





RESEARCH ARTICLE

REVISED A pan-genome method to determine core regions of the *Bacillus subtilis* and *Escherichia coli* genomes [version 2; peer review: 2 approved]

Granger Sutton ¹, Gary B. Fogel², Bradley Abramson¹, Lauren Brinkac³, Todd Michael ¹, Enoch S. Liu², Sterling Thomas³

¹J. Craig Venter Institute, Rockville, Maryland, 20850, USA

²Natural Selection, Inc., San Diego, CA, 92121, USA

³Noblis, Inc., Reston, VA, 20191, USA

V2 First published: 13 Apr 2021, 10:286
<https://doi.org/10.12688/f1000research.51873.1>
 Latest published: 02 Sep 2021, 10:286
<https://doi.org/10.12688/f1000research.51873.2>

Abstract

Background: Synthetic engineering of bacteria to produce industrial products is a burgeoning field of research and application. In order to optimize genome design, designers need to understand which genes are essential, which are optimal for growth, and locations in the genome that will be tolerated by the organism when inserting engineered cassettes.

Methods: We present a pan-genome based method for the identification of core regions in a genome that are strongly conserved at the species level.

Results: We show that the core regions determined by our method contain all or almost all essential genes. This demonstrates the accuracy of our method as essential genes should be core genes. We show that we outperform previous methods by this measure. We also explain why there are exceptions to this rule for our method.

Conclusions: We assert that synthetic engineers should avoid deleting or inserting into these core regions unless they understand and are manipulating the function of the genes in that region. Similarly, if the designer wishes to streamline the genome, non-core regions and in particular low penetrance genes would be good targets for deletion. Care should be taken to remove entire cassettes with similar penetrance of the genes within cassettes as they may harbor toxin/antitoxin genes which need to be removed in tandem. The bioinformatic approach introduced here saves considerable time and effort relative to knockout studies on single isolates of a given species and captures a broad understanding of the conservation of genes that are core to a species.

Keywords

pan-genome, pan-genome graph, core genes, essential genes

Open Peer Review

Reviewer Status  

Invited Reviewers

1 2

version 2

(revision)
02 Sep 2021

version 1

13 Apr 2021




report



report

1. **Kaleb Abram**, University of Arkansas for Medical Sciences, Little Rock, USA

David Ussery , University of Arkansas for Medical Sciences, Little Rock, USA

2. **Christos Ouzounis** , Aristotle University of Thessalonica, Thessalonica, Greece
Centre for Research & Technology Hellas, Thessalonica, Greece

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Granger Sutton (GSutton@jvvi.org)

Author roles: **Sutton G:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Fogel GB:** Data Curation, Formal Analysis, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Abramson B:** Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Brinkac L:** Data Curation, Formal Analysis, Investigation, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Michael T:** Conceptualization, Funding Acquisition, Project Administration, Writing – Review & Editing; **Liu ES:** Data Curation, Investigation, Validation, Writing – Review & Editing; **Thomas S:** Conceptualization, Data Curation, Funding Acquisition, Methodology, Project Administration, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research is based upon work supported [in part] by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under Finding Engineering Linked Indicators (FELIX) program contract #N6600118C-4506. The principal investigator for the award is Sterling Thomas. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Sutton G *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Sutton G, Fogel GB, Abramson B *et al.* **A pan-genome method to determine core regions of the *Bacillus subtilis* and *Escherichia coli* genomes [version 2; peer review: 2 approved]** F1000Research 2021, 10:286 <https://doi.org/10.12688/f1000research.51873.2>

First published: 13 Apr 2021, 10:286 <https://doi.org/10.12688/f1000research.51873.1>

REVISED Amendments from Version 1

Version 2 attempts to address all comments from the reviewers for version 1. This includes a new introductory paragraph explaining the overall thrust of the article before diving into a detailed background. A new paragraph better explaining the advantages of a pan-genome graph (PGG) over a simple pan-genome. A new table showing how complete genomes from RefSeq were filtered to arrive at the final set. A new supplementary table showing the 34 genes in *MiniBacillus* but not present in all 108 *B. subtilis* genomes. And new concluding paragraphs for the results and discussion section.

Any further responses from the reviewers can be found at the end of the article

Introduction

The primary focus of this paper is a new pan-genome method to determine core regions of a genome shared by all or almost all strains of the same species or subspecies. We evaluate the performance of this approach relative to other methods using experimentally determined essential genes under the hypothesis that all or at least most essential genes should be core across a species. This hypothesis implies that methods for determining core regions/genes are likely to be more accurate if they identify more essential genes as core genes. The paper reveals the potential usefulness of pan-genome analysis for synthetic engineering and genome analysis more broadly through the analysis of core regions in *Bacillus subtilis* and *Escherichia coli*.

Over the last decade, considerable interest has been directed towards the determination of a minimal bacterial cell, making use of a short genome consisting of only essential genes for viability. The *Mycoplasma mycoides* JCVI-syn3.0 is a case example of synthetic engineering to design and build a genome that contains a streamlined gene set essential for cell viability and cell replication.¹ Multiple genome reduction projects have been undertaken.²⁻⁴ More targeted genomic deletions of genomic loci have been performed to characterize essential genes, but generally targeted approaches are too laborious to perform on a whole genome.^{5,6} However, the identification of “essential” genes - those genes that are critical for cell viability and replication - takes considerable time and effort in a laboratory setting and is usually determined with respect to one reference genome under one set of specific growth conditions. For instance, Kobayashi *et al.*⁷ and Koo *et al.*⁸ experimentally and computationally determined the minimal gene set in the Gram-positive bacterium *Bacillus subtilis*. Koo *et al.*⁸ used a strictly experimental approach but Kobayashi *et al.*⁷ used three forms of evidence for their essential genes as given in their Table 4: RB: previous experimental work in *B. subtilis*; RO: previous experimental work in other bacteria; and TW: their experimental work. The RO evidence used is a mix of experimental and computational as the determination of orthologs is computational and essentiality of those orthologs was not experimentally confirmed: “Through predictions we propose that 79 other genes are essential, whereas 106 are not (Table 3)”.⁷ Of the ~4,100 genes of the type strain, a total of 271 genes for Kobayashi *et al.*⁷ and 257 genes for Koo *et al.*⁸ were shown to be essential. These essential genes were further categorized in terms of cell metabolism and enzymatic capability. Additionally, for the ~4400 genes in the Gram-negative bacterium *Escherichia coli*, Goodall *et al.*,⁹ Baba *et al.*,¹⁰ and Yamazaki *et al.*,¹¹ it was determined that 414 genes were essential to strain K-12.

Reuß *et al.*³ completed extensive further experimental and computational work to determine a minimal *B. subtilis* genome they call *MiniBacillus*. They present a list of 523 protein coding and 119 RNA genes necessary for a minimal *B. subtilis* cell growing in complex medium at 37°C. While many of these genes are not essential under single deletion experimental conditions, they are required for survival because a cell needs certain essential functions which may be carried out independently by more than one gene. As noted by Reuß *et al.*,³ the choice of which functionally isologous genes to choose for a minimal cell depends upon minimization goals and gene choices for different functions are not independent of one another. One criterion used by Reuß *et al.*³ is the conservation of the gene: “More strongly conserved genes were preferred over less conserved genes. In this respect, gene conservation and essentiality in genome-reduced *Mycoplasma* and other mollicutes species and the inclusion of genes in the genome of *M. mycoides* JCVI-syn3.0 had a high priority”. Reuß *et al.*³ do not explicitly use gene conservation at the species/subspecies pan-genome level but this seems in spirit with their criteria.

Reuß *et al.*³ extended their computational prediction of *MiniBacillus* by building on previous work to generate *B. subtilis* strains with large genome reductions.² They started by constructing the delta 6 strain¹² (~8% genome reduction). This reduction removed: “two prophages (SPβ, PBSX), three prophage-like regions, and the largest operon of *B. subtilis* (*pks*).” The phage/prophage regions were identified in part by GC content and codon usage as a method to identify probable horizontally transferred regions. Pan-genome analysis was not used. Next, strain IIG-Bs20 was constructed from delta 6¹³ by removing “all nine prophages, seven antibiotic biosynthesis gene clusters and two sigma factors for sporulation” in part to have a strain that would “not produce spores, antibiotics or bacteriocins”. A direct descendant of IIG-Bs20, strain IIG-Bs27-47-24 (~31% genome reduction), was then used to generate more reductions in two

independent strains, PG10 (~35% genome reduction) and PS38 (~36% genome reduction)² with the goal of removing genes “not necessary for the survival of the cell (*e.g.*, sporulation, antibiotic production, motility, metabolism of secondary carbon sources, and genes of unknown functions)”. For the strains IIG-Bs27-47-24, PG10, and PS-38, pan-genome analysis was not explicitly used but one of several criteria for deleting genes was a lack of conservation across broader taxonomic groups.

For *E. coli*, Kolisnychenko *et al.*⁵ generated an initial reduced genome in order to “serve both as a better model organism and as a more useful technological tool for genome science” by “deleting the largest K-islands of *E. coli*, identified by comparative genomics as recent horizontal acquisitions”. K-islands are regions unique to the K-12 strain MG1655 compared with the O157:H7 strain Sakai, and the uroseptic *E. coli* strain CFT073. This comparative analysis with a limited set of genomes is an obvious precursor to pan-genome analysis with a much larger set of genomes. Umenhoffer *et al.*¹⁴ generated the reduced *E. coli* strain MDS42 to be “free of mutation-generating IS elements”. This approach does not rely on comparison to other strains, just the ability to identify IS elements. Csorgo *et al.*¹⁵ further reduced the MDS42 strain by “constructing low-mutation-rate variants ... to lack most genes irrelevant for laboratory/industrial applications.” They targeted genes likely to be core and necessary for the species to adapt to the environment but detrimental in an industrial setting where strain stability is important.

Experimental studies to determine such essential genes are time consuming and often restricted to a single environmental condition using a single strain of the species. In addition, these approaches also knock out one gene at a time. As such, genes with multiple copies with redundant functions are often not considered as essential following knockout, as their additional copy is able to maintain cellular function. In other words, a viable organism would not result from deleting all but the experimentally determined essential genes from the genome. Another peculiarity of the single knockout essential genes is that pairs or cassettes of genes which can be removed and still have a viable organism are labeled essential because removal of just one gene is lethal. For example, removing the methylation gene(s) without removing the restriction digestion enzyme genes from the restriction mechanism results in cell death but the cell survives if the entire system is removed. This is likewise true for toxin/antitoxin systems.

While it is possible to define “essential” genes relative to viability, another larger question remains; which genes define a species? While specific phenotypes can vary across strains, in general a species seems to require some minimal set of genes to not only survive in the laboratory but to thrive in its natural environment. In contrast, some strains may have retained or acquired some genes which improve survival for specific niches. Comparing the genes from multiple diverse strains of a species can help answer these questions. We define the pan-genome for a species/subspecies to be the set of predicted orthologous gene clusters (OGC) across that set of strains. Others have allowed paralogs to be included in these gene clusters¹⁶⁻¹⁸ but here we do not. This constraint forces there to be at most one gene per genome in an OGC.

We further define a pan-genome graph (PGG) to be a graph with the pan-genome OGCs as nodes where an edge exists between two nodes if the respective genes for any genome from the two OGCs are adjacent in that genome. More precisely, an OGC node is represented as a dipole with 5' and 3' ends and the edges go between an end of one node (5' or 3') to an end of another node depending on the orientation of the genes which are adjacent. The edges primarily represent the order and orientation of OGCs in the pan-genome genomes. Secondly, the edges also represent the interstitial DNA sequences between the genes. A PGG edge has a weight equal to the number of genomes which contain the indicated adjacent gene ends. Core OGCs are defined to be those OGCs present in some large percentage of the strains in the pan-genome ($\geq 95\%$ in this work). Core edges are defined similarly. Core regions are defined to be the coordinates in a genome for each set of adjacent core genes in that genome provided the edges between the core genes are also core edges.

The PGG is important as it captures the structure of the pan-genome in ways that simply treating the pan-genome as a set of OGCs cannot. The inherent gene context in the PGG allows for more accurate annotation of a novel genome than OGCs alone which struggle to differentiate recent paralogs/repeats. The PGG allows core regions to be defined for any genome rather than just core OGCs/genes. The PGG indicates which OGCs occur in cassettes with implications for function, evolution, and synthetic engineering. As discussed later, the PGG allows for determination of probable orthologs not captured in the OGCs.

Core OGCs should determine the baseline phenotype (capabilities and traits) of a species. Previous pan-genome studies¹⁹ have shown that species tend to only tolerate the placement of noncore genes between core regions and not within those core regions. The reason an organism might constrain a core region rather than just core OGCs is that the region may include regulatory mechanisms such as operons, which allows for co-expression of multiple functionally associated genes, or regulons which would be disrupted with the insertion of other genes. We believe that conservation of core regions in species indicates resistance to insertion or deletion of genes in these regions through evolution or through human-mediated genetic engineering.

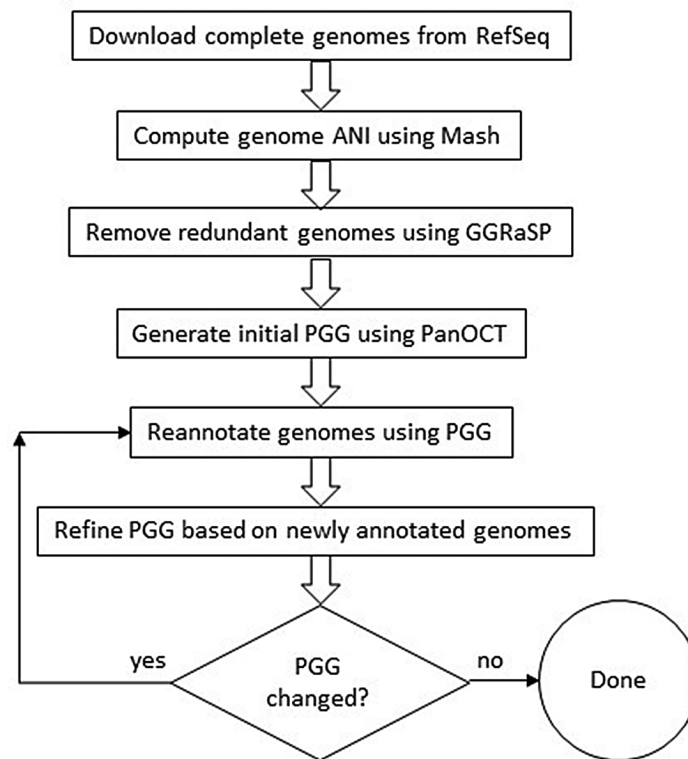


Figure 1. High-level overview of our method for generating a refined PGG.

Here we present a pan-genome based calculation of core regions for *B. subtilis* ssp. *subtilis* and for *E. coli*. These core regions are compared with previous experimentally determined essential genes from the literature. These core regions are not a replacement for experimentally determined essential genes, but rather provide complementary information about a much larger portion of the genome. We expect that all truly essential genes for the species/subspecies would be a subset of the core OGCs/regions, since core OGCs would encompass genes responsible for providing a fitness advantage in environmental conditions as well as being essential for viability. This approach automates computational prediction of core OGCs/regions which can be used to help guide the removal of genome regions not needed for species fitness and indicate which genome regions are amenable to engineered insertions. This approach is an incremental improvement over previous computational methods to aid genome engineering. Ortholog prediction^{17,18,20} and determination of genes essential for most bacteria has a long history.^{21,22} Computational prediction of nonessential genes *via* predicting prophage regions or other horizontal transfer events is also well established.^{23–25} Pan-genome tools, most of which at some level predict core genes, are also not new.^{26,27} Our method builds directly upon our previous pan-genome work^{28–30} and includes several improvements: 1) being able to use only complete high-quality genomes (this concept is not new, but we find it impacts the quality of the PGG and core region determination and is reasonable as more complete genomes become available); 2) checking for including the correct species/subspecies using average nucleotide identity (ANI); 3) reannotating gene features using homology and gene context to ensure consistency; and 4) generating a PGG for a rigorous definition of a core region. **Figure 1** shows the high-level view of our method with details provided in the Methods section.

Methods

Genome Selection

Reference *B. subtilis* ssp. *subtilis* and *E. coli* genomes were selected for pan-genome construction using a series of filtering steps resulting in high-quality, non-redundant genome datasets (**Table 1**). For *B. subtilis* ssp. *subtilis* and *E. coli*,

Table 1. Number of *B. subtilis* and *E. coli* genomes selected after each genome filtering step.

Organism	Text-based query RefSeq download	ANI classification	GGRaSP redundancy filtering	Final genome dataset
<i>B. subtilis</i>	143	132	109	108
<i>E. coli</i>	1097	1096	969	971

we selected strains with complete genomes in RefSeq.³¹ We restricted our analysis to complete genomes to ensure that missing genes due to incomplete genome sequencing/assembly did not affect the approach or results. We limited our choice to RefSeq for two reasons: RefSeq performs a series of quality checks to remove dubious genome assemblies, and the initial pan-genome construction depends upon reasonably consistent annotation which RefSeq provides. We extracted the genomes based on organism name: *Bacillus subtilis* (we did not specify subspecies, since for many RefSeq genomes a subspecies is not given) and *Escherichia coli* (we also specified *Shigella* since all *Shigella* species are actually considered to be the same species as *Escherichia coli*).^{32,33}

For each pan-genome, we then compared the genomes using a fast Average Nucleotide Identity (ANI) estimate generated using the MASH distance subtracted from 1 and multiplied by 100.³⁴ We used type strains and ANI to determine which of these genomes were the desired organism. We also used ANI to remove very closely related strains to reduce oversampling bias (for example for the *B. subtilis* type strain, 168, has at least eight genomes in RefSeq). We used GGRaSP²⁸ to choose a single medoid sequence from any complete linkage ANI cluster with a threshold of 0.01% or 1/10,000 base pair difference. We remove all other genomes besides the medoid as being redundant. Each removed redundant genome would be $\geq 99.99\%$ ANI to the retained medoid genome. The strain 168 medoid genome is the Entrez reference genome for the *B. subtilis* type strain (GenBank sequence AL009126.3, BioSample SAMEA3138188, Assembly ASM904v1/GCA_000009045.1) which can be used to map the Kobayashi *et al.*⁷ and Koo *et al.*⁸ results.

Using this approach, for *B. subtilis*, 143 genomes were downloaded from RefSeq. Of these, 132 genomes were determined to be *B. subtilis ssp. subtilis* based on type strains and ANI. The minimum ANI between any pair of the 132 *B. subtilis ssp. subtilis* genomes was 97.28% whereas the maximum ANI of any of the 11 other genomes to the 132 genomes was 95.73%, providing good separation between the other subspecies. By sorting the pairwise ANI matrix rows based on the ANI values in the type strain column it was clear there was a punctate threshold at $\sim 96.5\%$ ANI which divided *B. subtilis ssp. subtilis* genomes from other genomes. This means the 11 removed genomes all have $\leq 95.73\%$ ANI to the type strain well below the 96.5% ANI threshold. The 132 genomes were further reduced to 109 genomes after removing redundant strains (using GGRaSP as discussed above). Finally, we removed strain delta6 (BioSample SAMN05150066) because it is known to have been engineered to remove multiple genes. Thus, we were left with 108 *B. subtilis* genomes (Table 1). For *E. coli* (and *Shigella*) we downloaded 1097 complete genomes from RefSeq. Of these, 1096 were determined using ANI to be *E. coli*. The non-*E. coli* genome was clearly mislabeled as its maximum ANI to any other genome was 82.27%.

The minimum pairwise ANI of any of the 1096 genomes was 95.53% which is not as tight as for *B. subtilis ssp. subtilis* which is to be expected given that *E. coli* is a species grouping not a subspecies grouping. One could arbitrarily try to choose a tighter grouping around the K-12 reference genome but the pairwise ANI values of the other genomes compared with the K-12 reference genome vary continuously from 96.22% to 100% with no punctate break in the values. After removing redundant genomes (using GGRaSP as discussed above), 969 *E. coli* genomes remained. We added back in two redundant genomes: The K-12 Entrez *E. coli* reference strain MG1655 (BioSample SAMN02604091) and the K-12 strain BW25113 (GenBank sequence accession CP009273.1, GenBank Assembly accession ASM75055v1/GCA_000750555.1, GenBank BioSample accession SAMN03013572) used by Goodall *et al.*⁹ These two redundant genomes were added back in so that we could map the PGG OGCs to these genomes for comparison to the established literature resulting in 971 genomes in the PGG (Table 1). By using a 95% threshold for the number of genomes an OGC must be in to be considered core, some small number of the 971 genomes could be engineered to remove what are normally core OGCs and not affect the assignment of core OGCs.

Pan-genome and PGG construction

For *B. subtilis ssp. subtilis* and *E. coli*, initial pan-genomes were based on the RefSeq annotation of these genomes. The pan-genome was generated using the pan-genome pipeline at the J. Craig Venter Institute (JCVI) at the nucleotide level using default parameters with the exception that a minimum of 90% identity and 90% length for pairwise BLAST matches were used to prevent possible clustering of non-orthologous genes.²⁹ This produced OGCs using gene context³⁰ as well as a PGG.¹⁹ The PGG has two main components: nodes representing OGCs, and edges representing the sequence between OGCs and the order and orientation of the OGCs in the genomes. We updated the code repository for the JCVI pan-genome pipeline with a script: `iterate_pgg_graph.pl`, which calls `pgg_annotate.pl` for the genomes in the existing PGG in order to ensure consistent annotation of the genomes and iterates until the PGG stabilizes. The script `pgg_annotate.pl` uses an existing PGG to assign regions of a genome to nodes of the graph. This is done by searching the medoid sequence using BLAST for the OGC the node represents against the genome and then uses Needleman – Wunsch³⁵ to extend the alignment if needed. If there are conflicting BLAST matches, then the matches are resolved based on which matches are consistent with the structure of the PGG which encapsulates gene context across the entire pan-genome. Once the nodes of the PGG are mapped to each of the genomes in the pan-genome a new version of the PGG is intrinsic and then explicitly extracted. This process is iterated to stability. This ensures that each genome is consistently annotated so that genes

missing from the original annotation of some genomes will be consistently annotated across all genomes. A user manual for this new functionality is available at <https://github.com/JCVenterInstitute/PanGenomePipeline>.

Core regions were determined based on the PGG. Nodes in the PGG were OGCs. Edges in the PGG represented adjacency of genes (contained in the OGCs) in the underlying genomes. The definition of which OGCs were or were not considered “core” was determined relative to a threshold criterion. We used a criterion for core such that 95% or more of the underlying genome had to contain the OGC or edge. Considering that we used only complete genomes it might have been possible to use a 100% threshold. However, we opted for a 95% threshold based on prior experience and an abundance of caution to not under call core OGCs/edges which might result in false negatives. Each core region began with a core OGC followed by a core edge (if possible, otherwise the core region comprises a single OGC) to another core OGC and so on until a core edge cannot be found to continue the core region. A core region is just a path in the PGG which was then mapped onto each genome to determine the core region coordinates. When the core threshold was below 100% any genome may be missing an OGC (gene) or edge along this path which results in the path being broken into its remaining constituent parts.

Comparison to essential genes

In order to compare core regions to experimentally determined essential genes we needed a common base of reference. For each of the experimental studies, the genes are specified based on a reference strain that was used for the experiments and has a complete genome in RefSeq. For Kobayashi *et al.*,⁷ only gene symbols/names were given which we mapped to Entrez GeneIDs using Entrez search. GeneIDs with no matches were manually curated to estimate the best matching gene symbol listed in the literature. For Koo *et al.*,⁸ locus IDs were provided giving direct access to the gene coordinates for RefSeq accession NC_000964.3 (BioSample SAMEA3138188, Assembly GCF_000009045.1). For Goodall, we used the data from three studies in Table S2 from Goodall *et al.*⁹ Gene symbols/names again were all that was available but these were consistent with the GenBank annotation downloadable in gff format for the K-12 BW25113 reference genome (GenBank accession CP009273.1) used by Goodall *et al.*⁹ (BioSample SAMN03013572). This gave us coordinates for all essential genes on RefSeq genomes which were annotated with a PGG which produces a file with coordinates for OGCs and edges mapped to the genome. These coordinates allow us to affiliate essential genes to OGCs.

Results

The original and refined PGG statistics for *B. subtilis* and *E. coli* are provided in Table 2. The major goal of refining the PGG using reannotation and iteration until stabilization was to achieve consistent annotation across all genomes in the PGG leading to a more comprehensive and cohesive PGG. While the RefSeq annotations of these genomes tends to be highly consistent, many small genes are often arbitrarily called from genome to genome and even some common longer genes can occasionally be missed. There are three obvious points of improvement in the refined PGG for both the OGC and edge stats: the number of size 1 OGCs/edges significantly decreased due to some dubious RefSeq gene calls being eliminated and some becoming shared with other genomes; the number of core OGCs/edges significantly increased showing an improvement in the consistency of annotation across all genomes; and the number of genes/edge instances in OGCs/edges greatly increased again indicating a much more consistent annotation. We have included core OGC statistics for three threshold definitions of core: 95%, 99%, and 100%. In part, this is for comparison to previous studies but it also

Table 2. Pan-genome graph statistics for *B. subtilis* and *E. coli*.

PGG Statistic	<i>B. subtilis</i> original PGG	<i>B. subtilis</i> refined PGG	<i>E. coli</i> original PGG	<i>E. coli</i> refined PGG
Size 1 OGCs	4434	3231	87423	27273
Shared (size>1) OGCs	8174	8204	68970	48129
# of genes in shared OGCs	463311	487562	5039502	5610683
Core OGCs (95%)	3558	3778	2968	3631
Core OGCs (99%)	3356	3604	2168	2992
Core OGCs (100%)	3072	3419	713	1501
Size 1 edges	7282	5479	153199	67566
Shared edges	9755	9433	99284	67248
Edge instances in shared edges	460452	485177	4970823	5567497
Core edges (95%)	3230	3520	2218	3124

illustrates the relative larger impact of consistency as the threshold increases. For example, in *E. coli*, the refined PGG gives an increase of 22% in core OGCs at a 95% threshold but an increase of 111% in core OGCs at a 100% threshold. When even a single misannotated gene drops an OGC below core at the 100% threshold consistent annotation is crucial.

The *B. subtilis* ssp. *subtilis* refined PGG annotates 4654 OGCs for the reference genome (GenBank sequence AL009126.3, BioSample SAMEA3138188, Assembly ASM904v1/GCA_000009045.1) (Supplementary Table 1): 876 (18.8% of OGCs) noncore (<95% of genomes 102 or less), 359 (7.71%) core but not present in all genomes ($\geq 95\%$ and <100% of genomes 103–107), and 3419 (73.5%) core and present in all 108 genomes. For the union of the Koo *et al.*⁸ and Kobayashi *et al.*⁷ essential gene data sets there are 305 genes (Supplementary Table 2): 16 (5.25%) noncore (≤ 102 genomes), 2 (0.656%) core but not all (103–107 genomes), and 287 (94.1%) core all (108 genomes). This shows that most essential genes in *B. subtilis* ssp. *subtilis* are encompassed by core OGCs/regions. There are 258 core regions for *B. subtilis* (Supplementary Table 3). The 289 essential genes which are core OGCs are contained in only 63 of these regions. These 289 essential genes are not evenly distributed in these 63 regions (*e.g.* 46 are in core region 3). Similarly, the 16 essential genes in non-core regions (the regions between core regions) are contained in only seven non-core regions with eight genes in the non-core region between core regions 206 and 207 (Figure 2). A table of all *B. subtilis* genes mapped to the reference genome is provided in Supplementary Table 1.

The Reuß *et al.*³ data set for *MiniBacillus* has 523 protein coding and 119 RNA genes predicted to be necessary for a minimal *B. subtilis*. For the 523 protein coding genes: 18 are noncore (≤ 102 genomes), 16 are core but not in all genomes (one in 105, one in 106, 14 in 107 genomes), and 489 are in all 108 genomes (Supplementary Table 1). They include all 30 rRNA and 86 tRNA genes from the reference genome as well as three “misc” RNA genes in *MiniBacillus*. The three misc RNA genes are present in all 108 genomes. In all likelihood, the 10 copies of the 16S-23S-5S RNA operon are not required but it is safer for robust growth not to delete any of them. Likewise, for the tRNA genes where many are redundant. For the 30 rRNA genes: six are noncore (92–102 genomes), 16 are core but not in all genomes (103–107), and eight are in all 108 genomes. It is clearly possible that some of these strains are dispensing with some of the RNA operons but at most this is happening rarely reinforcing the decision not to remove any from *MiniBacillus*. In addition, some of the missing RNA operon genes may be due to incorrect assembly of the two sets of tandem RNA operons (one a two-unit tandem and one a three-unit tandem) as large tandem repeats can be problematic for assemblers. All the rRNA genes in fewer than 106 genomes are in the tandem rRNA operons (Supplementary Table 4). Of course, the tandem rRNA operons are the most likely to be deleted via recombination as well. For the 86 tRNA genes: 13 are noncore (100–102 genomes), 18 are core but not in all genomes (103–107), and 55 are in all 108 genomes. Retaining all the tRNA genes in *MiniBacillus* also seems to be the correct decision as strains rarely dispose of the tRNA genes.

Both the experimentally determined essential genes and the predicted core OGCs/regions are important data for genome engineering. They both indicate regions that should not be deleted without careful consideration. The noncore regions also indicate where the bacterium is more likely to tolerate engineered insertions. As a validation of our method and how to interpret the results our method produces it is important to understand why 16 essential genes are in noncore regions.

For *B. subtilis*, both Kobayashi *et al.*⁷ and Koo *et al.*⁸ used similar single knockout methods to determine “essential” protein-coding genes when grown in LB at 37°C. Koo *et al.*⁸ identified 257 essential genes while Kobayashi *et al.*⁷ identified 271 essential genes. The union of these two sets results in 305 essential genes (Supplementary Table 2). The Koo *et al.*⁸ data set has 257 genes. The Kobayashi *et al.*⁷ data set has 271 genes. There are 223 genes in common between the two data sets. 48 genes are only in the Kobayashi *et al.*⁷ data set. 34 genes are only in the Koo *et al.*⁸ data set. The Kobayashi *et al.*⁷ data set has been refined with time:³⁶ “Of the original 271 genes, 31 were shown to be non-essential in recent studies. Moreover, 21 new genes (19 protein-coding genes and two RNA-coding genes) were added to the list. Thus, 261 genes encoding 259 proteins and two RNAs are regarded as being essential today”. This list of 259 protein-coding genes is more consistent with the more recent Koo *et al.*⁸ data set. The 305 genes found in either data set were mapped to the PGG OGCs using the RefSeq genome NC_000964.3 (BioSample SAMEA3138188). Interestingly through this mapping, 16 of the essential genes were not identified as core OGCs (two more essential genes were core OGCs but not present in all 108 genomes). For the 18 essential genes not present in all 108 genomes (Supplementary Table 5), 12 are in both data sets and six are only in the Koo *et al.*⁸ data set. We believe only 11 of the 18 genes are truly essential. Gene *wapI/yyxG* (OGC 4769 present in 39 of 108 genomes) is an antitoxin for the *wapA* toxin gene which is adjacent to it (present in 85 of 108 genomes).³⁷ Gene *rttF/ycqF* (OGC 4590 present in 46 of 108 genomes) and gene *rtbE/yyxD* (OGC 4772 present in 53 of 108 genomes) are also the antitoxin of a cognate toxin-antitoxin pair.³⁸ Gene *yezG* (OGC 4411 present in 43 of 108 genomes) is also the toxin for a cognate toxin-antitoxin pair.³⁹ Gene *sknR/yyqAE* (OGC 4643 present in 34 of 108 genomes) is part of a phage-like region which, if removed would still allow *B. subtilis* to remain viable¹² possibly because it is another antitoxin or similar mechanism. Genes *bsuMA/ydiO* (OGC 4838 present in 24 of 108 genomes) and *bsuMB/ydiP* (OGC 4839 present in 24 of 108 genomes) are part of a prophage region of about 15 genes

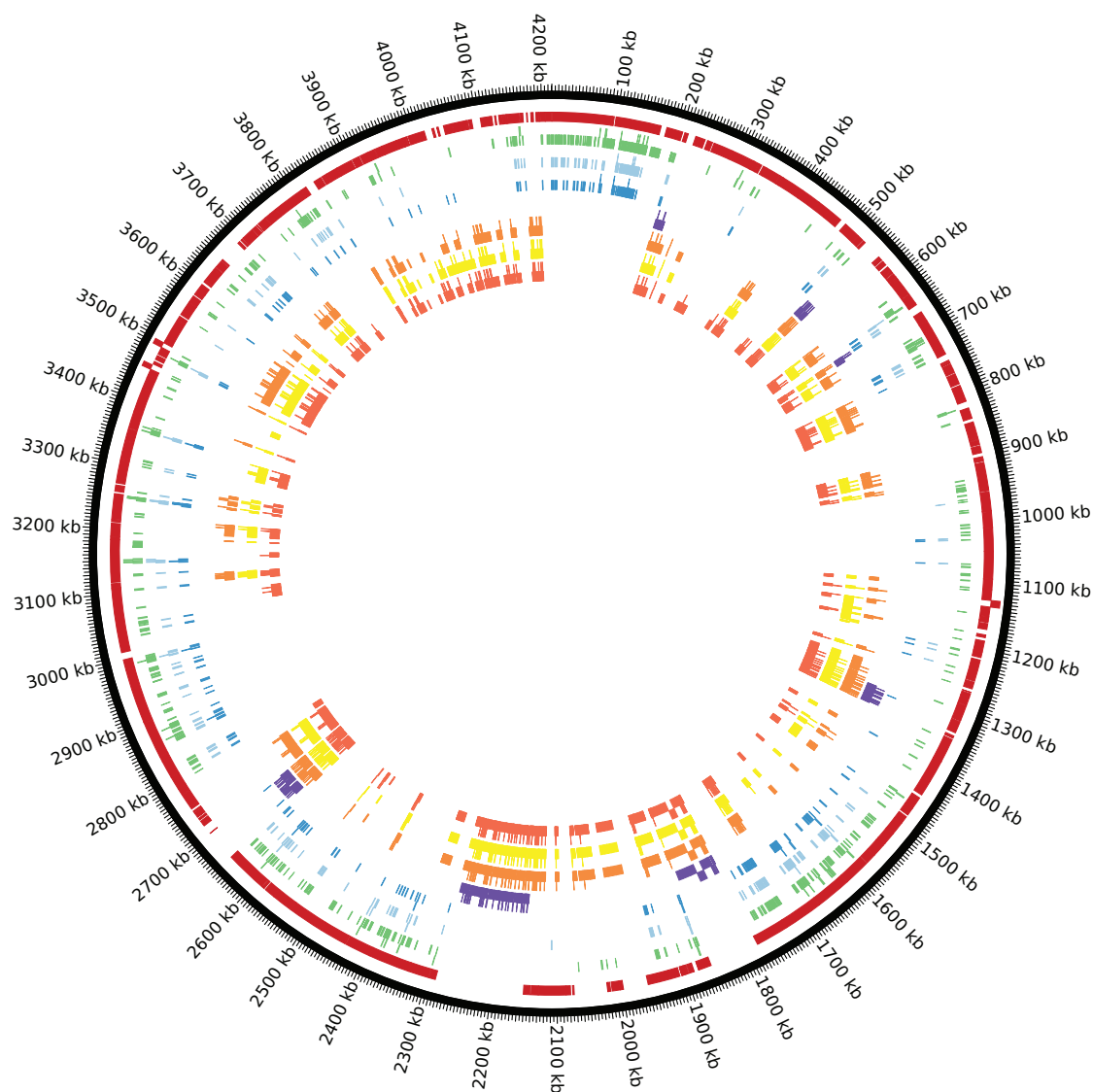


Figure 2. There are eight tracks mapped to the *B. subtilis* reference genome in this Circos figure. Going from the outside to the inside: track 1) core regions (dark red), 2) Minibacillus genes (green), 3) Koo *et al.*⁸ essential genes (light blue), 4) Kobayashi *et al.*⁷ essential genes (medium blue), 5) deleted genes in strain delta 6 (dark blue), 6) deleted genes in strain IIG-Bs27-47-24 (orange), 7) deleted genes in strain PG10 (yellow), and 8) deleted genes in strain PS38 (red).

in 48 genomes which includes *ydiR* and *ydiS* which are type-2 restriction enzymes. These are not essential genes, but they are essential if the restriction enzymes are present.⁴⁰ We are not the first to notice these issues with experimentally determined essential genes indicated by our references above. In their review, Commichau *et al.*³⁶ referred to these as "protective essential genes." In fact, Koo *et al.*⁸ also addressed this in their paper: "Of the 257 genes essential in LB medium, 30 are not essential in some other growth condition or genomic context...LB may have an insufficient amount of particular compounds; *e.g.*, the *ylaN* mutant requires a higher amount of iron than that present in LB ... or may lack a compound that could bypass the need for that gene product; *e.g.*, *eno*, *pgm*, *gapA*, and *alrA* ... Some gene products are essential only at high growth rates typical of LB at 37°C (*smc* and *scpA* ...), and these may not be essential in the natural soil environment where *B. subtilis* grows slower. Finally, some genes are non-essential in specific genetic backgrounds, *e.g.*, antitoxins can be deleted in strains lacking their cognate toxin gene".

Another eight essential non-core genes are involved in wall teichoic acid (WTA) biosynthesis: Genes *tuaB* (OGC 4729 present in 85 of 108 genomes), *mnaA/yvyH* (OGC 4735 present in 84 of 108 genomes), *tagH* (OGC 4744 present in 84 of 108 genomes), *tagG* (OGC 4745 present in 35 of 108 genomes), *tagF* (OGC 4746 present in 35 of 108 genomes), *tagD*

(OGC 4748 present in 35 of 108 genomes), *tagA* (OGC 4749 present in 35 of 108 genomes) and *tagB* (OGC 4750 present in 35 of 108 genomes). The WTA genes are involved in production of anionic glycopolymers required for consistent cell shape and division.⁴¹ The WTA genes are part of a 31 gene region which has been shown to be dispensable⁴² but results in malformed cells with poor growth properties. Gene *rodA* (OGC 3994 present in 97 of 108 genomes) appears to be the exception as it is asserted to be essential for maintaining a rod shape and preventing spherical cells which lyse.⁴³ Kobayashi *et al.*⁷ stated: “Ten essential genes are involved in cell shape and division. Septum formation requires seven (*ftsA*, *L*, *W*, and *Z*, *divIB* and *C*, and *pbpB* ...), whereas cell shape requires three (*rodA*, and *mreB* and *C*).” Interestingly, genes *ftsZ* (OGC 1675 present in 105 of 108 genomes) and *pbpB* (OGC 1662 present in 107 of 108 genomes) while considered core, using our 95% of genomes definition are the only core OGCs not present in all 108 genomes. We investigated these 11 genes further to understand why essential genes did not appear to be core OGCs. By examining the PGG we discovered that alternate OGCs with homology to the essential genes had replaced the essential genes. Gene *pbpB* (OGC 1662 in 107 genomes) is replaced in the one remaining genome by OGC 7120 which is also annotated as *pbpB*. Gene *ftsZ* (OGC 1675 in 105 genomes) is replaced in three genomes by a four gene insertion of OGCs 8068, 8300, 8069, and 8070 where both 8068 and 8070 are annotated as *ftsZ*. Gene *rodA* (OGC 3994 in 97 genomes) is replaced by either: OGC 8718 (two genomes) or OGCs 10492, 6436, and 6437 (one genome) or OGCs 6436 and 6437 (eight genomes) where 8718 and 6436 are annotated as *rodA*. As an illustrative example for *rodA*, Figure 3 shows how this is represented in the PGG. The medoid sequences for OGCs 6436 (A4A60_RS20560), and 8718 (C7M30_RS12210) have full length homology to the medoid sequence for *rodA* (OGC 3994, ETA10_RS20040) with 66% nucleotide /65% peptide and 83% nucleotide/85% peptide identity respectively. For *B. subtilis* ssp. *spizizenii* strain W23, poly (ribitol phosphate) is the main teichoic acid⁴⁴ and this was thought to distinguish ssp. *spizizenii* from ssp. *subtilis* whose type strain 168 has poly (glycerol phosphate) as the main teichoic acid. Further study found that the ribitol/glycerol distinction does not distinguish between *spizizenii* and *subtilis* subspecies⁴⁵ but rather either subspecies can contain one or the other. Our PGG confirms this and in fact finds six distinct variants of the WTA region. For example the *tagD* gene (OGC 4748 in 35 genomes) has been replaced by multiple orthologs with the same annotation: OGC 3746 (23 genomes), OGC 5431 (43 genomes), OGC 6915 (two genomes), OGC 7624 (three genomes), and OGC 8731 (one genome). The variation of the WTA region in *B. subtilis* will be the focus of a future paper.⁴⁶

For the 34 protein coding genes from *MiniBacillus*³ which were not in all 108 genomes (Supplementary Table 6), 10 were already discussed above as to why they were essential but not core. The seven essential genes previously shown to be protective essential genes are as expected not in the *MiniBacillus* data set. The noncore *tuaB* gene was essential in both data sets but not included in *MiniBacillus*. This leaves 24 *MiniBacillus* protein coding genes which are noncore and unexplained. The *tagU* and *gtaB* genes are part of the WTA cassette discussed above. The four *fecC-F* (also called *yfmC-F*) genes form a cassette and are in 98 genomes. From Reuß *et al.*:³ “For iron uptake, the minimal cell should possess the EfeUO system for elemental iron uptake and the iron-citrate ABC transporter YhfQ-YfmCDEF (136, 137).” (*yhfQ* is present in all 108 genomes) but no alternate mechanism is specified. The seven *purEKBCSQL* genes form a

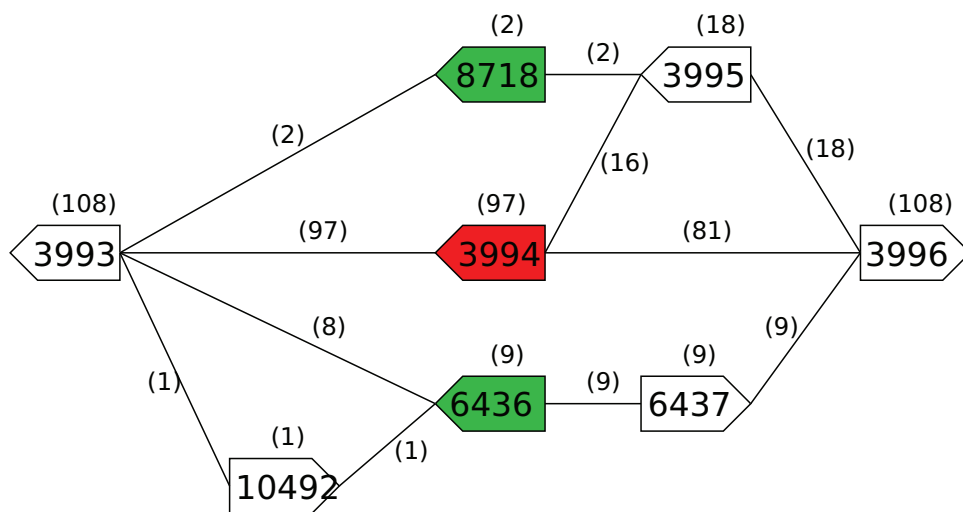


Figure 3. Region of the *B. subtilis* refined PGG encompassing the variation in the *rodA* gene across the pan-genome. OGC 3994 (red) contains the *rodA* gene from the reference strain. The medoid sequences of OGCs 6436 and 8718 (green) have RefSeq annotations of *rodA* and full-length homology below our 90% threshold to the medoid sequence for OGC 3994. The arrow boxes represent OGCs with gene directionality indicated by the 5' end being flat and the 3' end being pointed. Numbers above boxes and edges are the number of genomes the OGC or edge are in.

cassette and are in 107 genomes. These genes are involved in purine biosynthesis (see Figure 5 in Reuß *et al.*³) and it is not clear what alternative could be used. The *guaA* gene is involved in nucleotide biosynthesis downstream of purine biosynthesis (see Figure 5 in Reuß *et al.*³) present in 107 genomes. The *mntH* gene is a manganese transporter (see Figure 2 in Reuß *et al.*³) present in 106 genomes. The *rlmCD* gene is an rRNA methyltransferase present in 107 genomes. The *lytE*, and *ponA* genes are in 107 genomes. The *pbpB* gene was essential as discussed above and in 107 genomes. The *pbpA* gene is in 50 genomes. From Reuß *et al.*³ “For the minimal cell, we have selected penicillin-binding proteins 1 (PonA), 2B (PbpB), and 2A (PbpA) and the autolysins LytE and LytF. As outlined above, this selection was made according to their expression profiles and the dependence on other proteins. As an example, there is a functional paralog of LytE, CwlO. For the activity of CwlO, *B. subtilis* also needs the ABC transporter FtsEX and the small protein Mbl. Thus, the choice of LytE allowed a smaller number of genes.” Interestingly, genes *cwlO*, *ftsE*, *ftsX*, and *mbl* are in all 108 genomes. The *yitI* gene is in 107 genomes. From Reuß *et al.*³ “Moreover, based on our own experimental data and those of colleagues, YitI, YitW, and YqhY are important for viability (P. Dos Santos, personal communication; our unpublished results).” The *yoaE* gene is a formate dehydrogenase present in 89 genomes. The *thyB* gene is thymidylate synthase B present in 70 genomes. The *rpoE* gene is in 107 genomes. From Reuß *et al.*³ “Moreover, we have included the RNA polymerase-interacting protein HelD and the nonessential delta subunit (RpoE). HelD binding stimulates transcription in an RpoE-dependent manner, suggesting that these two accessory proteins are important to allow rapid growth (59, 60).” The *hutM* gene is a histidine permease present in 90 genomes. *MiniBacillus* does not include the adjacent *hutPHUIG* genes which are in 88-91 genomes probably indicating a cassette of genes which interact.

We looked at how our OGGs intersected with the gene deletions from *B. subtilis* strains delta 6, IIG-Bs27-47-24, PG10, and PS38 from Reuß *et al.*³ Supplemental Table S1. Strains PG10 and PS38 were derived from strain IIG-Bs27-47-24 which in turn was derived from strain delta 6. This means all deletions in delta 6 are present in the other strains, and all deletions in IIG-Bs27-47-24 are present in PG10 and PS38. For delta 6, most of the deleted genes are noncore which would be expected since most of the deleted regions were phage/prophage regions (Table 3). For additional deletions to IIG-Bs27-47-24, almost a quarter of the deleted genes are noncore which would again be expected as more prophage and horizontally transferred regions were intentionally targeted but now more core genes were deleted based on core functionality deemed not to be essential for laboratory growth such as sporulation (Table 3). For additional deletions to PG10 and PS38, most deleted genes were core as most of the obviously horizontally transferred regions had already been deleted (Table 3). While pan-genome analysis was not used to select the deleted regions, we believe it could have provided strong evidence to support the deletion of the noncore genes/regions which were deleted. In addition, it could be used to suggest further deletions. There are nine noncore regions which contain seven or more noncore genes which have not yet been deleted in any of these strains (Table 4). The largest of these regions contains the WTA genes cassette we discussed above and is not a good candidate for deletion. By examining the refined PGG at these regions it is straightforward to determine if there are alternate OGC choices for the region that in sum designate the region as likely to be core as we showed in Figure 3.

To show that our method produces significantly different results than previous methods we compared our pan-genome analysis to the very recent work on a *B. subtilis* pan-genome by Wu *et al.*⁴⁷ While the focus of Wu *et al.*⁴⁷ was on determining which genomes should be excluded from a species/subspecies pan-genome based on “incorrectly classified *Bacillus* subspecies strains, phylogenetically distinct strains, engineered genome-reduced strains, chimeric strains, strains with a large number of unique genes or a large proportion of pseudogenes, and multiple clonal strains”, their analysis focused on how this affected the determination of core OGCs. We compared our core OGC set to theirs for the reference genome. Wu *et al.*⁴⁷ discussed two pan-genome data sets: “old (89 strains) and new (153 strains)”. We compared to the new data set which is more recent and more comparable to our pan-genome of 108 strains (Supplementary Table 5). After removing “confounding” strains the new data set had 128 strains. From their Table 1 compared to our Table 2, Wu *et al.*⁴⁷ have many fewer core OGCs whether defined at 95%, 99%, or 100% both for our original and refined PGGs. We compared their methods to ours to attempt to account for the difference. They also apparently restricted genomes to those available

Table 3. The number of deleted genes from *B. subtilis* reduced strains which are noncore versus core.

Strains	Number of noncore deleted genes	Number of core deleted genes
delta 6, IIG-Bs27-47-24, PG10, PS38	340	46
IIG-Bs27-47-24, PG10, PS38	232	792
PG10, PS38	7	57
PG10	14	78
PS38	15	129

Table 4. Large noncore regions which have not been deleted from any of the strains delta 6, IIG-Bs27-47-24, or PG10, PS38.

First gene	Last gene	Number of noncore genes in region	Number of core genes in region	Alternate genes in refined PGG
BSU04270, <i>epsJ</i> , OGC4339	BSU04320, <i>kimA</i> ,OGC4344	7	0	no
BSU05040, <i>yddN</i> , OGC4348	BSU05110, <i>sufLC</i> ,OGC579	10	0	yes
BSU07440, <i>yfmK</i> , OGC4418	BSU07550, <i>yfIT</i> , OGC4427	10	0	no
BSU11910, <i>yjcM</i> , OGC4457	BSU11990, <i>yjdB</i> , OGC4465	10	1	yes
BSU18940, <i>yobHm</i> , OGC4568	BSU19000, <i>rttL</i> , OGC2078	10	0	yes
BSU29280, <i>ytnM</i> , OGC4678	BSU29400, <i>ascR</i> ,OGC4689	13	0	no
BSU35550, <i>tuaG</i> , OGC4724	BSU35770, <i>tagC</i> ,OGC4751	28	0	yes
BSU37220, <i>ywjB</i> , OGC3904	BSU37320, <i>narK</i> ,OGC3915	11	0	yes
BSU39850, <i>yxbF</i> , OGC4783	BSU39920, <i>asnH</i> ,OGC4790	9	0	yes

from RefSeq since they mention a RefSeq ID. They did not require the genomes to be considered complete by RefSeq as we did but instead used these criteria: “Among these *B. subtilis* strains, we removed strains whose N base content was greater than 1% of the genomic size (FB6-3,GS 188, SR1), and we removed the chimeric genome BEST7613 with a genome size of 7.6 Mb.” We used the RefSeq annotation which is generated by a consistent NCBI annotation pipeline. They also tried to ensure consistent annotation: to “ensure the consistency and reliability of the annotation and gene prediction of the genome, we used the program Prokaryotic Genome Annotation System (Prokka)”. We doubt the different annotations from these two established pipelines accounts for many differences in core OGCs. Both methods used a whole genome ANI method to discard outlier genomes. There are multiple differences in our pan-genome approach. First, we used PanOCT and they used Roary. Second, we used all annotated gene features: gene (protein coding), pseudogene, miscRNA, rRNA, and tRNA, whereas they used only protein-coding genes. Finally, and we think most importantly, we iterated over annotating the genomes and PGG refinement to ensure consistent annotation and they did not. To see what impact our choice of all gene features versus just protein-coding genes had we looked at the annotation of core OGCs on the reference genome (Supplementary Table 1). Luckily all 3778 core (95% threshold) OGCs are present in the reference genome. Of these, 3473, 3334, and 3189 are protein coding OGCs at thresholds 95%, 99%, and 100% respectively. All these numbers are still much higher than those reported by Wu *et al.*⁴⁷ We should note that even though we did not count the 25 core OGCs annotated as pseudogenes in the reference genome, some of the core protein-coding OGCs in the reference genome might be annotated as pseudogenes in other genomes which could impact the Wu *et al.*⁴⁷ numbers. Roary tends to require near full length gene matches which is why we required PanOCT to only use 90% or longer length matches. The authors chose to limit Roary to 95% identity or higher matches which we think is much too high since the species ANI threshold is 95% and even subspecies ANI threshold of 98% is too close to this threshold given that some genes are more rapidly evolving than others so we used a threshold of 90% or higher identity for matches. Even with our 90% identity threshold some genes such as *rodA*, discussed above, drop below this threshold generating possibly unnecessary branching in the PGG. Of the 128 strain pan-genome from Wu *et al.*⁴⁷ that we compared to our 108 strain pan-genome, 92 strains were in common with 16 being exclusive to our pan-genome and 36 being exclusive to theirs. Of the 36 strains exclusive to theirs 23 were removed as being redundant at the ANI level by us, 9 were in RefSeq but not complete genomes, and 4 either were never in RefSeq (they do not have RefSeq IDs in their Supplementary Table 3) or no longer are. Interestingly, while 15 of the 16 genomes exclusive to ours are just more recent strains to RefSeq, one strain, D12-5, was used by us but discarded by them. They discarded D12-5 because “BS155 and D12-5 possess the largest proportion of pseudogenes (37.96% and 11.32%) among the *B. subtilis* strains” and for D12-5 they indicated this was due to a large number of frameshifts. Pseudogenes due to frameshifts are often an indication of lower quality assembly consensus sequence from using only long reads at lower coverage. Our pan-genome method is resilient to this kind of error profile in the genome due to reannotation of the genomes and PGG refinement whereas other pan-genome methods are not. We believe our higher counts for core

protein coding OGCs is correct. To validate this, we compared how many of the 305 essential *B. subtilis* genes are core for both methods. For the 18 genes we discussed above that are essential but not in all 108 genomes of our PGG, 2 are core at 95% and 1 is core at 99%; whereas, for Wu *et al.*⁴⁷ 2 are core at 95%, 2 are core at 99%, and 1 is core at 100%. The only significant difference for these 18 genes is that *ftsZ* is in 95% (105) of our pan-genome and 100% of theirs. The Wu *et al.*⁴⁷ pan-genome misses many additional essential genes which ours does not: 28, 39, and 47 for 95%, 99%, and 100% thresholds respectively.

For *E. coli*, Goodall *et al.*⁹ determined *E. coli* essential genes using an analysis of transposon insertion events (TraDIS). The results of their study and two other studies, the Keio collection¹⁰ and the Profiling of the *E. coli* Chromosome (PEC)¹¹ were captured in Table S2 of Goodall *et al.*⁹ Of the 414 genes with overlap between these studies, the 248 essential genes in common for all three studies are all core OGCs (Figure 4 and Supplementary Table 7). This set of 248 essential genes should be the highest quality predictions as determined by all three studies and confirms our assertion that essential genes should almost always be core OGCs. The next highest quality set of essential gene predictions is the 45 essential genes where two of the three studies agree which 41 are core OGCs: for Keio–PEC, 15 of 16 are core OGCs; for TraDIS–Keio, eight of 11 are core OGCs; and for TraDIS–PEC, 18 of 18 are core OGCs

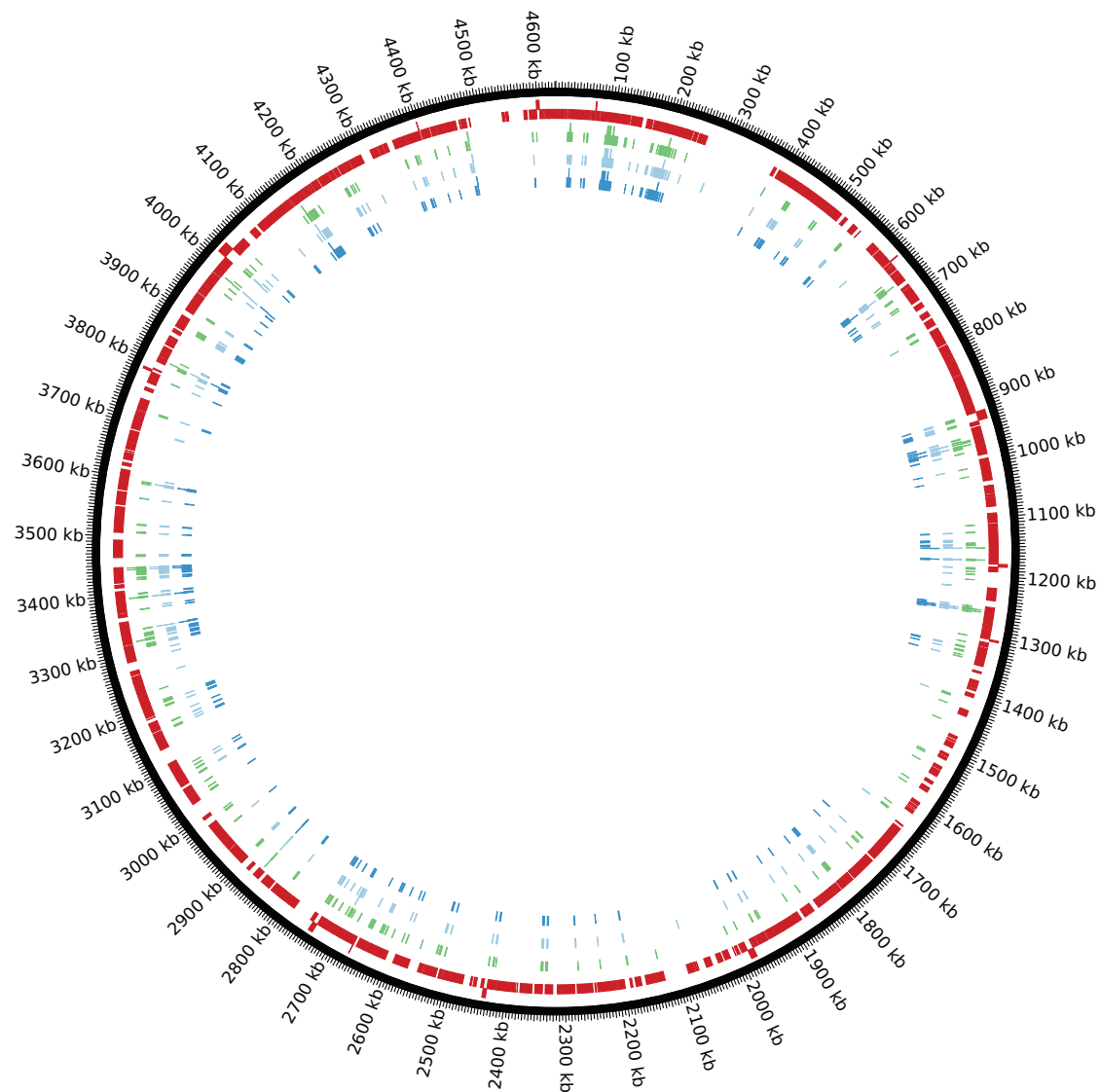


Figure 4. There are four tracks mapped to the *E. coli* reference genome in this Circos figure. Going from the outside to the inside: track 1) core regions (dark red), 2) TraDis essential genes (green), 3) Keio essential genes (light blue), and 4) PEC essential genes (medium blue).

(Supplementary Table 7). The lowest quality set of essential gene predictions is the 121 essential genes where only one study agrees which 89 are core OGCs: for Keio only, 12 of 22 are core OGCs; for PEC only, 18 of 18 are core OGCs; and for TraDIS only, 59 of 81 are core OGCs (Supplementary Table 7). One of the noncore essential genes present in two studies (TraDIS–Keio), *racR*, is probably a toxin suppressor which is not essential in the absence of the toxins. Bindal *et al.*⁴⁸ noted, “We further show that both YdaS and YdaT can act independently as toxins and that RacR serves to counteract the toxicity by tightly downregulating the expression of these toxins”. The *racR* gene is found in only 106 of the 971 genomes in the *E. coli* PGG, whereas *ydaS* and *ydaT* are found in 106 and 150 genomes respectively, perhaps arguing that *ydaS* is the key toxin gene. This recapitulates the pattern we observed in *B. subtilis* where toxin suppressor genes are only essential in the presence of toxin genes. Similarly, the *dicA* gene (TraDIS–Keio) can be deleted if the *dicB* gene is also deleted. Kato *et al.*⁴⁹ noted: “The *dicA* gene encoding a repressor of a cell division inhibitor was deleted in our study with the *dicB*, the inhibitor gene”. There are 521 core regions for *E. coli* (Supplementary Table 8). The 378 essential genes which are core OGCs are contained in only 133 of these regions. These 378 essential genes are not evenly distributed in these 133 regions (*e.g.*, 27 are in core region 362). Similarly, the 36 essential genes in non-core regions (the regions between core regions) are contained in only 23 non-core regions with four in the non-core region between core regions 152 and 153. A table of all *E. coli* genes mapped to the reference is provided in Supplementary Table 9.

Yang *et al.*⁵⁰ presented a similar pan-genome analysis for 491 *E. coli* strains. There were 420 strains in common between the Yang *et al.*⁵⁰ 491 strain pan-genome and our 971 strain pan-genome (Supplementary Table 10). Our pan-genome included *Shigella* species (see Methods) which Yang *et al.*⁵⁰ did not. This added diversity of our pan-genome should reduce the number of core OGCs. Likewise, the much larger number of strains in our pan-genome should reduce the number of core OGCs. Yang *et al.*⁵⁰ report 867 core protein-coding genes presumably at a 100% threshold although this is not explicitly stated. For our refined PGG, we had 1501 core OGCs at the 100% threshold. We include all genes in our OGCs but 1234 of the 1501 core OGCs are protein coding at the 100% threshold. Yang *et al.*⁵⁰ did not provide a table of their core genes for sake of comparison, however we expect for the same reasons as for our more detailed analysis of the *B. subtilis* pan-genome that our set of core OGCs is more complete. Yang *et al.*⁵⁰ reported that their core genes included 243 essential genes from the DEG database⁵¹ which contains essential genes from many studies but did not provide a table of these genes. Yang *et al.*⁵⁰ also reference two essential gene studies one by Gerdes *et al.*⁵² and one by Baba *et al.*¹⁰ which was one of the three studies we used (Keio). In the DEG database the Gerdes *et al.*⁵² study has 609 essential genes, and the Baba *et al.*¹⁰ study has 296 essential genes. Our version of the Baba *et al.*¹⁰ study we called Keio had 297 essential genes of which 218 were core OGCs at the 100% threshold. For the union of the three studies we compared against, we had 289 essential genes out of 414 which were core OGCs at the 100% threshold. It is unclear whether the 243 core essential genes Yang *et al.*⁵⁰ reported were from the Baba *et al.*¹⁰ study, the Gerdes *et al.*⁵² study, or the union of the two studies. Given the much lower number of core genes for the Yang *et al.*⁵⁰ core genes compared with our core OGCs, we believe that Yang *et al.*⁵⁰ used the union of essential genes from the Baba *et al.*¹⁰ and Gerdes *et al.*⁵² studies.

There is of course no “gold standard” that provides a 100% correct set of core regions/genes for a pan-genome/species. When comparing our method to others, this leaves only indirect measures of accuracy. We compared our method versus two other recent core gene determinations for *Bacillus subtilis* and *Escherichia coli* and showed that our method was superior using coverage of essential genes by core genes as an indirect measure. We also showed that the PGG allowed for a detailed analysis of exceptions such as when an OGC is replaced by a more distant ortholog.

Discussion

For the purpose of biological engineering, determining the set of core regions for a given species is critical as changes to these regions should be expected to reduce fitness or be lethal. Core regions indicate parts of the genome that are conserved across evolution within a species. These regions are not necessarily required for survival but presumably confer a fitness advantage and define the characteristic core genotype which produces the core phenotype (lifestyle). Since most essential gene studies are carried out under specific static laboratory growth conditions, genes which would normally be essential for a species across a diverse set of dynamic environmental conditions might not be discovered (*e.g.*, necessary for fluctuating temperatures). Correspondingly, genes required to out compete rival organisms through increased fitness or to evade immune responses might not be found under laboratory conditions are considered facultative essential.³¹ Core regions, therefore, should be a superset of essential genes in most cases but exceptions might occur for genes which are not needed in a species’ natural niche but are required in a laboratory setting. Another exception would be for genes which are essential for a particular strain but not for other strains due to the presence of compensating non-core genes.

Noncore OGCs/regions which are determined by pan-genome analysis are often horizontally transferred elements, such as phage, prophage, or mobile elements. For industrial applications these regions are dispensable and can even be sources

of genome instability.^{3,12–14,53} While there are other methods for identifying these regions, pan-genome analysis is a reliable complimentary tool. Pan-genome analysis can also reveal enzymatic and other systems/pathways that are present in some strains but not others⁵³ which indicates they can likely be removed. When choosing between retaining alternate systems for essential functions, biological engineers have looked at conservation of those systems across broad taxonomic levels³ as an indication of utility and we believe conservation across the pan-genome should also be considered. When specific genes/systems of known function are being targeted for removal pan-genome analysis is less useful but still good information to have. For instance, Reuß *et al.*² tried to delete region BSU07710-07820 from *B. subtilis* which was lethal. In this region, six of the 11 OGCs are core but the five noncore genes are adjacent so perhaps region BSU07750-07790 could have been successfully deleted.

Given that we believe pan-genome analysis is a useful complimentary tool for biological engineers, it is important that the pan-genome analysis used be as accurate and helpful as possible. We showed by comparing with other recent pan-genome studies for *B. subtilis* and *E. coli* that our method is more accurate for determining core OGCs/regions as validated by coverage of essential genes. Further, we believe that the PGG is valuable for confirming when noncore OGCs may be compensated for with alternate homologous OGCs at the same relative genomic location performing the same function as we showed in [Figure 3](#). The function of these noncore OGCs may be essential and should be considered appropriately.

Pan-genome studies often capture the diversity of sequenced species but fail to compare gene lists to experimentally validated essential genes lists or the results are confusing. Interestingly in *Mycoplasma*, fewer essential genes were determined with the pan-genome method compared with the laboratory experimental approach.⁵⁴ In *Pseudomonas*, only one-third of the pan-genome single copy genes had overlap with the essential genes from experimentally reduced genomic studies.⁵⁵ We showed that the core OGCs/regions from our refined PGG encompass 91% and 95% of the *E. coli* and *B. subtilis* experimentally determined essential gene lists, respectively. Both model bacterial species *E. coli* and *B. subtilis* have had many genome reduction studies performed and reviewed elsewhere.⁵⁶

Experimental verification of the essentiality of computationally predicted core OGCs or regions requires that each strain of the pan-genome study be minimized. However, it is cost prohibitive to do knockout studies on all strains of a pan-genome. One must carefully choose a single genome as a representative of the entire pan-genome for the purpose of verifying the essentiality of core regions and/or the non-essentiality of noncore regions by experimental validation. However, given the diversity of most bacterial species it is unlikely that any one strain completely captures the capabilities of the species in all environmental conditions. Further, while there are clearly core OGCs/regions associated with viability for a species, other core regions probably contribute to a lesser degree to cell viability. For example, for the purpose of biological engineering, changes in these locations may reduce fitness by slowing cell growth.

The use of a PGG for identifying core regions of a bacterium is an automatable, low-cost, rapid, and effective way to evaluate both Gram-negative and Gram-positive bacteria. This method compliments and expands upon the experimental knockout approach by including environmental diversity as a measure of what regions and OGCs are conserved across the species. The approach also overcomes the limitations of knockout studies that are specific to the strains and growth conditions used.

The *B. subtilis* WTA region provides a cautionary note for relying entirely upon core regions to determine what is safe to remove. While most non-core regions involve cassettes of genes which are entirely absent from some strains such as phage regions, sometimes orthologous replacement possibly due to homologous recombination can have functionally equivalent genes appearing to be non-core. A closer examination of the PGG can determine if a region is simply missing from some strains versus being replaced in which case further study may be needed before removal of the region. Of course, in some cases the orthologous replacement does not need to occur at the same location in the genome but that was the case for all instances we examined in *B. subtilis*.

While we showed that almost all essential genes are core OGCs and most are OGCs at the 100% threshold, the exceptions are interesting. We discussed issues such as “protective essential genes”³⁶ (such as toxin/anti-toxin gene pairs) and more distant orthologs not captured in OGCs. We did not discuss genes which might be undergoing gene loss.⁵⁷ The PGG is well suited to looking at which subset of genomes have suffered a gene loss and possible mechanisms such as gene replacement. The PGG has been used to show which genomic regions tend not to allow insertions of horizontally transferred genes¹⁹ and where metabolic cassettes can be swapped.⁵³

Acknowledgements

The authors would like to thank IARPA for sponsoring this research and would like to thank Derren Barken for his assistance in table generation.

Data availability

Underlying data

Figshare: Underlying data for 'A pan-genome method to determine core regions of the *Bacillus subtilis* and *Escherichia coli* genomes', <https://doi.org/10.6084/m9.figshare.15129636.v1>.⁵⁸

This project contains the following underlying data:

- **Table 1.** Selection of complete genomes for *B. subtilis* and *E. coli* PGGs.
- **Table 2.** Pan-genome graph statistics for *B. subtilis* and *E. coli*.
- **Table 3.** The number of deleted genes from *B. subtilis* reduced strains which are noncore versus core.
- **Table 4.** Large noncore regions which have not been deleted from any of the strains delta 6, IIG-Bs27-47-24, or PG10, PS38.

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC BY 4.0).

Extended data

Figshare: Extended data for 'A pan-genome method to determine core regions of the *Bacillus subtilis* and *Escherichia coli* genomes', <https://doi.org/10.6084/m9.figshare.15129636.v1>.⁵⁸

This project contains the following extended data:

- Supplementary Table 1
- Supplementary Table 2
- Supplementary Table 3
- Supplementary Table 4
- Supplementary Table 5
- Supplementary Table 6
- Supplementary Table 7
- Supplementary Table 8
- Supplementary Table 9
- Supplementary Table 10

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC BY 4.0).

References

- Hutchison CA, Chuang RY, Noskov VN, *et al.*: **Design and synthesis of a minimal bacterial genome.** *Science.* 2016; **351**: aad6253.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Reuß DR, Altenbuchner J, Mäder U, *et al.*: **Large-scale reduction of the *Bacillus subtilis* genome: consequences for the transcriptional network, resource allocation, and metabolism.** *Genome Res.* 2017; **27**(2): 289–299.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Reuß DR, Commichau FM, Gundlach J, *et al.*: **The Blueprint of a Minimal Cell: MiniBacillus.** *Microbiol Mol Biol Rev.* 2016; **80**(4): 955–987.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mario J, Reuß DR, Zhu B, *et al.*: ***Bacillus subtilis* and *Escherichia coli* essential genes and minimal cell factories after one decade of genome engineering.** *Microbiol.* 2014; **160**(11): 2341–2351.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kolisnychenko V, Plunkett G 3rd, Herring CD, *et al.*: **Engineering a reduced *Escherichia coli* genome.** *Genome Res.* 2002; **12**(4): 640–647.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang L, Maranas CD: **MinGenome: An *In Silico* Top-Down Approach for the Synthesis of Minimized Genomes.** *ACS Synth Biol.* 2018; **7**(2): 462–473.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kobayashi K, Ehrlich SD, Albertini A, *et al.*: **Essential *Bacillus subtilis* genes.** *Proc Natl Acad Sci U S A.* 2003; **100**: 4678–4683.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Koo BM, Kritikos G, Farelli JD, *et al.*: **Construction and Analysis of Two Genome-Scale Deletion Libraries for *Bacillus subtilis*.** *Cell Syst.* 2017; **4**: 291–305.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goodall ECA, Robinson A, Johnston IG, *et al.*: **The Essential Genome of *Escherichia coli* K-12.** *mBio.* 2018; **20**: e02096–17.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Baba T, Ara T, Hasegawa M, *et al.*: **Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection.** *Mol Syst Biol.* 2006; **2**: 2006.0008.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yamazaki Y, Niki H, Kato J: **Profiling of *Escherichia coli* Chromosome database.** *Methods Mol Biol.* 2008; **416**: 385–389.
[PubMed Abstract](#) | [Publisher Full Text](#)

12. Westers H, Dorenbos R, van Dijk JM, *et al.*: **Genome engineering reveals large dispensable regions in *Bacillus subtilis***. *Mol Biol Evol*. 2003; **20**: 2076–2090.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Wenzel M, Altenbuchner J: **Development of a markerless gene deletion system for *Bacillus subtilis* based on the mannose phosphoenolpyruvate-dependent phosphotransferase system**. *Microbiology*. 2015; **161**(10): 1942–1949.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Umenhoffer K, Fehér T, Balikó G, *et al.*: **Reduced evolvability of *Escherichia coli* MDS42, an IS-less cellular chassis for molecular and synthetic biology applications**. *Microb Cell Fact*. 2010; **9**: 38.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Csörgo B, Fehér T, Timár E, *et al.*: **Low-mutation-rate, reduced-genome *Escherichia coli*: an improved host for faithful maintenance of engineered genetic constructs**. *Microb Cell Fact*. 2012; **1**: 11.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Tettelin H, Massignani V, Cieslewicz MJ, *et al.*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”** *Proc Natl Acad Sci U S A*. 2005; **102**(39): 13950–13955. [published correction appears in *Proc Natl Acad Sci U S A*. 2005 Nov 8;102(45):16530].
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons**. *J Mol Biol*. 2001; **314**(5): 1041–1052.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: Identification of ortholog groups for eukaryotic genomes**. *Genome Res*. 2003; **13**(9): 2178–2189.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Chan AP, Sutton G, DePew J, *et al.*: **A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of *Acinetobacter baumannii***. *Genome Biol*. 2015; **16**: 143.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Tatusov RL, Galperin MY, Natale DA, *et al.*: **The COG database: a tool for genome-scale analysis of protein functions and evolution**. *Nucleic Acids Res*. 2000; **28**(1): 33–36.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Gil R, Silva FJ, Peretó J, *et al.*: **Determination of the core of a minimal bacterial gene set**. *Microbiol Mol Biol Rev*. 2004; **68**(3): 518–537.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Jordan IK, Rogozin IB, Wolf YI, *et al.*: **Essential genes are more evolutionarily conserved than are nonessential genes in bacteria**. *Genome Res*. 2002; **12**(6): 962–968.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Podell S, Gaasterland T: **DarkHorse: a method for genome-wide prediction of horizontal gene transfer**. *Genome Biol*. 2007; **8**(2): R16.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification**. *Annu Rev Microbiol*. 2001; **55**: 709–742.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Fouts DE: **Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences**. *Nucleic Acids Res*. 2006; **34**(20): 5839–5851.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Page AJ, Cummins CA, Hunt M, *et al.*: **Roary: rapid large-scale prokaryote pan genome analysis**. *Bioinformatics*. 2015; **31**(22): 3691–3693.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Vernikos GS: **A Review of Pangenome Tools and Recent Studies**. In: Tettelin H, Medini D, eds. *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Cham (CH): Springer; 2020: 89–112.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Clarke TH, Brinkac LM, Sutton G, *et al.*: **GGRASP: a R-package for selecting representative genomes using Gaussian mixture models**. *Bioinformatics*. 2018; **34**: 3032–3034.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Inman JM, Sutton GG, Beck E, *et al.*: **Large-scale comparative analysis of microbial pan-genomes using PanOCT**. *Bioinformatics*. 2019; **35**: 1049–1050.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Fouts DE, Brinkac L, Beck E, *et al.*: **PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species**. *Nucleic Acids Res*. 2012; **40**: e172.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. O’Leary NA, Wright MW, Brister JR, *et al.*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res*. 2016; **44**: D733–45.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Lan R, Reeves PR: ***Escherichia coli* in disguise: molecular origins of *Shigella***. *Microbes Infect*. 2002; **4**: 1125–1132.
[PubMed Abstract](#) | [Publisher Full Text](#)
33. Meier-Kolthoff JP, Hahnke RL, Petersen J, *et al.*: **Complete genome sequence of DSM 30083(T), the type strain (U5/41(T)) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy**. *Stand Genomic Sci*. 2014; **8**: 9: 2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Ondov BD, Treangen TJ, Melsted P, *et al.*: **Mash: fast genome and metagenome distance estimation using MinHash**. *Genome Biol*. 2016; **17**: 132.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol*. 1970; **48**: 443–453.
[PubMed Abstract](#) | [Publisher Full Text](#)
36. Commichau FM, Pietack N, Stülke J: **Essential genes in *Bacillus subtilis*: a re-evaluation after ten years**. *Mol Biosyst*. 2013; **9**(6): 1068–1075.
[PubMed Abstract](#) | [Publisher Full Text](#)
37. Koskiniemi S, Lamoureux JG, Nikolakakis KC, *et al.*: **Rhs proteins from diverse bacteria mediate intercellular competition**. *Proc Natl Acad Sci U S A*. 2013; **110**: 7032–7037.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
38. Holberger LE, Garza-Sánchez F, Lamoureux J, *et al.*: **A novel family of toxin/antitoxin proteins in *Bacillus* species**. *FEBS Lett*. 2012; **586**(2): 132–136.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Brantl S, Müller P: **Toxin-Antitoxin Systems in *Bacillus subtilis***. *Toxins*. 2019; **11**: pii: E262.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Ohshima H, Matsuoka S, Asai K, *et al.*: **Molecular organization of intrinsic restriction and modification genes *Bsu*M of *Bacillus subtilis* Marburg**. *J Bacteriol*. 2002; **184**: 381–389.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Brown S, Santa Maria Jr JP, Walker S: **Wall teichoic acids of gram-positive bacteria**. *Annu Rev Microbiol*. 2013; **67**: 313–336.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. D’Elia MA, Millar KE, Beveridge TJ, *et al.*: **Wall teichoic acid polymers are dispensable for cell viability in *Bacillus subtilis***. *J Bacteriol*. 2006; **188**: 8313–8316.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
43. Henriques AO, Glaser P, Piggot PJ, *et al.*: **Control of cell shape and elongation by the *rodA* gene in *Bacillus subtilis***. *Mol Microbiol*. 1998; **28**: 235–247.
[PubMed Abstract](#) | [Publisher Full Text](#)
44. Lazarevic V, Abellan F-X, Möller SB, *et al.*: **Comparison of ribitol and glycerol teichoic acid genes in *Bacillus subtilis* W23 and 168: Identical function, similar divergent organization, but different regulation**. *Microbiology*. 2002; **148**: 815–824.
[PubMed Abstract](#) | [Publisher Full Text](#)
45. Ahn S, Jun S, Ro H-J, *et al.*: **Complete genome of *Bacillus subtilis* subsp. *subtilis* KCTC 3135^T and variation in cell wall genes of *B. subtilis* strains**. *J Microbiol Biotechnol*. 2018; **28**: 1760–1768.
[PubMed Abstract](#) | [Publisher Full Text](#)
46. Sutton G, Fogel G, Abramson B, *et al.*: **Horizontal transfer and evolution of wall teichoic acid gene cassettes in *Bacillus subtilis* [version 1; peer review: awaiting peer review]**. *F1000Res*. 2021.
[Publisher Full Text](#)
47. Wu H, Wang D, Gao F: **Toward a high-quality pan-genome landscape of *Bacillus subtilis* by removal of confounding strains**. *Brief Bioinform*. 2020; bbaa013.
[PubMed Abstract](#) | [Publisher Full Text](#)
48. Bindal G, Krishnamurthi R, Seshasayee ASN, *et al.*: **CRISPR-Cas-mediated gene silencing reveals RacR to be a negative regulator of YdaS and YdaT toxins in *Escherichia coli* K-12**. *mSphere*. 2017; **2**: e00483–17.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Kato J, Hashimoto M: **Construction of consecutive deletions of the *Escherichia coli* chromosome**. *Mol Syst Biol*. 2007; **3**: 132.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
50. Yang ZK, Luo H, Zhang Y, *et al.*: **Pan-genomic analysis provides novel insights into the association of *E. coli* with human host and its minimal genome**. *Bioinformatics*. 2019; **35**(12): 1987–1991.
[PubMed Abstract](#) | [Publisher Full Text](#)
51. Luo H, Lin Y, Gao F, *et al.*: **DEG 10, an update of the database of essential genes that includes both protein-coding genes and**

- noncoding genomic elements.** *Nucleic Acids Res.* 2014; 42(Database issue): D574–D580.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
52. Gerdes SY, Scholle MD, Campbell JW, *et al.*: **Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655.** *J Bacteriol.* 2003; 185(19): 5673–5684.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
53. Chavda KD, Chen L, Fouts DE, *et al.*: **Comprehensive Genome Analysis of Carbapenemase-Producing *Enterobacter* spp.: New Insights into Phylogeny, Population Structure, and Resistance Mechanisms.** *mBio.* 2016; 7(6): e02093–16.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
54. Liu W, Fang L, Li M, *et al.*: **Comparative genomics of *Mycoplasma*: analysis of conserved essential genes and diversity of the pan-genome.** *PLoS One.* 2012; 7(4): e35698.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
55. Koehorst JJ, van Dam JC, van Heck RG, *et al.*: **Comparison of 432 *Pseudomonas* strains through integration of genomic, functional, metabolic and expression data.** *Sci Rep.* 2016; 6: 38699.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
56. Juhas M, Reuß DR, Zhu B, *et al.*: ***Bacillus subtilis* and *Escherichia coli* essential genes and minimal cell factories after one decade of genome engineering.** *Microbiology.* 2014; 160(Pt 11): 2341–2351.
[PubMed Abstract](#) | [Publisher Full Text](#)
57. Kunin V, Ouzounis CA: **The balance of driving forces during genome evolution in prokaryotes.** *Genome Res.* 2003 Jul; 13(7): 1589–1594.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
58. Sutton G: **PGG Core Genes - Tables F1000 version 2.xlsx.** *figshare.* Dataset. 2021.
[Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 26 May 2021

<https://doi.org/10.5256/f1000research.55083.r84280>

© 2021 Ouzounis C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Christos Ouzounis 

¹ Department of Computer Science, Aristotle University of Thessalonica, Thessalonica, Greece

² Centre for Research & Technology Hellas, Thessalonica, Greece

This extensive, complex report provides details about a new methodological approach for the detection of 'core' regions of *Bacillus subtilis* and *Escherichia coli*. Core regions are defined within the pan-genome context of conserved genomic loci for the two bacterial species, as a case study. An underlying assumption and implicit goal of the study is that the detected core regions largely correspond to 'essential' genes, as those have been determined by independent experimental methodology, with implications for synthetic engineering of bacteria. For both purposes, namely the detection of core regions and the correspondence of those to essential genes, this report is an important contribution, especially as it resolves the connection of core to essential genes. It brings to the forefront the use of pangenome analysis for synthetic biology – a factor that so far has been, to our amazement (!), ignored by biotechnologists. Solid work and a significant contribution to the field.

Major comments:

1. A general stylistic observation is that the manuscript is dense, in particular the Introduction and Methods are quite extensive and discursive, the Introduction containing multiple quotes from previous works. While this is not necessarily a bad thing, some details ("in their table 4, etc." and other quoted phrases from cited papers) could be avoided or better summarized. This level of detail is welcome for experts, but non-experts are at risk to miss the main point and the motivations for this study. A more standard style, perhaps for the first paragraph might be useful, in order to address a wider audience.
2. "While it is possible to define "essential" genes": the definition of 'essential' genes is problematic as it refers to the growth medium and general environmental conditions, as the authors correctly point out. Therefore, 'essentiality' is a functional definition. Core (conserved, species-defining) genes, on the other hand, do not rely on environmental factors but evolutionary history, therefore 'conservation' is a structural definition. Coupling those is always tricky, however as the authors state early on in their paper, the equivalence

between core and essential genes is indeed their primary hypothesis (“We expect that all truly essential genes for the species/subspecies would be a subset of the core OGCs/regions”). This should be more explicitly stated, perhaps in the first paragraph of the Introduction.

3. What advantage is provided by keeping the directionality of OGCs in the PGG? Is this purely a methodological checkpoint, i.e. improve the detection capability by reducing the number of false positive or negative hits, or is it further used in the analysis and interpretation of the results? Needs to be clarified, as it increases the complexity of the pan-genome turning a set into a graph. There is a passage “PGG refinement to ensure consistent annotation”, which alludes to the actual role of PGG.
4. Another general comment connected to the above, esp. major comment 2: the report serves a dual role as a software announcement (update) of JCVI’s pan-genome pipeline software suite, with additional elements and certain conceptual advances, as well as the comparison of the core-vs-essential sets for two of the best studied/sampled species pangenomes. This should be a bit more clearly explained perhaps. The correspondence of core to essential genes is a welcome contribution but may not be the main topic of the manuscript, just a conclusion drawn from the analysis.
5. Following major comment 4: the method does well in identifying core regions and indeed makes a convincing case for an improvement over other methods. Yet, the comparison with essential genes is an addition, but not a comparison against other methods that define core regions. As the authors decided to take this direction, as they improve over their own previous methodology, this point should be qualified appropriately. In other words, the ‘improvement’ can be shown as an incremental step over a previous protocol and explicitly shown that it is validated against ‘essential’ gene sets. If this point is not emphasized, the analysis will be seen as lacking a comparison to another ‘gold-standard’ method (experts know that there is no such thing, yet). Pages 11-12 have some elements of a comparison to another approach, this could be extended by a couple of concluding sentences. A good spot where some concluding remarks can be made might be a short paragraph before the Discussion.

Minor comments:

1. In Introduction: “We further define a pan-genome graph (PGG) to be a graph”, this should probably follow the paragraph starting “Here we present a pan-genome based calculation...” ?
2. “For *E. coli* (and *Shigella*) we downloaded 1097 complete genomes”, start a new paragraph? Using subtitles for Methods might also be a good idea, to break down the dense text into digestible sections.
3. Following minor comment 2: a mini table with three columns (filtering step, *B. subtilis*, *E. coli*) and as many rows as the filtering steps used with the number of genomes at each step might be helpful.
4. “used by Goodall” (reference 9? missing).
5. “This is done by blasting” - executing BLAST etc. / “conflicting blast” -> conflicting BLAST...

6. "to not under call core OGCs/edges", i.e. to reduce the number of potentially false negatives. Or, increase coverage.
7. for *B. subtilis*: "3419 (73.5%) core and present in all 108 genomes": this row in Table 1 should be somehow highlighted, perhaps by color or other means -- it is an important part of the study and a key result.
8. a word for missing genes in the context of potential gene loss and the possibility of including them in future steps (see PMID: 12840037¹); this is something we (and possibly others) have been trying to implement for pangenome data, without much success. Something to discuss as a partial explanation for 'key' (essential?) missing genes in certain lineages within the species pedigree, perhaps?
9. "For the 34 protein coding genes"... good yet incredibly dense paragraph, a (supplementary) table might help here.
10. Would the PGG implementation also help future studies in synteny analysis/conservation? Maybe a minor point that can be included in the discussion, with appropriate (1-2) references. A concluding short paragraph following the current one with the WTA region might be a good way to wrap up.

References

1. Kunin V, Ouzounis CA: The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* 2003; **13** (7): 1589-94 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational Biology, Biological Computation, Systems Biomedicine, Bioinformatics, Protein Structure

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Aug 2021

Granger Sutton, J. Craig Venter Institute, Rockville, USA

We thank the reviewer for the thoughtful comments and have tried to respond to all of the suggestions including the new table and supplementary table in version two of our manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 April 2021

<https://doi.org/10.5256/f1000research.55083.r83244>

© 2021 Ussery D et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Kaleb Abram

Programming Associate, DBMI, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

David Ussery 

Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Introduction:

The authors provide an adequate background literature detailing attempts by others in the field to produce minimal genomes. A variety of approaches are covered and drawbacks to these attempts are mentioned. The authors also provide sufficient explanation of the need for their approach in addition to experimental approaches. A good overview of their pan-genome graph approach is presented, along with the reasoning for their design choices for the graph.

Towards end of first paragraph in Introduction - should be *_G_*ram-negative (name for the Danish microbiologist, Hans Christian Gram)

Last sentence in last paragraph in Introduction - "Our method builds directly upon our previous pan-genome work and includes several improvements: 1) being able to *_automatically_* use only complete high-quality genomes..." Surely the previous methods could also have used only complete high-quality genomes as input? My understanding is that the advantage of this new

method is that it's now taking steps to ensure that 'bad genomes' are filtered out, and only the 'high-quality' ones are left....

Methods:

Figure 1, 2nd line: "Compute genome ANI using Mash". This doesn't make sense, as ANI and Mash are different approaches. Mash does not estimate ANI (as it is a distance). Unless the authors took the distance and subtracted it from 1 before multiplying by 100, they do not have an approximate ANI (see fastANI paper [PMID: 30504855]¹ or the Mash paper [their ref. 34] where Mash and ANI methods are compared). Further, the authors state they use type strains and ANI (presumably using the Mash derived approximation which is not ANI) to remove very closely related strains but do not specify what criteria/value was used to determine very closely related strains. Since the authors chose to use a program (GGRaSP) that uses ANI matrixes as the input, it can be assumed that either ANI values were calculated by an unspecified method, or they used transformed Mash values to approximate ANI and need to specify this transformation earlier in the methods. Either way, this should be clearly stated in the methods section, and not leave the reviewer to guess how this might have been done.

"The 132 genomes were reduced to 109 after removing..." – it is unclear what the condition for removal was. It would be helpful if this was explicitly stated (presumably an approximate ANI value between 95.73% and 97.28%). Also the authors state the minimum ANI between *B. subtilis* was 97.28% and the maximum ANI of any of the 11 other genomes to the 132 was 95.73%. The 11 genomes referenced here are unclear and the maximum ANI for the 132 is not provided. It is important to clearly bound their values, in order to enable comparison to other studies. For *E. coli* the parameters used to remove redundancy need to be explicitly stated and how the groups are collapsed (i.e. genomes A to genome B has 99% ANI value, which genomes is removed and which genome is retained?). The authors should explicitly state why they added 2 redundant genomes to the *E. coli* dataset but did not do similar additions for *B. subtilis*. While the PGG approach seems fairly good, the heavy reliance on RefSeq annotations could be problematic for other species.

Results:

The results shown in Table 1, and the bottom line is that for both *B. subtilis* and *E. coli*, the refined cores are a bit larger (and contain a larger fraction of 'essential genes' for the species). The *E. coli* core is about a third larger, going from 2200 to 3100. The latter number (3100) seems to be more consistent with what's expected for *E. coli*, based on many different experiments - historically, there has always been roughly 3000 *E. coli* genes. So from this perspective, 2218 genes seems a bit too small (and also some of the 'essential genes' were missing from the core.)

I'm curious as to whether a non-RefSeq gene annotation tool (for example, Prokka) be utilized to improve the consistency of gene calls? The specific results with number breakdowns are very confusing to read on a first pass and require very careful reading to understand the somewhat odd notation being used. This should be cleaned up to enhance readability.

Figure 2 should have a color key containing color to corresponding track to increase readability of this figure. (The same thing for Figure 4 for consistency.)

Discussion:

The discussion surrounding the issue of lab conditions and core regions is a good. In addition, the discussion around noncore OGCs/regions also shows how the proposed pan-genome analysis

could be used to identify noncore regions that could be removed that experimental results have been unable to identify. It might have been good to have a brief discussion of the phylogroup-specific cores in *E. coli* [see PMID: 33500552² - disclaimer - this is a recent publication from our group.]

The discussion section overall provides a good wrap up to the paper and summarizes how the PGG approach can be leveraged and the benefits from utilizing this approach.

References

1. Jain C, Rodriguez-R L, Phillippy A, Konstantinidis K, et al.: High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*. 2018; **9** (1). [Publisher Full Text](#)
2. Abram K, Udaondo Z, Bleker C, Wanchai V, et al.: Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups. *Commun Biol*. 2021; **4** (1): 117 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Comparative genomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Aug 2021

Granger Sutton, J. Craig Venter Institute, Rockville, USA

We thank the reviewers for their thoughtful comments and have attempted to address all of the suggestions in version 2 of our manuscript.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research