

Loyola University Chicago [Loyola eCommons](https://ecommons.luc.edu/)

[Biology: Faculty Publications and Other Works](https://ecommons.luc.edu/biology_facpubs) Faculty Publications and Other Works by **Department**

4-2017

Exploring the Diversity of Bacillus Whole Genome Sequencing Projects Using Peasant, the Prokaryotic Assembly and Annotation Tool

Jonathon Brenner Loyola University Chicago

Laurynas Kalesinskas Loyola University Chicago

Catherine Putonti Loyola University Chicago, cputonti@luc.edu

Follow this and additional works at: [https://ecommons.luc.edu/biology_facpubs](https://ecommons.luc.edu/biology_facpubs?utm_source=ecommons.luc.edu%2Fbiology_facpubs%2F85&utm_medium=PDF&utm_campaign=PDFCoverPages)

C Part of the Biology Commons

Author Manuscript

This is a pre-publication author manuscript of the final, published article.

Recommended Citation

Brenner, Jonathon; Kalesinskas, Laurynas; and Putonti, Catherine. Exploring the Diversity of Bacillus Whole Genome Sequencing Projects Using Peasant, the Prokaryotic Assembly and Annotation Tool. BioRxiv, , : 1-13, 2017. Retrieved from Loyola eCommons, Biology: Faculty Publications and Other Works, <http://dx.doi.org/10.1101/132084>

This Article is brought to you for free and open access by the Faculty Publications and Other Works by Department at Loyola eCommons. It has been accepted for inclusion in Biology: Faculty Publications and Other Works by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/) © The Authors 2017

peer-reviewed) is the author/funder. It is made available under a [CC-BY-NC 4.0 International license.](http://creativecommons.org/licenses/by-nc/4.0/) bioRxiv preprint first posted online Apr. 28, 2017; doi: [http://dx.doi.org/10.1101/132084.](http://dx.doi.org/10.1101/132084) The copyright holder for this preprint (which was not

Exploring the Diversity of Bacillus whole genome sequencing projects using Peasant, the Prokaryotic Assembly and Annotation Tool

しょくこく Jonathon Brenner⁺, Laurynas Kalesinskas^{2,3}, Catherine Putonti^{2,2,3,4}*
1 Department of Computer Science, Loyola University Chicago, 820
60611
2 Department of Biology, Loyola University Chicago, 1032 W Sherida
3 Bioinf

1 Department of Biology, Loyola University Chicago, 1032 W Sheridan Road, Chicago, IL USA 60660
1 Department of Biology, Loyola University Chicago, 1032 W Sheridan Road, Chicago, IL USA 60660
1 Department of Microbiology a 2 Depa
3 Bioin
4 Depa
Maywo
* Corre 2 Department of Microbiology and University Chicago, 1032 W Sheridan Road, Chicago, IL USA 606604
4 Department of Microbiology and Immunology, Stritch School of Medicine, Loyola University Chica
Maywood, IL, USA 60153
* Co 3 Biology and Immunology, Stritch School of Medicine, Loyola University Chicago, 1032 Maywood, IL, USA 60153
Maywood, IL, USA 60153
* Corresponding Author
Author Email Addresses: Maywood, IL, USA 60153
* Corresponding Author
Author Email Addresses:
JB: ibrenner@luc.edu

Maynology

* Corresponding Author

Author Email Addresses:

JB: jbrenner@luc.edu ,
|
|
|

Author Email Addresses:
B: jbrenner@luc.edu
LK: lkalesinskas@luc.edu
CP: cputonti@luc.edu ノ
」
(

Author Email Addresser
JB: jbrenner@luc.edu
LK: lkalesinskas@luc.edu
CP: cputonti@luc.edu

LK: Ikalesinskas@luc.edu
LK: Ikalesinskas@luc.edu
CP: cputonti@luc.edu LA: ANDERSINSKAS CHALES
CP: cputonti@luc.edu CP: cputonti@luc.edu

ABSTRACT

Background: The persistent decrease in cost and dimetally or whole genome sequencing or incrobial
organisms has led to a dramatic increase in the number of species and strains characterized from a w
variety of environments variety of environments. Microbial genome sequencing can now be conducted by small laboratories and
as part of undergraduate curriculum. While sequencing is routine in microbiology, assembly, annotation
and downstream anal as part of undergraduate curriculum. While sequencing is routine in microbiology, assembly, annotation
and downstream analyses still require computational resources and expertise, often necessitating
familiarity with progr

and downstream analyses still require computational resources and expertise, often necessitating
familiarity with programming languages. To address this problem, we have created a light-weight, user-
friendly tool for the familiarity with programming languages. To address this problem, we have created a light-weight,
friendly tool for the assembly and annotation of microbial sequencing projects.
Results: The Prokaryotic Assembly and Annot from NCBI, assembled and annotated on a single desktop computer. From the assemblies and **Results:** The Prokaryotic Assembly and Annotation Tool, Peasant, automates the quality control, genome assembly, and annotation for microbial sequencing projects. To all annotations can be generated by Peasant without the Results: The Frokaryotic Assembly and Annotation Tool, Feasant, automates the processes of read
quality control, genome assembly, and annotation for microbial sequencing projects. High-quality
assemblies and annotations ca The assemblies and annotations can be generated by Peasant without the need of programming experible.

high-performance computing resources. Furthermore, statistics are calculated so that users can evertheir sequencing pro high-performance computing resources. Furthermore, statistics are calculated so that users can evaluate
their sequencing project. To illustrate the computational speed and accuracy of Peasant, the SRA
records of 322 Illumi their sequencing project. To illustrate the computational speed and accuracy of Peasant, the SRA
records of 322 Illumina platform whole genome sequencing assays for *Bacillus* species were retrieved
from NCBI, assembled an records of 322 Illumina platform whole genome sequencing assays for *Bacillus* species were retrictiom NCBI, assembled and annotated on a single desktop computer. From the assemblies and annotations produced, a comprehensi

records of 322 manning platform whole genome sequencing assays for *buchins* species were retrieved
from NCBI, assembled and annotated on a single desktop computer. From the assemblies and
annotations produced, a comprehen From NCBI, and an anti-time and annualism and annotations producted, a comprehensive analysis of the diversity of over 200 high-quality samp
conducted, looking at both the 16S rRNA phylogenetic marker as well as the *Bacil* annotated, looking at both the 16S rRNA phylogenetic marker as well as the *Bacillus* core genome.
 Conclusions: Peasant provides an intuitive solution for high-quality whole genome sequence assembly

and annotation for Conducted, looking at both the 16S rivik phylogenetic marker as well as the *buchits* one genome.
 Conclusions: Peasant provides an intuitive solution for high-quality whole genome sequence assem

and annotation for user Conclusions: Peasant provides an intuitive solution for high-quality whole genome sequence assembly
and annotation for users with limited programing experience and/or computational resources. The
analysis of the *Bacillus* analysis of the *Bacillus* whole genome sequencing projects exemplifies the utility of this tool.
Furthermore, the study conducted here provides insight into the diversity of the species, the largest
such comparison conduc analysis of the Bacillus whole genome sequencing projects exemplines the utility of this tool.
Furthermore, the study conducted here provides insight into the diversity of the species, the
such comparison conducted to date Furthermore, the study conducted to date.
Furthermore, the study comparative study of the study of the specifical study of the specific section of the specific of the specific section of the specific specific section of th

Keywords: genome assembly, genoments
Keywords: genome assembly, genoments \overline{a} Keywords: genome assembly, genome annotation, automated pipeline, Bacillus comparative genomics

BACKGROUND

wast array of species, significantly expanding our view of genetic diversity on Earth [1]. With the adven
of second- and third-generation sequencing platforms, the capabilities of DNA sequencers far surpasse
that of the d of second- and third-generation sequencing platforms, the capabilities of DNA sequencers far surpassed
that of the decades prior. The ever-decreasing cost of sequencing has spurred the transition of genome
sequencing proje that of the decades prior. The ever-decreasing cost of sequencing has spurred the transition of genome
sequencing projects from a select few facilities to individual laboratories, and even curricular activities
(e.g. [2,3]

sequencing projects from a select few facilities to individual laboratories, and even curricular activities (e.g. [2,3]). With this transition in technology and increase in throughput, however, a major overhaul in the bioi (e.g. [2,3]). With this transition in technology and increase in throughput, however, a major overhaul in
the bioinformatic solutions for genome assembly was required.
Software tools for the processes of genome assembly an (e.g. [2,3]). This interactions for genome assembly was required.
Software tools for the processes of genome assembly and annotation are numerous. In addition to
reference guided assembly strategies, *de novo* assemblers h Software tools for the processes of genome assembly and annot
reference guided assembly strategies, *de novo* assemblers have
Velvet [5], SOAPdenovo [6], and SPAdes [7]. Additionally, newer
Nanopore, have incited the devel reference guided assembly strategies, *de novo* assemblers have been created including, e.g. Abyss [Velvet [5], SOAPdenovo [6], and SPAdes [7]. Additionally, newer long read platforms, such as PacBid Nanopore, have incited reference guided assembly strategies, at novo assemblers have been created including, e.g. Abyss [4],
Velvet [5], SOAPdenovo [6], and SPAdes [7]. Additionally, newer long read platforms, such as PacBio ar
Nanopore, have in Nanopore, have incited the development of a new class of genome sequence assemblers, such as Canu

[8] and miniasm [9]. In parallel, additional software solutions have been developed for read quality

filtering (e.g. FASTQ [8] and miniasm [9]. In parallel, additional software solutions have been developed for read quality
filtering (e.g. FASTQC [10], Trimmomatic [11], and Sickle [12]), assembly quality assessment (e.g. QUAST
[13] and REAPR [[8] and REAPR [10], Trimmomatic [11], and Sickle [12]), assembly quality assessment (e.g. QU

[13] and REAPR [14]), and genome scaffolding (e.g. SSPACE [15] and SCARPA [16]). In contrast to the

previously mentioned assemb Filtering (Fig. 1947), and genome scaffolding (e.g. SSPACE [15] and SCARPA [16]). In contrast to the
previously mentioned assembly tools, which are largely local command line solutions, annotation
solutions are more preval [13]), and assembly tools, which are largely local command line solutions, annotation
solutions are more prevalent as web-based solutions, e.g. NCBI's Prokaryotic Genome Annotation
Pipeline [17], BASys [18], RAST [19], IGS previously are more prevalent as web-based solutions, e.g. NCBI's Prokaryotic Genome Annotation
Pipeline [17], BASys [18], RAST [19], IGS [20], IMG [21], and Genix [22]. Nevertheless, the standalo
solution Prokka [23] is i Pipeline [17], BASys [18], RAST [19], IGS [20], IMG [21], and Genix [22]. Nevertheless, the standalor solution Prokka [23] is increasingly popular; the ability to annotate genomes locally has the benefit protecting sensiti solution Prokka [23] is increasingly popular; the ability to annotate genomes locally has the benefit of protecting sensitive data and often completes annotations more rapidly than web-based tools.
Moreover, a local soluti protecting sensitive data and often completes annotations more rapidly than web-based tools.
Moreover, a local solution provides users greater control on how annotations are performed. While the act salist of solution pro

protection and inclusive list of software spreater control on how annotations are performed. Whis a far from an inclusive list of software solutions that have emerged, it is a testament to the famolecular biology is increa is a far from an inclusive list of software solutions that have emerged, it is a testament to the fact that
molecular biology is increasingly integrating computational approaches [24].
While DNA sequencing is currently mor is a factor and provided approaches [24].

While DNA sequencing is currently more accessible, well-trained individuals that can conduct the

bioinformatics workflow of a genome sequencing project are not as plentiful. Seve molecular biology is increasingly analysing computational approaches [24].
While DNA sequencing is currently more accessible, well-trained individuals t
bioinformatics workflow of a genome sequencing project are not as ple bioinformatics workflow of a genome sequencing project are not as plentiful. Several pipelines ha
been developed to facilitate the process of assembly and quality control (QC), including A5 [25] ar
RAMPART [26]. Recently, been developed to facilitate the process of assembly and quality control (QC), including A5 [25] and
RAMPART [26]. Recently, tools have also been built to automate the entire processes of genome
assembly and annotation, e. RAMPART [26]. Recently, tools have also been built to automate the entire processes of genome
assembly and annotation, e.g. MEGAnnotator [27], MyPro [28], iMetAMOS [29], and the web-based
pipeline PATRIC [30]. These all-in assembly and annotation, e.g. MEGAnnotator [27], MyPro [28], iMetAMOS [29], and the web-bas
pipeline PATRIC [30]. These all-in-one solutions integrate existing software (such as the assemble
tools, and annotation tools pre pipeline PATRIC [30]. These all-in-one solutions integrate existing software (such as the assemblers, tools, and annotation tools previously listed) in a single tool. Thus, users only need knowledge concerning the single t phenols, and annotation tools previously listed) in a single tool. Thus, users only need knowledge
concerning the single tool; the requirements and terminology of the individual software within are
concealed from the user. concerning the single tool; the requirements and terminology of the individual software within
concealed from the user. While easy to use, installation of the individual components package
pipeline can present significant

concealed from the user. While easy to use, installation of the individual components packaged in t
pipeline can present significant challenges and/or require access to high performance computing
resources.
Herein, we pres pipeline can present significant challenges and/or require access to high performance computing
resources.
Herein, we present a new automated pipeline for bacterial genome assembly and annotation called
Peasant (available phenomental genome assembly and annotation callenges are therein, we present a new automated pipeline for bacterial genome assembly and annotation callend Peasant (available at https://github.com/jlbren/peasant). Raw reads Herein, we
Peasant (av
for quality
control of t
and rRNAs
First, we sa
automated
ease of use Peasant (available at <u>https://github.com/jlbren/peasant</u>). Raw reads, supplied by the user, are filter
for quality and assembled by Peasant; users can select from available filters to automate the quality
control of their for quality and assembled by Peasant; users can select from available filters to automate the quality
control of their assemblies. This assembly is then annotated, identifying protein coding genes, tRNAs,
and rRNAs using a control of their assemblies. This assembly is then annotated, identifying protein coding genes, tRNAs
and rRNAs using a local database. The motivation behind the development of this tool is three-fold.
First, we saw a need and rRNAs using a local database. The motivation behind the development of this tool is three-fold.
First, we saw a need for a robust, yet user-friendly solution for assembly, annotation, and reporting w
automated QC optio First, we saw a need for a robust, yet user-friendly solution for assembly, annotation, and reporting
automated QC options. As such, available assembly and annotation software were evaluated for the
ease of use and precisi First, We say a robust of a robust, yet user-friendly solution for a root ware were evaluated for their ease of use and precision. Peasant integrates several QC steps throughout the process. Second, while ease of use and p and a unit of the United States of use and precision. Peasant integrates several QC steps throughout the process. Second, while
ease of use and precision. Peasant integrates several QC steps throughout the process. Second, ease of use and precision. Peasant integrates several QC steps throughout the process. Second, while

the user's ability to fine-tune the process for their study. Thus, Peasant was designed to include
flexibility while necessitating only resources now commonplace in laboratories. Third, raw sequence
data provides an unbias flexibility while necessitating only resources now commonplace in laboratories. Third, raw seque
data provides an unbiased representation of the organism sequenced: published assemblies may
produced by outdated tools or ge data provides an unbiased representation of the organism sequenced: published assemblies may be
produced by outdated tools or generated to meet criteria often unknown to the downstream user.
Comparative genomic studies may produced by outdated tools or generated to meet criteria often unknown to the downstream user.
Comparative genomic studies may thus benefit by returning to raw data. Peasant provides a means t
feasibly process numerous gen comparative genomic studies may thus benefit by returning to raw data. Peasant provides a means
feasibly process numerous genomes in a uniform manner. To illustrate the utility of Peasant, the SR
records for 322 *Bacillus* Feasibly process numerous genomes in a uniform manner. To illustrate the utility of Peasant, the SRA
records for 322 *Bacillus* whole genome sequencing projects were retrieved from NCBI [31]. Each record
was analyzed by Pe Fecords for 322 *Bacillus* whole genome sequencing projects were retrieved from NCBI [31]. Each records for 322 *Bacillus* whole genome sequencing projects were retrieved from NCBI [31]. Each records for 322 *Bacillus* who

IMPLEMENTATION

records for 322 *Bacillus* whole genome sequencing projects were retrieved from NCBI [31]. Each record
was analyzed by Peasant, facilitating downstream analysis of the diversity of *Bacillus* species.
IMPLEMENTATION
The Pr was analyzed by Peasant, facilitating downstream analysis of the diversity of Buchlus species.
IMPLEMENTATION
The Prokaryotic Assembly and Annotation Tool - Peasant - automates assembly and annotation
existing tools as wel $\begin{array}{c} 1 \\ 1 \\ 2 \end{array}$

Assembly Steps: Sequencing reads, either
 Assembly Steps: Sequencing reads, either

conducted using Sickle [12]. Assembly is

it is (1) frequently used in WGS studies,

projects (e.g. [27]), and (3) performs we

SPAdes ノくードミセ Assembly Steps: Sequencing reads, either single or paired-end, are mist processed and read QC is
conducted using Sickle [12]. Assembly is next performed using SPAdes [7]. SPAdes was selected he
it is (1) frequently used in it is (1) frequently used in WGS studies, (2) it often outperforms other assemblers on microbial genomic
projects (e.g. [27]), and (3) performs well with feasible demands on RAM. Moreover, current versions of
SPAdes includ it is (e.g. [27]), and (3) performs well with feasible demands on RAM. Moreover, current versions of
SPAdes include the capability to conduct hybrid assemblies of reads from long and short read
technologies. The code was d projects include the capability to conduct hybrid assemblies of reads from long and short read
spAdes include the capability to conduct hybrid assemblies of reads from long and short read
technologies. The code was develop SPAD technologies. The code was developed to easily accommodate other assemblers, and future vertilis tool are anticipated to reflect new/additional tools in the field if they can provide quality and short reads from the f this tool are anticipated to reflect new/additional tools in the field if they can provide quality assemblies
this tool are anticipated to reflect new/additional tools in the field if they can provide quality assemblies this tool are anticipated to reflect new/additional tools in the field if they can provide quality assemblies at low cost (in the word size k=33, 55, 77, 99 and 127. Alternatively, users can supply their own genome
assembly file, removing the assembly process. Assemblies can then be filtered based upon user supplied
criteria, inc

Minimum SPAdes cov value
(SPAdes assemblies only)
ined, identifying rRNA, tRN,
through BLASTn queries to 5
cluded with the download on RNAmmer Server [33] +RP -cov, --min_SPAdes_cov Minimum SPAdes cov value (Spanning Contribution

(Spanning Tangle State

(Spanning BLAST n queries to

(Spanning Glimmer Server [33]

(Spanning Glimmer Server [33]

(Spanning Glimmer Server State)

(Spanning Glimmer Server State) **Annotation Steps:** Contigs are next exa
16S, and 23S rRNA regions are detected
databases. The three rRNA databases, i
github repository), were curated from t
tRNAscan-SE [34]. Coding regions are p
script creates a trainin しょうしょう Annotation Steps: Contigs are next examined, identifying river, there, and procent coung regions. 5S, 16S, and 23S rRNA regions are detected through BLASTn queries to 5S, 16S, and 23S blast sequence databases. The three rR databases. The three rRNA databases, included with the download of Peasant (available through the github repository), were curated from the RNAmmer Server [33]. tRNA sequences are predicted usir tRNAscan-SE [34]. Coding re github repository), were curated from the RNAmmer Server [33]. tRNA sequences are predicted using
tRNAscan-SE [34]. Coding regions are predicted by running Glimmer with the g3-iterated script [35]; t
script creates a train github repository). The symptom of the galaxies are predicted by running Glimmer with the g3-iterated script [35]; the script creates a training set from the genome assembly, builds an Interpolated Context Model (ICM) from script creates a training sequences, runs Glimmer, creates a position weight matrix from the predicted (ICM) from the training sequences, runs Glimmer, creates a position weight matrix from the predicted sequences, and run From the training sequences, runs Glimmer, creates a position weight matrix from the predicted
sequences, and runs Glimmer again to generate the final coding region predictions. These predicted
protein coding sequences are sequences, and runs Glimmer again to generate the final coding region predictions. These prediction coding sequences are subsequently assigned functionality by BLAST queries to an annotagene database. Users can specify the Fraction coding sequences are subsequently assigned functionality by BLAST queries to an annotated gene database. Users can specify the threshold for ascertaining homologous genes by specifying the minimum percent identity provided a sequence database. Users can specify the threshold for ascertaining homologous genes by specifying the minimum percent identity, query coverage, and/or blast bitscore; otherwise, default values are used 70%, 70% gene database, or blast bitscore; otherwise, default values are used
70%, 70%, and 50, respectively. A precomputed gene database, created from all archived RefSeq
bacterial genomes [36,37], is available from the github rep mow, 70%, 70%, and 50, respectively. A precomputed gene database, created from all archived RefSeq
bacterial genomes [36,37], is available from the github repository. Users can also create their own
custom database for ann Facterial genomes [36,37], is available from the github repository. Users can also create their own
custom database for annotations by using the script make_Peasant_db.py (also available through
Peasant github repository).

bacterial genomes for annotations by using the script make_Peasant_db.py (also available through t
Peasant github repository). This script will produce Peasant formatted databases from user-supplie
sequence (ffn format) an Peasant github repository). This script will produce Peasant formatted databases from user-supplied
sequence (ffn format) and annotation information (ptt format).
Outputs: In addition to the assembled genome file, severa sequence (ffn format) and annotation information (ptt format).
 Outputs: In addition to the assembled genome file, several other files are generated by the tool, as

indicated in Figure 1. The log file includes assembly **Outputs:** In addition to the assembled genome file, several othe indicated in Figure 1. The log file includes assembly statistics ger parameters used in the analysis, sequence file names, and commexternal programs. Identi **Culture:** In addition to the assembled genome file, several other files are generated by the tool, as indicated in Figure 1. The log file includes assembly statistics generated by Peasant, as well as the parameters used i parameters used in the analysis, sequence file names, and commands implemented when calling
external programs. Identified rRNA sequences, tRNA predictions, and protein coding sequences (a
nucleotide and amino acid sequence parameter and programs. Identified rRNA sequences, tRNA predictions, and protein coding sequences (and analy analy interest) for the assembled genome are also written to file. Finally, if the is generated listing annotatio

external and an animo acid sequences) for the assembled genome are also written to file. Finally, a CSV
file is generated listing annotation details, including, the protein coding region's location within the
contig sequen File is generated listing annotation details, including, the protein coding region's location within the contig sequence, the predicted gene name, and product information.
Tool specifics: This tool was developed in Python file is generated the predicted gene name, and product information.
 Tool specifics: This tool was developed in Python 2.7 and can be run on a UNIX/Mac OSX system thre terminal. As existing software tools are integrated Tool specifics: This tool was developed in Python 2.7 and can be run of the terminal. As existing software tools are integrated into the tool, it installed and included in the system's PATH environmental variable. In is au Tool specifics. This tool was developed in Fython 2.7 and can be run on a UNIX/Mac OSX system through
the terminal. As existing software tools are integrated into the tool, it is required that these tools are
installed and installed and included in the system's PATH environmental variable. Inclusion of all necessary package
is automatically checked by Peasant upon execution and the user will be notified of any missing
distant that the tools is automatically checked by Peasant upon execution and the user will be notified of any missing

is automatically checked by Peasant upon execution and the user will be notified of any missing

is automatically checked by is automatically checked by Peasant upon execution and the user will be notified of any missing

Climmer [35], BLAST+ [39] and BBMAP [32]. The documentation for Peasant includes links to these
and their respective installation instructions and is available with the aforementioned scripts at
https://github.com/jlbren/p and their respective installation instructions and is available with the aforementioned scripts at https://github.com/jlbren/peasant. In addition to parameters required for executing the assembly and annotation of the sequ https://github.com/jlbren/peasant. In addition to parameters required for executing the asseminanotation of the sequencing reads, Peasant includes a parameter in which the user can specify number of threads to use during B montation of the sequencing reads, Peasant includes a parameter in which the user can specify the
number of threads to use during BLAST searches. This parameter is integrated into calls to both the
SPAdes assembler and ann mumber of threads to use during BLAST searches. This parameter is integrated into calls to both the
SPAdes assembler and annotation process and thus can significantly expedite the execution of Peasant
RESULTS AND DISCUSSIO

RESULTS AND DISCUSSION

The Peasant software

SPAdes assembler and annotation process and thus can significantly expedite the execution of Peas:
RESULTS AND DISCUSSION
The Peasant software
Peasant is specifically tailored for whole genome assembly and annotation of an SPADES AND DISCUSSION
The Peasant software
Peasant is specifically tailored for whole genome assembly and annotation of an isolated bacterium.
Execution of Peasant is through the command line in which users indicate their 「「「「「」 「」 「」 「」 「」 「」 Execution of Peasant is through the command line in which users indicate their read files, database for annotation, and output location. This automated process provides feedback to the user via the log fil which can indica annotation, and output location. This automated process provides feedback to the user via the log file
which can indicate a poor-quality sequencing run and/or DNA prep. In addition, users can specify
parameter values to fi which can indicate a poor-quality sequencing run and/or DNA prep. In addition, users can specify
parameter values to filter their assembly (Table 1). These filters can remove low coverage contigs whic
may represent sequenc parameter values to filter their assembly (Table 1). These filters can remove low coverage contigs
may represent sequences of contaminants. Similarly, the existing pipelines MEGAnnotator [27] ar
PATRIC [30] include assembl may represent sequences of contaminants. Similarly, the existing pipelines MEGAnnotator [27] and
PATRIC [30] include assembly filters for size and coverage; neither MyPro [28] nor iMetAMOS [29]
include such functionality. may payring and the simulation of the simple pipeline. The simple pipeline include such functionality. The simplicity of Peasant provides mechanisms for even non-technical us to analyze their data: users can easily test di Phical de such functionality. The simplicity of Peasant provides mechanisms for even non-technical u
to analyze their data: users can easily test different filters on assembled contigs and evaluate the q
of their sequencin include such a may be their data: users can easily test different filters on assembled contigs and evaluate the qualit
of their sequencing data. Once an assembly is generated, users do not need to run the tool from start t of their sequencing data. Once an assembly is generated, users do not need to run the tool from start to finish to test filters or rerun annotations; Peasant can also accept a FASTA or multi-FASTA format assembly as a user finish to test filters or rerun annotations; Peasant can also accept a FASTA or multi-FASTA format
assembly as a user input and thus bypass the assembly step of the pipeline. Peasant's run-time is
dependent upon the number assembly as a user input and thus bypass the assembly step of the pipeline. Peasant's run-time is
dependent upon the number of input contigs and the size of the annotation database. Using the r
threading parameter, this ca dependent upon the number of input contigs and the size of the annotation database. Using the r
threading parameter, this can be expedited (dependent upon the number of threads supported b
user's machine).
Case Study: Asse

Case Study: Assembly and annotation of publicly available Bacillus spp. whole genome sequencing projects

generated using an Illumina instrument were selected and downloaded. Metadata for each sample was threading parameter, this can be expedited (dependent upon the number of threads supportions), the
user's machine).
**Case Study: Assembly and annotation of publicly available Bacillus spp. whole genome sequencing
projects** *user Emmerine).*
 **Case Study: Asse

projects**

All *Bacillus* SRA fi

generated using :

manually inspect

lists the 322 proje All Bacillus STAR files were identified using NCBI's STAR Kuni Selector [40]. Projects sen-identified as 'WGS'
generated using an Illumina instrument were selected and downloaded. Metadata for each sample was
manually insp manually inspected to verify it was for a *Bacillus* species and from single isolates. Supplemental Table 1
lists the 322 projects retrieved. Peasant then processed each sample with a single filter specified, which
removed manually inspected to verify it was for a Bacilius species and from single isolates. Supplemental Table 1
lists the 322 projects retrieved. Peasant then processed each sample with a single filter specified, which
removed c removed contigs less than 1000 bp in length. As part of the Sickle read QC process, we restricted
trimmed reads to be greater than or equal to 100 nucleotides in length. This read length threshold
automatically removed 14 Firmmed reads to be greater than or equal to 100 nucleotides in length. This read length threshol
automatically removed 14 samples (as these studies generated reads <100) as well as an addition
samples that did not meet th automatically removed 14 samples (as these studies generated reads <100) as well as an additional
samples that did not meet this threshold post-trimming. Read QC is a critical step in genome assem
only the existing pipelin samples that did not meet this threshold post-trimming. Read QC is a critical step in genome assembly;
only the existing pipeline MEGAnnotator [27] includes read QC as part of their automated process. Thu
in total, 249 *B* only the existing pipeline MEGAnnotator [27] includes read QC as part of their automated process. Thus
in total, 249 *Bacillus* isolates were fully processed. Figure 2 provides an overview of the data sets, which
included in total, 249 *Bacillus* isolates were fully processed. Figure 2 provides an overview of the data sets, which
included both single and paired-end reads through QC and assembly. Full statistics for each individual
included included both single and paired-end reads through QC and assembly. Full statistics for each individual
included both single and paired-end reads through QC and assembly. Full statistics for each individual included both single and paired-end reads through QC and assembly. Full statistics for each individual

וו
ו
t Figure 2. Overview of the 249 Bachnus samples processed through reduced of and assembly. Each dot
represents a single sample and the median number of reads or contigs across all samples is shown l
bar.
He uniform processin Free uniform processing of the *Bacillus* samples facilitates subsequent comparative analyses. For each sample processed, Peasant generates a file containing the rRNA (5S, 16S and 23S) gene sequences identified. The 16S rR The
The
sam
iden
less
not d
subs Teilres The uniform processing of the Bacilius samples facilitates subsequent comparative analyses. For each sample processed, Peasant generates a file containing the rRNA (55, 16S and 23S) gene sequences identified. The 16S rRNA identified. The 16S rRNA gene sequences were extracted from each sample; sequences with a leng
less than 1000 nucleotides were omitted from further analysis. It is worth noting that some sample
not contain a recognized 16S identified. The 16S rRNA gene control of the treating in the transformation is the 16S rRNA sequence and/or a partial sequence and thus were not included in subsequent analyses. In total 218 16S rRNA gene sequences were al not contain a recognized 16S rRNA sequence and/or a partial sequence and thus were not included in
subsequent analyses. In total 218 16S rRNA gene sequences were aligned with 30 *Bacillus* RefSeq [36]
16S rRNA gene sequenc subsequent analyses. In total 218 16S rRNA gene sequences were aligned with 30 *Bacillus* RefSeq [36]
16S rRNA gene sequences (retrieved from NCBI, listed in Supplemental Table 3) and two out-group
sequences – *Paenibacill* subsequent analyses. In total 218 163 nKM gene sequences were aligned with 30 Bachlus Refseq [36]
16S rRNA gene sequences (retrieved from NCBI, listed in Supplemental Table 3) and two out-group
sequences – *Paenibacillus p* sequences – *Paenibacillus polymyxa* SC2 (NC_014622) and *Lactobacillus casei* BLS23 (NC_010999). Talignment was used to derive a phylogenetic tree using the tool FastTree [41]. *L. casei* was specified the root of the tre sequences – Paenibacillus polymyxa SC2 (NC_014622) and Lactobacillus case/ BLS23 (NC_010999). This alignment was used to derive a phylogenetic tree using the tool FastTree [41]. *L. casei* was specified as the root of the alignment was used to derive a phylogenetic tree using the tool FastTree [41]. L. casei was specified as
the root of the tree shown in Figure 3. Branches without a label are representative of individual sample
assembled in the root of the tree shown in Figure 3. Branches with a label are representative of individual sampled in this study. assembled in this study.

Figure 3. 16S rivide sequence phylogeny of Bacillus sequencing projects and Bacillus spp. Refsequences (indicated in blue). The 16S rRNA gene sequences for *L. casei* (red) and *P. polymyxa* (green) are included as outgrou representatives (indicated in blue). The 16S rRNA gene sequences for L. caser (red) and P. polymyxa
(green) are included as outgroups. The tree was derived using FastTree [41], an approximate maxim
likelihood method, with clade shown in Figure 3 between the two outgroup species labels. This branch is separate from both the outlier
outlier L. casei and the rest of the bacilli sequences. The 16 sequences within this clade were selected As previous studies have found, distinguishing bacilli species via molecular methods (such as
is not possible for some taxa, e.g. the *B. cereus* sensu lato group which includes *B. cereus*, *B.*
thuringiensis, *B. anthrac* ノーゼ じょう is not possible for some taxa, e.g. the *B. cereus* sensu lato group which includes *B. cereus*, *B.*
thuringiensis, *B. anthracis*, *B. mycoides*, *B. pseudomycoides*, and *B. weihenstephanensis* [43-46]. These
individual is not possible for some taxa, e.g. the B. cereus sensu lato group which includes B. cereus, B.
thuringiensis, B. anthracis, B. mycoides, B. pseudomycoides, and B. weihenstephanensis [43-4
individual species were not found thuringiensis, B. anthracis, B. mycones, B. pseudomycones, and B. wemenstephanensis [43-46]. These individual species were not found to be monophyletic in the 16S rRNA tree of Figure 3. Of interest is the clade shown in Fi clade shown in Figure 3 between the two outgroup species labels. This branch is separate from both the outlier *L. casei* and the rest of the bacilli sequences. The 16 sequences within this clade were selected and queried outlier *L. casei* and the rest of the bacilli sequences. The 16 sequences within this clade were selected
and queried via NCBI's BLAST web interface against the '16S ribosomal RNA sequences (Bacteria and
Archaea)' databas outier L. cuser and the rest of the bacilli sequences. The 16 sequences within this clade were selected
and queried via NCBI's BLAST web interface against the '16S ribosomal RNA sequences (Bacteria and
Archaea)' database. Archaea)' database. All 16 sequences produced hits to taxa other than *Bacillus* (Supplemental Table 4 including hits to 16S rRNA gene sequences from *Clostridium*, *Francisella*, and *Methylobacterium* specfive of these 1 Archaea)' database. All 16 sequences produced hits to taxa other than Bachlus (Supplemental Table 4),

Fincluding hits to 16S rRNA gene sequences from *Clostridium*, *Francisella*, and *Methylobacterium* species

Five of t including hits to 16S rivide gene sequences from clostridium, Francisem, and Methylobacterium species.
Five of these 16 sequences are from samples in which there is only one recognized 16S rRNA sequence;
thus, based on 16S Five of thus, based on 16S rRNA sequence alone, we conclude that the isolate sampled is not a *Bacillus* species.
The Peasant annotation from the remaining 11 sequences within this clade are from samples in which
more than thus, based on 16S rRNA sequence alone, we conclude that the isolate sampled is not a Bachilas species.
The Peasant annotation from the remaining 11 sequences within this clade are from samples in which
more than one 16S r more than one 16S rRNA gene sequence was identified. This 16S sequence was *Bacillus* in origin (detain listed in Supplemental Table 4) suggesting the presence of a contaminant within the samples sequence This is certainly more than one 16S rRNA gene sequence was identified. This 16S sequence was Bachins in origin (details listed in Supplemental Table 4) suggesting the presence of a contaminant within the samples sequenced This is certainly This is certainly the case for SRA record SRR2120204 with one 16S sequence producing a hit to

This is certainly the case for SRA record SRR2120204 with one 16S sequence producing a hit to

This is certainly the case for S This is certainly the case for SRR2120204 with one 16S sequence production one 16S sequence production and to
The cord SRR2120204 with one 16S sequence production of the cord SRR2120204 with the cord SRR2120204 with the
Th

Cereus ATCC 14579 in Figure 3. The assembly produced for the SRR2120204 read set exceeds 13Mk
which well exceeds the genome size of either taxa.
While there are presently >1500 assemblies in GenBank for *Bacillus* specie cereus ATCC 14579 in Figure 3. The assembly produced for the SRR2120204 read set exceeds 13Mbp,
which well exceeds the genome size of either taxa.
While there are presently >1500 assemblies in GenBank for *Bacillus* specie While there are presently >1500 assemblies in Genl
(https://www.ncbi.nlm.nih.gov/assembly/?term=tx
genome sequencing projects for *Bacillus* species executation
data quickly, easily, and uniformly from many studion
inivalu While there are presently 21500 assembles in Genbank for Buchus species
(https://www.ncbi.nlm.nih.gov/assembly/?term=txid1386[Organism:expl]), igenome sequencing projects for *Bacillus* species exceeds 2000. The ability to (https://www.ncbi.nlm.nih/2010). The ability to process raw sequend
data quickly, easily, and uniformly from many studies, such as the *Bacillus* spp. examined here, pr
an invaluable tool for researchers conducting compar genome sequencing projects for *bacillus* species exceeds 2000. The ability to process raw sequencing
data quickly, easily, and uniformly from many studies, such as the *Bacillus* spp. examined here, provid
an invaluable t data quickly, easily, and uniformly from many studies, such as the Bachlus spp. examined here, provides
an invaluable tool for researchers conducting comparative genomics studies. Given the wealth of
genomic sequences avai genomic sequences available we can begin to explore the evolutionary history of the *Bacillus* gent
While this can be approximated via analyses of, e.g., the 16S rRNA gene (Figure 4), identification a
phylogenetic analysis genomic sequences available we can begin to explore the evolutionary instory of the Bachina genus.
While this can be approximated via analyses of, e.g., the 16S rRNA gene (Figure 4), identification and
phylogenetic analysi phylogenetic analysis of the 'core genome' provides a significantly more robust measure. Prior studie
have found at least 600 genes within the core genome of the *B. cereus* sensu lato group [47,48]. From
the examination o phylogenetic analysis of the *B. cereus sensulato group* [47,48]. From
the examination of 20 *Bacillus* genomes, including taxa outside of the *B. cereus sensulato group*, 814
orthologous genes were identified as the core have found at least 600 genes within the core genome of the B. cereus sensu lato group [47,40]. From
the examination of 20 *Bacillus* genomes, including taxa outside of the B. cereus sensu lato group, 814
orthologous genes the examination of 20 Buchinas genomes, including taxa outside of the B. cereus sensu lato group, our-
orthologous genes were identified as the core genome of the genus [49]. Looking here at a larger
collection of sequence collection of sequences than previously considered, we identified the *Bacillus* core genome from t
predicted coding regions generated by Peasant. (Core genome analysis considered 231 high-qualit
assemblies; samples were i concertion of sequences than previously considered, we identified the Bacillus core genome from the
predicted coding regions generated by Peasant. (Core genome analysis considered 231 high-quality
assemblies; samples were predicted as 'high-quality' assemblies based upon their 16S rRNA sequent
number of predicted genes, and genome size (Supplemental Table 1).) Thirty genes were identified
within all assembled isolates and 310 genes were ide number of predicted genes, and genome size (Supplemental Table 1).) Thirty genes were identified
within all assembled isolates and 310 genes were identified within over 95% of the isolates
(Supplemental Table 5). As shown within all assembled isolates and 310 genes were identified within over 95% of the isolates
(Supplemental Table 5). As shown in Figure 4, there is a large core genome (5023 genes) found with
~70% of the genomes examined; (Supplemental Table 5). As shown in Figure 4, there is a large core genome (5023 genes) for
 \sim 70% of the genomes examined; this reflects the over-representation of taxa within the dat

belonging to the *B. cereus* sens ~70% of the genomes examined; this reflects the over-representation of taxa within the dataset
belonging to the *B. cereus* sensu lato group. Inclusion of more distant relatives to this group reduces the
number of genes a

CONCLUSIONS

ว
!|
! Figure 4. Evaluation of the core genome within 231 Buchus isolates.
CONCLUSIONS
Peasant provides an expedient way to take raw sequencing reads and
lightweight construction makes it feasible to conduct whole genome
computin C
F
I
C
I Peak lightweight construction makes it feasible to conduct whole genome studies without high performance
computing resources or programming expertise. The modular design of the tool permits the addition of
new tools easily computing resources or programming expertise. The modular design of the tool permits the addition of
new tools easily and future development of Peasant will expand options of assembly, QC, and
the violation of peasant will new tools easily and future development of Peasant will expand options of assembly, QC, and
new tools easily and future development of Peasant will expand options of assembly, QC, and
 $\frac{1}{2}$ new tools easily and future development of Peasant will expand options of Peasant will expand options of assem
The C, and future development of assembly, QC, and future development of assembly, QC, and future development
T

perform uniform processing provides an unbiased platform for bioinformatic analysis. The case student presented here, looking at over 200 individual *Bacillus* genomes, exemplifies the use of Peasant for locale analyses. O presented here, looking at over 200 individual *Bacillus* genomes, exemplifies the use of Peasant for lare analyses. Our uniform processing of samples has identified samples with concerns of contamination and/or mixed comm presented here, looking at over 200 individual *Bachinas* genomes, exempines the use of Peasant for large
scale analyses. Our uniform processing of samples has identified samples with concerns of
contamination and/or mixed scale analyses. Our analysing of samples the accuracy contamination and/or mixed communities as well as putative mislabeling. Furthermore, the of these *Bacillus* genomes is the largest study to date into the core genome o

ACKNOWLEDGEMENTS

of these *Bacillus* genomes is the largest study to date into the core genome of this genus.
ACKNOWLEDGEMENTS
The authors would like to thank Dr. Alan Wolfe and Ms. Krystal Thomas-White for their feedback during
developm of these *Bachins* genomes is the largest study to date into the core genome of this genus.
ACKNOWLEDGEMENTS
The authors would like to thank Dr. Alan Wolfe and Ms. Krystal Thomas-White for their fer
development of this too

REFERENCES

The authors would be also would be a main of this tool.
The authors would be also would be their formation.
Thomas Alan Wolfe and Microbiol. 2016;1:16048.
The also wo

REFERENCES
1. Hug LA, Baker BJ, Anant
1. Hug LA, Baker BJ, Anant
life. Nat Microbiol. 2016;1
2. Drew JC, Triplett EW. W
lessons from a pilot cours

1. Hug LA, Baker Baker
1. Drew J.C, Triplett EW. Whole genome sequencing in the undergraduate classroom: outcomes and
1. Hugsons fro 2. Drew JC, Triplett EW. Whole ger
lessons from a pilot course. J Microsons from a pilot course. J Microsons
3. Bertelli C, Aeby S, Chassot B, Clu
genome of Estrella lausannensis as
Front Microbiol. 2015;6:101.
4. Simpson lessons from a pilot course. J Microbiol Biol Educ. 2008;9:3–11.
3. Bertelli C, Aeby S, Chassot B, Clulow J, Hilfiker O, Rappo S, et al. Sequencing and characterizing the
genome of Estrella lausannensis as an undergraduate lessons from a pilot controller that course. There is a pilot course.

19. Bertelli C, Aeby S, Chassot B, Clulow J, Hilfiker O, Rappo S, et a

19. Front Microbiol. 2015;6:101.

19. Simpson JT, Wong K, Jackman SD, Schein JE 3. Bertelli Causannensis as an undergraduate project: training students and biological insight

5. Bertelli Crobiol. 2015;6:101.

4. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assemble

genome Microbiol. 2015;6:101.
4. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short
read sequence data. Genome Res. 2009;19:1117–23.
5. Zerbino DR, Birney E. Velvet: algori

4. Simpson JT, Wong K, Jackm
read sequence data. Genome
5. Zerbino DR, Birney E. Velve
Genome Res. 2008;18:821–9.
6. Luo R, Liu B, Xie Y, Li Z, Hua
efficient short-read de novo a read sequence data. Genome Res. 2009;19:1117–23.
5. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
Genome Res. 2008;18:821–9.
6. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, read sequence data. Exception interaction, 2012.

S. Zerbino DR, Birney E. Velvet: algorithms for de nove

Genome Res. 2008;18:821–9.

6. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAI

efficient short-read de nov

5. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-
6. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-
efficient short-read de

Samme Research, 2008
19: 2008, Liu B, Xie Y, Li Z, Huarefficient short-read de novo as
19: Bankevich A, Nurk S, Antipo
19: Samme Res. 2008
18: Koren S, Walenz BP, Berlin P
19: Tread assembly via adaptive k
19: Pint] doi: 1 Protect short-read at the novel account of agreement spectrum.

7. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulik

assembly algorithm and its applications to single-cell sequencing

8. Koren S, Walenz BP, Be assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 2012;19:455–77.
8. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate lon
read assembly via ad assembly via adaptive k -mer weighting and repeat separation. Ganu: scalable and accurate loft read assembly via adaptive k -mer weighting and repeat separation. Genome Res. 2017: [Epub aheaprint] doi: 10.1101/gr.215087.11

9. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. print] doi: 10.1101/gr.215087.116.

Bioinformatics. 2014; 30: 2114-20. print] doi: 10.1101/gr.215087.116.
9. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.
Bioinformatics. 2016;32:2103–10.
10. FastQC. <u>http://www.bioinformatics.babraham.ac.uk/projects/f</u> 9. Bioinformatics. 2016;32:2103–10.
9. Hindmartics. 2016;32:2103–10.
10. FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina seq 10. FastQC. <u>http://www.bioinform</u>
11. Bolger AM, Lohse M, Usadel B.
Bioinformatics. 2014;30:2114–20. 11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illu
Bioinformatics. 2014;30:2114–20. Bioinformatics. 2014;30:2114–20.

11. Bioinformatics: a flexible trimmoment for Illumina sequence data.

2014:30:2114–20. Bioinformatics. 2014;30:2114–20.

12. Sickle: A sliding-omains-organizations-button: Attps://github.com/najoshi/sickle.
13. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
Bioinformatics. 2013;29:1072–5.
14. Hunt M, Kikuchi T

Bioinformatics. 2013;29:1072–5.
14. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome
assembly evaluation. Genome Biology. 2013;14:R47.
15. Boetzer M, Henkel CV, Jansen HJ, Bu

13. Gurevich A, Saveliev V, Vyahhi I
Bioinformatics. 2013;29:1072–5.
14. Hunt M, Kikuchi T, Sanders M, M
assembly evaluation. Genome Biolo
15. Boetzer M, Henkel CV, Jansen F
SSPACE. Bioinformatics. 2011;27:57 14. Hunt M, Kikuchi T, Sanders M
assembly evaluation. Genome Bio
15. Boetzer M, Henkel CV, Jansen
SSPACE. Bioinformatics. 2011;27:
16. Donmez N, Brudno M. SCARP.
2013;29:428–34.

assembly evaluation. Genome Biology. 2013;14:R47.
15. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using
SSPACE. Bioinformatics. 2011;27:578–9.
16. Donmez N, Brudno M. SCARPA: sc assembly evaluation. Entertainment and the USA 2013.
15. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirova
SSPACE. Bioinformatics. 2011;27:578-9.
16. Donmez N, Brudno M. SCARPA: scaffolding reads
2013;29:428-34.
17. Tatuso 16. Boetzer M, Brudno M. SCARPA: scaffolding reads with practical algorithms. Bioinformatics.
16. Donmez N, Brudno M. SCARPA: scaffolding reads with practical algorithms. Bioinformatics.
17. Tatusova T, DiCuccio M, Badretd 16. Donmez N, Brudno M. SCARPA: scaff
2013;29:428–34.
17. Tatusova T, DiCuccio M, Badretdin A,
genome annotation pipeline. Nucleic Aci
18. Van Domselaar GH, Stothard P, Shriv.
automated bacterial genome annotation 2013;29:428–34.
17. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI pro
genome annotation pipeline. Nucleic Acids Res. 2016;44:6614–24.
18. Van Domselaar GH, Stothard P, Shrivastava 2013;2013
2013;2014
2013;2019 annotation
2013;2014
2013;2014
2019;2014
2019;2014
2019;2014

17. Tatuson, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2017, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 201
18. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, et al. BASys: a web server for
autom automated bacterial genome annotation. Nucleic Acids Research. 2005;33:W455–9.
19. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotation
using subsystems technology. BMC Genom

genome annotation pipeline. Nucleic Acids Research. 2
18. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, D
automated bacterial genome annotation. Nucleic Acids Research. 2
19. Aziz RK, Bartels D, Best AA, DeJ automated backerial genome and SI-2008; The dwards RA, et al. The RAST Serve
using subsystems technology. BMC Genomics. 2008;9:75.
20. Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, White O, et al. The IGS Stan
Pro using subsystems technology. BMC Genomics. 2008;9:75.
20. Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, White O, et al. The IGS Standard Operating
Procedure for Automated Prokaryotic Annotation. Stand Genomic Sci. using subsystems technology. Binks substrates technology.
20. Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, V
Procedure for Automated Prokaryotic Annotation. Stand G
21. Markowitz VM, Chen I-MA, Palaniappan K, Chu

21. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, et al. IMG 4 vers
integrated microbial genomes comparative analysis system. Nucleic Acids Res. 2014;42
22. Kremer FS, Eslabão MR, Dellagostin OA, Pinto

Procedure for Automated Prokaryotic Annotation. Stand Genomic Sci. 2011;4:244–51.
21. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, et al. IMG 4 version of the
integrated microbial genomes comparative a integrated microbial genomes comparative analysis system. Nucleic Acids Res. 2014;42:D560-567.
22. Kremer FS, Eslabão MR, Dellagostin OA, Pinto L da S. Genix: a new online automated pipeline f
bacterial genome annotation. integrated micropolicial genome annotation. FEMS Microbiol. Lett. 2016;363.

23. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.

24. Markowetz F. All biology is computational biolog

Bacterial genome annotation. FEMS Microbiol. Lett. 2016;363.
23. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
24. Markowetz F. All biology is computational biology. PLOS Biology. Bacterial general anno 1810 in 2010 (and 1810 in 2016)
24. Markowetz F. All biology is computational biology. PLOS Bio
25. Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline
Illumina MiSeq data. Bioinformatics. 20 23. Coil D, Jospin G, Darling AE. A5-mised: an updated pipeline to assemble microbial genomes f
25. Coil D, Jospin G, Darling AE. A5-mised: an updated pipeline to assemble microbial genomes f
Illumina MiSed data. Bioinform

25. Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline to assemble microbial gen
Illumina MiSeq data. Bioinformatics. 2015;31:587–9.
26. Mapleson D, Drou N, Swarbreck D. RAMPART: a workflow management system for d 25. Liao Miseq data. Bioinformatics. 2015;31:587–9.
26. Mapleson D, Drou N, Swarbreck D. RAMPART: a workflow management system for de novo geno
27. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. MEGAnnotator: 26. Mapleson D, Drou N, Swarbreck D. RAMPART: a w
assembly. Bioinformatics. 2015;31:1824–6.
27. Lugli GA, Milani C, Mancabelli L, van Sinderen D, \
for microbial genomes assembly and annotation. Ore
2016;363:fnw049.
28. Li assembly. Bioinformatics. 2015;31:1824–6.
27. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. MEGAnnotator: a user-friendly pipeline
for microbial genomes assembly and annotation. Oren A, editor. FEMS Microbio assembly. Biomannic Collection 2016; Lugli GA, Milani C, Mancabelli L, van Sind
for microbial genomes assembly and annota
2016;363:fnw049.
28. Liao Y-C, Lin H-H, Sabharwal A, Haase EN
automated prokaryotic genome assembly

for microbial genomes assembly and annotation. Oren A, editor. FEMS Microbiology Letters.
2016;363:fnw049.
28. Liao Y-C, Lin H-H, Sabharwal A, Haase EM, Scannapieco FA. MyPro: A seamless pipeline for
automated prokaryotic 2016;363:fnw049.
28. Liao Y-C, Lin H-H, Sabharwal A, Haase EM, Scannapieco FA. MyPro: A seamless pipeline fo
automated prokaryotic genome assembly and annotation. Journal of Microbiological Method
2015;113:72–4. 28. Liao Y-C, Lin H-
automated prokary
2015;113:72–4. 28. Liao Y-C, Lin H-H, Sabharwal A, Haase EM, Scannapieco FA. MyPro: A seamless pipeline for automated prokaryotic genome assembly and annotation. Journal of Microbiological Methods. $\frac{1}{2}$ 2015;113:72–4.

29. Koren S, Treangen T, 20. Koren S, 2014;15:126.
29. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-
bacterial Bioinformatics Database and Analysis Resource Center. Nuc microbial genomes. Bramicrobial gradier and Analysis Resource Batterial Bioinformatics Database and Analysis Resource 31. Leinonen R, Sugawara H, Shumway M, Internationa
The sequence read archive. Nucleic Acids Res. 2011;3

31. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration.
The sequence read archive. Nucleic Acids Res. 2011;39:D19-21.
32. BBMap. https://sourceforge.net/projects/bbmap/.
33. Lagesen

The sequence read archiverence read archives.

32. BBMap. https://sourceforge.net/projects/bbmap/.

33. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Us

rapid annotation of ribosomal RNA genes. Nucleic Acids

34. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA The sequence read archive. Nucleic Acids Res. 2011;39:D19-21.
32. BBMap. <u>https://sourceforge.net/projects/bbmap/</u>.
33. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and
rapid an

33. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rograpid annotation of ribosomal RNA genes. Nucleic Acids
34. Lowe TM, Chan PP. tRNAscan-SE On-line: integratin
genes. Nucleic Acids Res. 2016;44:W54-57.
35. Delcher AL, 33. Lawe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer R
34. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer R
35. Delcher AL, Harm

rapid annotation of the search and context fo
genes. Nucleic Acids Res. 2016;44:W54-57.
35. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial ge
GLIMMER. Nucleic Acids Res. 1999;27:4636–41.
36. O'Lear genes. Nucleic Acids Res. 2016;44:W54-57.
35. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with
GLIMMER. Nucleic Acids Res. 1999;27:4636–41.
36. O'Leary NA, Wright MW, Brister generation and the member of your control.

35. Delcher AL, Harmon D, Kasif S, White O,

GLIMMER. Nucleic Acids Res. 1999;27:4636

36. O'Leary NA, Wright MW, Brister JR, Ciuf

(RefSeq) database at NCBI: current status, t
 35. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence
36. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence
(RefSeq) database at NCBI: curre

FERRERR MARRETHER. 1999-1999-1999

G. O'Leary NA, Wright MW, Brister JR, Ciufo S, F

(RefSeq) database at NCBI: current status, taxon

Acids Res. 2016;44:D733-745.

37. RefSeq Bacterial Genome Archive. ftp://ftp.n

38. Coc (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nuclides Res. 2016;44:D733-745.
Acids Res. 2016;44:D733-745.
37. RefSeq Bacterial Genome Archive. <u>ftp://ftp.ncbi.nlm.nih.gov/genom</u> (RefSeq Bacterial Genome Archive. <u>ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacte</u>
18. Cock PA, Antao T, Chang JT, Bradman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
Wilczynski B and de Hoon MJL B Actor Actor 2017, Antao T, Chang JT
38. Cock PA, Antao T, Chang JT
Wilczynski B and de Hoon MJL
biology and bioinformatics. Bio
39. Camacho C, Coulouris G, Arapplications. BMC Bioinformat 38. Cock PA, Antao T, Chang JT, Bradman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
Wilczynski B and de Hoon MJL Biopython: freely available Python tools for computational molecular
biology and bioinformatics. Wilczynski B and de Hoon MJL Biopython: freely available Python tools for computational molecology and bioinformatics. Bioinformatics. 2009;25:1422-1423.
Biology and bioinformatics. Bioinformatics. 2009;25:1422-1423.
39. C

biology and bioinformatics. Bioinformatics. 2009;25:1422-1423.
39. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture applications. BMC Bioinformatics. 2009;10:421.
40. SRA Run S Biology

Biology and biology and N, Ma N, Papadopoulos J, E

applications. BMC Bioinformatics. 2009;10:421.

40. SRA Run Selector. https://www.ncbi.nlm.nih.gov/Traces/stud

41. Price MN, Dehal PS, Arkin AP. FastTree 2--app

applications. BMC Bioinformatics. 2009;10:421.
40. SRA Run Selector. https://www.ncbi.nlm.nih.gov/Traces/study/
41. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large
alignments. PLo approximations. 2008.

40. SRA Run Selector. https://www.ncbi.nlm.nih

41. Price MN, Dehal PS, Arkin AP. FastTree 2--ap

alignments. PLoS ONE. 2010;5:e9490.

42. Jordan GE, Piel WH. PhyloWidget: web-base

2008;24:1641-2.
 41. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximu
alignments. PLoS ONE. 2010;5:e9490.
42. Jordan GE, Piel WH. PhyloWidget: web-based visualizations for th
2008;24:1641-2.
43. Rasko D, Altherr M, Han C, Rav

alignments. PLoS ONE. 2010;5:e9490.
42. Jordan GE, Piel WH. PhyloWidget: web-based visualizations for the tree of life. Bioinformatics
2008;24:1641–2.
43. Rasko D, Altherr M, Han C, Ravel J. Genomics of the group of organi alignments. Piel WH. PhyloWidget:
1908;24:1641–2.
43. Rasko D, Altherr M, Han C, Ravel J.
Reviews. 2005;29:303–29.
44. Økstad OA, Kolstø A-B. Evolution o
thuringiensis Biotechnology. Dordrech 2008;24:1641–2.
43. Rasko D, Altherr M, Han C, Ravel J. Genomics of the group of organisms. FEMS Microbiology
Reviews. 2005;29:303–29.
44. Økstad OA, Kolstø A-B. Evolution of the *Bacillus cereus* Group. In: Sansinenea E, 2008;29
2008;29
2008;29
24. *Økstad OA, Kc
thuringiensis Biot*
25. Wang A, Ash (2015;5:13644. Reviews. 2005;29:303–29.
A4. Økstad OA, Kolstø A-B. Evolution of the *Bacillus cereus* Group. In: Sansinenea E, editor. Bacill
thuringiensis Biotechnology. Dordrecht: Springer Netherlands; 2012. p. 117–29.
45. Wang A, Ash 44. Økstad OA, Kolstø A-B.
thuringiensis Biotechnolog
45. Wang A, Ash GJ. Whole
2015;5:13644.
46. Okinaka RT, Keim P. Th
The Bacterial Spore: from M

44. Økstad OA, Kolstø A-B. Evolution of the Buchlas Eccels Group. In: Sansinenea E, Editor: Bachlas
thuringiensis Biotechnology. Dordrecht: Springer Netherlands; 2012. p. 117–29.
45. Wang A, Ash GJ. Whole genome phylogeny 15. Wang A, Ash GJ. Whole genome phylogeny of *Bacillus* by Feature Frequency
2015;5:13644.
46. Okinaka RT, Keim P. The Phylogeny of *Bacillus cereus* sensu lato. In: Eichenber
The Bacterial Spore: from Molecules to System 45. Wang A, Ash GJ. Whole genome phylogeny of Bacillus by Feature Frequency Fromes (FFF). Scribep.
2015;5:13644.
46. Okinaka RT, Keim P. The Phylogeny of *Bacillus cereus* sensu lato. In: Eichenberger P, Driks A, editors
T 2020
16. Okinaka RT
The Bacterial S 46. Okinaka RT, Keim P. The Phylogeny of Buchus cereus sensu lato. In: Eichenberger P, Driks A, editors.
The Bacterial Spore: from Molecules to Systems. American Society of Microbiology; 2016. p. 239–51.
. The Bacterial Spore: from Molecules to Systems. American Society of Microbiology; 2016. p. 239–51.

19. 2012;22:1512–24.
2012;22:1512–24.
2012;22:1512–24.
48. Bazinet AL. Pan-genome and phylogeny of *Bacillus cereus* sensu lato. bioRxiv. 2017; doi:
10.1101/119420.
49. Alcaraz L, Moreno-Hagelsieb G, Eguiarte LE, Souza V, of the Bacillus cereus sensu lato species: Backdrop to the evolution of Bacillus antifiatis. Genome Res.
2012;22:1512–24.
48. Bazinet AL. Pan-genome and phylogeny of *Bacillus cereus* sensu lato. bioRxiv. 2017; doi:
10.110 48. Bazinet AL. Pan
10.1101/119420.
49. Alcaraz L, More
the evolutionary re
Genomics. 2010;11 48. Bazinet AL. Pan-genome and phylogeny of Buchlus cereus sensu lato. bioNxiv. 2017, doi.
10.1101/119420.
49. Alcaraz L, Moreno-Hagelsieb G, Eguiarte LE, Souza V, Herrera-Estrella L, Olmedo G. Unde
the evolutionary relati 10.1101/119420. 49. Alcores Line and Moreover Line and Moreover Line and Moreover Line and Moreover Channel L, Moreover Channel
Genomics. 2010;11:332.
Supplemental Table Legends
Supplemental Table 1. List of the 322 *Bacillus* sequencing the evolutionary relationships and major traits or *Bacillus* through comparative genomics. BMC
Genomics. 2010;11:332.
Supplemental Table 1. List of the 322 *Bacillus* sequencing SRA projects retrieved.
Supplemental Table

Sanamar 2020,2002.
Supplemental Table Lege
Supplemental Table 1. Li
Supplemental Table 2. A
projects. $\frac{1}{2}$ $\frac{1}{2}$ c. c. c. c. c. c. Supplemental Table 1. List of
Supplemental Table 2. Assem
projects.
Supplemental Table 3. NCBI R
assembled *Bacillus* isolates.
Supplemental Table 4. Details

Supplemental Table 2. Assembly and annotation statistics for processed *Bacillus* s
projects.
Supplemental Table 3. NCBI RefSeq 16S rRNA gene sequences included in phyloge
assembled *Bacillus* isolates.
Supplemental Table Supplemental Table 2. Assembly and annotation statistics for processed *Buchnas* sequenting SRA
projects.
Supplemental Table 3. NCBI RefSeq 16S rRNA gene sequences included in phylogenetic analysis of
assembled *Bacillus* Supplem
Supplem
assemble
Supplem
Bacillus w
Supplem Supplemental Table 3. NCBI RefSeq 105 FMM gene sequences included in phylogenetic analysis of
assembled *Bacillus* isolates.
Supplemental Table 4. Details regarding samples from which 16S rRNA sequences from taxa other
Ba assembled Bacillus isolates.
Supplemental Table 4. Deta
Bacillus were identified.
Supplemental Table 5. Gene
and within 95% (Sheet 2) of

Supplemental Table 4. Details regarding samples from which 16S rRNA sequences from taxa other than
Bacillus were identified.
Supplemental Table 5. Genes identified within all (Sheet 1) 231 high-quality Bacillus genome asse Buchas were identified.
Supplemental Table 5. G
and within 95% (Sheet 2) Supplemental Table 5. Genes identified within all (Sheet 1) 231 high-quality Bacillus genome assembles.
and within 95% (Sheet 2) of the high-quality *Bacillus* genome assemblies. and within 95% (Sheet 2) of the high-quality Bacillus genome assemblies.