Heliyon 9 (2023) e13314

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

StrainSelect: A novel microbiome reference database that disambiguates all bacterial strains, genome assemblies and extant cultures worldwide

Todd Z. DeSantis [a,b,*], Cesar Cardona [a], Nicole R. Narayan [a], Satish Viswanatham [a], Divya Ravichandar [a], Brendan Wee [a], Cheryl-Emiliane Chow [a], Shoko Iwai [a]

[a] *Second Genome, Inc., 1000 Marina Blvd, Suite 500, Brisbane, CA, 94005, USA*

[b] *Environmental Metagenomics, Research Center One Health Ruhr of the University Alliance Ruhr, Faculty of Chemistry, University of Duisburg-Essen, Germany*

ARTICLE INFO

ABSTRACT

Motivation: Microbial metagenomic profiling software and databases are advancing rapidly for development of novel disease biomarkers and therapeutics yet three problems impede analyses: 1) the conflation of "genome assembly" and "strain" in reference databases; 2) difficulty connecting DNA biomarkers to a procurable strain for laboratory experimentation; and 3) absence of a comprehensive and unified strain-resolved reference database for integrating both shotgun metagenomics and 16S rRNA gene data.

Results: We demarcated 681,087 strains, the largest collection of its kind, by filtering public data into a knowledge graph of vertices representing contiguous DNA sequences, genome assemblies, strain monikers and bio-resource center (BRC) catalog numbers then adding inter-vertex edges only for synonyms or direct derivatives. Surprisingly, for 10,043 important strains, we found replicate RefSeq genome assemblies obstructing interpretation of database searches. We organized each strain into eight taxonomic ranks with bootstrap confidence inversely correlated with genome assembly contamination. The StrainSelect database is suited for applications where a taxonomic, functional or procurement reference is needed for shotgun or amplicon metagenomics since 636,568 strains have at least one 16S rRNA gene, 245,005 have at least one annotated genome assembly, and 36,671 are procurable from at least one BRC. The database overcomes all three aforementioned problems since it disambiguates strains from assemblies, locates strains at BRCs, and unifies a taxonomic reference for both 16S rRNA and shotgun metagenomics.

Availability: The StrainSelect database is available in igraph and tabular vertex-edge formats compatible with Neo4J. Dereplicated MinHash and fasta databases are distributed for sourmash and usearch pipelines at http://strainselect.secondgenome.com.

Contact: todd.desantis@gmail.com.

Supplementary information: Supplementary data are available online.

* Corresponding author at: Second Genome, Inc., 1000 Marina Blvd, Suite 500, Brisbane, CA, 94005, USA.
E-mail addresses: todd.desantis@gmail.com, todd@secondgenome.com (T.Z. DeSantis).

https://doi.org/10.1016/j.heliyon.2023.e13314
Received 7 September 2022; Received in revised form 1 December 2022; Accepted 26 January 2023

Available online 4 February 2023

1. Statement of significance

Problem: Although clinical microbiome data is being evaluated for both precision biomedical decision support and therapeutic discovery, three problems impede translation of data into beneficial products and services: 1) the conflation of "genome assembly" and "strain" in reference databases; 2) difficulty mapping microbiome DNA biomarkers to an extant strain for purchase and experimentation; and 3) absence of a unified comprehensive strain-resolved reference database for integrating both shotgun data and 16S rRNA gene data.

What is Already Known: Reference databases, such as RefSeq, are currently available for organizing microbiome data at the strain-level resolution. Unfortunately, novices are unaware these databases contain multiple genome records generated from a single strain but deposited as separate strains. For instance, sequence data labeled as *Mesorhizobium loti* HAMBI 1129, *M. loti* DSM 2626, *M. jarvisii* ATCC 700743, and *M. jarvisii* ATCC 33669 are all from the same source strain isolate. As another example, four different RefSeq genome assemblies, GCF_001571425, GCF_001652705, GCF_001678855, and GCF_003628755, are all derived from the same source strain isolate. Most reference databases improperly assume that each name bestowed to an organism and each genome assembly equates to a unique strain.

What This Paper Adds: We describe a method to identify 681,087 unique strains that are represented by over 8 million synonymous monikers in public records. We constructed a database that overcomes all three aforementioned problems since it disambiguates strains from assemblies, maps which strains are available for procurement and experimentation from a culture collection, even if those strains are named differently in the respective catalogs, and allows integration of both shotgun and 16S rRNA gene data against a single organized taxonomy which is a key utility for comprehensive meta-analyses and robust biomarker applications.

2. Introduction

Both shotgun metagenomics and 16S rRNA gene amplicon marker gene publications exhibit year-over-year growth (Fig. 1) due to broad applications in clinical, agricultural, and environmental data sciences. Depending on the experimental design, molecular microbiologists process the raw data to determine, as examples, which genera are significantly elevated in the colons of one group of patients relative to another [16], which combination of bacterial species predicts a beneficial response to a pharmaceutical agent [32], or which novel chromosomes from yet-to-be-cultured bacteria can be reconstructed from 0.1 to 8.0 kilo-base sequencing reads *in silico* into metagenomic assembled genomes (MAGs) [1,39,52] to discover novel CRISPR-Cas systems [6]. But other investigators will go beyond descriptive analytics and will conduct follow-up experiments to establish causation linking certain strains or their products to a particular outcome in an animal model of disease [49,53] or an agricultural field trial [35]. To accomplish this, microbiome data would need to be interpreted with methods to reveal the individual **strains** associated with the outcome of interest. Then in an efficient manner those strains would need to be grown in the lab and tested against controls in experiments structured to prove/disprove the causal hypothesis. A data engineer tasked to determine the set of strains within a metagenomics data set that significantly associate with an experimental variable and then to map those strains against worldwide bio-resource centers (BRCs) from which individual strains can be purchased, will need to first settle on a definition of a "strain" that fits this endeavor. Then, the engineer must overcome three challenges which motivated this work: 1) ambiguity between a "genome assembly" and a "strain" in reference databases; 2) difficulty connecting observations in the metagenomic analysis to a procurable strain for laboratory experimentation; and, if confirmation of findings among different library techniques was desired as in Tessler et al. [54], 3) integration of both shotgun and 16S rRNA data against a single reference.

An investigator will need to be clear about their operational definition of "strain" for the investigation and they may favor the MAG definition, where each unique MAG is one strain, or the isolate-and-propagate definition where an isolate and its descendants are one strain. If the investigator adopts the definition that any chromosomal variant among any MAG from any biospecimen is a distinct strain, then a reference database is not required nor is a BRC connection valued. Instead, a multitude of isolates would need to be directly cultured from the biospecimens, each sequenced and assembled until the desired MAGs were matched exactly before proceeding to the causative experiments. On the other hand, to accelerate procurement of a live strain for a causative experiment, we suggest the second more traditional and tractable strain definition. In this definition, a single strain encapsulates all the descendants of a single colony isolation in pure (axenic) culture and is disseminated among microbiologists by a succession of cultures [4,19]. It is appreciated that the initial process of isolation from a living community is itself a selection event which captures one point-in-time of a mutable genome [13]. Nonetheless, these isolated and propagated strains are important tools for experimental microbiology and provide necessary points of reference for scientific communication and intellectual property delineations.

Heterogeneity exists among the methods of naming and bio-banking the descendants of a single isolate among microbiologists and this has led to downstream confusion for the bioinformatician. Oftentimes microbiologists, after isolating and naming a single strain from clinical or environmental material, will send replicate sub-cultures to multiple BRCs, such as ATCC (http://www.atcc.org), DSMZ (http://www.dsmz.de) and JCM (http://jcm.brc.riken.jp) or dozens of other worldwide centers. These BRCs then assign their own catalog numbers. DNA sequencing institutes throughout the international scientific community procure strains from various BRCs, extract and sequence the DNA then upload single genes or whole genome assemblies to public databases, such as GenBank [2], who assigns an identifier for each assembly received. Because this is a decentralized international activity, there has been persistent uncertainty about what data belongs to each strain [3,21]. A prime example of the need for unification can be seen in a strain isolated from a healthy Japanese male in 2011 [37]. The research team bestowed novel genus and species level nomenclature for the isolate which they publicized as *Christensenella minuta* YIT 12065. Two independent BRCs (DSM and JCM)

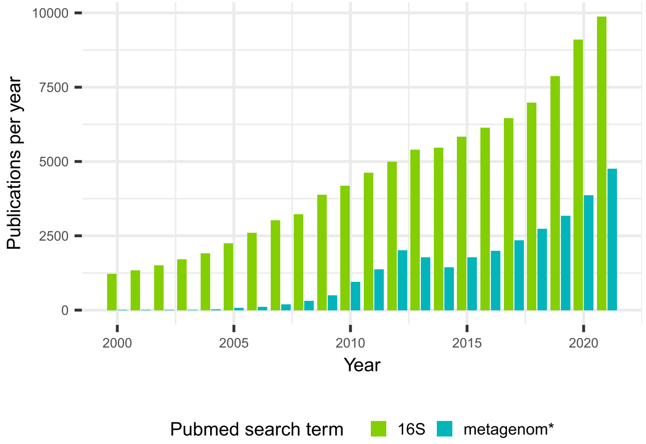

Fig. 1. Quantity of Pubmed indexed publications found with search terms `16S` or `metagenom*` (where * represents a wildcard) have increased throughout the last two decades. Publications leveraging metagenomics are less frequent than those leveraging 16S rRNA gene amplicons. A reference database that enables integration of both technologies is ideal.

also propagated sub-cultures of this strain with their own unique catalog numbers, DSM 22607 and JCM 16072. The University of California at Davis, Beijing Genome Institute, Washington University, and South China University of Technology each procured the strain from one of the BRCs then separately sequenced the extracted DNA and submitted their optimal assembly to public databases which are now downloadable from RefSeq under four different assembly identifiers: GCF_001571425, GCF_001652705, GCF_001678855, and GCF_003628755. Novice users of these public databases could easily misinterpret these four assembles as four different genomes from four different strains. In contrast, we see these as technical replicates. In building the StrainSelect database, we sought to overcome confusion by tracing through the synonymous identifiers for sub-cultures and genomic data records and assign a consistently formatted identifier for the strain, which in this example is “StrainSelectID:t_520”, and connect all the genomic records together.

Now if the investigator decides to match metagenomic data to strains according to the isolate-and-propagate aforementioned definition, the bioinformatician will need to build or acquire a reference database with three properties to overcome three challenges. **First**, the database will appropriately label each gene and genome assembly by the strain of origin carefully avoiding conflation of genome assemblies as strains. Unfortunately, NCBI, a central foundational database, has announced cessation in efforts to organize data in this fashion [20]. What is needed in a reference database is reliable linkage of clandestine technical replicates, those genome assemblies from the same strain published from two or more institutes using dissimilar monikers. A recent study on of the deleterious effects of duplicate sequence records in bioinformatics reference databases demonstrated inefficiency, obviously in computational search load, and less obvious but more severe, in the manual or scripted assessment of the results of a search [8]. As a simple example of the problem, consider a single query DNA sequence matching the set of database subjects *Mesorhizobium loti* HAMBI 1129, *M. loti* DSM 2626, *M. jarvisii* ATCC 700743, as well as *M. jarvisii* ATCC 33669, with zero matches outside this set. The inexperienced bioinformatician would likely interpret these match results as a non-strain-specific “hit” since the names share only the genus. But since these are all synonyms for the same strain it would be accurate to conclude that the hit was in fact strain-specific. **Second**, the database will need a schema to relate each genomic record to zero or one extant procurable strain cultures distributed by one or more BRCs. In other words, users should know if a genomic record is not only linked to a strain but if that strain is available in a BRC. **Third**, since microbiome meta-analysis provides opportunity to find concordant observations among cohorts often profiled with differing lab technologies [51], a single reference database should enable integration of metagenomic shotgun and the more popular 16S rRNA gene amplicon data (Fig. 1) into a single taxonomic ontology. StrainSelect was built to overcome all three challenges and is available as a reference database (http://strainselect.secondgenome.com) describing 681,087 strains for use in standalone pipelines. The R code to reproducibly generate all tables, figures and text for this manuscript is provided, as well.

2.1. Other notable resources

Over the last decade, several data curators have attempted to solve these problems however each effort has either been abandoned or lacks key features to support current data analysis needs. StrainInfo [55], the early inspiration for StrainSelect, endeavored to build a database that would include both genome assemblies and 16S rRNA genes apart from assemblies, but is no longer maintained. BacDive [48] organizes genome assemblies, 16S rRNA genes and functional attributes via an informative interactive web tool. It contains a small number of the known strains (89,545 strains) and does not provide a downloadable database for high-throughput data pipelines. GOLD [38] appeared more comprehensive representing 395,286 bacterial and archaeal “organisms” but in some cases one strain has multiple organism identifiers as exemplified in Sup. Fig. 1 so the actual strain count is likely less. The Genome Taxonomy Database (GTDB) [43] contains 258,406 genome assemblies taxonomically organized from domain to species but does not attempt to categorize the assemblies by strain and only includes 16S rRNA genes if they are embedded into genome assemblies of pure cultures or connected to a MAG. GTDB has fully disclosed its methods for placing assemblies into species and distributes useful

Table 1
Vertex types in the graph schema.

Vertex type	Description	Examples
contig	Contiguous DNA sequence	JF079054, NZ_FJOC01000002, NC_013353
g16	16S rRNA gene	g16_4602054
wgs_master_pre	NCBI WGS master record prefix	wgs_AADD, wgs_FJOC, wgs_CAADNE
gb_assembly	Genbank genome assembly	GCA_000155415
rs_assembly	RefSeq genome assembly	GCF_000001635
kegg_genome	KEGG genome	gn_ebw, gn_ecok
patric_genome	Patric genome	pat_1131286.3, pat_1123738.3
biocyc_pgdb	BioCyc PGDB	bc_LLAC1295826, bc_GCF_000001635
si_culture_id	Culture recorded by StrainInfo	ci_119674
si_grouping_id	Group of replicate cultures recorded by StrainInfo	gr_2, gr_171641
brc_cat_id	Bioresource center catalog identifier	ATCC 700598, DSM 2281, CCUG 38580
gold_org	GOLD organism	Go0516098, Go0000004
gss	genus species strain string	escherichia.coli.k.12.dh10b

files and software for species-level classification. StrainSelect expands on the esteemed work from StrainInfo, BacDive, GOLD, and GTDB by including more than double the number of strains than previous resources, resolving synonymous organism names for the same strain, and building a unified taxonomy for use with both shotgun or amplicon techniques.

3. Approach

Various known monikers of the isolated and published strains as well as the identifiers for the public genomic records attached to each were collected from relevant sources. Genomic records gathered were either full genome assemblies or 16S rRNA gene assemblies covering eight of the nine hyper-variable regions and both types were filtered by standardized procedures. All monikers and sequence identifiers were placed as vertices (nodes) of a network knowledge graph and inter-vertex edges (connections) were created to represent direct material derivatives. The graph was decomposed into components, where one component is a connected sub-graph of vertices that is disjointed from any other sub-graph. Each component defined exactly one archaeal or bacterial strain and each strain was assigned a StrainSelectID identifier.

Where possible, taxonomic nomenclature for seven levels from domain to species was adapted from GTDB with the additional and relevant constraint that one strain can belong to only one species. For strains with 16S rRNA genes available but without a genome assembly, taxonomic placement was estimated by k-mer similarity. Where formal taxonomic names were not yet coined for demarcated genera-level and species-level groups, provisional identifiers were assigned. The stability of both formally-named and provisionally-named species-level groupings was measured by bootstrapping prompting a subset of provisionally-named species to be merged into formally-named siblings.

Because all data was organized by species and by strain, intrastrain versus intraspecies genomic similarity was contrasted. We present a new estimate of variation among related but distinct strains as well as an estimate of technical variation of genome assemblies from the same strain sequenced and assembled at different institutes.

4. Methods and results

4.1. Software

R, https://www.R-project.org, [47] was used for the majority of the graph construction pipeline with Python, http://www.python.org, used to download and filter NCBI data. The R libraries, data.table, https://CRAN.R-project.org/package=data.table [14] and kableExtra, https://CRAN.R-project.org/package=kableExtra [62] were used for tabular operations and ggplot2 [60], ggnetwork, https://CRAN.R-project.org/package=ggnetwork, [5], and ggbreak [61] for data visualizations. Additional software packages for specific steps are cited in subsequent sections.

4.2. Input data

Monikers (i.e. published names, abbreviated names, machine readable identifiers and synonyms) for strains and their associated genomic data were collected from PATRIC [56] on 2021-11-23, GOLD [38] on 2021-11-23, GTDB [42] on 2021-12-26, BioCyc [30] on 2021-08-05, KEGG [29] on 2021-10-31, RefSeq [22] on 2021-11-28 with the NCBI Type-Strain Report, https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/ on 2021-11-29, WGS, https://www.ncbi.nlm.nih.gov/genbank/wgs/ on 2021-11-23, and StrainInfo [55] on 2013-10-20. Custom parsers are maintained for each data source and require adjustments as source formats evolve.

Input data was categorized into 13 vertex types as shown in Table 1 and denoted in `fixed width font` in this description. The StrainSelect data model follows the common data types created as information is generated. An institute generates and assembles sequencing reads from an isolated strain into contiguous DNA sequences (`contig`) identified by GenBank accession numbers and can encode a single gene, as in the case of the 16S rRNA gene (`g16`), or encode many genes when assembled from of a whole genome shotgun read library. These shotgun projects are registered at NCBI and assigned a 4 or 6 letter string that becomes the master prefix (`wgs_master_pre`) for all the project's `contigs`. A set of one or more `contigs` representing a genome assembly effort is distributed from GenBank (`gb_assembly`) and if the set meets certain quality thresholds for completeness and purity will additionally be distributed from RefSeq (`rs_assembly`). When KEGG or PATRIC annotate an assembly, they re-distribute the data and StrainSelect includes those vertices as `kegg_genome` or `patric_genome`, respectively. If BioCyc creates a specially formatted database from an assembly for interactive pathway analyses, then a `biocyc_pgdb` vertex was included. StrainInfo recognized that one strain can exist as cultures at multiple institutes and established separate culture identifiers for each (`si_culture_id`) and a list of the disseminated cultures from the same strain defines the `si_grouping_id`. Bio-resource centers (BRCs), sometimes known as culture collections, will receive live strains then store, propagate and ship the strains under their own catalog numbers (`brc_cat_id`). The GOLD organism identifier was captured as `gold_org`. The `gss` vertex type was established for both human- and machine-readable processes and encodes the genus-species-strain concatenation, as described below.

4.3. Genus-species-strain vertices

Due to differing database entry conventions, strains have been dubbed with slight variations in the formatting of character strings for genus, species and strain names. For example, one strain classified within the species *Comamonas terrigena* can be found as "R. Hugh 247", "R.Hugh 247", and "R Hugh 247". To prevent the creation of multiple vertices that are only slight deviations in string content, all alphabetical characters are converted to lowercase and each series of non-alpha-numeric characters are converted to a single period. Thus, the genus-species-strain (`gss`) vertex in each of these cases would be unified to "comamonas.terrigena.r.hugh.247". Since this same strain has also been referenced as "Vron 31", a distinct vertex of "comamonas.terrigena.vron.31" is also included. To avoid insufficient vertex name complexity resulting from this process, a `gss` vertex was not formed when less than three words were available for the concatenation or when the `gss` would be less than 10 characters thereafter.

4.4. Genome assembly quality control

Genome assemblies in RefSeq are assumed to be more reliable than those only in GenBank since, as the documentation at https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq attests, each has at least one copy of a 16S rRNA gene and are not contaminated with DNA sequence from multiple strains. StrainSelect further scans the set of contigs of each RefSeq assembly using profile hidden Markov models (HMMs) with nhmmer [58]) to obtain the count, lengths, coordinates and taxonomic domains of origin for the 16S and 23S rRNA genes. In total, 250,511 RefSeq assemblies were processed and 4,882 (1.9%) were rejected due to rRNA genes found from more than one domain within the same assembly, suggesting contamination. In other words, after discarding potentially problematic genomes StrainSelect provided 98.1% coverage of RefSeq. Surprisingly, in 5,723 (2.3%) RefSeq assemblies nhmmer failed to find any archaeal nor bacterial rRNA and in 25,938 (10.4%) when a 16S rRNA gene was encountered it was incomplete (under 1,250 nt where ~1,500 nt is expected) or it contained over 1% non-ACGT characters. These assemblies were still retained as vertices but their 16S rRNA genes were not.

4.5. 16S rRNA gene assembly quality control

An NCBI search for 16S rRNA genes ≥1250 nt that were derived from isolated strains and not from clones, unculturable materials, nor PCR libraries resulted in 511,858 records. These records are of the type `contig` in the StrainSelect graph schema since they are contiguous DNA sequences. Contigs can exist independent from or belonging to one genome assembly. All contigs were processed by nhmmer (described above) to reject those with regions from more than one taxonomic domain or containing under 1,250 nt matching the 16S rRNA gene model or if that span contained over 1% non-ACGT characters. In total 11,279 were rejected, which resulted in 500,579 `contigs` remaining. A separate vertex was formed from each 16S rRNA gene instance within each `contig` totaling 917,079 `g16` type vertices. Although many 16S rRNA gene sequences are identical across genome assemblies [44], no sequence dereplication was applied at this step.

4.6. Graph composition, component discovery, and component filtering

All vertices resulting from parsing input records and filtering sequence data were loaded into the network-based data management software package, iGraph, https://igraph.org [10], with each having exactly one vertex "type" attribute from Table 1. Graph edges, where each edge is a link between exactly two vertices were defined by an equality represented by an input data source or from the HMM analysis. As an example of vertices connected by edges consider the case in Fig. 2. Parsing data from GOLD equated *Alistipes senegalensis* JC50 [36] (vertex(id = alistipes.senegalensis.jc50, type = `gss`)) to the organism identifier Go0014227 (vertex(id = Go0014227, type = `gold_org`)). Identifying a 16S rRNA gene (vertex(id = g16_0018901, type = `g16`)) spanning positions 4 to 1528 within the sequence NZ_CAHI01000040 (vertex(id = NZ_CAHI01000040, type = `contig`)) established an edge. Since this `contig` was from the set of contigs defining one RefSeq genome assembly (vertex(id = GCF_000312145, type = `rs_assembly`)) submitted in 2012, an edge was established for this relationship, as well. Fig. 2 displays how these edges and others connect all the monikers for this strain. The graph integration of information reveals that RefSeq genome assembly GCF_000312145 was derived from alistipes.senegalensis.jc50 which is available for procurement at two different BRCs and under another synonym,

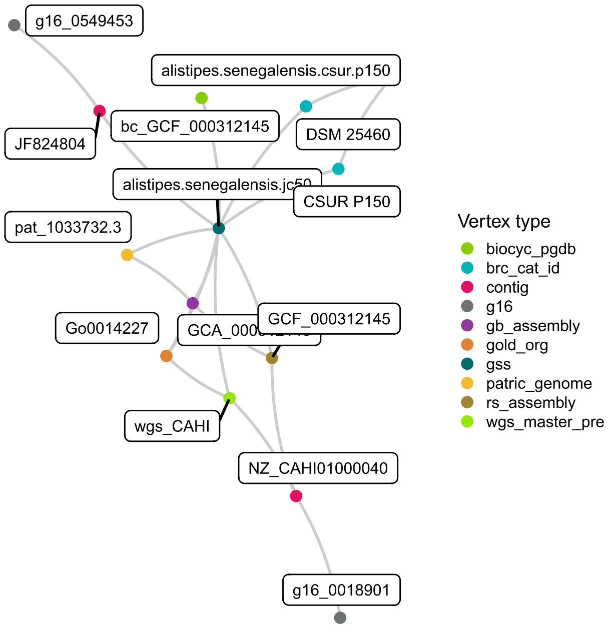

Fig. 2. An exemplary subgraph of the vertices comprising one strain, StrainSelectID t_117676. This subgraph is a disjointed component within the entire knowledge graph and connects vertices of various types shown by color. All data, although distributed from different sources such as NCBI, KEGG, BioCyc. Patric, and GOLD, was derived from an isolate originally named *Alistipes senegalensis* str. jc50 which is available for procurement at two different BRCs (blue). The subgraph conveys that two high quality 16S rRNA genes (grey) are available for this strain, one from a RefSeq genome assembly (brown) containing a contig (pink) encoding a 16S rRNA gene (grey) and the other is independent of a genome assembly project but was deposited as contig JF824804 (pink).

alistipes.senegalensis.csur.p150. The GOLD organism identifier is attached to the gb_assembly and wgs_master_pre. Also notice that two contigs carry high quality 16S rRNA genes, one as described above from a multigene contig and the other from a single gene contig, (vertex(id = JF824804, type = contig)), submitted to NCBI in 2011. For a more complex example see Sup. Fig. 1.

The entire graph was initialized with 8,412,126 vertices connected by 7,603,203 edges. Vertices with degree = 0, in other words vertices with zero edges, were dropped leaving 8,339,151. These vertices are not useful for our purpose since they represent an assembly without a name for the isolate nor a BRC entry or these vertices are cultures in a BRC with no public sequence data available. The remaining graph was decomposed into components, where one component was a connected sub-graph of vertices that is disjointed from any other sub-graph.

Components removed were those encompassing zero gss vertices or zero g16 and rs_assembly vertices, a condition formalized in Eq. (1).

$$\sum(V_{gss}) = 0 \vee \sum(V_{g16} + V_{rs\_assembly}) = 0 \tag{1}$$

Each of the 681,087 remaining components defined exactly one StrainSelect strain and each strain was assigned an integer identifier prefixed with a t_. For example, the single component in Fig. 2 is StrainSelect strain t_117676 and a vertex-rich component in Sup. Fig. 1 is StrainSelect strain t_47740. The single lowercase letter plus two underscores prefix format [34] was popularized when integrated into the Greengenes database [12] to disambiguate which taxonomic rank was referenced by a term (for example, p_Firmicutes, c_Bacilli, o_Staphylococcales, as the phylum, class and order names, respectively). Since s_ is already the prefix for species rank, t_ was used for the strain rank prefix. This final graph of 4,002,309 vertices and 4,380,302 edges in 681,087 strain components can be obtained in a single R iGraph formatted file as StrainSelect_iGraph.rds or in two tsv files StrainSelect_vertices.tab.txt and StrainSelect_edges.tab.txt.

The component-producing procedure did not presume nor constrain that each rs_assembly should belong to a different strain and therefore revealed that 10,043 strains have more than one high-quality assembly (Fig. 3) and, surprisingly, 19 strains have over 25. The general membership of vertex types among components was examined with multiple intersection analysis [9] in Fig. 4. Components containing brc_cat_id, rs_assembly, g16 and a gss vertices represent a large proportion of all components. Components containing g16, gss and brc_cat_id vertices without rs_assembly vertices were the most common. Overall, RefSeq only covers 36% of the strains in StrainSelect since most strains do not yet have high quality assemblies publicly available.

4.7. Taxonomy adaptation

The opportunity to apply a single taxonomic ontology to all the sequence records in StrainSelect to create a single ontology to encompass both the 16S rRNA genes and the genome assemblies was challenging. Some strains are missing either a high-quality 16S rRNA gene or a genome assembly, while others have multiples of each. In the StrainSelect schema, a contig belongs to exactly one

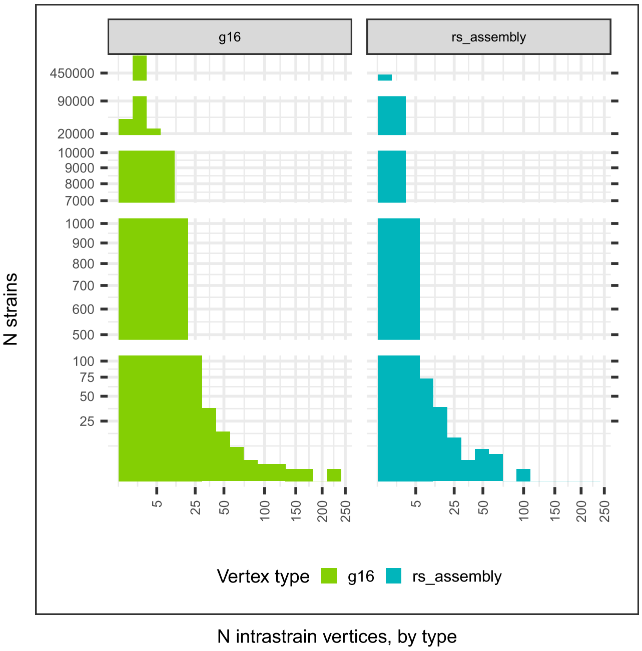

Fig. 3. Distribution of strains binned by count of available 16S rRNA gene records (`g16`) or RefSeq genome assembly (`rs_assembly`) records derived from the strain. Most strains have less than 25 of either type but a minority of strains have over 100 of these vertex types suggesting many technical replicates exist in the public databases.

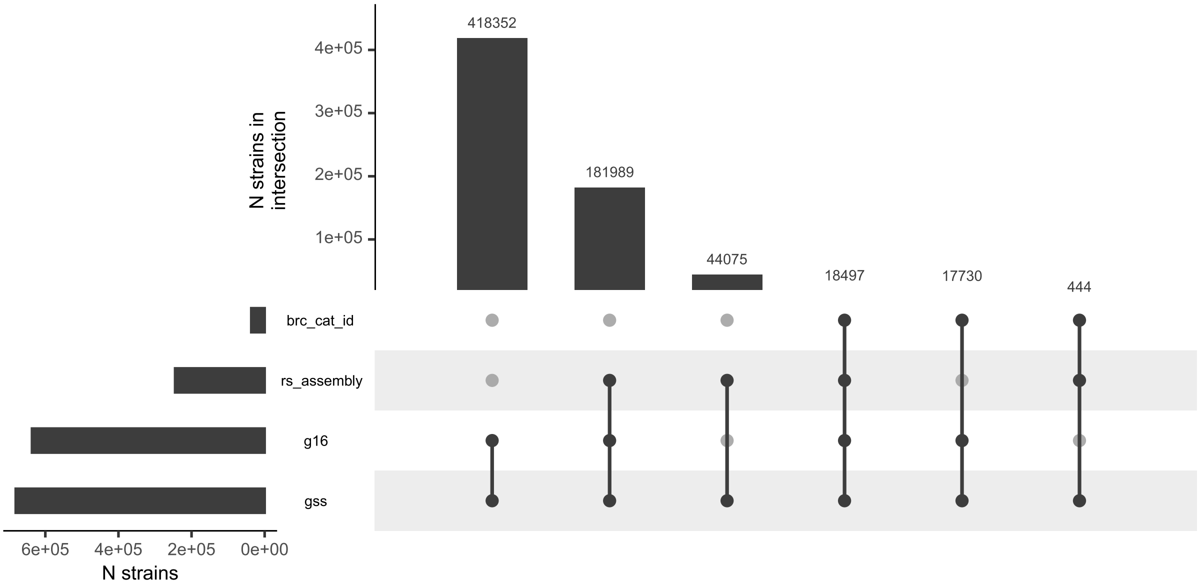

Fig. 4. Component counts based on presence or absence of four types of vertices. For the majority of strains, names and 16S rRNA genes are known but they have not been deposited in a BRC nor has a genome assembly been entered into RefSeq.

strain and a strain can belong to only one species, therefore the StrainSelect taxonomy is the first, to our knowledge, to ensure that `contig`s from the same strain do not end up in different taxonomic lineages. GTDB was conscripted as the base taxonomy because it has balanced traditional microbiological nomenclature with modern tree construction based on similarity across multiple genes [42] and has placed the majority of the RefSeq assemblies into categories from domain to species. The adaptation of GTDB taxonomy to satisfy the schema constraints of StrainSelect was accomplished for 236,992 strains as described below.

In a first step, taxon names in GTDB that are not in Latin form but instead take a variety of formats as placeholder strings used until agreement in the nomenclatural literature emerges, were identified. To these, StrainSelect assigned a consistently formatted provisional identifier using the characters “PROV” for reliable machine reading/parsing. For instance, the name “s_PROV_95247” indicates a species level taxon without a formal Latin name.

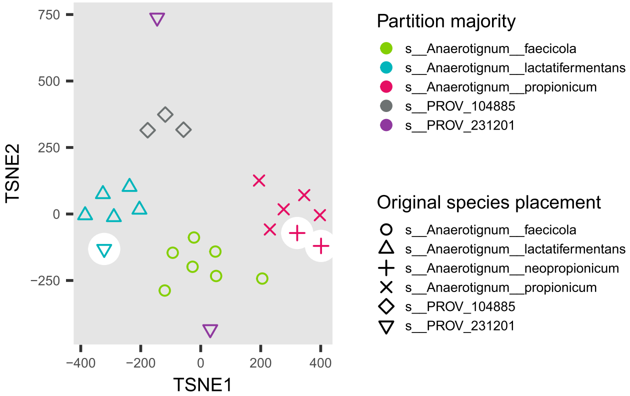

Fig. 5. Example of species placements of 16S rRNA genes from one bootstrap cycle within one family. In the Anaerotignaceae family, the DNA sequence distance matrix between the known high-quality 16S rRNA genes is partitioned around medoids (PAM) and visualized with t-distributed stochastic neighbor embedding (TSNE). Each point represents one 16S rRNA gene and the symbol represents their membership to one of six species before the initiation of the bootstrapping process. In this family, four formally and two provisionally (PROV) named species were available. After partitioning, the majority species within each partition is determined as represented by the color. The points highlighted with white circles are 16S rRNA genes that were affixed with new species names in this cycle. One gene within s_PROV_231201 (inverted triangle) was affixed with the species name s_Anaerotignum_lactatifermentans (blue) and two genes within s_Anaerotignum_neopropionicum (plus symbol) were affixed with the species name s_Anaerotignum_propionicum. After 100 cycles, only PROV assignments were adjusted in the final taxonomy when >50% of the bootstrap cycles were concordant. Also noteworthy are the three genes from s_PROV_231201 (inverted triangles) separated on both axes indicating, at least in one bootstrap cycle comparing these 16S rRNA genes, the instability of this taxon group.

A GTDB taxonomic placement was available for at least one genome assembly from 185,872 strains. Because the StrainSelect data model recognizes that some strains have replicate genome assemblies, we had to examine if GTDB had placed replicates in different lineages. We found, for small percentage (429 strains, 0.2%), the replicate assemblies were spread into more than one GTDB species. The discordance was minor. For example, two assemblies from one strain, t_104183, were placed by GTDB in distinct but sister species, GCF_001490875 in *Listeria monocytogenes* and GCF_001711055 in *Listeria monocytogenes_B*. Of the 185,443 strains without this discrepancy, 165,904 have one or more 16S rRNA genes, useful for anchor points for taxonomic estimation where only a 16S rRNA gene is available without a RefSeq genome assembly.

To classify all the BRC deposited strains not yet placed into a single GTDB lineage but with available 16S rRNA genes, the kmer-based sintax algorithm of usearch [18] was applied to each 16S sequence to make an initial placement for each gene. For strains with multiple 16S rRNA genes split to multiple species placements due to dissimilarities, preference was given for the Latin-named, non-provisional species placement with the greatest sintax confidence score and that preferred species was applied to all 16S rRNA genes of that strain. To test the stability of these initial strain-to-species memberships, 100 bootstrap cycles were performed where each 16S gene was compared against up to 200 randomly chosen intrafamily 16S genes and one randomly chosen 16S gene from a near-neighbor taxon outside the family (out-group). A multiple sequence alignment was solved by muscle [17], the hamming symmetric distance matrix [23] was calculated then partitioned by pamk, https://CRAN.R-project.org/package=fpc [24] as visualized with t-distributed stochastic neighbor embedding (TSNE) in Fig. 5. Partitions were created purely from the distance matrix without any added parameters for mutational rates nor tree-constructions since phylogeny was not the objective. Each gene in each bootstrap was affixed with the species name that comprised the majority of its partition. The percentage of bootstrap cycles where a gene was affixed to the same species was the gene-to-species bootstrap support score.

To then summarize the support from all genes from a strain, the strain-to-species bootstrap support score was the average observed among its 16S genes. Bootstrap support varied among strains and was compared against attributes of genome assemblies reported by GTDB. An inverse correlation ($p < 1e\text{-}90$) between bootstrap support and various metrics of genome size, G+C percentage and contamination was observed (Fig. 6, Sup. Fig. 3).

All strains originally placed in provisionally-named species but whose bootstrap support was >50% for an alternate species were re-assigned. Of the 429 strains with GTDB discrepancies described above, 211 had 16S rRNA genes available and were placed into a single species using this same method. In total, 236,992 strains were placed into a structured taxonomy with specific ranks from domain to strain.

4.8. Knowledge graph quality control

To verify the reliability of the final information linkage within the StrainSelect graph it was compared to pre-existing knowledge. Two highly dissimilar sources of pre-existing knowledge were used in the comparisons: NCBI's Prokaryote Type Strain Report (PTSR) and previously reported sequence similarity within taxonomic boundaries. The PTSR contains a map between `brc_cat_id`s that are replicate cultures of the same strain and one or more of the synonymous `gss` names. In this file was 9,267 edges between `brc_cat_id` and `gss` vertices and 8,953 of those edges have vertices that met all criteria for StrainSelect inclusion (96.6% coverage). If the construction of the StrainSelect graph introduced errant linkages, we should find cases where the two vertices connected by these 8,953 PTSR edges ended-up in different StrainSelectIDs as different strains. We observed zero of these errors. Thus, based

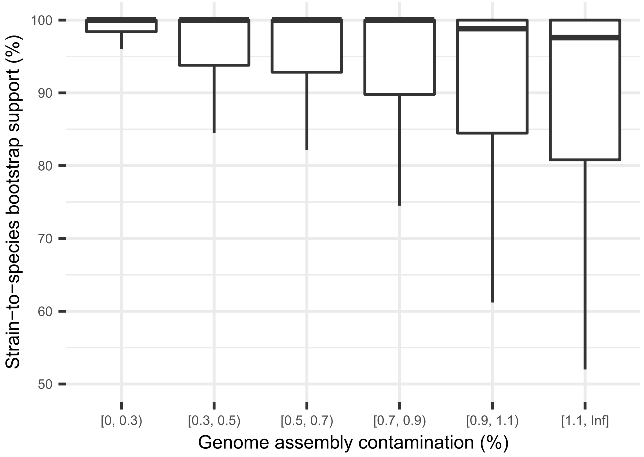

Fig. 6. Bootstrap support for species assignments from 16S rRNA analysis inversely correlates with assembly contamination. For 161,094 strains, all three of the following were available: RefSeq assemblies, GTDB-reported assembly contamination and 16S rRNA genes. Where multiple assemblies for a strain were available, the mean assembly contamination was calculated. A significant inverse relationship (Spearman correlation coefficient = -0.16, p < 1e-90) between the magnitude of a strain's genome assembly contamination and the likelihood that the strain's 16S rRNA genes come from the same species was observed.

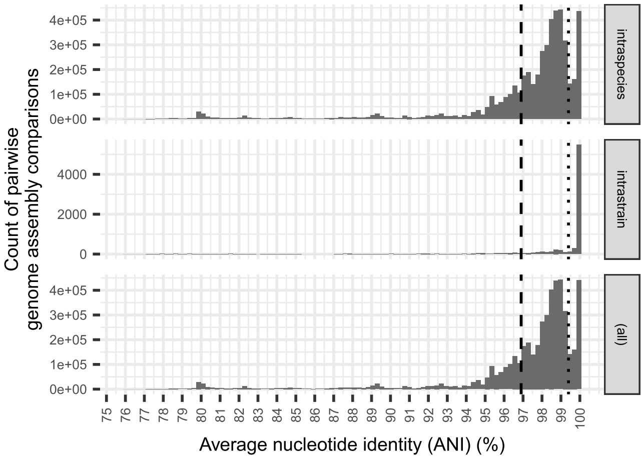

Fig. 7. Distribution of pairwise identities between genome assemblies within the same species (intraspecies) and within the same strain (intrastrain). Displayed are the observations from a set 80,741 assemblies within 75,894 strains from 5,436 species. Assemblies were compared pairwise for their average nucleotide identity (ANI). 75% of ANIs between assemblies from the same species but from different strains were greater than 96.9% (dashed line) while 75% of intrastrain ANIs was greater than 99.4% (dotted line).

on the PTSR comparison, the StrainSelect database includes nearly all type strains and, when passing all filters, reliably connects synonymous information into the same component.

The second comparison of the final graph to pre-existing knowledge was based on DNA sequence comparisons. Since the vertices of the knowledge graph are not connected by edges defined by genome sequence identity, the validity of components was evaluated by this metric as a *post hoc* analysis. If a meaningful demarcation among strains existed in the graph, we would expect the majority of components with more than one assembly to have low DNA divergence between those assemblies explained by technical variation expected when independent institutes sequence the same strain. Conversely, if the graph-building methods resulted in a poor-quality over-connected graph generating components that unintentionally merged assemblies derived from different strains, we would expect high divergence among intracomponent assemblies. To test this, a sampling of 80,741 assemblies within 75,894 strains from 5,436 species where >2 and <700 intraspecies genome assemblies were available were compared with FastANI [26] to determine the average nucleotide identity (ANI). Limiting the sampling to species with under 700 assemblies held-out species within *Escherichia* and *Shigella* which contain large numbers of assemblies with species boundaries under debate for likely reorganization [25] but included well-studied species such as *Yersinia pestis*, *Haemophilus influenzae*, and *Bacillus subtilis*. In previously published observations, intraspecies ANI among genome assembles is typically >95% [26,31,43]. Since strains are a finer taxonomy rank than species, we expected that most intrastrain ANIs should be at least this high. We observed that of the 3,621 strains investigated, only 264 (7.3%) contained a pair of assemblies <95% ANI, indicating that nearly all components had avoided over-merging vertices belonging to different species. This analysis also allowed a systematic estimate, for the first time to our knowledge, of the technical variation observed among assemblies from the same strain to be approximately 0.6% (Fig. 7). Consequently, we determined the distributions of intraspecies identities without the bias of repetitive intrastrain comparisons in Fig. 7 and observed an intraspecies ANI of $>96.9\%$.

To identify hub vertices potentially over-connecting genome assemblies that do not belong to the same strain, the betweenness centrality for each vertex (BC_V) was calculated to find the vertices acting as frequent bridges between divergent assemblies $<95\%$ ANI within the same component. Eq. (2) defines BC_V where P_S is the number of possible shortest paths from one `rs_assembly` to another and P_V are the count of those paths passing through vertex, V.

$$BC_V = \sum_{i=1}^{P_V} (1/P_s) \tag{2}$$

The vertex types accumulating the greatest BC_V were `gss` and `brc_cat_id` indicating that genome assemblies submitted to NCBI that share a genus-species-strain name and/or a BRC catalog identifier can, in rare cases, have divergent DNA sequence `contig`s. As a case study, we investigated the vertex with the greatest BC_V, the `gss` vertex, serratia.marcescens.cdc.813.60 from strain t_19847. This hub is perhaps reflective of the experimental design (NCBI BioProject: PRJEB40306) to produce many assemblies from isolates generated by thermal mutagenesis of a culture grown from ATCC 13880. In this case, it is not obvious that all these assemblies are still representative of a single strain even though the annotations attached to the assemblies asserted that they were. For all 264 strains containing divergent genome assemblies ($<95\%$ ANI), the strain was removed from the set of strains with taxonomy placements although it remains in the graph. This results in 236,992 strains with taxonomy placements of which 219,349 strains (92.6%) have ≥1 genome assembly and 217,454 strains (91.8%) have ≥1 16S rRNA gene.

4.9. *Reduced MinHash and fasta files*

Although all the DNA sequence data encompassed by StrainSelect can be downloaded from NCBI, we have provided users with reduced dimensionality reference files for taxonomic classification of shotgun metagenomic reads in sourmash's MinHash sketch format [45] with parameters -k 51 –scaled 5000. Sketches were attempted for 219,349 strains but 166 of these strains were omitted as only deprecated RefSeq assemblies were available. Of the remaining 219,183 strains, the single assembly with the lowest GTDB-reported contamination was included to represent the strain. The sum of contigs from all these assemblies is 873 Gb but after sketch formatting, aggregation, and compression all signatures fit into one 5.3 Gb file. The StrainSelect21_README.txt file accompanying the downloadable sourmash reference file contains example commands to assist informatitians in building computational pipelines.

A 16S rRNA gene reference fasta file was prepared using sintax-formatted taxonomy headers containing only intrastrain dereplicated sequences meaning that `g16` sequences which are an exact sub-sequence within another from the same strain were not included in the file. After compression the final file is 61 Mb in size and contains 333,204 16S rRNA gene sequences from 217,454 strains preserving intrastrain diversity helpful for training classifiers.

5. Discussion

The conceptual approach of connecting synonymous monikers for each strain sourced from a variety of data sources was overall successful and produced a knowledge graph with a variety of utility. It allowed us to run component discovery to find the boundaries around the data records pertaining to each of 681,087 strains. It facilitated calculations of betweenness centrality to prioritize, for manual inspection, the hub vertices potentially over-connecting identifiers, such as in *Serratia marcescens*. With the graph we could connect known 16S rRNA records and genome assemblies for each strain and discover cases where technical replicates are available. This empowered dissimilarity analysis among replicate genome assemblies and bootstrap support scoring for the taxonomic placement of species using 16S rRNA genes. Overall, we found the graph methodology to be appropriate for this application and able to cover a large portion of the graph with a structured taxonomy. Surprises that were encountered during the database build are worth consideration as they have implications on the future of microbial genomics and the adept usage of StrainSelect.

5.1. *On graph methodological validation*

There is a valid concern that bioinformatic creation of mega-graphs from public resources can over-connect information that domain experts would find disagreeable. For examples of problematic false or spurious edges in the domain of protein-protein interaction graphs see López et al. [33]. Since we combined large quantities of relationships from multiple public sources we benefited from a emphpost hoc test to measure the frequency of improbable connections, namely genome assemblies connected within the same strain but with divergence beyond what is likely from technical variation. We demonstrated that the StrainSelect graph building method defined reasonable boundaries between strains by component decomposition and revealed intracomponent (intrastrain) ANI was over 99.6% for 75% of the comparisons. In a second test of component integrity, we verified that 96.6% of type strain synonyms published by NCBI were included in StrainSelect and of those none were improperly separated into different components by any methodological step in the graph construction method. These observations, one using sequence comparison method independent of how StrainSelect constructed and the other using a knowledge preservation test reveals minor limitations of StrainSelect but provides evidence that the components, which are simply groups of data and monikers from a single strain, are generally reliable.

5.2. On comprehensive taxonomy

In building the taxonomic ontology for StrainSelect, we valued the work of Greengenes which implemented consistent data filtering, DNA similarity based taxa and consistent machine-friendly taxonomic ranks for all tree leaves and, even more so, GTDB which has carried the burden of balancing traditional microbiological taxonomic nomenclature with hierarchical incongruities revealed in multi-gene tree construction. Therefore, the basic ontology for StrainSelect will be familiar to users of either. Only 429 strains had multiple assemblies split between different GTDB species and those were either resolved to a single species based on the 16S rRNA genes (211 strains) or withheld from the structured taxonomy. Thus, 99.8% agreement exists between GTDB taxonomy and StrainSelect. The larger future endeavor will be to incorporate into the taxonomy the over 400,000 strains known only by a `gss` name and a 16S rRNA gene (Fig. 4). In the current version, these strains were left out of the taxonomy but with the steady reduction in DNA sequencing costs many of these strains' genomes are likely to become publicly available. The group of over 17,000 strains with 16S rRNA genes as well as cultures deposited at a BRC but without a RefSeq assembly (Fig. 4) are possibly queued for laboratory or bioinformatics progression for eventual broadcast via RefSeq. It's likely that at any point in time there will be a set of strains at this stage and StrainSelect includes them to build a more comprehensive taxonomy based on available 16S rRNA genes. Overall, the taxonomy includes 236,992 strains of which 219,349 (92.6%) have a genome assembly and 217,454 (91.8%) have a 16S rRNA gene. We expect both of these percentages to increase in future versions.

5.3. On implications for the field of microbial genomics

As a by-product of constructing this database and overcoming challenges in DNA sequence contamination, clandestine technical replicate records, and incorrect metadata, we formed some remarks on the general state of the field.

In this work 1.9% of RefSeq assemblies were eliminated from entering the knowledge graph due to interdomain contamination which means we were more permissive compared to EMBL's estimate that 5.2% of RefSeq genomes are impure [41]. EMBL may be correct because even after our RefSeq filter, we observed that even minor genome assembly contamination levels (Fig. 6) were inversely correlated to the bootstrap confidence of a strain's placement into a species. These observations are unsettling to the assumption that RefSeq is a pristine reference database for any genomics inquiry. It holds valuable data and has been a dependable resource with consistent availability for international collaborative research. But, until sequencing facilities or RefSeq editors can optimize the identification and elimination of contaminant contig regions, users should be aware that taxonomic placement, and more broadly, phylogenetic conclusions are subject to improvements.

Since duplicates and redundant information exist in biological databases, any database maintainer should assist their users by documenting how these cases are identified and handled. The presence of duplicate sequence records in bioinformatics reference databases creates inefficiency in computational search load, and in the assessment of the results of a search [8]. Most users would agree that clear duplicates, for instance an assembly from the same strain sequenced once at one institute but deposited at NCBI twice under different accession identifiers ought to be removed. But these types of duplicates don't appear to be the problem. Instead, we counted that for 10,043 strains submitters created genome assemblies from the same strain in different sequencing projects usually at different institutes. Whether these repeats produced slightly different results or identical results, these data observations should be made public to enable measurements of technical variation, for instance, but should be clearly labeled as such. Since NCBI does not attempt this after genome submissions but instead allows rich metadata to accompany a submission [20], it is up to the user to either determine from the metadata which assemblies are technical replicates and which are from distinct strains or to use a resource such as StrainSelect. In the StrainSelect graph, our findings of technical replicates derived from the same strain are documented and all are labeled with the same StrainSelectID.

Despite the capacity for NCBI data contributors to include metadata to describe one or a collection of assemblies, we uncovered a problem in naming isolates created from a mutagenesis experiment (NCBI BioProject: PRJEB40306) by re-applying the same name as the origin (parental) strain. This led to a graph component connecting a set of assemblies that were $<95\%$ ANI. Thus, it is recommended in these cases if a mutagen was applied and the genome content changed then the new isolate should be given a separate name from the parent strain. Otherwise the mutant genome assemblies would be assumed to be taken from a single strain and the casual data consumer would attribute divergence to technical artifacts/errors instead of the intentional experimental design.

We observed a genetic discontinuity between strains (Fig. 7) at 99.4% ANI. We contemplated an interassembly ANI exceeding this threshold as an edge in the graph construction process in future versions of StrainSelect. To add these ANI-based edges would result in fewer overall components but would merge genetic information where subject matter experts would keep them discriminated due to critical genes. For example, strains within the pathogenic species *Corynebacterium bovis* such as t_915 (synonyms: str. DSM 20582, str. Evans, str. CIP 54-80T, and 14 others) and t_254639 (synonym: str. MI 82-1021) have $>99.7\%$ ANI but have dissimilar virulence genes [7] warranting their distinction. Future research could involve weighting edges more when direct culture sharing is known and less if ANI is the only connection, then implementing a more sophisticated component boundary definition that would resolve a set of training cases such as within *C. bovis*, but in the current version no ANI edges were created in the knowledge graph.

5.4. On future directions for StrainSelect

In addition to potential future improvements in leveraging ANI, we also foresee opportunities in leveraging MAGs and consensus assemblies. It's conceivable that identical MAGs will be observed in multiple biospecimens as is suggested by a clinical study where $>50\%$ of a MAG can be $>99.999\%$ similar in two different stool samples [40]. Once metagenomic technology advances to enable entire MAGs to be found as nearly identical in among biospecimens, and the recurring MAGs are dissimilar to known isolates then StrainSelect should recognize them as yet-to-be-isolated strains. In the meantime, if a research team has ample resources to culture isolates matching MAGs from their metagenomic sequencing, then the isolates and corresponding genome assemblies should be submitted to NCBI and BRCs, respectively, to increase the diversity available.

As the number of MAGs and assemblies grow, the number of strains with technical replicate assemblies will also grow. In our current build, when technical replicates were found, we selected the least contaminated for inclusion in the MinHash (sourmash) database. Alternatively, one could create a single consensus assembly before the MinHash is derived. A potential tool to implement this process would be Trycycler [59] although in its current implementation requires subjective steps in post-processing, or polishing, which would introduce a non-reproducible step in the StrainSelect build. Once a validated automated process is available, StrainSelect will likely focus MinHashes to regions harmonious across technical replicates.

5.5. *On usage of StrainSelect*

The first published usage of StrainSelect was described for organizing raw fecal 16S rRNA gene sequencing data to identify a composite biomarker for colorectal cancer [51]. In the manuscript, binning the reads by unique matches to one strain, where possible, was compared to binning the reads by a popular operational taxonomic unit (OTU) method. The StrainSelect method produced biomarkers that outperformed the OTU method in accurately classifying patients. Other notable examples are the use of StrainSelect to pinpoint the strains phagocytized by specific macrophage types in Crohn's Disease patients [50], and to enable a strain-level meta-analysis across 21 Inflammatory Disease datasets [46] and across 10 Autism Spectrum Disorder data sets [57].

With the description of StrainSelect herein, biologists can now choose to organize microbiome profiling data from shotgun or amplicon techniques for strain-level analyses. Compared to shotgun metagenomic laboratory techniques, it is expected that a fewer number of individual strains will be uniquely detected from short 16S rRNA gene amplicons covering only 1 or 2 of the 9 hyper-variable regions of the 16S rRNA gene [28]. The short NGS reads covering 1 or 2 hyper-variable regions of the gene often align equally well to sequences from multiple taxa [27], limiting the ability to pinpoint specific strains. Longer amplicons that span all 9 regions [15] are preferable and can be assessed by Sanger sequencing, probe arrays such as the PhyloChip [11] and now also possible with long read high-throughput sequencing [28]. As with all bioinformatic sequence reference databases, as knowledge of new strains expands, we expect two changes to previously published shotgun metagenomic and 16S rRNA amplicon findings. First, additional reads from the raw data will match to the newly isolated and sequenced strains and, second, some reads that were perceived as evidence of unique strain hits in the past will be determined to be non-unique in the future. Data platforms will need to be developed that can easily remap all public raw data into strain bins in a cost-effective manner with each update of StrainSelect. These platforms will need stable funding and resources since the growth of this data is not exhibiting deceleration (Fig. 1).

6. Conclusion

StrainSelect is a reference database of archaeal and bacterial genomic identifiers organized by strain (see Graphical abstract for a visual summary). StrainSelect assigns a consistently formatted identifier for known strains that have been isolated and have had their genome assembled or at least their 16S rRNA gene assembled and shared publicly. StrainSelect has three important properties. First, the database appropriately labels each contig and genome assembly by the strain of origin carefully avoiding conflation of genome assemblies as strains and in doing so identified over 10,000 strains with at least two technical replicate assemblies. Second, each strain is mapped to the bio-resource centers where the live strain can be procured if extant. Third, a single comprehensive domain to strain taxonomic ontology is included integrating both 16S rRNA genes and genome assemblies as points of reference so meta-analysis sourced from both technologies are possible. StrainSelect, with 681,087 strains demarcated, is the largest collection of its kind. The database can be inspected in graph or tabular formats in its entirety allowing mapping between StrainSelect strain identifiers, genome assemblies, 16S rRNA genes, international bio-resource center catalog identifiers for strain procurement, and genome function-focused databases.

With the StrainSelect foundation, research teams can annotate microbial community data into strain-level biomarkers, accelerate translational research after biomarker discovery into *in vivo* laboratory experiments with those strains to establish causality and confirm findings with meta-analyses across a growing public data warehouse containing 16S rRNA and shotgun metagenomics data.

StrainSelect database is available for download at http://strainselect.secondgenome.com.

Funding

This work was supported in part by the National Institutes of Health's (NIH) National Institute on Drug Abuse (NIDA) [R44DA043954].

CRediT authorship contribution statement

Todd Z. DeSantis: Conceptualization, Software, Formal analysis, Data Curation, Writing, Visualization. **Cesar Cardona:** Software, Investigation. **Nicole R. Narayan:** Methodology, Investigation. **Satish Viswanatham:** Resources, Supervision. **Divya Ravichandar:** Methodology. **Brendan Wee:** Software. **Cheryl-Emiliane Chow:** Methodology. **Shoko Iwai:** Methodology.

Declaration of competing interest

All authors were employed by Second Genome, Inc. during the course of the work. Second Genome, Inc. is a biomarkers and therapeutics company with products in development derived from bacterial polypeptides to treat cancer, ulcerative colitis and other human diseases. A publication announcing the availability of StrainSelect for academic use will allow greater transparency on the contents of our reference database but will not affect the value of our therapeutic products.

Acknowledgements

The authors thank Dr. Peter Karp at SRI International, Dr. Phil Hugenholtz at the University of Queensland, Dr. Nikos Krypides at the Joint Genome Institute, and Drs. Wim De Smet and Peter Dawyndt at Ghent University for their helpful discussions regarding the databases they initiated and curated. We also appreciate the guidance of Dr. Alex Probst in forming the discussion on MAGs as future entities in StrainSelect.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e13314.

References

[1] A. Almeida, A.L. Mitchell, M. Boland, S.C. Forster, G.B. Gloor, A. Tarkowska, T.D. Lawley, R.D. Finn, A new genomic blueprint of the human gut microbiota, Nature 568 (7753) (2019) 499–504.

[2] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, Nucleic Acids Res. 41 (Database issue) (2013) D36–42.

[3] M.I. Bidartondo, Preserving accuracy in GenBank, Science 319 (5870) (2008) 1616.

[4] D.R. Boone, R.W. Castenholz, Bergey's Manual of Systematic Bacteriology, 2nd edition, Springer, New York, 2001.

[5] F. Briatte, ggnetwork: Geometries to Plot Networks with ggplot2, 2021.

[6] D. Burstein, L.B. Harrington, S.C. Strutt, A.J. Probst, K. Anantharaman, B.C. Thomas, J.A. Doudna, J.F. Banfield, New CRISPR-Cas systems from uncultivated microbes, Nature 542 (7640) (2017) 237–241.

[7] C. Cheleuitte-Nieves, C.A. Gulvik, J.R. McQuiston, B.W. Humrighouse, M.E. Bell, A. Villarma, V.A. Fischetti, L.F. Westblade, N.S. Lipman, Genotypic differences between strains of the opportunistic pathogen corynebacterium bovis isolated from humans, cows, and rodents, PLoS ONE 13 (12) (2018) e0209231.

[8] Q. Chen, J. Zobel, K. Verspoor, Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study, Database (Oxford) 2017 (2017).

[9] J.R. Conway, A. Lex, N. Gehlenborg, UpSetR: an R package for the visualization of intersecting sets and their properties, Bioinformatics 33 (18) (2017) 2938–2940.

[10] G. Csardi, T. Nepusz, The Igraph Software Package for Complex Network Research, 2006.

[11] T.Z. DeSantis, I. Dubosarskiy, S.R. Murray, G.L. Andersen, Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA, Bioinformatics 19 (12) (2003) 1461–1468.

[12] T.Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E.L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, G.L. Andersen, Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, Appl. Environ. Microbiol. 72 (7) (2006) 5069–5072.

[13] L. Dijkshoorn, B.M. Ursing, J.B. Ursing, Strain, clone and species: comments on three basic concepts of bacteriology, J. Med. Microbiol. 49 (5) (2000) 397–401.

[14] M. Dowle, A. Srinivasan, data.table: Extension of data.frame, 2021.

[15] L.M. Durso, G.P. Harhay, T.P.L. Smith, J.L. Bono, T.Z. Desantis, D.M. Harhay, G.L. Andersen, J.E. Keen, W.W. Laegreid, M.L. Clawson, Animal to animal variation in fecal microbial diversity among beef cattle, Appl. Environ. Microbiol. 76 (14) (2010) 4858–4862.

[16] C. Duvallet, S.M. Gibbons, T. Gurry, R.A. Irizarry, E.J. Alm, Meta-analysis of gut microbiome studies identifies disease-specific and shared responses, Nat. Commun. 8 (1) (2017) 1784.

[17] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (5) (2004) 1792–1797.

[18] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26 (19) (2010) 2460–2461.

[19] S. Federhen, Type material in the NCBI taxonomy database, Nucleic Acids Res. 43 (Database issue) (2015) D1086–1098.

[20] S. Federhen, K. Clark, T. Barrett, H. Parkinson, J. Ostell, Y. Kodama, J. Mashima, Y. Nakamura, G. Cochrane, I. Karsch-Mizrachi, Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records, Stand. Genom. Sci. 9 (3) (2014) 1275–1277.

[21] D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M.J. Allen, S.V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. DePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F.O. Glöckner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S.E. Lewis, K. Li, A.L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S.-A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzel, I. San Gil, G. Wilson, A. Wipat, The minimum information about a genome sequence (MIGS) specification, Nat. Biotechnol. 26 (5) (2008) 541–547.

[22] D.H. Haft, M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin, K. O'Neill, W. Li, F. Chitsaz, M.K. Derbyshire, N.R. Gonzales, M. Gwadz, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, R.A. Yamashita, C. Zheng, F. Thibaud-Nissen, L.Y. Geer, A. Marchler-Bauer, K.D. Pruitt, RefSeq: an update on prokaryotic genome annotation and curation, Nucleic Acids Res. 46 (D1) (2018) D851–D860.

[23] R.W. Hamming, Error detecting and error correcting codes, Bell Syst. Tech. J. 29 (2) (1950) 147–160.

[24] C. Hennig, fpc: Flexible Procedures for Clustering, 2020.

[25] D. Hu, B. Liu, L. Wang, P.R. Reeves, Living trees: high-quality reproducible and reusable construction of bacterial phylogenetic trees, Mol. Biol. Evol. 37 (2) (2020) 563–575.

[26] C. Jain, L.M. Rodriguez-R, A.M. Phillippy, K.T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries, Nat. Commun. 9 (1) (2018) 5114.

[27] J. Jeong, K. Yun, S. Mun, W.-H. Chung, S.-Y. Choi, Y.-d. Nam, M.Y. Lim, C.P. Hong, C. Park, Y.J. Ahn, K. Han, The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology, Sci. Rep. 11 (1) (2021) 1727.

[28] J.S. Johnson, D.J. Spakowicz, B.-Y. Hong, L.M. Petersen, P. Demkowicz, L. Chen, S.R. Leopold, B.M. Hanson, H.O. Agresta, M. Gerstein, E. Sodergren, G.M. Weinstock, Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis, Nat. Commun. 10 (1) (2019) 5029.

[29] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (D1) (2017) D353–D361.

[30] P.D. Karp, R. Billington, R. Caspi, C.A. Fulcher, M. Latendresse, A. Kothari, I.M. Keseler, M. Krummenacker, P.E. Midford, Q. Ong, W.K. Ong, S.M. Paley, P. Subhraveti, The BioCyc collection of microbial genomes and metabolic pathways, Brief. Bioinform. 20 (4) (2019) 1085–1093.

[31] M. Kim, H.-S. Oh, S.-C. Park, J. Chun, Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes, Int. J. Syst. Evol. Microbiol. 64 (Pt 2) (2014) 346–351.

[32] K.A. Lee, A.M. Thomas, L.A. Bolte, J.R. Björk, L.K. de Ruijter, F. Armanini, F. Asnicar, A. Blanco-Miguez, R. Board, N. Calbet-Llopart, L. Derosa, N. Dhomen, K. Brooks, M. Harland, M. Harries, E.R. Leeming, P. Lorigan, P. Manghi, R. Marais, J. Newton-Bishop, L. Nezi, F. Pinto, M. Potrony, S. Puig, P. Serra-Bellver, H.M. Shaw, S. Tamburini, S. Valpione, A. Vijay, L. Waldron, L. Zitvogel, M. Zolfo, E.G.E. de Vries, P. Nathan, R.S.N. Fehrmann, V. Bataille, G.A.P. Hospers, T.D. Spector, R.K. Weersma, N. Segata, Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma, Nat. Med. 28 (3) (2022) 535–544.

[33] Y. López, K. Nakai, A. Patil, HitPredict version 4: comprehensive reliability scoring of physical protein-protein interactions from more than 100 species, Database (Oxford) 2015 (2015), bav117.

[34] D. McDonald, M.N. Price, J. Goodrich, E.P. Nawrocki, T.Z. DeSantis, A. Probst, G.L. Andersen, R. Knight, P. Hugenholtz, An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea, ISME J. 6 (3) (2012) 610–618.

[35] R. Mendes, M. Kruijt, I. de Bruijn, E. Dekkers, M. van der Voort, J.H.M. Schneider, Y.M. Piceno, T.Z. DeSantis, G.L. Andersen, P.A.H.M. Bakker, J.M. Raaijmakers, Deciphering the rhizosphere microbiome for disease-suppressive bacteria, Science 332 (6033) (2011) 1097–1100.

[36] A.K. Mishra, G. Gimenez, J.-C. Lagier, C. Robert, D. Raoult, P.-E. Fournier, Genome sequence and description of Alistipes senegalensis sp. nov., Stand. Genom. Sci. 6 (3) (2012) 1–16.

[37] M. Morotomi, F. Nagai, Y. Watanabe, Description of Christensenella minuta gen. nov., sp. nov., isolated from human faeces, which forms a distinct branch in the order Clostridiales, and proposal of Christensenellaceae fam. nov, Int. J. Syst. Evol. Microbiol. 62 (Pt 1) (2012) 144–149.

[38] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, H.Y. Katta, A. Mojica, I.-M.A. Chen, N.C. Kyrpides, T. Reddy, Genomes OnLine database (GOLD) v. 7: updates and new features, Nucleic Acids Res. 47 (D1) (2019) D649–D659.

[39] S. Nayfach, Z.J. Shi, R. Seshadri, K.S. Pollard, N.C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome, Nature 568 (7753) (2019) 505–510.

[40] M.R. Olm, A. Crits-Christoph, K. Bouma-Gregson, B.A. Firek, M.J. Morowitz, J.F. Banfield, inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains, Nat. Biotechnol. 39 (6) (2021) 727–736.

[41] A. Orakov, A. Fullam, L.P. Coelho, S. Khedkar, D. Szklarczyk, D.R. Mende, T.S.B. Schmidt, P. Bork, GUNC: detection of chimerism and contamination in prokaryotic genomes, Genome Biol. 22 (1) (2021) 178.

[42] D.H. Parks, M. Chuvochina, D.W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life, Nat. Biotechnol. 36 (10) (2018) 996–1004.

[43] D.H. Parks, M. Chuvochina, C. Rinke, A.J. Mussig, P.-A. Chaumeil, P. Hugenholtz, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy, Nucleic Acids Res. 50 (D1) (2022) D785–D794.

[44] A.Y. Pei, W.E. Oberdorf, C.W. Nossa, A. Agarwal, P. Chokshi, E.A. Gerz, Z. Jin, P. Lee, L. Yang, M. Poles, S.M. Brown, S. Sotero, T.Z. DeSantis, E. Brodie, K. Nelson, Z. Pei, Diversity of 16S rRNA genes within individual prokaryotic genomes, Appl. Environ. Microbiol. 76 (12) (2010) 3886–3897.

[45] N.T. Pierce, L. Irber, T. Reiter, P. Brooks, C.T. Brown, Large-scale sequence comparisons with sourmash, F1000Res. 8 (2019) 1006.

[46] J.D. Ravichandar, E. Rutherford, C.-E.T. Chow, A. Han, M.L. Yamamoto, N. Narayan, G.G. Kaplan, P.L. Beck, M.J. Claesson, K. Dabbagh, S. Iwai, T.Z. DeSantis, Strain level and comprehensive microbiome analysis in inflammatory bowel disease via multi-technology meta-analysis identifies key bacterial influencers of disease, Front. Microbiol. 13 (2022) 961020.

[47] R_Core_Team, R: A Language and Environment for Statistical Computing, 2021.

[48] L.C. Reimer, J. Sardà Carbasse, J. Koblitz, C. Ebeling, A. Podstawka, J. Overmann, BacDive in 2022: the knowledge base for standardized bacterial and archaeal data, Nucleic Acids Res. 50 (D1) (2022) D741–D746.

[49] M.G. Rooks, P. Veiga, A.Z. Reeves, S. Lavoie, K. Yasuda, Y. Asano, K. Yoshihara, M. Michaud, L. Wardwell-Scott, C.A. Gallini, J.N. Glickman, N. Sudo, C. Huttenhower, C.F. Lesser, W.S. Garrett, QseC inhibition as an antivirulence approach for colitis-associated bacteria, Proc. Natl. Acad. Sci. USA 114 (1) (2017) 142–147.

[50] Y. Sekido, J. Nishimura, K. Nakano, T. Osu, C.-E.T. Chow, H. Matsuno, T. Ogino, S. Fujino, N. Miyoshi, H. Takahashi, M. Uemura, C. Matsuda, H. Kayama, M. Mori, Y. Doki, K. Takeda, M. Uchino, H. Ikeuchi, T. Mizushima, Some Gammaproteobacteria are enriched within CD14+ macrophages from intestinal lamina propria of Crohn's disease patients versus mucus, Sci. Rep. 10 (1) (2020) 2988.

[51] M.S. Shah, T.Z. DeSantis, T. Weinmaier, P.J. McMurdie, J.L. Cope, A. Altrichter, J.-M. Yamal, E.B. Hollister, Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer, Gut 67 (5) (2018) 882–891.

[52] I. Sharon, M. Kertesz, L.A. Hug, D. Pushkarev, T.A. Blauwkamp, C.J. Castelle, M. Amirebrahimi, B.C. Thomas, D. Burstein, S.G. Tringe, K.H. Williams, J.F. Banfield, Accurate, multi-kb reads resolve complex populations and detect rare microorganisms, Genome Res. 25 (4) (2015) 534–543.

[53] A. Sivan, L. Corrales, N. Hubert, J.B. Williams, K. Aquino-Michaels, Z.M. Earley, F.W. Benyamin, Y.M. Lei, B. Jabri, M.-L. Alegre, E.B. Chang, T.F. Gajewski, Commensal bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy, Science 350 (6264) (2015) 1084–1089.

[54] M. Tessler, J.S. Neumann, E. Afshinnekoo, M. Pineda, R. Hersch, L.F.M. Velho, B.T. Segovia, F.A. Lansac-Toha, M. Lemke, R. DeSalle, C.E. Mason, M.R. Brugler, Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing, Sci. Rep. 7 (1) (2017) 6589.

[55] B. Verslyppe, W. De Smet, B. De Baets, P. De Vos, P. Dawyndt, StrainInfo introduces electronic passports for microorganisms, Syst. Appl. Microbiol. 37 (1) (2014) 42–50.

[56] A.R. Wattam, T. Brettin, J.J. Davis, S. Gerdes, R. Kenyon, D. Machi, C. Mao, R. Olson, R. Overbeek, G.D. Pusch, M.P. Shukla, R. Stevens, V. Vonstein, A. Warren, F. Xia, H. Yoo, Assembly, annotation, and comparative genomics in PATRIC, the all bacterial bioinformatics resource center, Methods Mol. Biol. 1704 (2018) 79–101.

[57] K.A. West, X. Yin, E.M. Rutherford, B. Wee, J. Choi, B.S. Chrisman, K.L. Dunlap, R.L. Hannibal, W. Hartono, M. Lin, E. Raack, K. Sabino, Y. Wu, D.P. Wall, M.M. David, K. Dabbagh, T.Z. DeSantis, S. Iwai, Multi-angle meta-analysis of the gut microbiome in Autism Spectrum Disorder: a step toward understanding patient subgroups, Sci. Rep. 12 (1) (2022) 17034.

[58] T.J. Wheeler, S.R. Eddy, nhmmer: DNA homology search with profile HMMs, Bioinformatics 29 (19) (2013) 2487–2489.

[59] R.R. Wick, L.M. Judd, L.T. Cerdeira, J. Hawkey, G. Méric, B. Vezina, K.L. Wyres, K.E. Holt, Trycycler: consensus long-read assemblies for bacterial genomes, Genome Biol. 22 (1) (2021) 266.

[60] H. Wickham, ggplot2: Elegant Graphics for Data Analysis. Use R!, 2nd ed. 2016 edition, Springer International Publishing: Imprint: Springer, Cham, 2016.

[61] S. Xu, M. Chen, T. Feng, L. Zhan, L. Zhou, G. Yu, Use ggbreak to effectively utilize plotting space to deal with large datasets and outliers, Front. Genet. 12 (2021) 774846.

[62] H. Zhu, kableExtra: Construct Complex Table with kable and Pipe Syntax, 2021.