Large scale function-based genome prospecting for industrial traits applied to 1,3-propanediol production.

Jasper J. Koehorst, Nikolaos Strepis, Sanne de Graaf, Alfons J. M. Stams, Diana Z. Sousa, and Peter J. Schaap

Introduction

Next generation sequencing (NGS) technologies have turned the publicly available genome repositories into data-rich scientific resources that currently contain structural and functional genome information of hundreds of thousands bacterial genomes (1) available for further exploitation for medical, environmental and biotechnological purposes. While the biotechnology field has embraced the use of omics technologies in biotech innovation in multiple ways, functional screening of genomic resources for industrially relevant traits at a large scale remains challenging. NGS data generation has caused manual curation to be outpaced rapidly and currently more than 99% of the functional predictions in UniProt are based on automated computational predictions (2). The accuracy of these computationally inferred functional genome annotations is largely unknown due to inconsistent naming of protein functions, use of annotation pipelines of various quality, over-prediction due to the use of generic, non-standardised, annotation acceptance thresholds (3), and lack of data provenance in general. This results in a degree of interoperability lower than required for efficient function based genome prospecting (4). As the predicted gene and protein sequences, underlying the functional annotations are available in the standardised FASTA format, 'bottom-up' approaches that start with a gene or protein sequence in an effort to find a corresponding function in genome(s) of interest are normally used. However, over larger phylogenetic distances, strategies that use sequence-based clustering algorithms are hampered by accumulation of evolutionary signals, lateral gene transfer, gene fusion/fission events and domain expansions (5). Furthermore, when applied at larger scales, such strategies suffer from high computational time and memory requirements that scale quadratically with the number of genome sequences to be compared (6). A robust function-based genome screening requires a sufficiently high degree of interoperability meaning that functional information can be directly compared on the basis of a pre-established syntactic interoperable genome annotation and that any computational functional prediction is directly linked to its provenance. To accomplish this, we have developed SAPP, a semantic annotation infrastructure supporting FAIR computational genomics (7). SAPP uses the Genome Biology Ontology Language (GBOL) as syntax (8) and automatically predicts, tracks and stores structural and functional genome predictions and associated data- and element-wise provenance in RDF, the W3C standard for representing linked data (9). SAPP uses protein domain architectures to systematically describe protein functions (10). Demonstrating a high level of scalability, this set up has already been successfully used in an integrated analysis of the functional landscape of 432 Pseudomonas strains (4). In this study, interoperable genome annotations are used in a systematic function-based silico screening for bacterial species that potentially can convert the bio-refinery by-product glycerol in 1,3-propanediol (1,3-PD) (11, 12). 1,3-PD is an important precursor of biodegradable plastics, such as poly(trimethylene terephthalate) (11). 1,3-PD is a typical product of anaerobic glycerol fermentation and only a limited number of species, mostly enterobacteria, are known to form it. The first step, dehydration of glycerol to 3-hydroxypropionaldehyde (3-HPA) is mediated by a vitamin B12-dependent glycerol dehydratase although an oxygen sensitive, B12-independent alternative enzyme has also been reported (13). The second step reduces the toxic 3-HPA intermediate to 1,3-PD using the (NADH)+H+ dependent 1,3-propanediol-oxydoreductase (PDOR) regenerating NAD+ (12).

For large scale genome prospecting a collection of 83,308...
publicly available bacterial genomes were loaded in the SAPP semantic framework, systematically structurally and functionally annotated and subsequently mined for candidate 1,3-PD producers using a knowledge discovery in databases approach (KDD) (14). Overall, the systematic analysis increased our knowledge on the genetic architecture of this metabolic trait, in terms of the overall domain composition, domain order, distribution, persistence and essentiality. The approach suggested that, in addition to the some 30 producers reported in literature (table 1), at least 187 other genome sequenced species potentially have this trait.

Results

Development of the genome mining workflow. To cope with the influx of biological input data, we followed the Knowledge Discovery in Databases multi step process model (14), which was adjusted to the specific needs of being able to prospect genome data at a large scale. The KDD process includes multiple steps with the first step being to capture relevant prior knowledge. In the next step, a selection of data is used as a training set to develop a screening procedure followed by data mining at a large scale and finally the evaluation and interpretation of the results.

Incorporation of prior knowledge. In 1,3-PD production, two pathways, an oxidative and reductive pathway, are used in parallel for dissimilation of glycerol. In the oxidative pathway, glycerol is first dehydrogenated to dihydroxyacetone by a NAD+-linked glycerol dehydrogenase, and then converted to dihydroxyacetone phosphate by an ATP-dependent dihydroxyacetone kinase (15). Additionally, two alternative reductive pathways, one dependent on vitamin B12, the other not, are used for regenerating NAD+ with 1,3-PD as byproduct. For validation of the search strategy and to gain further insights into the genetics and the domains involved in 1,3-PD formation a list of known 1,3 PD producing strains was obtained for cross-validation. From literature some 30 strains have been reported to produce 1,3-PD (Table 1). For training purposes genome sequences from 13 different strains could be obtained from the EBI-ENA data warehouse (1) as not for all of them genome sequences were available. For Citrobacter freundii ATCC8090 two draft genome sequences of comparable quality were available and initially both were added to the training set increasing the total number of genomes.

Data transformation. To obtain a high degree of interoperability the selected genome sequences were de novo structurally and functionally annotated using the Semantic annotation Platform with Provenance (SAPP) (7) using Prodigal (16) for gene prediction and InterProScan (17) for protein annotation. Protein domain architectures were used to systematically describe the encoded protein functions (10). Structural and functional genome annotations were subsequently transformed into Linked Data fragments using the Resource Description Framework (RDF) as a data-metadata model (7, 18).

A function-based search strategy for the oxidative branch. The first two enzymes essential for the oxidative
pathway involved in the production of 1,3 PD are glycerol dehydrogenase which is dependent on NADH and produces dihydroxyacetone (DHA) and DHA-kinase, which subsequently phosphorylates DHA. These key reactions are represented by four Pfam domains; Fe-ADH (PF00465) for glycerol dehydrogenase and DAK1, (PF02734) or DAK1_2, (PF13684) both capturing the kinase domain of the dihydroxyacetone kinase family in combination with DAK2, (PF2040325) capturing the phosphatase domain of the dihydroxyacetone kinase family (Figure 1). A pattern recognition analysis was performed (see methods section for details) for DAK2 domain neighbours in the training set database followed by a Boolean DAK1/DAK2 or DAK1_2/DAK2 proximity search in the search results. Pattern recognition showed that the 14 genomes of the training data set encoded at least one dihydroxyacetone kinase family protein with either a DAK1/DAK2 or DAK1_2/DAK2 protein architecture; 13 strains coded for DAK1/DAK2, 8 strains for DAK1_2/DAK2 and 7 strains encoded for both configurations (Table 1). In addition, other domains were identified that clustered together with the DAK configuration such as an Alkaline shock protein, a phosphotransferase system component and an alcohol dehydrogenase domain. Another copy of this domain is an important factor for the 1,3 PD pathway (Table 2).

### Table 1. Overview of known 1,3-PD producers.

| Organism                      | Genome assembly ID | B12 (in silico prediction) | DAK configuration | 1,3-PD (mol/mol) | Acetate (mol/mol) | Reference                        |
|-------------------------------|--------------------|----------------------------|-------------------|-----------------|-----------------|----------------------------------|
| *Clostridium pefringens* sta. 13 | GCA_000009685.1    | Dependent                  | DAK1/DAK1_2       | 0.58            | 0.03            | Petidemange et al., 1995         |
| *Clostridium butyricum* E4    | GCA_000182605.1    | Independent                | DAK1/DAK1_2       | 0.41            | 0.03            | Yang et al., 2007                |
| *Klebsiella michignensis* strain M5al ==oxygena | GCA_000030735.2 | (in-) dependent            | DAK1             | 0.65            | 0.22            | Barbirato et al., 1998           |
| *Citrobacter freundii* ATCC 8090 | GCA_00012465.1    | Dependent                  | DAK1             | 0.65            | 0.22            | Barbirato et al., 1998           |
| *Citrobacter freundii* ATCC 8090 | GCA_000073405.1    | Dependent                  | DAK1             | 0.65            | 0.12            | Taconi et al., 2009              |
| *Halanaeroebium saccharolyticum* DSM 6643 | GCA_000350165.1 | Dependent                  | DAK1             | 0.41            | 0.26            | Kivisto et al., 2012             |
| *Clostridium diol DSM* 15410 | GCA_000406955.1    | Independent                | DAK1             | 0.58            | 0.12            | Otte et al., 2009                |
| *Klebsiella pneumoniae* ATCC 25955 | GCA_000049751.1 | Dependent                  | DAK1             | 0.65            | 0.22            | Barbirato et al., 1998           |
| *Clostridium butyricum* VPI3266 = DSM 10702 | GCA_000497551.1 | Dependent                  | DAK1             | 0.6             | 0.01            | Gonzalez-Pajuelo et al., 2004    |
| *Raoultella planticola* DR3 = DSM ATCC 33531 | GCA_000735435.1 | Independent (no alc)       | DAK1             | N.D.            | 0.15            | Jarvis et al., 1997              |
| *Klebsiella pneumoniae* DSM 2026 = ATCC 15380 | GCA_000949515.1 | Dependent                  | DAK1             | 0.62            | 0.09            | Mu et al., 2006                  |
| *Clostridium beijerincki* NBRL B-593 | GCA_00200325.1 | Independent (no alc)       | DAK1_2           | 0.59            | 0               | Guzman-Mendez et al., 2011       |
| *Trichosporum strain* E5S | GCA_900067165.1 | Dependent                  | DAK1/DAK1_2       | 0.66            | 0.21            | Streips 2020                     |
| *Caloramator vierebens* DSM 13723T |                      |                            |                   |                 |                 |                                  |
| *Clostridium butyricum* AKR012a |                      |                            |                   |                 |                 |                                  |
| *Clostridium butyricum* CNCM 1211 |                      |                            |                   |                 |                 |                                  |
| *Clostridium butyricum* DSM 5343 |                      |                            |                   |                 |                 |                                  |
| *Clostridium butyricum* E5 |                      |                            |                   |                 |                 | Rehman et al., 2008              |
| *Clostridium butyricum* F2b |                      |                            |                   |                 |                 | Papanikolaou et al., 2004        |
| *Clostridium butyricum* VPI 1718 |                      |                            |                   |                 |                 | Chazaratzkou et al., 2011        |
| *Clostridium multifermentans* DSM 5431 |                      |                            |                   |                 |                 |                                  |
| *Klebsiella pneumoniae* DSM4799 |                      |                            |                   |                 |                 |                                  |
| *Klebsiella pneumoniae* JS-26 |                      |                            |                   |                 |                 |                                  |
| *Klebsiella pneumoniae* JS-Li |                      |                            |                   |                 |                 |                                  |
| *Lactobacillus brevis* NV1.3 |                      |                            |                   |                 |                 |                                  |
| *Lactobacillus buchneri* B 190 |                      |                            |                   |                 |                 |                                  |
| *Lactobacillus panis* PM1 |                      |                            |                   |                 |                 |                                  |
| *Pantoea agglomerans* CNCM 1210 |                      |                            |                   |                 |                 |                                  |
| *Shinwella blattae* ATCC 33430 |                      |                            |                   |                 |                 |                                  |
| *Clostridium beijerincki* DSM 791 |                      |                            |                   |                 |                 |                                  |

### Development of a function-based search strategy for the reductive branch.

For NAD+ regeneration and 1,3-PD production two alternative pathways are known, a well-studied B12-dependent pathway and a lesser studied oxygen sensitive B12-independent pathway (figure 1). Pattern recognition analysis that was used to obtain the DAK configuration was used to obtain the domain persistency and copy number of key domain classes in the two alternative pathways among the genomes of the training set. The iron containing aldehyde reductase/alcohol dehydrogenase domain (PF00465), is present in both the oxidative and reductive branch but functions in the reductive branch as an aldehyde reductase. The three dehydratase Pfam domains signifying the B12-dependent pathway, PF02287 (small subunit), PF02288 (medium subunit), and PF02286 (large subunit) have a domain class persistency of 0.6 with, when present, on average per genome 1.7 (large and small subunit) or 3.7 copies (medium subunit) in the training set. The diol dehydratase reactive ATPase-like domain, PF08841 has a persistency of 0.6 and a copy number of 1. Note that only four strains in the training data set were molecularly characterized. The degree of persistency of these key domains of the B12-dependent pathway in the training set therefore suggests that multiple 1,3-PD-producing strains of the training set may actually not use the B12-dependent pathway for 1,3 PD-production. Alternatively, a B12-independent glycerol dehydratase function is reported to be involved in 1,3-PD production (13) and consists of a glycine radical (PF01228) and a pyruvate formate lyase-like domain (PF02901). Both domains are present in all 13 genomes of the training data set with a relative high copy number of 6. The persistency of the radical SAM superfamily domain, a key part of the B12-independent glycerol dehydratase reactive enzyme (19), was also 1 with an average copy number of 32 indicating that these three domains are ubiquitous and promiscuous and normally also used to support other functionalities (Table 3). As the high copy number of persistent signifying domains of both the B12-dependent and B12-independent reductive branch suggests that they play roles in multiple processes, a Boolean multi-compact proximity search was developed for the reductive branch. In this approach, the physical co-localization...
of the persistent signifying domain classes in the respective genomes is included in the search. From the list of natural producers with a molecularly characterized 1,3-PD operon, the size of 1,3-PD operon was estimated to encompass approximately 18,000 bases. Furthermore, signifying domains could be found on both strands due to the use of internal promoters. For the B12-independent pathway search criteria were based on prior knowledge obtained from a single characterized strain (13). Here four domains were specified (PF00465, PF01228, PF02901, PF04055) extending 20,000 base pairs up and downstream of the alternative dehydratase domains. Furthermore, taking internal promoters into account, domains were allowed to be present on both strands when genes are facing in the opposite direction in case of bidirectional promoters. The B12-dependent pathway should contain at least five signature domains (PF00465, PF02286, PF02287, PF02288, PF08841) and were analyzed with the same restrictions as applied to the independent pathway. Including the signifying domains used in the query, the core of the B12-dependent 1,3-PD operon may consist of eight domains. Other highly persistent domains (Support level > 0.5) enriched within the syntenic region are ‘Cobalamin adenosyltransferase’, (PF01923), ‘Haem-degrading’ (PF03928), ‘Alcohol dehydrogenase, iron-type’ (PF00465), and the ‘Micro-compartment protein’ (PF00936), (see Table 4).

Data preparation. To be able to mine a large set of publicly available genomes for the presence of the vitamin B12 variant 1,3-PD operon, 84,308 bacterial genomes containing 51 phyla, 64 classes, 145 orders, 335 families, 1,126 genera and 2,661 species were downloaded from the EBI-ENA data warehouse. The SAPP annotation platform was subsequently used for de novo structural and functional annotation of these genomes resulting in a semantic database of 300,196,266 predicted protein encoding genes linked to the corresponding protein sequences, predicted protein domain architectures and provenance.

Data Mining for candidate 1,3 PD producers. Two parallel pathways must be present to convert glycerol to 1,3-PD. Data mining for candidate 1,3-PD producers was therefore split in two steps. First, the search space was reduced by searching for the presence of a possible oxidative branch. In the second step, the reduced search space was searched for the presence of a possible reductive branch. As the molecular knowledge obtained for B12-independent pathway is based on the characterization of a single Clostridium butyricum strain, data mining for the reductive branch focused on finding candidate B12-dependent 1,3 PD producers. Oxidative branch: Following the function-based search strategy outlined above and summarised in figure 1, a proximity search with the two DAK configurations in parallel resulted in a 41% reduction of the search space. The thus reduced search space contained 49,110 genomes assigned to 3,526 species from 976 genera. The most abundant species were Listeria monocytogenes (5521), Streptococcus pneumoniae (6305), Escherichia coli (7577) and Staphylococcus aureus (10232). The reduced search space was used as an input for a proximity search for the presence of the B12-dependent reductive pathway.

Reducive branch Identification of B12-dependent strains according to the criteria obtained from the training set and based upon the results obtained from the oxidative branch, resulted in a list of 247 species (5,538 strains). Strains of Klebsiella pneumoniae (1,637) and Listeria monocytogenes (1,788), both pathogenic species were over-represented as well as Escherichia coli (1,067). An overview of identified species can be found in Supplementary table S1. The most frequently observed additional syntenic domains showing a persistency of 0.75 and above are shown in Table 6. For further analysis, strains were collapsed in species groups and each species group was treated as one. At species level, domain connectivity (figure 2) and persistency of the B12-dependent 1,3-PD trait showed a high degree of similarity with the training set (Table 6).

Additional constraints. As the candidate species are depending on vitamin B12 for the production of 1,3-PD it could be beneficial to additionally select for the ability to synthesize vitamin B12. Secondly, the best known natural B12-dependent 1,3-PD producers are considered to be pathogenic or opportunistic pathogenic microorganisms such as Klebsiella pneumonia (20). Following the strategy outlined above a function-based screening was developed for the distribution of the trait for vitamin B12 synthesis among the candidate species. Using a multi-domain approach containing all domain signatures required for vitamin B12 synthesis a vitamin B12 (21) biosynthesis pathway was detected in 86 of the 187 candidate species using a multi-domain approach containing all domain signatures required for vitamin B12 synthesis. In order to obtain the bio-safety level of each candidate species, the BacDive database was queried (22). Forty-seven candidate species were classified at bio-safety level 2 and higher.

| Domain | PatternFrequency | support | JaccardIndex | Domain description |
|--------|------------------|---------|--------------|-------------------|
| PF02734 32 | 1 | 1 | DhaL domain |
| PF02733 19 | 0.59 | 0.93 | DhaK domain |
| PF13684 8 | 0.25 | 0.75 | Uncharacterised protein, DhaK domain |
| PF03780 8 | 0.25 | 0.41 | Alkaline shock protein Asp23 |
| PF03610 16 | 0.5 | 0.11 | Phosphotransferase system, mannose-type IIA component |
| PF00465 8 | 0.25 | 0.05 | Alcohol dehydrogenase, iron-type |
Table 3. B12-independent

| Domain      | PatternFrequency | Support | JaccardIndex | Domain description                                   |
|-------------|------------------|---------|--------------|-----------------------------------------------------|
| PF00465     | 4                | 1       | 0.03         | Alcohol dehydrogenase, iron-type                    |
| PF01288     | 4                | 1       | 0.07         | Glycine radical domain                             |
| PF02901     | 4                | 1       | 0.11         | Pyruvate formate lyase domain                       |
| PF04055     | 4                | 1       | 0.01         | Radical SAM                                         |
| PF00465     | 4                | 1       | 0.03         | Alcohol dehydrogenase, iron-type                    |
| PF13353     | 3                | 0.75    | 0.03         | 4Fe-4S single cluster domain                        |
| PF03319     | 1                | 0.25    | 0.08         | Ethanolamine utilization protein EutN/carboxysome structural protein Ccml |
| PF06130     | 1                | 0.25    | 0.12         | Phosphate propanoyltransferase                      |

Table 4. Frequent pattern analysis on reference genomes for B12-dependent 1,3 PD and a minimum support value of 0.5

| Domain      | PatternFrequency | Support | JaccardIndex | Domain description                                           |
|-------------|------------------|---------|--------------|-------------------------------------------------------------|
| PF02288     | 12               | 1       | 0.86         | Diol/glycerol dehydratase/dehydratase reactivating factor   |
| PF08841     | 12               | 1       | 1            | Diol dehydratase reactivase ATPase-like domain              |
| PF02286     | 12               | 1       | 1            | Diol/glycerol dehydratase, large subunit                    |
| PF02287     | 12               | 1       | 1            | Propanediol/glycerol dehydratase, small subunit             |
| PF01923     | 8                | 0.67    | 0.4          | Adenosylcobalamin biosynthesis, ATP: cob(I)alamin adenosyltransferase-like |
| PF03928     | 8                | 0.67    | 0.59         | Haem-degrading                                              |
| PF00465     | 8                | 0.67    | 0.08         | Alcohol dehydrogenase, iron-type                           |
| PF00936     | 6                | 0.5     | 0.35         | Microcompartment protein, bacteria                          |

Table 5. Oxidative on genome resource

| Domain      | PatternFrequency | Support | JaccardIndex | Domain description |
|-------------|------------------|---------|--------------|-------------------|
| PF02734     | 74929            | 1       | 1            | DhaL domain       |
| PF02733     | 38482            | 0.51    | 0.91         | DhaK domain       |
| PF13684     | 32142            | 0.43    | 0.97         | Uncharacterised protein, DhaK domain                      |
| PF03610     | 27381            | 0.37    | 0.15         | Phosphotransferase system, mannose-type IIA component     |
| PF03780     | 23828            | 0.32    | 0.34         | Alkaline shock protein Asp23                               |

Table 6. Pattern frequency analysis for the reductive branch on genomes obtained from the oxidative analysis. Protein domains in bold are reoccurring from the training set.

| Domain      | PatternFrequency | Support | JaccardIndex | Domain description |
|-------------|------------------|---------|--------------|-------------------|
| PF08841     | 8193             | 1       | 0.48         | Diol dehydratase reactivase ATPase-like domain            |
| PF03928     | 8188             | 1       | 0.16         | Haem-degrading                                             |
| PF02288     | 8193             | 1       | 0.62         | Diol/glycerol dehydratase/dehydratase reactivating factor |
| PF02287     | 8193             | 1       | 0.61         | Propanediol/glycerol dehydratase, small subunit           |
| PF02286     | 8193             | 1       | 0.62         | Diol/glycerol dehydratase, large subunit                  |
| PF01923     | 8178             | 1       | 0.13         | Adenosylcobalamin biosynthesis, ATP: cob(I)alamin adenosyltransferase-like |
| PF00465     | 8193             | 1       | 0.03         | Alcohol dehydrogenase, iron-type                          |
| PF15953     | 8098             | 0.99    | 0.76         | Microcompartment protein PduM                             |
| PF06130     | 8127             | 0.99    | 0.44         | Phosphate propanoyltransferase                            |
| PF03319     | 8126             | 0.99    | 0.26         | Ethanolamine utilization protein EutN/carboxysome structural protein Ccml |
| PF00936     | 8135             | 0.99    | 0.32         | Microcompartment protein, bacteria                         |
| PF00171     | 8125             | 0.99    | 0.01         | Aldehyde dehydrogenase domain                             |
| PF00871     | 6278             | 0.77    | 0.06         | Aliphatic acid kinase, short-chain                        |
| PF13375     | 6152             | 0.75    | 0.16         | RnfC Barrel sandwich hybrid domain                         |
| PF10662     | 6119             | 0.75    | 0.22         | Ethanolamine/propanediol utilisation protein, EutP/PduV   |
| PF10531     | 6158             | 0.75    | 0.03         | Soluble ligand binding domain                             |
| PF01512     | 6179             | 0.75    | 0.07         | NADH-ubiquinone oxidoreductase 51kDa subunit, FMN-binding domain |
| PF00288     | 6118             | 0.75    | 0.03         | GHMP kinase N-terminal domain                             |

and 48 species were considered to be non-pathogenic (biosafety level 1) while 92 species remained unclassified. The bio-safety level 1 label was combined with the ability for vitamin B12 synthesis with resulted in the following candidate species: Anaeromusa acidaminophila, Anaerosalibacter massiliensis, Blautia obeum, Clostridium drakei, Clostridium magnum, Geobacillus thermoglucosidasius, Intestimonas butyriciproducens and Thermincola ferriacetica, (see Table suppl S1).

Discussion

Bioprospecting entails the systematic search for economically valuable genetic and biochemical resources from nature (23). For genome prospecting, the in silico mining of sequenced genomes and metagenomes for new biotechnologically relevant proteins and enzymes is a continuously evolving field. Two approaches can be applied, a “top-down” approach that begin by searching for a function and is followed by identification of the corresponding gene(s) and a

Koehorst et al. | genome prospecting for industrial traits | bioRxiv | 5
“bottom-up” approach that start with a gene of interest in an effort to find a corresponding function in genome(s) of interest. For bottom-up approaches many sequence similarity tools such as Blast (24) exist. Bottom-up studies usually start with a selection of (meta)genome sequences followed by sequence similarity-based clustering and selection of candidate sequences and re-annotation of interesting candidates thereby avoiding ambiguity related problems in current functional annotations. For bacterial species however, a priori, gene fusion-fission events can be expected (25) hampering sequence similarity-based detection and clustering of multi-domain proteins encoding genes, while the same domains may also be present in multiple proteins. When searching for polygenic traits these problems are aggravated and thus a function-based approach searching for key functions may be more effective especially when there is insufficient understanding of the genetic architecture of trait in terms of the minimal number of genes required for the trait and function of the domains encoded by the different genes. In order to overcome ambiguity related problems in functional annotations, SAPP identifies and annotates protein function based on domain architecture. Protein domains have been shown to provide an accurate representation of the functional capabilities of a protein (10). To overcome ambiguity related problems in functional annotations SAPP identifies and annotates protein function based on domain architecture. As profile hidden Markov Models (HMM), which favour in their scoring functionally important sites, are used for the identification of protein domains, statistically robust annotation profiles can be obtained over large phylogenetic distances. KDD is a multi-step process. Crucia steps are data standardisation and feature extraction for pattern searching. By using protein domain architectures as proxy for protein functions, a high degree of standardisation is obtained. As protein domain architectures can be directly transformed in highly interoperable strings of Pfam domain identifiers, a “top-down” functional screening can be done efficiently. Applying the KDD approach on Linked Data allows for validation of initial results in multiple ways and for iteration after modification. Driven by a strong need to be able to integrate and analyse biological datasets across databases, there has been a considerable increase in the adoption of Semantic Web technologies in the life-sciences (26). However, a SPARQL endpoint for phenotype data is only available via WikiData (27). Other resources for phenotype data such as BacDive, here used to obtain bio-safety levels of candidate species, do provide API’s to mine their data but currently cannot be directly queried using SPARQL. With the growing importance of Semantic Web technologies for the life sciences interoperability levels on all aspect will increase, enhancing mining possibilities, aiding the discovery of new traits in an unprecedented pace.

By using molecular knowledge obtained from molecular characterisations of the 1,3-PD operon of four species and validating and iterating the search pattern on a training data set of genome sequences of twelve known 1,3-PD producers, a large collection of 84,308 publicly available bacterial could be efficiently mined in a top-down approach yielding 187 new candidate species some of which could be confirmed by literature (28–32).

**Conclusions**

Through transformation of genomic data into a FAIR linked-data format, iterative function-based approaches can be de-
developed to mine the large genome repositories. By presenting functional annotation as unambiguous protein domain architectures a high level of interoperability is obtained allowing for the development of efficient function-based top-down searches not limited to supervised trait identification.

Materials and methods

Data annotation and mining. Bacterial genomes were downloaded from the ENA database using the enaBrowserTools (1). All downloaded microorganisms were converted into a semantic repository using the annotation platform based on functional analysis (SAPP) (7). All genomes were de novo structurally re-annotated using Prodigal (16) and functionally annotated using Pfam from InterProScan (17). Each genome was stored as an individual graph database using the HDT library and were processed using a domain pattern recognition software that was developed to aid in the data processing (33, 34). This application constructs operon-like structures around one or more protein domains from genomic data and discovers patterns in the present protein domains using a Frequent Pattern Mining and a Latent Dirichlet Allocation approach.

Pattern recognition. The application that was used for pattern recognition has been developed in Java 8 and uses Scala version 2.11. The annotated genomes are provided as RDF data in HDT format allowing to easily scale beyond the size limitations of current graph databases. Spark Core has been implemented to enable parallelization for the processing of the genomes. The Frequent Pattern Mining and Latent Dirichlet allocation algorithms from the Apache SPARK MLlib Machine Learning Library, are applied to analyse patterns in the domain architecture of operons Meng2016. The support value of a pattern with one domain is a measure for the local persistency of that specific domain. The support value is defined as the number of regions where the domain occurs divided by the total number of regions.

Oxidative branch identification. To identify strains containing the oxidative branch as shown in figure 1, DAK domains were processed using the domain pattern recognition software. Genes containing DAK1 (PF02733) or DAK1_2 (PF13684) in combination with DAK2, (PF02734), 95% gene confidence according to prodigal and neighbour linking was applied to identify strains capable for growth on glycerol.

Reductive branch identification. Strains that contained the oxidative branch were further investigated for the reductive branch using domains according to the B12 dependent or independent pathway. Region identification was performed using PF02287 and 20,000 base pairs up and downstream. These regions were modified to only retain genes that form an operon-like structure. The biggest intergenic region allowed between two genes on the same strand was 50 nucleotides. To also include the possibility of a bidirectional promoter, a gap of 500 (or less) nucleotides is allowed between to stretches of genes on the reverse and forward strand. To prevent identical regions created around a gene which contains multiple copies of the PF02287 or different genes containing PF02287, regions that (partly) overlap are merged into one. The operon-like structure was preserved.

B12 synthesis identification. Phylogenetic information was complemented with phenotype information through BacDive (22). The BacDive resource was parsed and each entry record was transformed into RDF allowing integration of genotypic and phenotype information. SPARQL queries were used to retrieve information with regards to pathogenicity and temperature.

ACKNOWLEDGEMENTS

This research was supported by the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement (323009) and by the Gravitation grant (024.002.002) of the Netherlands Ministry of Education, Culture and Science. The work conducted by the U.S. Department of Energy Joint Genome Institute (DOE-JGI), a DOE Office of Science User Facility, was supported by the Office of Science of the DOE under Contract No. DE-AC02-05CH11231. IBISBA 1.0 - Project ID: 730976 - Funded under: H2020-EU.1.4.1. - Integrating and opening existing national and regional research infrastructures of European interest. This work was carried out on the Dutch national e-infrastructure with the support of the SURF foundation.

Bibliography

1. Nicole Silvester, Blaise Alako, Clara Amid, Ana Cerdeño-Tarragona, Laura Clarke, Ian Clareland, Peter W. Harrison, Suran Jayathilaka, Simon Ray, Thomas Keane, Rasko Leinonen, Xin Liu, Josué Martínez-Villacorta, Manuela Menchi, Keith Reddy, Nima Pakerseresh, Jeena Rajan, Marc Rossollo, Dmitrij Smirnov, Ana L. Turibo, Daniel Vaughan, Vadim Zalunin, and Guy Cochrane. The European Nucleotide Archive in 2017. Nucleic Acids Research, 46 (D1):D26–D40, Jan 2018. ISSN: 13662492. doi: 10.1093/nar/gkx1125.

2. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Research, 46(5):2699–2709, mar 2018. ISSN: 0305-1048. doi: 10.1093/nar/gky992.

3. Alexandra M. Schnoes, Shoshana B. Brown, Igor Dodevski, and Patricia C. Babbitt. Annotation Error in Public Databases: Missattribution of Molecular Function in Enzyme Superfamilies. PLoS Computational Biology, 5(4):e1000605, may 2009. ISSN: 1553-7368. doi: 10.1371/journal.pcbi.1000605.

4. Jasper J. Koehorst, Jesse C.J. J Van Dam, Ruben G.A. A Van Heck, Edoardo Saccenti, Vitor A.P.Martins P. Martins Dos Santos, Maria Suarez-Diez, and Peter J. Schaap. Comparision of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data. Scientific Reports, 6(1):38699, 2016. ISSN: 2045-2322. doi: 10.1038/srep38699.

5. Sarah K Kummerfeld and Sarah A Teichmann. Protein domain organisation: adding order. BMC Bioinformatics, 10(1):39, jan 2009. ISSN: 1471-2105. doi: 10.1186/1471-2105-10-39.

6. Dennis P. Wall, Parul Kudtarkar, Vincent A. Fusaro, Rimma Pivovarov, Prasad Patil, and Peter J. Tonellato. Cloud computing for comparative genomics. BMC Bioinformatics, 11, 2010. ISSN: 1471-2105. doi: 10.1186/1471-2105-11-259.

7. Jasper J. Koehorst, Jesse C.J. J Van Dam, Edoardo Saccenti, Vitor A.P. Martins Dos Santos, Maria Suarez-Diez, and Peter J. Schaap. SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. Bioinformatics, 34(8):1401–1403, apr 2018. ISSN: 1367-4803. doi: 10.1093/bioinformatics/btx767.

8. J.C.J. van Dam, J.J. Koehorst, J.O. Vik, V.A.P. Martins dos Santos, P.J. Schaap, and M. Suarez-Diez. The Empusa code generator and its application to GBOL, an extendable ontology for genome annotation. Scientific Data, 6(1), 2019. ISSN: 20524463. doi: 10.1038/s41597-019-0263-7.

9. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data. In Interoperability and Web Applications, pages 205–227. IGI Global, 2011. doi: 10.4018/978-1-60960-593-3.ch008.

10. Jasper J. Koehorst, Edoardo Saccenti, Peter J. Schaap, Vitor A. P. Martins dos Santos, and Maria Suarez-Diez. Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. F1000Research, 5(1987):1987, jun 2017. ISSN: 2046-1402. doi: 10.12688/f1000research.9416.3.

11. R. K. Saxena, Priki Anand, Saurabh Saran, and Jasmine Isar. Monobial production of 1,3-propanediol: Recent developments and emerging opportunities. Biotechnology Advances, 27(6):895–913, 2009. ISSN: 07349750. doi: 10.1016/j.biotechadv.2009.07.003.
Supplementary

12. Wei Jiang, Shuichen Wang, Yuanpeng Wang, and Baishan Fang. Key enzymes catalyzing glycerol to 1,3-propanediol. Biotechnology for Biofuels, 9(1):57, dec 2016. ISSN 1754-6834. doi: 10.1186/s13092-016-0477-6.

13. C. Raynaud, P. Sarrabial, I. Meyniel-Salles, C. Croux, and P. Soucaill. Molecular characterization of the 1,3-propanediol (1,3-PD) operon of Clostridium butyricum. Proceedings of the National Academy of Sciences, 100(9):5010–5015, 2003. ISSN 0027-8424. doi: 10.1073/pnas.0734330019.

14. Petar Ristoski and Heiko Paulhein. Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics, 36(1–2):1–22, jan 2016. ISSN 15708288. doi: 10.1016/j.websem.2016.01.001.

15. R. G. Forage and E C Lin. DHA system mediating aerobic and anaerobic dissimilation of glycerol in Klebsiella pneumoniae NCIB 418. Journal of bacteriology, 151(2):591–9, aug 1982. ISSN 0022-5324.

16. Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1):119, jan 2010. ISSN 1471-2105.

17. Robert D. Finn, Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. G. Forage and E C Lin. DHA system mediating aerobic and anaerobic dissimilation of glycerol in Klebsiella pneumoniae NCIB 418. Journal of bacteriology, 151(2):591–9, aug 1982. ISSN 0022-5324.

18. Graham Klyne and Jeremy J Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 10(1):20–20, 2004.

19. Jonathan M. Demick and William N. Lanzilotta. Radical SAM activation of the B12-dependent glycerol dehydratase results in formation of 5'-deoxy-5'- (methylthio)adenosine and not 5'-deoxyadenosine. Biochemistry, 50(4):440–2, feb 2011. ISSN 1520-4995. doi: 10.1021/bi101255c.

20. Yu Wang, Fei Tao, C. Li, Lixiang Li, and Ping Xu. Genome Sequence of Klebsiella pneumoniae Strain ATCC 25955, an Oxygen-Insensitive Producer of 1,3-Propanediol. Genome Announcements, 1(4):4–5, aug 2013. ISSN 2169-8287. doi: 10.1128/gna00087-13.

21. E. Raux, H. L. Schubert, and M. J. Warren*. Biosynthesis of cobalamin (vitamin B12): a bacterial conundrum. Cellular and Molecular Life Sciences, 57(13):1880–1893, dec 2000. ISSN 1420-682X. doi: 10.1007/s000180050670.

22. Carola Söhngen, Adam Podstawka, Boyke Bunk, Dorothea Gleim, Anna Vetcininova, E. Raux, H. L. Schubert, and M. J. Warren*. Biosynthesis of cobalamin (vitamin B12): a bacterial conundrum. Cellular and Molecular Life Sciences, 57(13):1880–1893, dec 2000. ISSN 1420-682X. doi: 10.1007/s000180050670.

23. Anthony Artuso. Bioprospecting, benefit sharing, and biotechnological capacity building. World Development, 30(8):1355–1368, aug 2002. ISSN 0305750X. doi: 10.1016/S0305-750X(02)00040-2.

24. Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, sep 1997. ISSN 03051048. doi: 10.1093/nucleiconstructb41000.09608524. doi: 10.1016/j.biortech.2012.05.121.

25. Sophie Paszek, Jean-Loup Risler, and Pierre Bézéllec. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics, 22(12):1418–1423, 2006.

26. K.-H. Cheung, Eric Prud'hommeaux, Yimin Wang, and Susie Stephens. Semantic Web for Health Care and Life Sciences: a review of the state of the art. Briefings in Bioinformatics, 16(2):111–113, mar 2009. ISSN 1467-5463. doi: 10.1093/bib/bbp015.

27. Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller-Muehlbacher, Lynn M Schriml, Damiano Piovesan, Simon C. Potter, Neil D. Rawlings, Nicole Redaschi, Lorna Richardson, Natale, Marco Necci, Gift Nuka, Christine A. Orengo, Y oungmi Park, Sebastien Pesseat, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A. Natale, Marco Necchi, Wil H. Wu, Ioannis Xenarios, Lai Su Wh, Siew Yi Young, and Alex L. Mitchell. InterPro in 2017 beyond protein family and domain annotations. Nucleic Acids Research, 45(D1): D190–D199, 2017. ISSN 13624962. doi: 10.1093/nar/gkw1107.

28. Anniina Kivistö, Ville Santala, and Matti Karp. 1,3-Propanediol production and tolerance of a halophilic fermentative bacterium, Halanaerobium saccharolyticum subsp. saccharolyticum. Journal of Biotechnology, 158(4):242–247, apr 2012. ISSN 01681656. doi: 10.1016/j.jbti.2011.10.013.

29. K.-H. Cheung, Eric Prud'hommeaux, Yimin Wang, and Susie Stephens. Semantic Web for Health Care and Life Sciences: a review of the state of the art. Briefings in Bioinformatics, 16(2):111–113, mar 2009. ISSN 1467-5463. doi: 10.1093/bib/bbp015.

30. Stefan Pilatil, Hans Marx, Diethard Mattarovich, and Michael Sauer. 1,3-Propanediol production from glycerol with Lactobacillus delbruekii. Biosource Technology, 2012. ISSN 09608524. doi: 10.1016/j.biotecbio.2012.05.121.

31. Nurul H. Khan, Tae Sun Kang, Douglas A. S. Grahame, Monique C. Haakensen, Kom-sulee Ratanapariyanuch, Martin J. Heaney, Darren R. Korb, and Takuij Tanaka. Isolation and characterization of novel 1,3-propanediol-producing Lactobacillus panis PM1 from bioethanol thin stillage. Applied Microbiology and Biotechnology, 57(13):1880–1893, dec 2000. ISSN 1437-5193. doi: 10.1016/j.ejmech.2012.05.162.

32. K.-H. Cheung, Eric Prud'hommeaux, Yimin Wang, and Susie Stephens. Semantic Web for Health Care and Life Sciences: a review of the state of the art. Briefings in Bioinformatics, 16(2):111–113, mar 2009. ISSN 1467-5463. doi: 10.1093/bib/bbp015.

33. Javier D Fernández, Miguel A Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary RDF Representation for Publication and Exchange (HDT). Web Semantics: Science, Services and Agents on the World Wide Web, 19:22–41, 2013.

34. Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3C Recommendation, 2009(January):1–106, 2008. doi: urn:isbn:–article-id:2620949.
| Table S1. Overview of species identified to have the potential of producing 1,3 PD |
|---------------------------------------------------------------|
| Anaeromusa acidaminophila                                     |
| Anaerobalibacter massiliensis                                 |
| Aneurinibacillus migulanus                                    |
| Aneurinibacillus sp. UBA3580                                  |
| Bacillaceae bacterium MTCC 10057                              |
| Bacillus azotoformans                                         |
| Bacillus massiliosenegalensis                                 |
| Bacillus sp. 7504-2                                           |
| Bacteroides fragilis                                         |
| Blautia obeum                                                 |
| Blautia schinkii                                              |
| Blautia sp. UBA2945                                          |
| Carnobacteriaceae bacterium UBA7837                           |
| Carnobacterium alterfunditum                                  |
| Carnobacterium funditum                                       |
| Cetobacterium somerae                                         |
| Citrobacter amalonaticus                                     |
| Citrobacter freundii complex                                  |
| Citrobacter koseri                                            |
| Citrobacter rodentium                                         |
| Citrobacter sp. 30-2                                         |
| Citrobacter sp. A1                                           |
| Citrobacter sp. CFSAN044567                                  |
| Clostridia bacterium UC5.1-2H11                              |
| Clostridiaceae bacterium BRH_e20a                             |
| Clostridiales bacterium DRI-13                                |
| Clostridiales bacterium PH28_bin88                           |
| Clostridiales bacterium VE202-03                              |
| Clostridiales bacterium mt11                                  |
| Clostridium baratii                                           |
| Clostridium botulinum                                         |
| Clostridium drakei                                            |
| Clostridium estheteticum                                     |
| Clostridium lundense                                          |
| Clostridium magnus                                            |
| Clostridium pasteurianum                                     |
| Clostridium perfingens                                        |
| Enterobacter cloacae complex                                  |
| Enterobacter sp. 10-1                                         |
| Enterobacter sp. BispH2                                       |
| Enterobacter sp. GN02600                                      |
| Enterobacteriaceae bacterium UBA2606                          |
| Enterobacteriaceae bacterium UBA3163                          |
| Enterobacteriaceae bacterium UBA3516                          |
| Enterobacteriaceae bacterium UBA4747                          |
| Enterobacteriaceae bacterium UBA5935                          |
| Enterobacteriaceae bacterium UBA869                           |
| Enterobacteriaceae bacterium UBA898                           |
| Enterococcus avium                                            |
| Enterococcus malodoratus                                      |
| Enterococcus mundtii                                          |
| Enterococcus pseudoavium                                     |
| Enterococcus raffinosus                                       |
| Enterococcus sp. 3H8_DIV0648                                  |
| Enterococcus sp. HMSC066C04                                   |
| Enterococcus sp. HMSC072H05                                   |
| Enterococcus sp. HMSC29A04                                    |
| Escherichia coli                                              |
| Escherichia fergusonii                                        |
| Enterobacteriaceae bacterium CHKCI004                         |
| Eubacteriaceae bacterium sp. 14-2                             |
| Eubacteriaceae bacterium sp. An11                             |
| Eubacteriaceae bacterium sp. An3                              |
| Eubacteriaceae bacterium sp. UBA3279                         |
| Eubacteriaceae bacterium sp. UBA3326                         |
| Eubacteriaceae bacterium sp. UBA7134                         |
| Firmicutes bacterium UBA3567                                 |
| Firmicutes bacterium UBA3570                                 |
| Firmicutes bacterium UBA3573                                 |
| Flavonifractor plautii                                       |
| Flavonifractor sp. An306                                     |
| Flavonifractor sp. An82                                      |
| Fusobacteriae bacterium UBA2433                              |
| Fusobacterium hwasookii                                      |
| Fusobacterium nucleatum                                      |
| Fusobacterium sp. CM1                                        |
| Fusobacterium sp. CM22                                       |
| Fusobacterium sp. HMSC064B11                                 |
| Fusobacterium sp. HMSC073F01                                 |
| Fusobacterium sp. OBRC1                                      |
| Fusobacterium ulcerans                                       |
| Fusobacterium varium                                         |
| Geobacillus sp. (strain Y4.1MC1)                             |
| Geobacillus thermoglucosidans                                |
| Geosporobacter ferrireducens                                |
| Halanaerobium hydrogeniformans                              |
| Halanaerobium saccharolyticum                                |
| Ilyobacter polytropus                                        |
| Intestimononas butyriciproducens                             |
| Klebsiella aerogenes                                         |
| Klebsiella michiganensis                                    |
| Klebsiella oxytoca                                           |
| Klebsiella pneumoniae                                        |
| Klebsiella quasipneumoniae                                   |
| Klebsiella sp. 10982                                         |
| Klebsiella sp. 1_1_55                                        |
| Klebsiella sp. AA405                                         |
| Klebsiella sp. AS10                                          |
| Klebsiella sp. HMSC09D12                                      |
| Klebsiella sp. HMSC16A12                                      |
| Klebsiella sp. HMSC22F09                                     |
| Klebsiella sp. KGM-IMP216                                    |
| Klebsiella sp. KTE92                                         |
| Klebsiella sp. LTGPAF-6F                                     |
| Klebsiella sp. M5al                                          |
| Klebsiella sp. OBRC7                                         |
| Klebsiella sp. PO552                                         |
| Klebsiella sp. S1                                           |
| Klebsiella variicola                                         |
| Klebsiella variicola CAG:634                                 |
| Kluyvera intermedia                                          |
| Lachnospiraceae bacterium 3-1                                |
| Lachnospiraceae bacterium 7_1_58FAA                          |
| Lachnospiraceae bacterium A2                                 |
| Lactobacillus brevis                                         |
| Lactobacillus collinoides                                    |
| Lactobacillus coryniformis                                   |
| Lactobacillus curvatus                                       |
| Lactobacillus diolivorans                                    |
| Lactobacillus fuchuensis                                     |
| Lactobacillus ginsenosidimutans                              |
| Lactobacillus graminis                                       |
| Lactobacillus kimchicus                                      |
| Lactobacillus kisonensis                                     |
| Lactobacillus mellifer                                       |
| Lactobacillus namurensis                                     |
| Lactobacillus panis                                          |
| Lactobacillus paracalloloides                                |
| Lactobacillus pobuziihi                                      |
| Lactobacillus rapi                                           |
| Lactobacillus renmini                                       |
| Lactobacillus reuteri                                        |
| Lactobacillus rosiae                                         |
| Lactobacillus silagei                                        |
| Lactobacillus siliginis                                      |
| Lactobacillus similis                                        |
| Lactobacillus spicheri                                       |
| Lactobacillus versmoldensis                                 |
| Listeria innocua                                             |
| Listeria ivanovii                                            |
| Listeria monocytogenes                                       |
| Listeria seeliger                                            |
| Listeria welshimer                                           |
| Metakosakonia massiliensis                                   |
| Mycobacterium vaccae                                        |
| Paraclostridium benzoeltyicum                                |
| Paraclostridium bifermantans                                |
| Pediococcus acidilactici                                     |
| Pediococcus clausenii                                       |
| Pediococcus pentosaceus                                      |
| Pelosinus fermentans                                         |
| Pelosinus sp. HCF1                                          |
| Phytobacter urisingi                                         |
| Pluralibacter gergoviae                                      |
| Propionibacterium freundreichii                              |
| Propionimicrobium sp. BV2F7                                  |
| Pseudopropionibacterium propionicum                          |
| Quasibacillus thermotolerans                                |
| Salmonella choleraesuis                                     |
| Sebaldella termiditis                                       |
| Serratia marcescens                                         |
| Shigella boydii                                              |
| Shigella flexneri                                            |
| Shigella sonnei                                              |
| Sporomusa acidivorans                                        |
| Sporomusa phaeoides                                         |
| Sporomusaceae bacterium UBA5961                             |
| Sporomusaceae bacterium UBA5989                              |
| Sporomusaceae bacterium UBA7701                              |
| Sporomusaceae bacterium UBA864                               |
| Streptococcus australis                                      |
| Streptococcus sanguinis                                      |
| Streptococcus sp. AS14                                      |
| Streptococcus sp. F0442                                     |
| Streptococcus suis                                          |
| Thermocinola ferricetica                                    |
| Thermocinola potens                                         |
| Thermoanaerobacter sp. (strain X513)                         |
| Thermoanaerobacter sp. (strain X514)                         |
| Thermoanaerobacter sp. A7A                                   |
| Thermoanaerobacterium thermosaccharolyticum                  |
| Tissierella bacterium S5-A11                                 |
| Eubacterium hallii                                           |