and lesser number of cloud genes identified in the current study are expected to change in future with the increase in number of *Psi* genomes being added to the pan-genome. Much lesser number of accessory genes were observed as compared to our previous study. Lesser number of accessory genes in this study can be a result of HGT because gene exchanges due to HGT can lead to extensive gene repertoire differences among closely related species or within species [28]. This may partly explain as to why a smaller number of cloud genes were identified in three *Psi* pv. *cepacicola* (PNA 14-11, PNA 14-12 and PNA 14-9) as compared to rest of the *Psi* strains used in this study. Perhaps as more *Psi* pv. *cepacicola* genomes get included, number of core genes may decrease and accessory genes may increase as indicated in other bacterial pan-genomes (*Escherichia coli*) [29].

A dynamic pan-genome is also dependent on the frequent HGT events it encounters during evolution. We therefore studied HGT within *Psi* strains and among the four species of *Pantoea* complex (*P. ananatis*, *P. agglomerans*, *P. stewartii* subsp. *stewartii* and *P. allii*). The maximum number of HGT events occurred from PNA 03-3 (*Psi* pv. *cepacicola*) to PANS 07-4 (*Psi* pv. *setariae*), which may explain the phylogenetic clustering of PNA 03-3 with PANS 07-6 and PANS 07-4 using shell and cloud genes. As compared to three *Psi* pv. *cepacicola* strains namely PNA 14-11, PNA 14-12 and 14-9 that contained six, five and eighteen cloud genes, respectively, *Psi* pv. *cepacicola* strain PNA 03-3 contained 259 cloud genes, which is similar to the number of cloud genes exhibited by *Psi* pv. *setariae* strains (range: 107 to 382). It seems that HGT has deeply impacted the gene loss in the above mentioned three *Psi* pv. *cepacicola* strains.

4.2. An open pan-genome of Pantoea stewartii subsp. indologenes

Pan-genome approach is important for exploring the genomic repertoires of a phylogenetic lineage of microbes [30]. Pan-genome of *Psi* showed linear progression with >6,000 genes, with ~43 new genes adding on average to the

 doi:10.20944/preprints202107.0400.v1

pan-genome with each new *Psi* genome sequenced suggesting that it was open, attributed to frequent evolutionary changes mediated by gene(s) gain and loss resulted due to HGT events. The open feature of *Psi* is consistent with *P. ananatis* and *Geobacillus* spp. dataset [8, 17, 31] as opposed to other species, such as *Bifidobacterium breve* and *Staphylococcus lugdunensis*, which depicted a closed trend [32, 33].

4.3. Phylogenetic study of Pantoea stewartii subsp. indologenes and other Pantoea spp. complex

A comparative phylogenetic approach was undertaken to evaluate the phylogeny of both *Psi* pathovars compared to other *Pantoea* spp. We used PAVs, SNPs and wgMLST approaches to understand the differences in phylogeny of *Psi* strains along with four *Pantoea* spp. based on the underlying genomic variants. Core genome with conserved genes instead of accessory genome with PAVs may convey a true measure of phylogeny. We therefore inferred our phylogenetic analysis based on core genome analysis. We found ANI based on core genes clustered *Psi* pv. *cepacicola* strains close. However, soft-core genes ANI clustered one *Psi* pv. *cepacicola* strain (PNA 03-3) distantly from the other three strains. Similar, observations were made when shell and cloud genes were used to conduct the phylogenetic analysis. This suggests that onion pathogenic *Psi* pv. *cepacicola* strains and onion non-pathogenic *Psi* pv. *setariae* strains could be distinguished using core genes, which was not the case with accessory genes. Clustering based on the core and accessory genome ANI showed a slight difference indicating the evolutionary and pathogenicity relationship was better depicted with core genes than the accessory genes. However, number of input genomes may have a role to play and probably core genes-based phylogeny could change if the number of strains/genomes are further added. It could also be possible that core genome is impacted by HGT or homologous recombination and as a result phylogenetic relationship based on core genes is obscured or distorted [34].

SNPs are vertically inherited and are one of the dominant forms of evolutionary change that have become an indispensable tool for phylogenetic analyses [35-37]. Hence, core genes were used to classify onion pathogenic *Psi* pv. *cepacicola* vs. onion non-pathogenic *Psi* pv. *setariae* strains. This was done by identifying SNPs from the core genome and performing a phylogenetic analysis. Core SNPs clustered *Psi* pv. *cepacicola* strains together except for a strain, PNA 03-3. Our earlier study showed that *Psi* pv. *cepacicola* PNA 03-3 was highly aggressive on leek, foxtail millet and pearl millet whereas moderately aggressive on onion, chive, Japanese bunching onion and oat. Other *Psi* pv. *cepacicola* strains (PNA 14-9, PNA 14-11, PNA 14-12) were moderately-to-highly aggressive on onion, foxtail millet and pearl millet but were moderate-to-less aggressive on leek, chive, Japanese bunching onion and oat [7]. We believe that this difference in aggressiveness is truly represented by core SNPs. Also, *Psi* pv. *cepacicola* PNA 03-3 is more closely related to *P. ananatis* (LMG 2665^{T}), which is reflected in the core SNP-based phylogeny. The WgMLST, an extended concept of MLST is complementary to PAV based phylogenetic analysis [8]. As expected, we observed similar pattern of phylogenetic classification of strains both with PAVs- and wgMLST-based phylogeny. Particularly, *Psi* pv. *cepacicola* PNA 03-3 was clustered closely with *Psi* pv. *setariae* strains (NCPPB 2275 and NCPPB 1877) in both PAV and wgMLST based phylogenetic analysis.

4.4. Gene cluster identified from Pan-GWAS analysis

The PAVs identified using *Psi* strains (n=17) and the phenotyping data when subjected to pan-GWAS identified a gene cluster associated with pathogenicity in *Allium* species. The phenotyping data were utilized from our

 doi:10.20944/preprints202107.0400.v1

earlier study [7]. Pan-GWAS identified several genes that were associated with the pathogenicity on onion seedlings. Out of all the genes identified (*n*=154), a cluster of 11 well-annotated genes was identified and found to be strongly associated with pathogenicity in onion seedlings. Ten genes out of the 11 associated genes were found to be present in three out of the four pathogenic *Psi* pv. *cepacicola* strains (PNA 14-9, PNA 14-11 and PNA 14-12). However, there was one gene annotated as phosphoenolpuruvate mutase (pepM) that was present in all four pathogenic strains. The first gene in the cluster is pyridoxal 5'-phosphate synthase (pdxH_2), which catalyzes pentose and triose isomerizations, imine formation, amine addition, and ring formation, all in a single enzymatic system [38, 39]. It is involved in ammonia transport [40]. Second gene of the cluster codes for AMP binding protein. AMP binding proteins in bacteria are regarded as a global activator proteins that are required to regulate the gene transcription [41]. Third gene (ydeE) in the cluster is a efflux MFS transporter known to export peptides [42]. Pgk-tpi protein coding gene is next in cluster that is involved in the sub-pathway, which synthesizes D-glyceraldehyde 3-phosphate from glycerone phosphate (a part of the pathway glycolysis which is itself part of carbohydrate degradation). Pgk-tpi codes for enzymes phosphoglycerate kinase and triosephosphate isomerase that form a covalent bifunctional enzyme complex [43]. Next gene in the gene cluster codes for FAD/NAD(P)-binding protein. The NAD(P)-binding enzymes are involved in catalyzing redox or non-redox reactions [44]. Interestingly, the next gene in cluster (pepM) codes for phosphoenolpuruvate mutase, which was the first gene of HiVir cluster identified in *P. ananatis* (Asseline et al., 2018). It was identified as the first pathogenicity factor associated with the fitness of *P. ananatis* as well as with symptom development in infected onion leaves and bulbs [45]. PepM is involved in phosphonate biosynthesis. Organophosphonates are synthesized as secondary metabolites in certain prokaryotes to function as antibiotics, and can have specialized roles in pathogenesis or signaling [46]. Next gene in the cluster codes for NAD(P)-binding domain-containing protein, which catalyzes the NADPH-dependent reactions. For example, hydropxypyruvate reductase carries out reduction of glyoxylate and hydroxypyruvate into glycolate and glycerate, respectively [47, 48]. N-acetyl-gamma-glutamyl-phosphate reductase is coded by argC1 gene in the cluster. It catalyzes the NADPH-dependent reduction of N-acetyl-5-glutamyl phosphate to yield N-acetyl-L-glutamate 5-semialdehyde. This enzyme is involved in step 3 of the sub-pathway that synthesizes N(2)-acetyl-L-ornithine from L-glutamate. This sub-pathway itself is part of the L-arginine biosynthesis pathway [49]. Another gene in the cluster was annotated as alcohol dehydrogenase catalytic domain-containing protein. Alcohol dehydrogenases are the oxidoreductases that catalyse the reversible oxidation of alcohols to aldehydes or ketones, with the concomitant reduction of NAD^+ or $NADP^+$ [50]. Second last gene in the cluster codes for another alcohol dehydrogenase i.e. iron containing alcohol dehydrogenase (FeADH). FeADH family have been characterized exhibiting different catalytic activities. ADH are capable of catalyzing a wide variety of substrates (e.g. normal and branched-chain aliphatic and aromatic alcohols, both primary and secondary alcohols, corresponding aldehydes and ketones, polyols) and they are involved in an astonishingly wide range of metabolic processes; they are for instance involved in alcohol, alkane, sugar and lipid metabolism, and cell defense towards exogenous alcohols and aldehydes [51-56]. The last gene in the cluster codes for Lyse-family translocator. The physiological function of the exporter is to excrete excess L-Lysine and L-arginine as a result of natural flux imbalances or peptide hydrolysis. It also plays important roles in ionic homeostasis, cell envelope assembly, and protection from excessive cytoplasmic heavy metal/metabolite concentrations. [57, 58]. Overall, we found only two out of the eleven genes identified in this cluster

that were common with the HiVir cluster identified in *P. ananatis* [59]. These two common genes code for phosphoenolpyruvate phospomutase (PepM) enzyme and MFS transporter protein. These findings suggest a potential alternate set of onion pathogenicity-related genes in *Psi*, which is distinct from the known onion pathogenicity (HiVir) and virulence (allicin tolerance; *alt*) factors identified in *P. ananatis* [45, 59, 60]. We are currently trying to understand the role of this gene cluster in *Psi* pv. *cepacicola* on onion pathogenicity using traditional gene mutation studies.

5. Conclusion

Pan-GWAS approach predicted the genes associated with onion-pathogenicity in *Psi* strains particularly in *Psi* pv. *cepacicola*. We found a cluster of genes different from HiVir/PASVIL (identified in *P. ananatis*) cluster linked to onion pathogenicity in *Psi* pv. *cepacicola*. We conclude that there might be several pathogenicity factors involved in onion pathogenicity and to some extent these might be specific to some *Pantoea* spp. We also observed a large repertoire of accessory genes in *Psi* strains, which is suggestive of a potential for a broad and diverse niche-adaptation and host-range expansion capabilities. We observed HGT events as major contributing factor for PAVs resulting in diversification of *Psi* and other *Pantoea* species. In future, it would be interesting to assess if aggressiveness of *Psi* strains on Allium and Poacea species can be predicted using GWAS utilizing SNPs rather than PAVs. We expect that SNP based GWAS study will not only corroborate our current findings but may potentially lead to the development of SNP-based PCR markers that can distinguish the *Psi* pv. *cepacicola* and *Psi* pv. *setariae* strains.

Funding: This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia Office of the Vice President for Research and Office of the Vice President for Information Technology. This work is partially supported by the Specialty Crop Block Grant AWD00009682.

Authors' contribution: GA and BD conceived the project, GA performed the bioinformatics analyses, and complied the manuscript. BD designed and finalized the manuscript. BD planned the project, secured extramural funds, and revised and submitted manuscript.

Competing interest: The authors declare that they have no competing interests.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

References

1. Gitaitis, R.; Walcott, R.; Culpepper, S.; Sanders, H.; Zolobowska, L.; Langston, D., Recovery of Pantoea ananatis, causal agent of center rot of onion, from weeds and crops in Georgia, USA. *Crop Protection* **2002,** 21, (10), 983-989.
2. Stumpf, S.; Kvitko, B.; Gitaitis, R.; Dutta, B., Isolation and characterization of novel Pantoea stewartii subsp. indologenes strains exhibiting center rot in onion. *Plant disease* **2018,** 102, (4), 727-733.
3. Edens, D.; Gitaitis, R.; Sanders, F.; Nischwitz, C., First report of Pantoea agglomerans causing a leaf blight and bulb rot of onions in Georgia. *Plant disease* **2006,** 90, (12), 1551-1551.
4. Brady, C.; Cleenwerck, I.; Venter, S.; Vancanneyt, M.; Swings, J.; Coutinho, T., Phylogeny and identification of Pantoea species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). *Systematic and Applied Microbiology* **2008,** 31, (6-8), 447-460.
5. Kini, K.; Dossa, R.; Dossou, B.; Mariko, M.; Koebnik, R.; Silué, D., A semi-selective medium to isolate and identify bacteria of the genus Pantoea. *Journal of General Plant Pathology* **2019,** 85, (6), 424-427.
6. Mergaert, J.; Verdonck, L.; Kersters, K., Transfer of Erwinia ananas and Erwinia stewartii to the genus Pantoea and description of Pantoea stewartii ssp. indologenes. *Int J Sys Bacteriol* **1993,** 43, 162-173.
7. Koirala, S.; Zhao, M.; Agarwal, G.; Stice, S.; Gitaitis, R.; Kvitko, B.; Dutta, B., Identification of two novel pathovars of Pantoea stewartii subsp. indologenes affecting Allium sp. and millets. *Phytopathology* **2021**, (ja).
8. Agarwal, G.; Choudhary, D.; Stice, S.; Myers, B.; Gitaitis, R.; Venter, S.; Kvitko, B.; Dutta, B., Pan-genome-wide analysis of Pantoea ananatis identified genes linked to pathogenicity in onion. *bioRxiv* **2020**.

9. Agarwal, G.; Kavalappara, S. R.; Gautam, S.; Silva, A. d.; Simmons, A.; Srinivasan, R.; Dutta, B., Field Screen and Genotyping of Phaseolus vulgaris against Two Begomoviruses in Georgia, USA. *Insects* **2021,** 12, (1), 49.
10. Agarwal, G.; Clevenger, J.; Kale, S. M.; Wang, H.; Pandey, M. K.; Choudhary, D.; Yuan, M.; Wang, X.; Culbreath, A. K.; Holbrook, C. C., A recombination bin-map identified a major QTL for resistance to Tomato Spotted Wilt Virus in peanut (Arachis hypogaea). *Scientific reports* **2019,** 9, (1), 1-13.
11. Agarwal, G.; Clevenger, J.; Pandey, M. K.; Wang, H.; Shasidhar, Y.; Chu, Y.; Fountain, J. C.; Choudhary, D.; Culbreath, A. K.; Liu, X., High-density genetic map using whole-genome resequencing for fine mapping and candidate gene discovery for disease resistance in peanut. *Plant biotechnology journal* **2018,** 16, (11), 1954-1967.
12. Divya, C.; Gaurav, A.; Hui, W.; Pandey, M. K.; Culbreath, A. K.; Varshney, R. K.; Guo, B., Molecular markers and genomic resources for disease resistance in peanut-A review. *Legume Research-An International Journal* **2019,** 42, (2), 137-144.
13. Clevenger, J.; Chu, Y.; Chavarro, C.; Agarwal, G.; Bertioli, D. J.; Leal-Bertioli, S. C.; Pandey, M. K.; Vaughn, J.; Abernathy, B.; Barkley, N. A., Genome-wide SNP genotyping resolves signatures of selection and tetrasomic recombination in peanut. *Molecular plant* **2017,** 10, (2), 309-322.
14. Pandey, M. K.; Agarwal, G.; Kale, S. M.; Clevenger, J.; Nayak, S. N.; Sriswathi, M.; Chitikineni, A.; Chavarro, C.; Chen, X.; Upadhyaya, H. D., Development and evaluation of a high density genotyping 'Axiom_ Arachis' array with 58 K SNPs for accelerating genetics and breeding in groundnut. *Scientific Reports* **2017,** 7, (1), 1-10.
15. Khan, A. W.; Garg, V.; Roorkiwal, M.; Golicz, A. A.; Edwards, D.; Varshney, R. K., Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends in plant science* **2020,** 25, (2), 148-158.
16. Tettelin, H.; Masignani, V.; Cieslewicz, M. J.; Donati, C.; Medini, D.; Ward, N. L.; Angiuoli, S. V.; Crabtree, J.; Jones, A. L.; Durkin, A. S., Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* **2005,** 102, (39), 13950-13955.
17. Wang, M.; Zhu, H.; Kong, Z.; Li, T.; Ma, L.; Liu, D.; Shen, Q., Pan-Genome Analyses of Geobacillus spp. Reveal Genetic Characteristics and Composting Potential. *International journal of molecular sciences* **2020,** 21, (9), 3393.
18. Bosi, E.; Monk, J. M.; Aziz, R. K.; Fondi, M.; Nizet, V.; Palsson, B. Ø., Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proceedings of the National Academy of Sciences* **2016,** 113, (26), E3801-E3809.
19. Nowell, R. W.; Green, S.; Laue, B. E.; Sharp, P. M., The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome biology and evolution* **2014,** 6, (6), 1514-1529.
20. Contreras-Moreira, B.; Vinuesa, P., GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and environmental microbiology* **2013,** 79, (24), 7696-7701.
21. Chaumeil, P.-A.; Mussig, A. J.; Hugenholtz, P.; Parks, D. H., GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. In Oxford University Press: 2020.
22. Song, W.; Wemheuer, B.; Zhang, S.; Steensen, K.; Thomas, T., MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* **2019,** 7, (1), 1-14.
23. Laing, C.; Buchanan, C.; Taboada, E. N.; Zhang, Y.; Kropinski, A.; Villegas, A.; Thomas, J. E.; Gannon, V. P., Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC bioinformatics* **2010,** 11, (1), 1-14.
24. Liu, Y.-Y.; Chen, C.-C.; Chiou, C.-S., Construction of a pan-genome allele database of Salmonella enterica serovar enteritidis for molecular subtyping and disease cluster identification. *Frontiers in microbiology* **2016,** 7, 2010.
25. Al-Shahrour, F.; Díaz-Uriarte, R.; Dopazo, J., FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **2004,** 20, (4), 578-580.
26. Götz, S.; García-Gómez, J. M.; Terol, J.; Williams, T. D.; Nagaraj, S. H.; Nueda, M. J.; Robles, M.; Talón, M.; Dopazo, J.; Conesa, A., High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* **2008,** 36, (10), 3420-3435.
27. Willenbrock, H.; Hallin, P. F.; Wassenaar, T. M.; Ussery, D. W., Characterization of probiotic Escherichia coliisolates with a novel pan-genome microarray. *Genome biology* **2007,** 8, (12), R267.
28. Shapiro, B. J.; Friedman, J.; Cordero, O. X.; Preheim, S. P.; Timberlake, S. C.; Szabó, G.; Polz, M. F.; Alm, E. J., Population genomics of early events in the ecological differentiation of bacteria. *science* **2012,** 336, (6077), 48-51.
29. Lukjancenko, O.; Wassenaar, T. M.; Ussery, D. W., Comparison of 61 sequenced Escherichia coli genomes. *Microbial ecology* **2010,** 60, (4), 708-720.
30. Shin, J.; Song, Y.; Jeong, Y.; Cho, B.-K., Analysis of the core genome and pan-genome of autotrophic acetogenic bacteria. *Frontiers in microbiology* **2016,** 7, 1531.
31. Bezuidt, O. K.; Pierneef, R.; Gomri, A. M.; Adesioye, F.; Makhalanyane, T. P.; Kharroub, K.; Cowan, D. A., The Geobacillus pan-genome: implications for the evolution of the genus. *Frontiers in microbiology* **2016,** 7, 723.
32. Argemi, X.; Matelska, D.; Ginalski, K.; Riegel, P.; Hansmann, Y.; Bloom, J.; Pestel-Caron, M.; Dahyot, S.; Lebeurre, J.; Prévost, G., Comparative genomic analysis of Staphylococcus lugdunensis shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC genomics* **2018,** 19, (1), 1-16.
33. Bottacini, F.; Motherway, M. O. C.; Kuczynski, J.; O'Connell, K. J.; Serafini, F.; Duranti, S.; Milani, C.; Turroni, F.; Lugli, G. A.; Zomer, A., Comparative genomics of the Bifidobacterium breve taxon. *BMC genomics* **2014,** 15, (1), 1-19.

34. Straub, C.; Colombi, E.; McCann, H. C., Population genomics of bacterial plant pathogens. *Phytopathology®* **2021,** 111, (1), 23-31.
35. McNally, K. L.; Childs, K. L.; Bohnert, R.; Davidson, R. M.; Zhao, K.; Ulat, V. J.; Zeller, G.; Clark, R. M.; Hoen, D. R.; Bureau, T. E., Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences* **2009,** 106, (30), 12273-12278.
36. Faison, W. J.; Rostovtsev, A.; Castro-Nallar, E.; Crandall, K. A.; Chumakov, K.; Simonyan, V.; Mazumder, R., Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes. *Genomics* **2014,** 104, (1), 1-7.
37. Shakya, M.; Ahmed, S. A.; Davenport, K. W.; Flynn, M. C.; Lo, C.-C.; Chain, P. S., Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Scientific reports* **2020,** 10, (1), 1-15.
38. Burns, K. E.; Xiang, Y.; Kinsland, C. L.; McLafferty, F. W.; Begley, T. P., Reconstitution and biochemical characterization of a new pyridoxal-5 '-phosphate biosynthetic pathway. *Journal of the American Chemical Society* **2005,** 127, (11), 3682-3683.
39. Raschle, T.; Amrhein, N.; Fitzpatrick, T. B., On the two components of pyridoxal 5′-phosphate synthase from Bacillus subtilis. *Journal of Biological Chemistry* **2005,** 280, (37), 32291-32300.
40. Strohmeier, M.; Raschle, T.; Mazurkiewicz, J.; Rippe, K.; Sinning, I.; Fitzpatrick, T. B.; Tews, I., Structure of a bacterial pyridoxal 5′-phosphate synthase complex. *Proceedings of the National Academy of Sciences* **2006,** 103, (51), 19284-19289.
41. Botsford, J. L.; Harman, J. G., Cyclic AMP in prokaryotes. *Microbiology and Molecular Biology Reviews* **1992,** 56, (1), 100-122.
42. Hayashi, M.; Tabata, K.; Yagasaki, M.; Yonetani, Y., Effect of multidrug-efflux transporter genes on dipeptide resistance and overproduction in Escherichia coli. *FEMS microbiology letters* **2010,** 304, (1), 12-19.
43. Schurig, H.; Beaucamp, N.; Ostendorp, R.; Jaenicke, R.; Adler, E.; Knowles, J. R., Phosphoglycerate kinase and triosephosphate isomerase from the hyperthermophilic bacterium Thermotoga maritima form a covalent bifunctional enzyme complex. *The EMBO journal* **1995,** 14, (3), 442-451.
44. Hua, Y. H.; Wu, C. Y.; Sargsyan, K.; Lim, C., Sequence-motif detection of NAD (P)-binding proteins: discovery of a unique antibacterial drug target. *Scientific reports* **2014,** 4, (1), 1-7.
45. Asselin, J. A. E.; Bonasera, J. M.; Beer, S. V., Center rot of onion (Allium cepa) caused by Pantoea ananatis requires pepM, a predicted phosphonate-related gene. *Molecular Plant-Microbe Interactions* **2018,** 31, (12), 1291-1300.
46. Hilderbrand, R. L., *Role of phosphonates in living systems*. CRC Press: 1983.
47. NUÑEZ, M. F.; PELLICER, M. T.; BADIA, J.; AGUILAR, J.; BALDOMA, L., Biochemical characterization of the 2-ketoacid reductases encoded by ycdW and yiaE genes in Escherichia coli. *Biochemical Journal* **2001,** 354, (3), 707-715.
48. But, S.; Egorova, S.; Khmelenina, V.; Trotsenko, Y., Biochemical properties and phylogeny of hydroxypyruvate reductases from methanotrophic bacteria with different c 1-assimilation pathways. *Biochemistry (Moscow)* **2017,** 82, (11), 1295-1303.
49. Baich, A.; Vogel, H. J., N-acetyl-γ-glutamokinase and N-acetylglutamic γ-semialdehyde dehydrogenase: Repressible enzymes of arginine synthesis in Escherichiacoli. *Biochemical and biophysical research communications* **1962,** 7, (6), 491-496.
50. De Smidt, O.; Du Preez, J. C.; Albertyn, J., The alcohol dehydrogenases of Saccharomyces cerevisiae: a comprehensive review. *FEMS yeast research* **2008,** 8, (7), 967-978.
51. Park, D.-H.; Plapp, B., Isoenzymes of horse liver alcohol dehydrogenase active on ethanol and steroids. cDNA cloning, expression, and comparison of active sites. *Journal of Biological Chemistry* **1991,** 266, (20), 13296-13302.
52. Ma, K.; Loessner, H.; Heider, J.; Johnson, M. K.; Adams, M., Effects of elemental sulfur on the metabolism of the deep-sea hyperthermophilic archaeon Thermococcus strain ES-1: characterization of a sulfur-regulated, non-heme iron alcohol dehydrogenase. *Journal of bacteriology* **1995,** 177, (16), 4748-4756.
53. Tani, A.; Sakai, Y.; Ishige, T.; Kato, N., Thermostable NADP+-Dependent Medium-Chain Alcohol Dehydrogenase from Acinetobacter sp. Strain M-1: Purification and Characterization and Gene Expression inEscherichia coli. *Applied and environmental microbiology* **2000,** 66, (12), 5231-5235.
54. Burdette, D.; Jung, S.-H.; Shen, G.-J.; Hollingsworth, R.; Zeikus, J., Physiological function of alcohol dehydrogenases and long-chain (C30) fatty acids in alcohol tolerance of Thermoanaerobacter ethanolicus. *Applied and environmental microbiology* **2002,** 68, (4), 1914-1918.
55. Vangnai, A. S.; Arp, D. J.; Sayavedra-Soto, L. A., Two distinct alcohol dehydrogenases participate in butane metabolism by Pseudomonas butanovora. *Journal of bacteriology* **2002,** 184, (7), 1916-1924.
56. Yoon, S.-Y.; Noh, H.-S.; Kim, E.-H.; Kong, K.-H., The highly stable alcohol dehydrogenase of Thermomicrobium roseum: purification and molecular characterization. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **2002,** 132, (2), 415-422.
57. Vrljic, M.; Sahm, H.; Eggeling, L., A new type of transporter with a new type of cellular function: l-lysine export from Corynebacterium glutamicum. *Molecular microbiology* **1996,** 22, (5), 815-826.
58. Tsu, B. V.; Saier Jr, M. H., The LysE superfamily of transport proteins involved in cell physiology and pathogenesis. *PloS one* **2015,** 10, (10), e0137184.

 doi:10.20944/preprints202107.0400.v1

59. Polidore, A. L.; Furiassi, L.; Hergenrother, P. J.; Metcalf, W. W., A Phosphonate Natural Product Made by Pantoea ananatis is Necessary and Sufficient for the Hallmark Lesions of Onion Center Rot. *Mbio* **2021,** 12, (1).
60. Stice, S. P.; Thao, K. K.; Khang, C. H.; Baltrus, D. A.; Dutta, B.; Kvitko, B. H., Pantoea ananatis defeats Allium chemical defenses with a plasmid-borne virulence gene cluster. *BioRxiv* **2**