OrtSuite – a flexible pipeline for annotation of ecosystem processes and prediction of putative microbial interactions

João Pedro Saraiva[1], Marta Gomes[2], René Kallies[1], Carsten Vogt[1], Antonis Chatzinotas[1,3,4], Peter Stadler[5,6,7,8,9], Oscar Dias[2], Ulisses Nunes da Rocha[1*]

[1]Department of Environmental Microbiology, Helmholtz Centre for Environmental Research – UFZ, Permoserstraße 15, 04318 Leipzig, Germany
[2]Centre of Biological Engineering, University of Minho, R. da Universidade, 4710-057, Braga, Portugal
[3]Institute of Biology, Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany
[4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany
[5]Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, and Competence Center for Scalable Data Services and Solutions Dresden/Leipzig, University of Leipzig, PF 100920, 04009 Leipzig, Germany
[6]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
[7]Institute for Theoretical Chemistry, University of Vienna, Währinger Str. 17, 1090 Wien, Austria
[8]Facultad de Ciencias, Universidad Nacional de Colombia, Cra 45, Bogotá, Colombia
[9]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, U.S.A.

Author email addresses: Joao Saraiva (joao.saraiva@ufz.de), Marta Gomes (martalopesgomes@hotmail.com), Rene Kallies (rene.kallies@ufz.de) , Carsten Vogt (carsten.vogt@ufz.de), Antonis Chatzinotas (antonis.chatzinotas@ufz), Peter Stadler (studla@bioinf.uni-leipzig.de), Oscar Dias (odias@ceb.uminho.pt) , Ulisses Nunes da Rocha (ulisses.rocha@ufz.de)

*Corresponding author: ulisses.rocha@ufz.de

Abstract

Background

The exponential increase in high-throughput sequencing data and the development of computational sciences and bioinformatics pipelines has advanced our understanding of microbial community composition and functional and genetic diversity in complex ecosystems. Despite these advances, the identification of microbial interactions from genomic and metagenomics data remains a major bottleneck. To address this challenge, we present OrtSuite, a flexible workflow to predict putative microbial interactions based on genomic content.

Results

OrtSuite combines ortholog clustering strategies with genome annotation based on a user-defined set of functions allowing for hypothesis-driven data analysis such as assessing microbial interactions in specific ecosystems. OrtSuite allows users to install and run all workflow components and analyze the generated outputs using a simple pipeline consisting of 23 bash commands and one R command. Annotation is based on a two-stage process. First, only a subset of sequences from each ortholog cluster are aligned to all sequences in the Ortholog-Reaction Association database (ORAdb). Next, all sequences from clusters that meet a user-defined identity threshold are aligned to all sequence sets in ORAdb to which they had a hit. This approach results in a decrease in time needed for functional annotation. Further, OrtSuite identifies putative interspecies interactions based on their individual genomic content based on constraints given by the users. Additional control is afforded to the user at several stages of the workflow: 1) The construction of ORAdb only needs to be performed once for each specific process also allowing manual curation; 2) The identity and sequence similarity thresholds used during the annotation stage can be adjusted; and 3) Constraints related to pathway reaction composition and known species contributions to ecosystem processes can be defined.

Conclusions

OrtSuite is an easy to use workflow that allows for rapid functional annotation based on a user curated database. Further, this novel workflow allows the identification of interspecies interactions through user-defined constraints. Due to its low computational demands, for small datasets (e.g. maximum 100 genomes) OrtSuite can run on a personal computer. For larger datasets (> 100 genomes), we suggest the use of computer clusters. OrtSuite is an open-source software available at https://github.com/mdsufz/OrtSuite.

Keywords: orthologs, functional annotation, microbial interactions, partial genome-scale models, microbial modelling

Background

In environments where microorganisms play a key role, ecosystem processes are dependent on the functional potential of the associated microbiome and universe of possible interspecies interactions within the microbiome members [1,2]. For example, in environments rich in methane, microbial communities are dominated by species with genes encoding proteins involved in methanogenesis [3]. Soil microbes, especially those in the rhizosphere are genetically adapted to perform functions such as pathogen resistance and tolerance to stress [4]. Natural ecosystems are populated by an enormous number of microbes [5]. For example, soil environments can contain more than 10^{10} species per gram of soil making a global search for interspecies interactions unfeasible [6]. The exponential increase in high-throughput sequencing data and the development of computational sciences and bioinformatics pipelines has advanced our understanding of microbial community composition and distribution in complex ecosystems [7]. This knowledge increased our ability to reconstruct and functionally characterize genomes in complex communities, for example by the recovery of metagenome-assembled genomes (MAGs) [8–10]. While tools have been developed to improve the reconstruction of MAGs, the same cannot be said for predicting interspecies interactions [11]. The challenge of the latter increases due to the multitude of potential interactions

not only between species in microbial communities but also between microbes and their hosts (e.g., plants, animals and microeukaryotes) [12]. In order to decrease complexity and facilitate analysis, the search of interactions can be limited to groups of organisms (e.g. microbe-microbe or host-microbe).

To improve our understanding on how microorganisms work within their communities and contribute to different ecosystem processes, it is essential to assess their functional capacity. Accurate annotation of gene function is essential to predict, from sequencing data, ecosystem processes potentially performed by microbial communities, particularly in cases where an ecosystem process is performed by the synergy of two or more species.

Simple methods for the annotation of genomes rely, for instance, on the search for homologous sequences. Computational tools such as BLAST [13] and DIAMOND [14] allow the comparison of nucleotide or protein sequences to those present in databases. These approaches allow inferring the function of uncharacterized sequences from their homologous pairs whose function is already known. The degree of confidence in the assignment of biological function is increased if this has been validated by, for example, experimental data. Orthologs are homologous sequences that descend from the same ancestor separated after a speciation event retaining the same function [15]. Approaches based on orthology are increasingly used for genome-wide functional annotation [16]. For example, OrthoMCL [17], CD-HIT [18] and OrthoFinder [19,20] are just a few tools that identify homologous relationships between sequences using orthology.

OrthoFinder has been shown to be more accurate than several other orthogroup inference methods since it takes into account gene length in the detection of ortholog groups by introducing a score transformation step [19]. However, OrthoFinder, due to its all-versus-all sequence alignment approach, requires intensive computational resources resulting in long running times when using large data sets for clustering. Nevertheless, integrating ortholog clustering tools with functional annotation strategies can improve our ability to functionally characterize microbial communities and reduce the time needed to obtain results.

Genome-based modelling approaches have routinely been used to study single organisms as well as microbial communities [21]. For example, constraint-based models are highly employed in the study and prediction of metabolic networks [22]. These models are generated upon the premise that any given function is feasible as long as the protein-encoding gene is present. Although species may lack the genetic potential to perform all functions necessary to survive in a given ecosystem, in nature microbes do not exist in isolation and usually benefit from their interaction with other species. By assessing the genomic content of individual species we are able to identify groups of microbes whose combined content may account for complete ecosystem functioning.

In this study, we developed OrtSuite; a pipeline to perform rapid and accurate functional annotation of various microbial species based on the detection of orthologs. This pipeline integrates the use of a user-defined database – Ortholog-Reaction Association database (ORAdb) – with up-to-date ortholog clustering tools. Additionally, OrtSuite allows the search for putative microbial interactions by calculating the combined genomic potential of individual species in specific ecosystem processes. We evaluate this pipeline using a clearly defined set of reactions involved in the well-described benzoate-to-Acetyl-CoA (BTA) conversion. Further, we used this pipeline to functionally characterize a set of known benzoate degraders. OrtSuite's ability to identify putative interspecies interactions was evaluated on species whose interactions have been previously predicted under controlled conditions [23].

Implementation

The OrtSuite workflow, excluding installation, consists of four main tasks performed by the use of 10 bash commands and one R command (Figure 1): (1) generation of a user defined ortholog-reaction associated database (ORAdb) consisting of protein sequences associated with a set of reactions/pathway of interest; (2) generation of ortholog clusters; (3) functional annotation of sequences in the generated ortholog clusters using ORAdb, consisting of a two-stage process termed relaxed and restrictive search; and, (4) identification of putative microbial interactions based on the

genomic content. To assess the performance of OrtSuite, we selected the transformation of benzoate to acetyl-CoA as a model pathway and a set of previously characterized species known to be involved in this pathway. This set of species was divided into two groups. The first group contained species known to possess all or a subset of the necessary protein encoding genes involved in the model pathway and were selected as positive controls. The second group consisted of species known to lack all required protein encoding genes and were selected as negative controls. Using their genomes, we evaluated the performance of OrtSuite based on precision and recall rates for both, clustering and functional annotation of sequences. Also, a set of genomes from the species containing the genetic potential to degrade benzoate were artificially mutated at the nucleotide level at different rates in order to determine how different open reading frames (ORFs) need to be in order to be clustered in a different ortholog group. Putative microbial interactions were assessed based on known species interactions and their potential to degrade benzoate individually or as a result of their interaction [22].

Conversion of benzoate to acetyl-CoA as a model pathway

We selected three alternative pathways involved in the conversion of benzoate to acetyl-CoA (BTA) (Figure 2). Two pathways consisted in the anaerobic degradation of benzoate to acetyl-CoA via benzoyl-CoA differing only in the reactions required for transformation of glutaryl-CoA to crotonyl-CoA (hereafter, respectively, BTA_P1 and BTA_P2). BTA_P1 first converts glutaryl-CoA to glutaconyl-CoA and then to crotonoyl-CoA while BTA_P2 directly converts glutaryl-CoA to crotonoyl-CoA. One pathway consisted in the aerobic degradation of benzoate via catechol (hereafter BTA_P3). The complete number of reactions, enzymes, KO identifiers and KO-associated sequences in each alternative pathway is shown in the supplementary data (Additional file 1: Table S1).

Selection of species with known genetic potential to convert benzoate to acetyl-CoA

We designed OrtSuite to identify putative microbial interactions based on the combined genomic potential of species in a given set of enzymatic reactions. To test this feature, we used previously sequenced genomes of species whose ability to convert benzoate to acetyl-CoA has been demonstrated (Additional file 2: Table S2). These species were classified according to their genomic potential: *complete*, if all protein encoding genes required for a BTA pathway were present in their genome; *partial*, if not all protein encoding genes were present; and *absent*, if none of the required protein encoding genes were present in their genomes. In total, we selected 18 species. Seven of them have the genetic potential to perform the alternative BTA_P2 pathway; eight have the genetic potential to perform alternative path 3 (positive controls); and, none able to completely perform the alternative path 1 (for BTA pathways, see Figure 2). As negative controls, four of the selected organisms were unable to perform completely any of the BTA pathways. Note that species *Thauera sp.* MZ1T has the genetic potential to perform BTA_P2 and BTA_P3 pathways.

ORAdb – User defined ortholog-reaction association database (OrtSuite step 1)

The ORAdb used for functional annotation consists of sets of protein sequences involved in the enzymatic reactions that compose a pathway/function of interest defined by the user. This database is generated by running the command *download_kos* in OrtSuite. The user provides a text file with a list of identifiers (e.g. reaction (Additional file 3: Example_reaction_list), enzyme commission (EC) number (Additional file 4: Example_ec_list) or KEGG orthologs (Additional file 5: Example_ko_list). This file is used by OrtSuite to automatically retrieve the KEGG Ortholog identifiers (KO) and to download all their associated protein sequences [24]. OrtSuite makes use of the python library *grequests* which allows multiple queries in KEGG subsequently decreasing the time required for retrieving the ortholog associated sequence. The user-defined ORAdb will be composed of KO-specific sequence files in FASTA format associated with all reactions of interest. Users also have the opportunity to manually add or edit the sets of reactions and the associated protein sequences in the final ORAdb. This feature is of particular importance since many reactions

associated with ecosystem processes are constantly being discovered and updated and might not be included in the latest version of KEGG. In addition, the OrtSuite pipeline also allows the automated download of the gene-protein-reaction (GPR) rules from KEGG. Here, the user only needs to provide the location of the folder where the sequences of orthologs for each KO is stored. This feature is of particular importance since many reactions can be catalyzed by enzymes with a single (i.e., one protein) or multiple subunits (i.e., protein complexes).

Generation of protein ortholog clusters (OrtSuite step 2)

The second stage of OrtSuite, takes a set of protein sequences and generates clusters of orthologs. This set of protein sequences can originate from single isolates or from the complete set of protein sequences recovered from metagenomes or metagenome-assembled genomes. Indeed, the use of protein sequences from isolates, metagenome-assembled genomes and co-culture experiments will benefit greatly from OrtSuite's reduction of the universe of potential microbial interactions based on a user defined ORAdb. Orthology considers that phylogenetically distinct species can share functional similarities based on a common ancestor [25]. Potentially, genes with equal function will be grouped together. To perform this task the OrtSuite pipeline uses OrthoFinder [19]. Two sequence aligners are available in OrthoFinder – DIAMOND [14] and BLAST [13]. DIAMOND is used by default due to its improved trade-off between execution time and sensitivity [20].

Functional annotation of ortholog clusters (OrtSuite step 3)

The third stage of OrtSuite consists in the assignment of functions to protein sequences contained in the ortholog clusters. Functional annotation of these clusters consists of a two-step process termed relaxed and restrictive search, respectively. The goal of the relaxed search is to decrease the number of alignments required to assign functions to ortholog clusters. Here, 10% of the total number of sequences from each cluster are randomly selected and aligned to all sequences associated to each reaction present in the ORAdb. During this stage only sequence identity is evaluated. Ortholog

clusters where protein sequence identity meets a user-defined threshold to sequences in the ORAdb proceed to the restrictive search. The default value was set to 50% identity, as the main objective of the relaxed search is to capture as many sequences for annotation as possible whilst maintaining a reasonable number of sequence alignments. In the restrictive search, all sequences in the transitioned ortholog clusters are aligned to all the sequences in the reaction set(s) present in the ORAdb to which they had a hit during the relaxed search. The query sequence is only assigned to the function of a reference sequence if the combined score of sequence identity, positive matches, query sequence coverage and reference sequence coverage meets a user-defined threshold (Equation 1). The default value was set to 90% as an excessive restrictive search score can lead to a high number of sequences without annotation. Although we established default values for the relaxed and restrictive search, the user has the option to define the thresholds for all individual parameters. An explanation of how the different parameters might affect the results can be found in https://github.com/mdsufz/OrtSuite/README.txt.

Equation 1

$$score = \frac{\%\, identity + \%positive\ matches + \%query\ seq.coverage + \%reference\ seq.coverage}{4}$$

where, percentage of identity (*%identity*) measures the exact alignment of amino acids between two sequences in a range of alignment. Here, sequences of interest with unknown function are provided by the user and aligned to the sequences in the user defined ORA database. The default identity value is set to 95% to capture as many perfect amino acid alignments as possible without being too stringent. Many mutations consistently occur in related protein groups that do not necessarily alter the function of a gene. These mutations alter the sequence composition and can result in low identity scores. The use of scoring matrices [26] is an example of measuring similarity between two

or more sequences when sequences are not identical. The percentage of positive matches (*%positive matches*) using scoring matrices reflects the number of sequences that, although not identical, are similar to each other in a range of alignment. The default value is set to 99% only returning sequences where a single amino acid is different in the whole alignment. The percentage of the query sequence that is covered (*%query seq.coverage*) refers to the percentage of the query sequence that overlaps with the reference sequence in the alignment. The default value is set 90% so that the entire query sequence is not required during the alignment. The percentage of reference sequence that is covered (*%reference seq.coverage*) refers to the percentage of the sequence in the ORA database that overlaps with the query sequence in the alignment. The default value is set 90% so that the entire reference sequence is not required during the alignment.

Defining putative microbial interactions (OrtSuite step 4)

The identification of putative interactions between species is based on all combinations of bacterial isolates with the genomic content to perform the user-defined pathway defined in the ORAdb. The input for this task consists of: (1) a binary table generated at the end of the restrictive search for functional annotation which indicates the presence or absence of sequences annotated to each reaction in the ORA database in each species (Additional file 6: Table S3); (2) a set of Gene-Protein-Reaction (GPR) rules for all reactions considered (Additional file 7: Table S4); and (3) a user-defined tab-delimited file where the sets of reactions for complete pathways, subsets of reactions required to be performed by single species and transporter-associated genes (Additional file 8: Table S5) are described. The script *combinations.sh* included in OrtSuite (https://github.com/mdsufz/OrtSuite/blob/master/combinations.sh) is used to calculate all combinations of species capable of performing a set of reactions of interest to the user and results are stored in a text file.

To further reduce the vast amount of putative microbial interactions and to increase confidence in the results manual filtering can be performed to reflect available knowledge (e.g.

known cross-feeding relationship between species) or the likelihood of biologically feasible species interactions. The user also may have interest in assessing subsets of microbial interactions using specific criteria. Therefore, additional constraints can be applied to the list of putative microbial interactions to further reduce the search space. These include the degree of completeness of a pathway, the number of reactions expected to be performed by a single species, the presence or absence of transporter genes and the different number of species required to perform a complete pathway (e.g. only combinations of two or three species). The script *user_constraints.sh* included in OrtSuite (https://github.com/mdsufz/OrtSuite/blob/master/user_constraints.sh) allows the inclusion of these constraints.

Evaluation of ortholog clustering

The clustering of orthologs was evaluated by measuring the pairwise precision and recall. Clustering precision measures how many pairs of sequences associated with the same molecular function are grouped together and is calculated by dividing the number of correctly clustered sequences by the total number of clustered sequences (Equation 2).

Equation 2

$$clustering\ precision = \frac{correctly\ clustered\ sequences}{total\ number\ of\ clustered\ sequences}$$

where, *correctly clustered sequences* refers to the pairs of sequences that share the same function and are clustered together and *total number of clustered sequences* refers to all pairs of sequences that are clustered together irrespective of sharing the same function.

Clustering recall measures how many pairs of sequences with the same molecular function are not clustered together. Recall is calculated by dividing the number of correctly clustered sequences by the total true sequence clusters (Equation 3).

Equation 3

$$clustering\ recall = \frac{correctly\ clustered\ sequences}{total\ true\ sequence\ clusters}$$

where, *correctly clustered sequences* refers to the pairs of sequences that share the same function and are clustered together and *total true sequence clusters* refers to all pairs of sequences that have the same function.

Evaluation of sequence aligner used for clustering of orthologs

Changes of a single DNA base can result in the production of a different amino acid which might result in a different protein. To determine the impact of mutations on the clustering of orthologs a single gene from three species was artificially mutated at different rates. These mutations were introduced in the nucleotide sequences of each gene. Only substitutions were considered since these are the most commonly studied [27] and none of the mutations were allowed to occur on the first and last codon. When, during the mutation, new stop or/and start codons were introduced, the translation was made for all the possible proteins and the largest was selected.

Burkholderia vietnamiensis G4 was mutated on the gene K05783, *Azoarcus sp.* CIB on the gene K07537 and *Aromatoleum aromaticum* EbN1 on the gene K07538. Each gene was mutated at rates of 0.01, 0.03, 0.05, 0.1, 0.15 and 0.25. Each mutation rate resulted in an *in silico* strain of the original genome (e.g., *Burkholderia vietnamiensis* G4 strain K05783_25, where “K05783” is the KEGG ortholog identifier and “25” is the rate of mutation). A total of 18 strains were generated (six *in silico* mutated strains per genome). The complete set of original and artificially mutated genomes

is available in a compressed file (Additional file 9: Set_A_genomes.zip) and at https://github.com/mdsufz/OrtSuite/upload/master/Set_A_genomes.zip.

Evaluation of functional annotation

Similar to clustering of orthologs, performance of functional annotation was determined based on the correct assignment of function to sequences in the ortholog clusters. *Annotation precision* is calculated by dividing the number of correctly annotated sequences by the total number of annotated sequences (Equation 4).

Equation 4

$$Annotation\ precision = \frac{correctly\ annotated\ sequences}{total\ number\ of\ annotated\ sequences}$$

where, *correctly annotated sequences* refers to the number of sequences annotated with the correct function and *total number of annotated sequences* refers to the total number of sequences with annotation.

Annotation recall is calculated by dividing the number of correctly annotated sequences by total true annotated sequences (Equation 5).

Equation 5

$$Annotation\ recall = \frac{correctly\ annotated\ sequences}{total\ true\ sequence\ annotations}$$

where, *correctly annotated sequences* refers to the number of sequences annotated with the correct function and *total true sequence annotations* refers to the total number of sequences annotated with the correct function and sequences which were wrongfully not assigned a function.

Evaluation of microbial interaction predictions

We evaluated the prediction of microbial interactions by selecting species with the potential to interact in the transformation of benzoate to acetyl-CoA identified by Fetzer and collaborators [23] (hereafter defined as Test_genome_set). These isolates did not have their genome sequenced prior this study. Therefore, bacterial cryo-cultures of the different isolates were revived on LB agar plates. Single colonies were picked and grown overnight in 2 ml LB medium at 37°C. The cells were pelleted by centrifugation. Cells were lysed and genomic DNA was extracted using a Nucleospin Tissue Kit (Machery and Nagel). Approximately 150 to 1000 ng of DNA were used for fragmentation (insert size: 300 – 700 bp) and sequencing libraries were prepared following the NEB Ultra II FS Kit protocol (New Enland Biolabs). Libraries were quantified using a JetSeq Library Quantification Lo-ROX Kit (Bioline) and quality-checked by Bioanalyzer (Agilent). These libraries were sequenced on an Illumina MiSeq Instrument with a final concentration of 8 pM using the v3 600 cycles chemistry and 5% PhiX. The complete set of species and their genome sequences are available as compressed file (Additional file 10: Test_genome_set). According to Fetzer and collaborators [23], five of the twelve species (*C. testosteroni* ATCC 11996, *C. necator* JMP 134, *P. putida* ATCC 17514, *P. fluorescens* DSM 6290 and *R. ruber* BU3) were shown to grow as monocultures using benzoate as sole carbon and energy source (Additional file 11: Table S6). After their genomes were sequenced and assembled, we evaluated whether these species possessed the complete genomic content to encode all proteins required for each benzoate to acetyl-CoA conversion pathway. The remaining seven species were not able to grow as monocultures in media with benzoate as sole carbon source. Here, we evaluated whether the lack of growth was confirmed by lack of essential protein-encoding genes involved in conversion of benzoate to acetyl-CoA.

Additionally, Fetzer and collaborators [23] also observed that some of the species not able to grow as monocultures with benzoate as sole carbon and energy source could grow in medium with benzoate when being in a community with other species. Finally, their study also showed that, under nutrient specific conditions, total growth yields improved when communities contained species able to metabolize benzoate and several non-degrading species. Therefore, we evaluated whether putative species interactions identified by OrtSuite fit the results obtained by Fetzer et al. For the complete set of species combinations and benzoate degradation capabilities and effects identified by Fetzer and collaborators, see [23] (Additional file 12: Table S7).

Results and discussion

Ortsuite is a flexible and user friendly pipeline

The OrtSuite installation produces an environment on the user's computer with recently developed tools used for ortholog clustering [20] and sequence alignment [14]. Furthermore, OrtSuite contains several scripts that were developed to download sequence information, in bulk, and generate Gene-Protein-Reaction (GPR) rules from KEGG repositories in an automated manner by only requiring users to provide lists with identifiers (see https://github.com/mdsufz/OrtSuite/). Additional control is given to the user such as establishing thresholds in the minimum sequence identity (during sequence alignment of sequences in ortholog clusters to ORAdb) to restrict the number of putative microbial interactions based on constraints such as the presence of transporters and subsets of reactions to be performed by individual species (Additional file 8: Table S5). Since data in public repositories are frequently being added or updated and to include personal knowledge the user can manually curate the files in the ORAdb and GPR rules.

The entire workflow (excluding installation) consists of a total of 10 bash and 1 R easy to follow commands from the construction of the Ortholog-Reaction Association database (ORAdb), which will be used for the functional annotation of sequences and identification of potential microbial interactions.

Computing time of OrtSuite stages

The runtime of each OrtSuite step was evaluated on a set of genomes whose genomic content was known (Additional file 2: Table S2). The same set of genomes is used in the OrtSuite's GitHub tutorial page (https://github.com/mdsufz/OrtSuite/blob/master/OrtSuite_tutorial.md) with a total of 75.5 Megabytes of data. OrtSuite was used to analyze this data on a laptop with 4 cores and 16 Gigabytes of RAM. All OrtSuite steps were run on default settings, and the total runtime of each step was recorded (Additional file 13: Table S8). The total workflow was completed in 3 h 50 min and the longest single step runtime consisted of 2 h and 47 min which involved the construction of the ORAdb.

In this study, we ran OrtSuite on a dataset comprising 18 genomes and we postulate that up to 100 genomes can be analyzed in a customary laptop. To determine if this range would be within the number of genomes in regular microbiome studies we calculated the average number of metagenome-assembled genomes (MAGs) from metastudies where MAGs were recovered. A study performed by Parks and collaborators [8] analyzed sequencing data from 149 projects. Most projects (91%) consisted of less than 20 samples. On average, they recovered 5.3 metagenome-assembled genomes (MAGs) per metagenome. Work performed by Pasolli and collaborators [9] on microbial diversity in the human microbiome recovered, on average, 16 MAGs per metagenomic library. From the 46 studies used in their work, 30 consisted of less than 200 samples. Another study by Tully and collaborators focusing on marine environments [10] recovered 2631 MAGs from 234 samples (average of 11 MAGs per sample). Our analysis demonstrates that the average number of MAGs recovered from a metagenome currently range from five to 16. Therefore, performing targeted functional annotation and interspecies interactions predictions using OrtSuite in average sized metagenome samples is still feasible using a customary laptop.

Higher recall rates during clustering of orthologs with DIAMOND

OrthoFinder's [20] clustering of orthologs allows users to choose between DIAMOND [14] and BLAST [13] as sequence aligners. To test which sequence aligner yielded the best results we performed ortholog clustering of a dataset consisting of the original target genomes as well as a set of artificially mutated genomes (Additional file 9: Set_A_genomes) using both aligners. The results obtained (Table 1) do not show a striking difference between both aligners. However, an increase of eight percent in the recall rate (which measures how many pairs of sequences with the same molecular function are not clustered together) was observed when using DIAMOND.

Table 1 – Pairwise precision and recall rates of clustering of orthologs using OrthoFinder

	BLAST	DIAMOND
Precision	0.63	0.64
Recall	0.77	0.85

Pair-wise precision and recall results of OrthoFinder using BLAST and DIAMOND as an alignment search tool. Clustering was performed on the dataset Set_A_genomes plus the mutated genomes.

All artificially mutated sequences (even those with mutation rates of 25%) were clustered together with their non-mutated ortholog. In parallel, we also performed sequence alignment using NCBI's BLASTp [28] between the protein sequences of the DNA-mutated and un-mutated genes (Additional file 14: Table S9). E-values of sequence alignments in all species ranged from 0 to 5E-180 and percentage of identity from 61.32 to 98.84% which indicates that the mutated genes still shared enough sequence similarity to the original protein sequence. The results obtained during the clustering of orthologs combined with the protein sequence alignments suggest that mutation rates of up to 25% of single DNA base pairs will not have an observable effect on the clustering of orthologs. OrthoFinder's algorithm removes the gene length bias from the sequence alignment process, which may also explain why mutated genes were clustered with the original. Although it has been suggested that most genetic variations are neutral, changes in single base pairs can have a

drastic effect on protein function (e.g. depending on the location where the mutation occurs) [29]. To this purpose, experimental functional studies can be used to validate previously unannotated orthologs. Nevertheless, future studies increasing the rates of DNA base pair substitutions and other types of mutations as well as experiments targeting protein function in ortholog clusters are warranted.

For the subsequent steps of the OrtSuite workflow, clustering of protein orthologs was repeated using only the original genomes and the default aligner. A complete overview of the results generated during the clustering of orthologs (e.g. number of genes in ortholog clusters, number of unassigned genes and number of ortholog clusters) was also obtained (Additional file 15: Table S10 and at https://github.com/mdsufz/OrtSuite/blob/master/examples/ /Comparative_Genomics_Statistics/Statistics_Overall_test_set.tsv.

Precision and recall rates of annotation improve with relaxed identity thresholds

The third step of OrtSuite consists of performing cluster annotation in a two-stage process. In the first, only 10% of sequences are used in the alignment to the sequences ORAdb. Those with a minimum identity threshold proceed to the second stage where all sequences contained in the proceeded cluster are used. At the end, annotation of clusters will take into consideration additional parameters such as percentage of positive matches. To evaluate the thresholds used in each stage of the annotation of ortholog clusters we used two different relaxed (40 and 70%) and restrictive (90 and 95%) search identity cutoffs. An overview of the results (e.g. number of clusters containing orthologs from ORAdb, number of ortholog clusters with annotated sequences) is also generated (Additional file 16: Table S11). Precision rates for functional annotation were not strikingly different when switching from 40 to 70% relaxed identity cutoffs and 90 to 95% restrictive search cutoffs (Table 2). However, recall rates were, on average 15 to 25 % higher when using a 90% identity cutoff during the restrictive search compared to 95%. Furthermore, the difference in computing time between lower and higher identity thresholds was negligible (< 2 min).

Other annotation tools, such as NCBI's BLAST tool [13] and Prokka [30], can annotate full genomes, the latter at a relatively fast pace. On average, full genome annotation of our genomes in the Set_A_genomes dataset using Prokka required 12 mins on a customary laptop with 16 Gigabytes of RAM and four CPUs to complete. However, exploring the greater amount of data generated from full genome annotation of individual species from complex microbial communities is a daunting task. Studies aiming to address specific ecosystem processes or metabolic functions will require a manual inspection of each annotated genome to confirm the presence of the associated genes. The same drawbacks are observed when attempting to identify interspecies interactions based on the microbe's genomic potential. OrtSuite overcomes these challenges by limiting the genome annotation to user-defined functions of interest. The user-defined database coupled with the scripts for automated identification of interspecies interactions contained in OrtSuite decreases the time required to generate the data and facilitates its interpretation by the user.

Table 2 - Precision and recall rates of annotation for the Set_A_genomes dataset.

Pathway	BPA_P1				BTA_P2				BTA_P3			
Relaxed search identity cutoff (%)	40		70		40		70		40		70	
Restricted search identity cutoff (%)	90	95	90	95	90	95	90	95	90	95	90	95
Precision	0.89	0.92	0.89	0.90	0.90	0.93	0.89	0.92	1	1	1	1
Recall	0.66	0.42	0.57	0.33	0.67	0.44	0.59	0.37	0.89	0.75	0.91	0.59

Precision and recall rates of annotation using different combinations of identity and score cut-offs. The relaxed search identity cut-offs were set to 40 and 70%. The restrictive search score cut-offs were set to 90 and 95%. Annotation was performed on the genomes from the Set_A_genomes dataset for all three benzoate to acetyl-CoA conversion pathways (P1_BTA, P2_BTA and P3_BTA) used in this work.

Identifying genetic potential to perform a pathway

Ortsuite's ortholog cluster annotation approach (i.e. based on user-defined functions or pathways) not only decreases the time needed to obtain the functional potential of individual species but also facilitates analysis of the results for the whole community. To test Ortsuite's ability to identify

481 species with the genetic potential to perform a pathway individually we defined sets of reactions
482 that are used in three alternative pathways for the conversion of benzoate to acetyl-CoA (Additional
483 file 8: Table S5). Next we compared the results to the species' known genomic content in each
484 alternative pathway (shown in Additional file 2: Table S2). This reported species' individual
485 genomic potential to convert benzoate to acetyl-CoA in every alternative pathway was confirmed
486 by OrtSuite. Using OrtSuite we identified seven species capable of performing one of the anaerobic
487 conversion pathways (BTA_P2). A total of eight species were identified with the ability to perform
488 the aerobic conversion of benzoate to acetyl-CoA (BTA_P3). No single species was identified by
489 OrtSuite with the ability to convert benzoate to acetyl-CoA via the alternative anaerobic process
490 (BTA_P1). For these 12 species, cluster annotation using ORAdb required only 6 min using a
491 laptop with 16 Gigabytes of RAM and 4 CPUs (Additional file 13: Table S8). In this study, we only
492 defined three pathways for analysis but the user can increase this number without a substantial
493 increase in running times. Additionally, the results are tailored to the user-defined constraints (see
494 example, Additional file 8: Table S5) facilitating analysis.
495
496 *Using OrtSuite to predict interspecies interactions*
497 We designed OrtSuite to allow hypothesis-driven exploration of microbial interactions by limiting
498 the search by user-defined ecosystem processes or metabolic functions. In this study, we tested
499 these OrtSuite functions by identifying interspecies interactions involved in the aerobic and
500 anaerobic conversion of benzoate to acetyl-CoA where experimental data were available. OrtSuite's
501 ability to predict interspecies interactions was assessed on a set of sequenced isolates
502 (Test_genome_set) whose growth as mono- and mixed-cultures was analyzed previously [23] under
503 three different environmental conditions (low substrate:1g/L benzoate, high substrate: 6g/L
504 benzoate and high substrate+salt stress: 6g/L benzoate supplemented with 15g /L of NaCl). This
505 dataset contained 69,193 protein sequences distributed across the 12 species resulting in a total of
506 59 Megabytes of data. More than 84% of all genes were placed in 9,533 ortholog clusters. In

addition, 541 clusters were composed of sequences obtained from all 12 species (Additional file 15: Table S10.tsv). OrtSuite's annotation stage resulted in 44 ortholog clusters with annotated sequences from ORAdb (Additional file 16: Table S11.csv). The mapping of KOs to each species in the Test_genome_set is available as supplementary data (Additional file 17: Table S12). The genomic potential of each species for aerobic and anaerobic benzoate metabolizing pathways is shown in Figure 3. The complete mapping of reactions to each species is available in the supplementary data (Additional file 18: Table S13). Based on the 44 ortholog clusters and the Gene-Protein-Reaction (GPR) rules, four species (*Cupriavidus necator* JMP134, *Pseudomonas putida* ATCC17514, *Rhodococcus ruber* BU3 and *Sphingobium yanoikuyae* DSM6900) contained all protein-encoding genes required to perform aerobic conversion of benzoate to acetyl-CoA. Except for *S. yanoikuyae* all others were in agreement with Fetzer's observation of their ability to grow in medium containing 1g/L of benzoate as the carbon source. In Fetzer's study, in a medium containing 1g/L of benzoate, growth was considered when optical densities (OD) were above 0.094. The OD measured for *S. yanoikuyae* was 0.0916. The annotation of genes with the ability to perform the complete aerobic conversion of benzoate to acetyl-CoA combined with the small difference in OD to the minimum threshold suggests that *S. yanoikuyae* indeed can grow on low benzoate containing medium but at lower growth rates. The observed growth of *Comamonas testosteroni* ATCC11996 and *Pseudomonas fluorescens* DSM6290 in the low benzoate environment was not correlated with their genomic potential according to OrtSuite in any benzoate-to-acetyl-CoA conversion pathway. It is likely that these two species utilize reactions or pathways that were not included in the benzoate degradation pathways used in our study.

Based on the genomic content, any combination of species where one of them possesses the ability to perform a complete pathway will always result in a potential interspecies interaction. To account for this scenario all species with the complete genomic potential to perform a complete pathway are excluded when calculating interspecies interactions. In the aerobic benzoate degradation pathway (P3_BTA) the protein-encoding genes that are involved in the conversion of

2-Hydroxy-2,4-pentadienoate to 2-Hydroxymuconate semialdehyde (R02604) were lacking in all remaining species and therefore, no species interactions were obtained. Similarly, in the anaerobic degradation pathways (P1_BTA and P2_BTA) no species presented the genomic content to encode proteins involved in the reactions R02451, R03028, R05579, R05581, R05594 and R05597. Therefore, no species interactions were identified that would allow the complete anaerobic conversion of benzoate to acetyl-CoA. Next, we considered the interactions where species with complete conversion capabilities are present. In the low substrate environment, OrtSuite identified 776 of 830 (93.4%) species combinations showing growth. In the high substrate environment, OrtSuite predicted 614 of 646 (95%). In the high substrate+salt stress environment, OrtSuite predicted all 271 (100%) combinations of species exhibiting growth (Additional file 19: Table S14). According to the same study in the high benzoate concentrations, 237 combinations of species did not exhibit growth (OD < 0.2545) even in the presence of known benzoate degraders which occurred in 160 combinations. From this group, OrtSuite identified 185 interspecies interactions that had the potential to perform aerobic conversion of benzoate. A possible explanation as to why growth in Fetzer's study was not observed, despite the presence of benzoate degraders, is the lack of tolerance of these species to high benzoate concentrations. For example, *C. necator* growth was shown to be stimulated at low benzoic acid concentrations but inhibited at high concentrations [31]. To assess these effects, additional stress-resistance associated genes and reactions such as those involved in medium acidification [32] could be added as constraints. Similar results were obtained when using a high substrate+salt stress medium. Growth of species combinations was not observed despite the presence of benzoate degraders. Benzoate degradation has been shown to decrease in hyperosmotic environments [33] therefore, additional constraints such as genes that confer resistance to environmental stressors or adverse conditions could be included during the identification of interspecies interactions under different or changing environmental conditions.

No single species or combination of species possessed the complete genomic potential to anaerobically convert benzoate to acetyl-CoA via the two proposed pathways (P1 and P2). Since all

growth experiments were conducted in aerobic conditions, it is possible that the species in question are only capable of using benzoate as a carbon source in aerobic environments. To fully explore all the species potential to convert benzoate, additional degradation pathways could be tested in the future in a multi-omics context. Furthermore, the only constraints added were related to the reactions that composed each pathway. Additional constraints can be included in future studies, such as potential mandatory transport-associated reactions, to increase confidence in the proposed interspecies interactions. In summary, OrtSuite confirmed that most interspecies interactions (96.1%) identified by Fetzer and collaborators were possible via the aerobic conversion of benzoate to acetyl-CoA but were not possible at anoxic conditions.

Conclusions

The OrtSuite workflow allows hypothesis-driven exploration of potential interactions between microbial genomes by limiting the search universe to a user-defined set of ecosystem processes. This is achieved by rapidly assessing the genetic potential of a microbial community for a given set of reactions. The two-step annotation of clusters of orthologs with a personalized ORAdb decreases the overall number of sequence alignments that need to be computed. User-specified constraints, such as the presence of transporter genes, further reduces the search space for putative microbial interactions. Users have substantial control over several steps of OrtSuite: from manual curation of ORAdb, custom sequence similarity cutoffs to the addition of constraints for inference of putative microbial interactions. As long as links between genes and reactions exist, the flexibility and easy usage of OrtSuite allow its application to the study of any given ecosystem process.

Declaration Sections

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The genomes used to test the workflow are available at National Centre for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) under the accession identifiers CP029389-CP029397, GCF_000001735, AP012304, AP012305, CP021731, CP011072, CP007785-CP007787, CP000614-CP000621, CP003230, CP005996, CP003108, CR555306-CR5553068, GCF_000225785, LN997848-LN997849, CP022989-CP022996, CP024315, AP012547, CP022046-CP022047 and CP001281-CP001282. The genome assemblies used to predict interspecies interactions are available at the European Nucleotide Archive (ENA - https://www.ebi.ac.uk/ena) with the study accession PRJEB38476. All analysis results and scripts used to generate them are available at https://github.com/mdsufz/OrtSuite.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the Helmholtz Young Investigator grant VH-NG-1248 Micro ‘Big Data’.

Author’s contributions

JS, OD, PS and UNR developed the concept of OrtSuite. JS, MG and UNR developed the OrtSuite workflow. JS, MG and UNR performed the benchmarks. CV provided information and data for defining benzoate to acetyl-CoA conversion pathways. RK sequenced bacterial isolates provided by AC. JS and UNR wrote the manuscript. All authors read and approved the manuscript.

Acknowledgments

We thank the early users of OrtSuite Sandra Silva and Felipe Côrrea for their help with beta testing and for workflow suggestions. We also thank Diogo Lima for his assistance in the implementation of the script required to generate the Gene-Protein-Reaction (GPR) rules; and Nicole Steinbach for her work in the sequencing of the isolates used as the test set.

Availability and requirements

Project name: OrtSuite

Project home page: https://github.com/mdsufz/OrtSuite

Operating system: Linux64

Programming languages: Shell; Python3.6, R

Other requirements: Java, BLAST+ [34], BLAST [13], DIAMOND [14], OrthoFinder [20], mcl [35]

License**:** MIT

List of abbreviations

BTA: Benzoate-to-Acetyl-CoA

CPU: Central Processing Unit

EC: Enzyme Commission

GPR: Gene-Protein-Reaction

KEGG: Kyoto Encyclopedia of Genes and Genomes

KO: KEGG ortholog

OD: Optic Density

ORAdb: Ortholog-Reaction Association database

ORF: Open Reading Frame

RAM: Random Access Memory

References

1. Maestre FT, Castillo-Monroy AP, Bowker MA, Ochoa-Hueso R. Species richness effects on ecosystem multifunctionality depend on evenness, composition and spatial pattern: Community attributes and multifunctionality. J Ecol. 2012;100:317–30.

2. Mulder CPH, Uliassi DD, Doak DF. Physical stress and diversity-productivity relationships: The role of positive interactions. Proc Natl Acad Sci. 2001;98:6704–8.

3. Lyu Z, Shao N, Akinyemi T, Whitman WB. Methanogenesis. Curr Biol. 2018;28:R727–32.

4. Liu F, Hewezi T, Lebeis SL, Pantalone V, Grewal PS, Staton ME. Soil indigenous microbiome and plant genotypes cooperatively modify soybean rhizosphere microbiome assembly. BMC Microbiol. 2019;19:201.

5. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci. 2016;201521291.

6. Raynaud X, Nunan N. Spatial Ecology of Bacteria at the Microscale in Soil. PLoS ONE. 2014

7. Roh SW, Abell GCJ, Kim K-H, Nam Y-D, Bae J-W. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends Biotechnol. 2010;28:291–9.

8. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;1.

9. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019;176:649-662.e20.

10. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5:170203.

11. Morin M, Pierce EC, Dutton RJ. Changes in the genetic requirements for microbial interactions with increasing community complexity. eLife. 2018;7:e37072.

12. Slade EM, Kirwan L, Bell T, Philipson CD, Lewis OT, Roslin T. The importance of species identity and interactions for multifunctionality depends on how ecosystem functions are valued. Ecology. 2017;98:2626–39.

13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

14. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.

15. Koonin EV. Orthologs, Paralogs, and Evolutionary Genomics. Annu Rev Genet. 2005;39:309–38.

16. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol. 2017;34:2115–22.

17. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 2003;13:2178–89.

18. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma Oxf Engl. 2006;22:1658–9.

19. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.

20. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.

21. Gottstein W, Olivier BG, Bruggeman FJ, Teusink B. Constraint-based stoichiometric modelling from single organisms to microbial communities. J R Soc Interface. 2016;13:20160627.

22. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. 2019;14:639–702.

23. Fetzer I, Johst K, Schäwe R, Banitz T, Harms H, Chatzinotas A. The extent of functional redundancy changes as species' roles shift in different environments. Proc Natl Acad Sci U S A. 2015;112:14888–93.

24. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–D280.

25. Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 2013;14:360–6.

26. Pertsemlidis A, Fondon JW. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol. 2001;2:reviews2002.1.

27. Lynch M. Evolution of the mutation rate. Trends Genet. 2010;26:345–52.

28. Madden T. The BLAST Sequence Analysis Tool. National Center for Biotechnology Information (US); 2003.

29. Ng PC, Henikoff S. Predicting the Effects of Amino Acid Substitutions on Protein Function. Annu Rev Genomics Hum Genet. 2006;7:61–80.

30. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma Oxf Engl. 2014;30:2068–9.

31. Wang W, Yang S, Hunsinger GB, Pienkos PT, Johnson DK. Connecting lignin-degradation pathway with pre-treatment inhibitor sensitivity of Cupriavidus necator. Front Microbiol. 2014

32. Kitko RD, Cleeton RL, Armentrout EI, Lee GE, Noguchi K, Berkmen MB, et al. Cytoplasmic Acidification and the Benzoate Transcriptome in Bacillus subtilis. PLoS ONE. 2009.

33. Bazire A, Diab F, Jebbar M, Haras D. Influence of high salinity on biofilm formation and benzoate assimilation by Pseudomonas aeruginosa. J Ind Microbiol Biotechnol. 2007;34:5–8.

34. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

35. van Dongen S. Graph Clustering by Flow Simulation. University of Utrecht; 2000.

Figures legends

Figure 1. OrtSuite workflow. Protein sequences from samples supplied by the user are clustered using OrthoFinder **(a)**. OrtScraper takes a text file containing a list of identifiers for each reaction in the pathway of interest supplied by the user to retrieve all protein sequences from KEGG **(b)**. Sequences mapped to reactions are stored in ORAdb **(c)**. Functional annotation **(d)** consists of a two-stage process (relaxed and restrictive search). Relaxed search **(e)** performs sequence alignments between 10% of randomly selected sequences from each generated cluster. Clusters whose representative sequences share a minimum 50% identity to sequences in reaction set(s) in ORAdb transition to the restrictive search **(f)**. Here, all sequences from the cluster are aligned to all sequences in the corresponding reaction set(s) to which they had a hit. Finally, the annotated sequences are used to identify putative microbial interactions based on their functional potential **(g and h)**. Additional constraints can be added to reduce the search space of microbial interactions (rounded blue rectangles) (e.g. subsets of reactions required to be performed by single species, transport-related reactions required).

Figure 2. Benzoate to acetyl-CoA conversion pathways. Reactions in blue background illustrate the anaerobic degradation of benzoate to acetyl-CoA via benzoyl-CoA. Reactions in yellow background illustrate the aerobic degradation of benzoate via catechol.

Figure 3. Mapping of the genomic potential of each species from the Test_genome_set dataset to each reaction in aerobic (yellow) and anaerobic (blue) benzoate-to-acetyl-CoA conversion pathways.

Additional files

Additional file 1:

Table S1. Number of reactions, enzymes, KO groups and KO-associated sequences represented in each alternative benzoate to acetyl-CoA conversion pathway used (BTA_P1, BTA_P2 and BTA_P3).

Additional file 2:

Table S2. Species names, strain and abbreviation codes used to validate OrtSuite. The genomic potential, based on KEGG database, to completely encode all proteins involved in a BTA pathway is identified in the column “BTA pathway” (P1 – Anaerobic conversion of benzoate to acetyl-CoA 1; P2 – Anaerobic conversion of benzoate to acetyl-CoA 2; P3 – Aerobic conversion of benzoate to acetyl-CoA) . The column OrtSuite_result contains which BTA pathway(s) were identified as being completely performed by each species used for validation.

Additional file 3:

Example_reaction_list. Example file of a list of KEGG reaction identifiers used to generate ORAdb.

Additional file 4:

Example_ec_list. Example file of list of enzyme commission (EC) numbers used to generate ORAdb.

Additional file 5:

Example_ko_list. Example file of list of KEGG ortholog identifiers used to generate ORAdb.

Additional file 6:

Table S3. Example of the binary table with the mapping of KO identifiers (rows) to the species of interest (columns) (1 – indicates the presence of genes in the species; 0 – indicates the absence of genes in the species).

Additional file 7:

Table S4. Gene-Protein-Reaction (GPR) rules and metadata associated with all reactions present in the Ortholog Reaction-Associated database (ORAdb).

Additional file 8:

Table S5. User-defined constraints: pathway name, complete set of reactions present in each pathway, sets of reactions required to be performed by single species (each subset is described between parenthesis) and transport reactions. Transporter column describes the transport reaction (e.g. R00750) that is coupled to the reaction in the pathway (e.g. R02601). Thus, species that perform the latter must also contain the genes associated with the transport reaction.

Additional file 9:

Set_A_genomes. Compressed folder containing all the genomes (FASTA format) of the test species used to test the workflow. Species abbreviation codes are used as file names.

Additional file 10:

Test_genome_set. Sequenced genomes of the species used in a study by Fetzer and collaborators [23].

Additional file 11:

Table S6. Species used in the Fetzer study [23] and their ability to grow as monocultures in benzoate-containing medium.

Additional file 12:

Table S7. Growth results of microbial communities composed of individuals and groups of species from the Fetzer study [23] on three different environments. Environment 1 contained a benzoate concentration of 1g/L. Environment 2 contained a benzoate concentration of 6g/L. Environment 3 contained a benzoate concentration of 6g/L and 15 g/L of NaCl.

Additional file 13:

Table S8. OrtSuite workflow runtime. The total runtime of each OrtSuite step when analyzing the genomic potential of species in Set_A_genomes dataset in three pathways (P1, P2 and P3) for the conversion of benzoate to acetyl-CoA (BTA). Steps were performed with default parameters on a laptop with 4 cores and 16 GB of RAM.

Additional file 14:

Table S9. Sequence alignments of original and mutated sequences using BLAST [13].

Additional file 15:

Table S10. Statistics obtained during the clustering of protein orthologs using the Test_genome_set. Results include numbers of species, genes, clusters and genes per cluster.

Additional file 16:

Table S11. Overview of the number of clusters, sequences and KOs during the annotation of the Test_genome_set.

Additional file 17:

Table S12. Mapping of species annotated with KO identifiers from ORAdb.

Additional file 18:

Table S13. Potential of species in the Test_genome_set to perform reactions associated with benzoate degradation based on ORAdb. Species identifiers: (A) *Bacillus subtilis* ATCC, (B) *Paenibacillus polymyxa* ATCC 842, (C) *Brevibacillus brevis* ATCC 8246, (D) *Comamonas testosteroni* ATCC 11996, (E) *Cupriavidus necator* JMP 134, (F) *Pseudomonas putida* ATCC 17514, (G) *Pseudomonas fluorescens* DSM 6290, (H) *Variovorax paradoxus* ATCC 17713, (I) *Rhodococcus sp.* (isolate UFZ), (J) *Acidovorax facilis* (isolate UFZ), (K) *Rhodococcus ruber* BU3, (L) *Sphingobium yanoikuyae* DSM 6900. (1 – species with the complete genomic potential to perform the reaction; 0 – species without the complete genomic potential to perform the reaction).

Additional file 19:

Table S14. Growth of single species and in combination with others measured by Fetzer and collaborators in three different media (low substrate: 1g/L benzoate, high substrate: 6g/L benzoate and high substrate+salt stress: 6g/L benzoate supplemented with 15 g/L of NaCl). Species identifiers: (A) *Bacillus subtilis* ATCC, (B) *Paenibacillus polymyxa* ATCC 842, (C) *Brevibacillus brevis* ATCC 8246, (D) *Comamonas testosterone* ATCC 11996, (E) *Cupriavidus necator* JMP 134, (F) *Pseudomonas putida* ATCC 17514, (G) *Pseudomonas fluorescens* DSM 6290, (H) *Variovorax paradoxus* ATCC 17713, (I) *Rhodococcus sp.* (isolate UFZ), (J) *Acidovorax facilis* (isolate UFZ), (K) *Rhodococcus ruber* BU3, (L) *Sphingobium yanoikuyae* DSM 6900. Growth was considered when optical density (OD) was equal or above 0.094, 0.2545 and 0.0752 in environments with low substrate, high substrate and high substrate+salt stress benzoate, respectively.