 doi:10.20944/preprints201909.0070.v1

ReFDash - A Repository of Functional Dashboards providing comprehensive functional insights inferred from 16S microbiome data sets

Sunil Nagpal [1], Mohammed Monzoorul Haque [1] and Sharmila S. Mande [1]*

[1] Bio-Sciences R&D Division, TCS Research, Tata Consultancy Services Limited, 54-B, Hadapsar Industrial Estate, Pune, India 411013

*Corresponding Author: sharmila.mande@tcs.com

Keywords: Microbiome, Inferred functions, Database, 16S, Metagenomics, Comparative metagenomics

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2019 doi:10.20944/preprints201909.0070.v1

Abstract

Motivation: 16S rRNA gene amplicon based sequencing has significantly expanded the scope of metagenomics research by enabling microbial community analyses in a cost-effective manner. The possibility to infer functional potential of a microbiome through amplicon sequencing derived taxonomic abundance profiles has further strengthened the utility of 16S sequencing. In fact, a surge in 'inferred function metagenomic analysis' has recently taken place, wherein most 16S microbiome studies include inferred functional insights in addition to taxonomic characterization. Tools like PICRUSt, Tax4Fun, Vikodak and iVikodak have significantly eased the process of inferring function potential of a microbiome using the taxonomic abundance profile. A platform that can enable hosting of inferred function 'metagenomic studies' with comprehensive metadata driven search utilities (of a typical database), coupled with on-the-fly comparative analytics between studies of interest, can be a major improvement to the state of art. ReFDash represents an effort in the proposed direction.

Methods: This work introduces ReFDash - a Repository of Functional Dashboards. ReFDash, developed as a significant extension of iVikodak (function inference tool), provides three broad unique offerings in inferred function space - (i) a platform that hosts a database of inferred function data being continously updated using public 16S metagenomic studies (ii) a tool to search studies of interest and compare upto three metagenomic environments on the fly (iii) a community initiative wherein users can contribute their own inferred function data to the platform. ReFDash therefore provides a first-of-its-kind community-driven frame-work for scientific collaboration, data analytics, and sharing in this area of microbiome research.

Results: Overall, the ReFDash database is aimed at compiling together a global ensemble of 16S-derived Functional Metagenomics projects. ReFDash currently hosts close to 50 ready-to-use, re-analyzable functional dashboards representing data from approximately 18,000 microbiome samples sourced from various published studies. Each entry also provides direct downloadable links to associated taxonomic files and metadata employed for analysis.

Conclusion: The vision behind ReFDash is creation of a framework, wherein users can not only analyze their microbiome datasets in functional terms, but also contribute towards building an information base by submitting their functional analyses to ReFDash database.

ReFDash web-server may be freely accessed at https://web.rniapps.net/iVikodak/refdash/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2019 doi:10.20944/preprints201909.0070.v1

1. Introduction

16S rRNA gene-based profiling techniques are widely used for deciphering the structure of bacterial communities residing in varied ecological niches (Gill et al., 2006; Huse et al., 2012; Yatsunenko et al., 2012). In the context of clinical studies, a (cross-sectional/ longitudinal) comparison of community structures between the two or more states of health/ disease can help in preliminary identification of specific bacterial taxa (or groups of taxa) that demonstrate statistically significant association(s) between their presence (or abundance) and the state of health/ disease severity (Alekseyenko et al., 2013; Botero et al., 2014; Cui et al., 2012; Ganju et al., 2016; Griffen et al., 2012; Kirst et al., 2015; Tandon et al., 2018). In some cases, differences in community structure (quantified using alpha or beta diversity measures) of the studied microbial communities have also served as important cues to understand dynamics of specific clinical conditions (Haque et al., 2017; Tandon et al., 2019). Although from the perspective of health/ disease diagnostics, the mentioned applications hold clinical significance, it is important to understand whether (and how) various microbes (inhabiting a given ecological niche) 'functionally' contribute to a physiological state. Detection and quantification of microbial functions (at individual as well as at community level) is a logical next step with respect to generating hypotheses about the mechanistic aspects of a disease/ clinical condition.

Although shotgun metagenomic sequencing is the preferred technique for analysing microbiomes in terms of both taxonomy and function, the significantly lower sequencing (and downstream computational) costs associated with 16S rRNA gene amplicon-based sequencing has rendered the latter as the method of choice for microbial community analysis (Bose et al., 2015; Haque et al., 2015; Keegan et al., 2016; Narayanasamy et al., 2016; Quince et al., 2017; Reddy et al., 2014; Rossen et al., 2018; White et al., 2017). 16S rRNA gene amplicon-based sequencing for microbiome analysis has recently gained further traction due to the availability of several stand-alone computational methods that make it possible to predict/ infer (from 16S rRNA gene sequence counts) the repertoire of functions that are encoded by various microbes constituting the studied bio-specimen (Aßhauer et al., 2015; Langille et al., 2013; Nagpal et al., 2016). Although the mentioned methods (Tax4Fun, Picrust, Vikodak, etc.) vary with respect to the core methodology employed for generating functional inferences (from taxonomic input data), the outputs of these tools are typically in form of textual matrices indicating the abundances of predicted/ inferred functions.

Two recent tools (viz. Burrito and iVikodak) however go a step ahead and provide accompanying interactive visualizations corresponding to the generated textual outputs (McNally et al., 2018;

Nagpal et al., 2019). In particular, the tool 'iVikodak' provides end-users a comprehensive ensemble of functional inference and analysis tools, the results of which can be accessed via an intuitive visualization interface. The interface not only enables interactive visualization of predicted functions, but also allows users to download the entire visual compendium of results in the form of a personalized 'dashboard' file. The latter file (referred to as '.dash' file) can be re-uploaded (whenever required) to iVikodak to recreate the entire dashboard of results (Nagpal et al., 2019).

Overall, a dashboard file ('.dash file') generated by iVikodak represents a storable (and readily shareable/ retrievable) compilation of 'pre-analyzed, re-analyzable, and visualizable' functions that have been computationally inferred/ derived from 16S rRNA gene sequencing data corresponding to various bio-specimens sampled from a given environmental niche. If well-annotated, a compilation of such 'functional' dashboards would primarily be of immense use to researchers in the field of microbiome analysis. Such an organised compilation (comprising of 'functional information' corresponding to various sequenced microbial environments) is expected to complement the functionality of other existing databases providing microbiome sequence data and/ or associated taxonomic profiles.

In this study, we present 'ReFDash' (Repository of Functional Dashboards) - a compilation of 'Functional Dashboards' generated by iVikodak. ReFDash has been built with the following objectives –

1. Host pre-generated functional dashboards that have been generated from taxonomic profiles (and available metadata) corresponding to various microbial environments.
2. Create an interactive web-accessibe computational framework that facilitates (automated) 'on-demand' (re-processing and) 'comparison' of functional profiles of a subset of selected microbial environments that are available as pre-generated/ end user deposited dash-files in the ReFDash repository.
3. Enable end users to upload, deposit and share their functional dashboards (i.e. dash-files generated by iVikodak) with other members in the scientific community. The purpose behind this enablement is to encourage microbiome researchers to help expand the repertoire of environments (represented by dash files in the ReFDash database). This would ease the process of scientific data collaboration, sharing or even a peer-review process.

Overall, ReFDash represents a vision towards development of a community platform/ database that not only hosts 'functional profiles' corresponding to diverse microbial communities/ environments but also facilitates automated processing and cross-comparison of multiple (user selected)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2019 doi:10.20944/preprints201909.0070.v1

environments, in terms of the functions they encode.

2. Materials and Methods

Database entries in ReFDash are primarily 'functional profiles' i.e. dashboard files created using iVikodak. **Figure 1** provides a conceptual overview of how a dashboard is generated, the process of inclusion of a dashboard into ReFDash, and other details about database entries in the ReFDash repository.

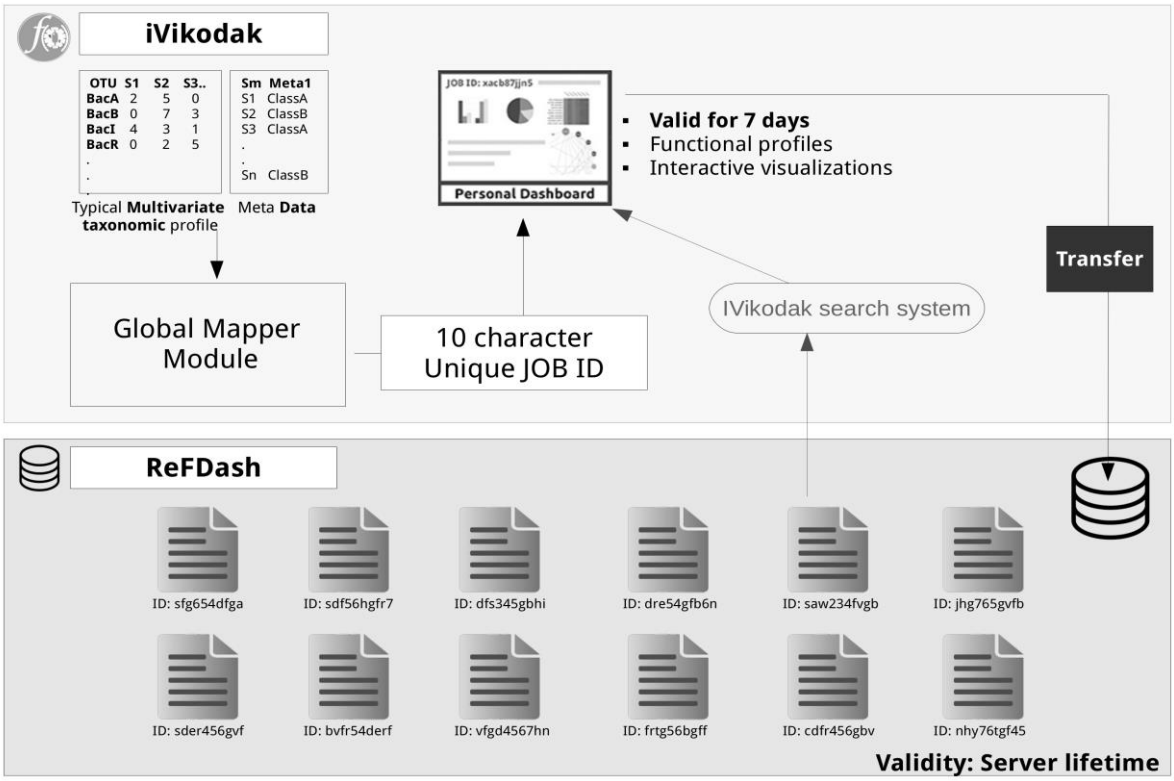


Figure 1: A conceptual overview of dashboard generation procedure and the workflow followed for dashboard inclusion in the ReFDash server. Functional dashboard entries, the primary database entries in ReFDash, are populated in two ways. Dashboard entries are either (a) 'pre-generated' by (database admin) through processing publicly-available microbiome taxonomic data, or are (b) deposited by scientific user community. In both cases, taxonomic abundance and related metadata information pertaining to multiple samples (corresponding to a particular microbial environment) is initially processed using Global Mapper module of iVikodak. Functional Dashboards generated by iVikodak get transferred as database entries in ReFDash after satisfying appropriate quality check criteria. Once transferred to ReFDash, a database entry obtains life-time validity.

2.1. Dashboard Generation

As mentioned previously, a plethora of functional inferences are generated by iVikodak by processing 16S rRNA taxonomic profiles using user specified algorithms and run-time parameters. The generated functional inferences reflect the functional potential of a given microbial environment and its constituent microbes. Besides enabling interactive visualization of the generated inferences, iVikodak's interface allows end-users to generated additional insights by providing automated overlay of metadata over the results. Users are provided a unique job id for

each such cycle of functional inference that they perform using iVikodak. End-users are ultimately provided the option to download and store the full ensemble of the results packaged in form of a dashboard file. Any user intending to access/ retrieve results corresponding to a given environment can either (a) provide the respective job-id in the iVikodak portal, or (b) upload the '.dash' file to the 'Recreator' module in iVikodak. Overall, a dashboard file represents a 'functional' equivalent of a 'taxonomic' profile. Multiple such dashboards can be created/ generated from taxonomic profiles corresponding to diverse environmental niches. However in contrast to a simple taxonomic or functional profile, dashboards generated by iVikodak have value addition in terms of the multiple types of interactive visualizations and additional analysis that this compilation enables.

2.2. Types of Dashboards in ReFDash

Dashboards (the primary database entries in ReFDash) are of the following two types (a) pre-generated dashboards (b) user-contributed dashboards. While the former dashboards have been 'pre-generated' using publicly available microbiome sequence data (or taxonomic profiles corresponding to the same), the latter represent functional dashboards contributed to the ReFDash repository by users who have opted to contribute/ share the functional inferences that were derived for a particular environmental niche with their peers, reviewers or the scientific community at large. However it may be noted that in ReFDash' present version, iVikodak remains the primary back-end driver for generation of both types of dashboards. The methodology adopted for building 'pre-generated' dashboards (including sources of data) and the curation/ quality-check protocol that has been put in place for adding 'user-contributed' dashboards to the ReFDash repository are detailed in the sections below.

2.2.1. Pre-generated Dashboards: Methodology and Data sources

The ReFDash repository currently hosts 50 pre-generated dashboards accounting for more than 20,000 microbiome samples. All existing 'pre-generated' entries in the database have been created by the administrators of the database (i.e. authors). Wherever possible, authors have explicitly ensured the use of popular open-access taxonomic abundance profiles data that accompany the publication corresponding to the specific microbiome study (Duvallet. 2017; Mitchell et al., 2018). RDP classifier v2.12 (Cole et al., 2014), executed at a boot-strap confidence threshold of 80%, was employed for taxonomic classification of sequence data of studies where pre-generated taxonomic abundance profiles were not available. Prior to taxonomic classification, sequence data was pre-processed using prinseq v0.20.4 (Schmieder and Edwards, 2011) maintaining a minimum quality score of 25. Metadata for all the studies was diligently compiled from NCBI Run selector. It is pertinent to note that the strain level taxonomic profiles were transformed to Genus level using in-house scripts. It is important to note that taxonomic profile and corresponding metadata information

for every study hosted (as Dashboard entries) on ReFDash is open access and freely available for download. **Supplementary Table 1** provides a comprehensive summary of source and journal (doi) of all the studies currently hosted on ReFDash.

2.2.2. User-contributed dashboards

The current version of iVikodak provides end-users the option to contribute/ share the functional inferences that were derived for a particular environmental niche with his/ her peers, reviewers or the scientific community at large. Alternatively, users can directly access the 'Contribute' widget on the ReFDash database to submit/ deposit taxonomic profiles, associated metadata, and the functional inferences (i.e. iVikodak's dashboards). Every such user-contribution undergoes a curation/ quality-check protocol (as depicted in **Figure 2**) prior to getting populated in the ReFDash database.

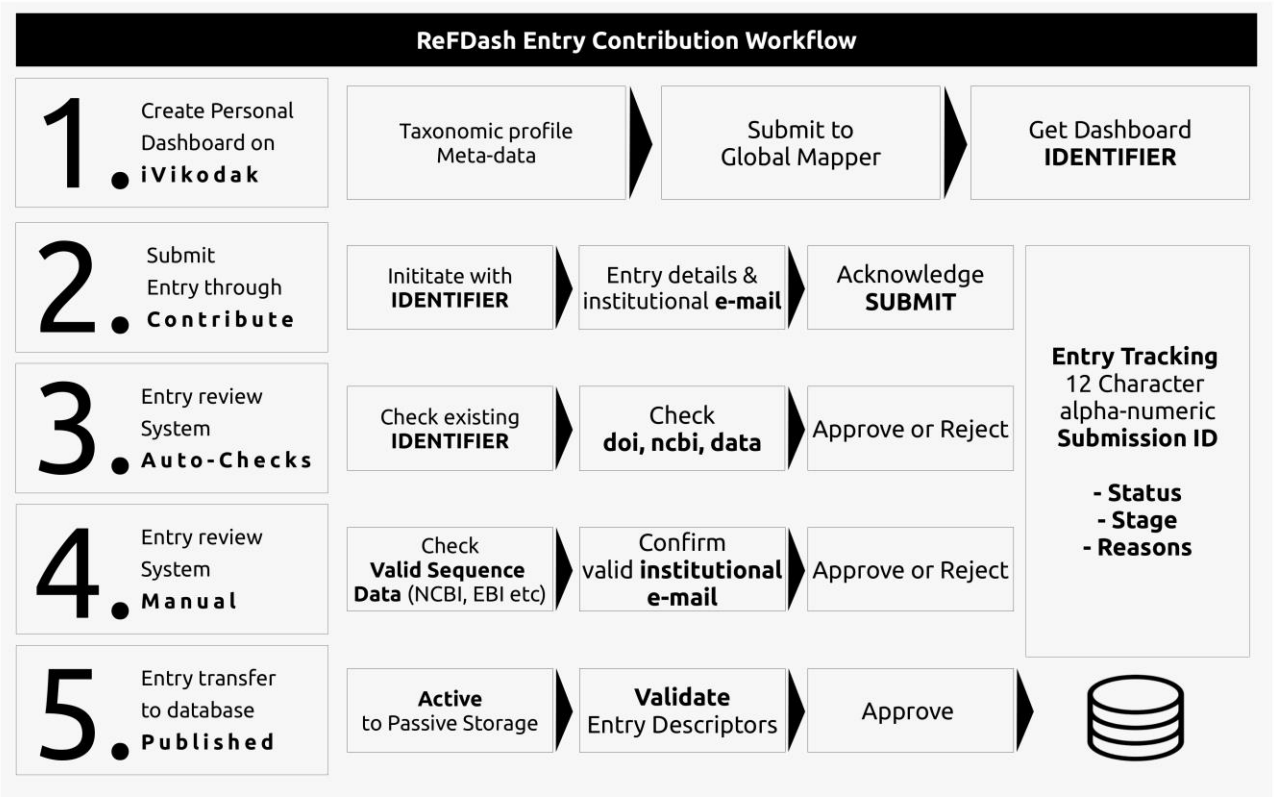


Figure 2: An overview of the process/ protocol followed for users 'contributing' taxonomic profiles, associated metadata, and the functional inferences to the ReFDash database. Once a user generates a Functional Dashboard using iVikodak (Step 1), he/ she can store the dashboard and retrieve it for re-analysis via a unique dashboard id (having 7 days validity). If intention is to share, he/ she may initiate the submission via the dashboard identifier along with a few preliminary details (Step 2). The database admin then confirms the authenticity of the entry via a check of the entered database identifiers (a valid Bioproject/ Study ID) along with study digital object identifier and the dashboard id (Steps 3 and 4). An entry clearing the above two steps is then processed by database admin for publishing it as a valid entry in RefDash.

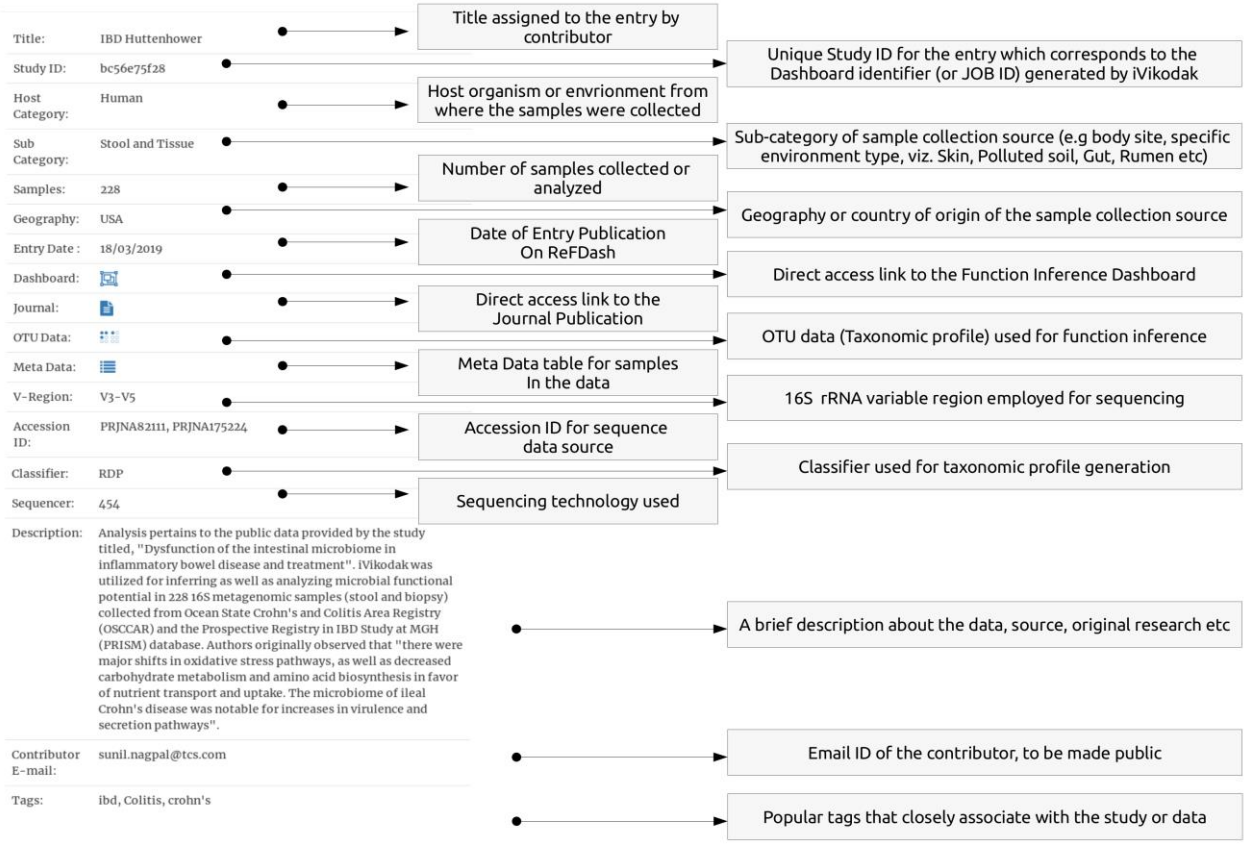
To prevent frivolous submissions, every contribution is assessed based on the following two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2019 doi:10.20944/preprints201909.0070.v1

inclusion criteria - (a) The sequence data corresponding to the study should have a valid Bioproject/ Study ID that is searchable on the NCBI Sequence Read Archive (accessible at https://www.ncbi.nlm.nih.gov/sra) or the European Nucleotide Archive (accessible at https://www.ebi.ac.uk/ena) (b) The need for a submission from a valid institutional email address. An email is sent to the institutional email-id (provided in the submission form) to confirm the credentials of the person making the submission. Redundancy checks are subsequently carried out for all submission requests that satisfy the above criteria.

2.3. Description of database entries

Figure 3 provides an overview of the kinds of information in each of the entries of the ReFDash



Database.

Figure 3: An overview of the kinds of information in each of the entries of the ReFDash Database. For facilitating easy, structured searches and instant access for end-users, all fields of information categories/ columns (described in the image) are 'sortable'. Information provided in all fields is also indexed and a separate customised search bar has been provided in the ReFDash main page to enable users to use one/ more keywords to dynamically filter and enlist database entries of interest.

At the outset, it is important to note that all columns of information categories/ columns (described below) in ReFDash have been rendered 'sortable' for facilitating easy, structured searches and

instant access. Information provided in all columns have also been indexed and a separate customised search bar has been provided to enable users to use one/ more keywords to dynamically filter and enlist database entries of interest to the user.

Each entry in the database provides the following information and (downloadable) data for end-users -

(1) Title and Study Id: A unique title and an alphanumeric code (of length 10) that serves as a study ID. End-users can (a) either click on the dashboard icon (as indicated in Figure 3) to directly access, visualize and analyse various functional inferences corresponding to a microbial environment, or (b) provide the study ID in iVikodak to retrieve the entire dashboard. In this case, the study ID code serves as a Job ID in iVikodak.

(2) Publication Access: The 'publication' corresponding to each study (accessed through its unique 'doi' i.e. digital object identifier) is also available to the end-user and forms a part of the information stored in each individual database entry.

(3) Microbiome habitat (as category/ sub-category): This indicates the source habitat category from which various microbiome bio-specimens were sampled from in the respective study. It indicates if the microbiome samples are host-associated (e.g. human, termite, etc.) or an environmental sample (e.g. marine, freshwater, acid-mine etc.). A sub-category field is also available that provides specific location details e.g. gut, skin, colon, cheese, rumen, etc.

(4) Sample Details: Multiple fields of information pertaining to the geographical location, number of samples, the 16S variable region (V-Region) that was targeted for amplicon sequencing, the sequencing technology employed, etc are also provided to end-users (as searchable, sortable, filterable information fields). The accession ID corresponding to the sequence data also constitutes one of the fields of information.

(5) **Downloadable (Annotated) Information:** A ReFDash database entry created for each individual microbiome study acts as a single-point resource for access to -

(a) **Taxonomic Abundance Tables:** Users can download (in tab-delimited format) the taxonomic abundance profiles corresponding to all samples belonging to a study entry. It may be noted that dashboards (i.e. collation of functional inferences) in ReFDash have been generated by providing the same abundance table as input to iVikodak. Information regarding the classification method/ tool e.g. RDP (Cole et al., 2014), QIIME (Kuczynski et al., 2011), SILVA (Pruesse et al., 2007), Mgnify (Mitchell et al., 2018), etc. that was employed while generating the respective taxonomic annotations is also provided with each database entry.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2019 doi:10.20944/preprints201909.0070.v1

(b) **Study Metadata**: Available meta-information fields (e.g. age, sex, disease status, disease severity, location, geography, BMI, medication status, etc.) for every study entry in ReFDash is also provided as a downloadable file for end-users of ReFDash.

(c) **Functional Dashboard:** The dashboard file for the study - essentially a compact version of pre-generated functional inferences that is storable and readily shareable, retrievable. Clicking on the dashboard icon retrieves the inferences and enables visualization and further re-analysis as required.

2.4. Functionalities for comparing microbial environments

An important objective in microbiome studies is to a comparative analysis of two or more study cohorts comprising of microbiome samples that differ in terms of state/ stage/ sampling time-point. The aim is to find (and quantify to the extent possible) overlap/ differences in taxonomic structure and the functional potential of constituent microbes. In a clinical context, such analyses have the potential in identifying candidate microbes (or functions) that account for a healthy or a diseased state.

Given the above, ReFDash incorporates functionalities that enable an 'automated' comparison of the functional potential of a pair (or three) microbial environments that have been represented as Dashboards within ReFDash. When a user chooses two (or three) dashboards for comparison, ReFDash then provides the end-users a listing of metadata field headers thereby enabling them to appropriately select metadata fields that are suitable for overlay and comparison. Once this information is selected by the end-user, ReFDash, at its backend, performs an automated merger/ compilation of the taxonomic abundance profiles corresponding to the chosen environments as well as combines information pertaining to the chosen metadata fields to create unified files. The latter files are subsequently processed by ReFDash (using iVikodak in the back-end) to provide a Dashboard that contains analysed functional information that can be now be interactively visualized by end-users in the same context.

2.5. Implementation and Technology

ReFDash uses a typical NoSQL architecture (Leavitt et al., 2010) for storage of approved dashboard jsons. The approved dashboards are passively stored using a key-value system, enabling a read-only access framework for the data in passive container (**Figure 4**). A 10-character unique alphanumeric identifier or key is tagged to each approved dashboard. On the other hand, dashboards created using comparison submission system are consequently part of an active storage system that is purged every 30 minutes for clearing any temporary dashboard created by end-users for comparison

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2019 doi:10.20944/preprints201909.0070.v1

purposes. It is therefore pertinent to note that results of comparisons stay online for a maximum period of 30 minutes from the time of their creation Dashboards in active storage system are accessible using a 12 character alphanumeric key. The active system purging is set in place to avoid storage bottle-necks possible due to frivolous use of the comparison functionality. The server side connections are made using PHP, while the front end is designed using bootstrap3 (Bostock et al., 2011), in-house java-scripts, datatables.js and plotly.js (Plotly Technologies Inc., 2015).

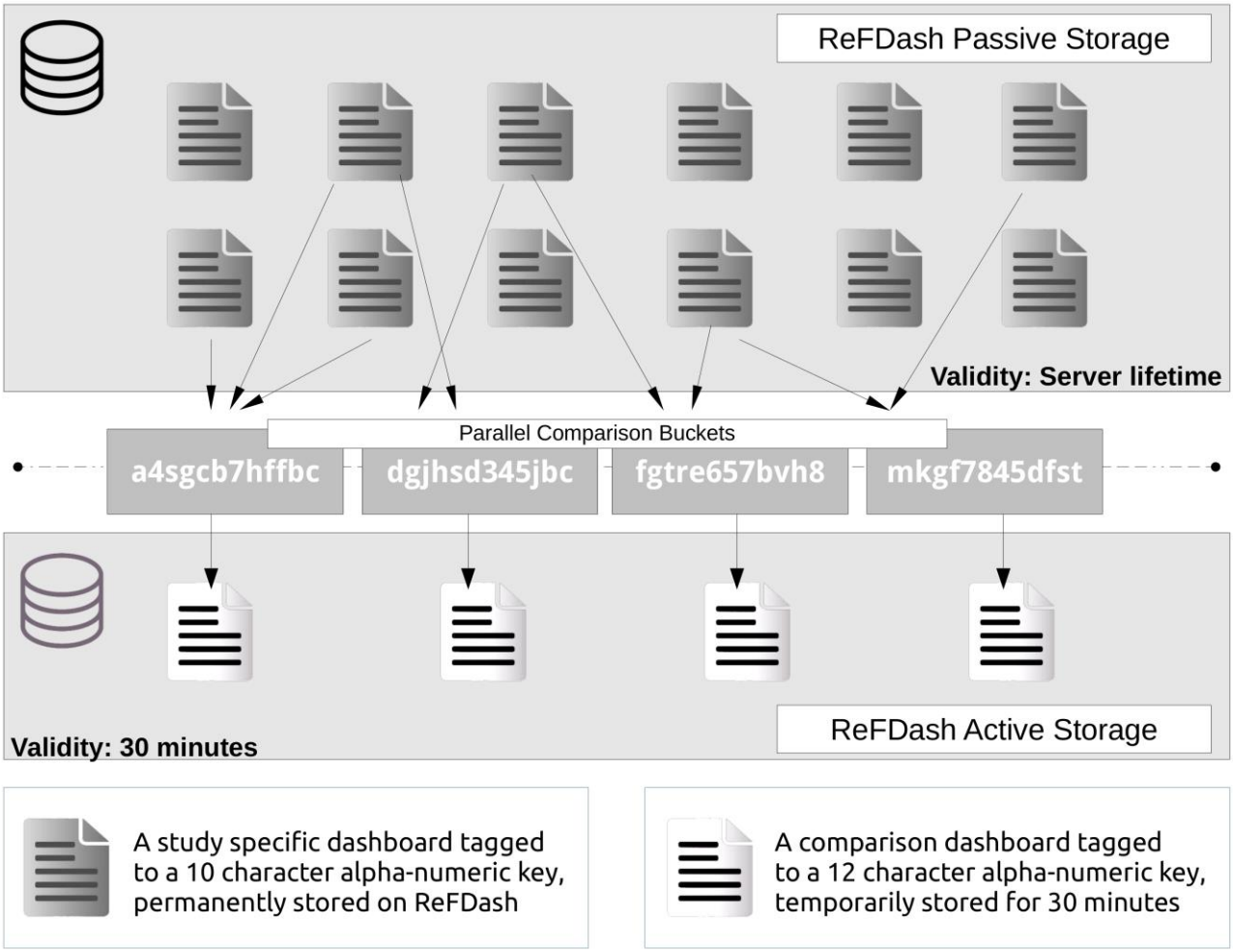


Figure 4: The architecture of ReFDash storage system. The passive storage system contains permanently stored dashboards which can be temporarily retrieved for comparative analysis inside the active storage system. Each dashboard inside the passive storage is tagged to a 10 character alpha-numeric key, whereas each temporary dashboard inside active storage system is tagged to a 12 character alpha-numeric key. Data (temporary comparison dashboards) stored in the active storage system has a life span of 30 minutes.

2.6. Limitations, future development and enhancements

Following are the key enhancements planned for ReFDash –

1. **Multi-algorithmic enabled Functional Inferences:** Currently, ReFDash hosts dashboards that have been generated using iVikodak (Nagpal et al., 2019) As a future enhancement, each database entry in ReFDash is planned to allow end-users to view functional dashboards generated not only using iVikodak but also Dashboards generated using functional

inferences obtained using other analogous algorithms like PICRUSt (Langille et al., 2013) and and Tax4Fun (Aßhauer et al., 2015).

2. **Information-rich Dashboards:** Future dashboard versions in ReFDash are planned to allow end-users to view both taxonomic and functional inferences via an interface that is enhanced with repect to the functionalities that are available in the current version of iVikodak.

3. **Comparison Module:** Currently, ReFDash allows comparion of at most 3 dashboards. This functionality is planned to be expanded using a suitable set of technologies.

3. Availability and Requirements

1. **Project name**: ReFDash
2. **Home page**: https://web.rniapps.net/iVikodak/refdash/
3. **Operating system(s):** Web platform compatible with all operating systems
4. **Browser requirement(s):** WebGl enabled for viewing 3D plots

For inquiries and general discussions, please contact sunil.nagpal@tcs.com, mm.haque@tcs.com or sharmila.mande@tcs.com.

4. Availability of data and materials

A summary of the source(s) of publicly available datasets that were used for building pre-generated dashboards in the ReFDash repository has been provided in Supplementary Table 1.

5. Author Contributions

SN, MH, and SM conceived idea for ReFDash. SN and MH mined data for ReFDash. SN designed and developed the web platform for ReFDash. MH, SN, and SM prepared the manuscript. All authors have reviewed and validated the platform and manuscript.

6. Funding

The authors declare that this study received funding in form of monthly remuneration (salary) for the authors, from TCS Ltd. The funder had the role of a promoter of fundamental research in Bio-Sciences Research Division of TCS Research. This study was an outcome of one of such fundamental research efforts. If anyone intends to use the outcomes of this research for commercial goals, TCS Ltd shall hold the rights for such commercial relationships with respect to iVikodak.

7. Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8. Acknowledgments

Authors would like to thank colleagues in Bio-Sciences Research group (of TCS Research) for their help in testing the ReFDash platform.

9. Abbreviations

ReFDash, Repository of Functional Dashboards; RDP, Ribosomal Data Project; NCBI, National Center for Biotechnology Information; EBI, European Bioinformatics Institute; DOI, Digital Object Identifier; V-Region, 16S rRNA gene hyper-variable region; QIIME, Quantitative Insights Into Microbial Ecology; No-SQL, not only SQL; PHP, Hypertext Pre-processor

10. References

1. Alekseyenko, A. V., Perez-Perez, G. I., De Souza, A., Strober, B., Gao, Z., Bihan, M., et al. (2013). Community differentiation of the cutaneous microbiota in psoriasis. Microbiome 1:31. doi: 10.1186/2049-2618-1-31
2. Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31, 2882–2884. doi: 10.1093/bioinformatics/btv287
3. Bose, T., Haque, M. M., Reddy, C., and Mande, S. S. (2015). COGNIZER: a framework for functional annotation of metagenomic datasets. PLoS ONE 10:e0142102. doi: 10.1371/journal.pone.0142102
4. Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 data-driven documents. IEEE Trans. Visual. Comput. Graph. 17, 2301–2309. doi: 10.1109/TVCG.2011.185
5. Botero, L. E., Delgado-Serrano, L., Cepeda, M. L., Bustos, J. R., Anzola, J. M., Del Portillo, P., et al. (2014). Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis. Microbiome 2:29. doi: 10.1186/2049-2618-2-29

6. Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42, D633–D642. doi: 10.1093/nar/gkt1244

7. Cui, Z., Zhou, Y., Li, H., Zhang, Y., Zhang, S., Tang, S., et al. (2012). Complex sputum microbial composition in patients with pulmonary tuberculosis. BMC Microbiol. 12:276. doi: 10.1186/1471-2180-12-276

8. Duvallet, C., Gibbons, S. B., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat. Commun. 8, 1784. doi: 10.1038/s41467-017-01973-8

9. Ganju, P., Nagpal, S., Mohammed, M. H., Kumar, P. N., Pandey, R., Natarajan, V. T., et al. (2016). Microbial community profiling shows dysbiosis in the lesional skin of Vitiligo subjects. Sci. Rep. 6:18761. doi: 10.1038/srep18761

10. Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. Science 312, 1355–1359.

11. Griffen, A. L., Beall, C. J., Campbell, J. H., Firestone, N. D., Kumar, P. S., Yang, Z. K., et al. (2012). Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. ISME J. 6, 1176–1185. doi: 10.1038/ismej.2011.191

12. Haque, M. M., Bose, T., Dutta, A., Reddy, C. V., and Mande, S. S. (2015). CS-SCORE: Rapid identification and removal of human genome contaminants from metagenomic datasets. Genomics 106(2):116-21. doi: 10.1016/j.ygeno.2015.04.005

13. Haque, M. M., Merchant, M., Kumar, P. N., Dutta, A., and Mande, S. S. (2017). First-trimester vaginal microbiome diversity: A potential indicator of preterm delivery risk. Sci. Rep. 7:16145. doi: 10.1038/s41598-017-16352-y

14. Huse, S. M., Ye, Y., Zhou, Y., and Fodor, A. A. (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS ONE 7:e34242. doi: 10.1371/journal.pone.0034242

15. Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol. 1399, 207–233. doi: 10.1007/978-1-4939-3369-3_13

16. Kirst, M. E., Li, E. C., Alfant, B., Chi, Y.-Y., Walker, C., Magnusson, I., et al. (2015). Dysbiosis and alterations in predicted functions of the subgingival microbiome in chronic

periodontitis. Appl. Environ. Microbiol. 81, 783–793. doi: 10.1128/AEM.02712-14

17. Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., and Knight, R. (2011). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr. Protoc. Bioinform. Chapter, Unit10.7. doi: 10.1002/0471250953.bi1007s36

18. Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31, 814–821. doi: 10.1038/nbt.2676

19. Leavitt, N. (2010). Will NoSQL databases live up to their promise? Computer. 43(2):12–14. doi: 10.1109/MC.2010.58

20. McNally, C. P., Eng, A., Noecker, C., Gagne-Maynard, W. C., and Borenstein, E. (2018). BURRITO: an interactive multi-omic tool for visualizing taxa-function relationships in microbiome data. Front. Microbiol. 9:365. doi: 10.3389/fmicb.2018.00365

21. Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., et al. (2018). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res. 46, D726-D735. doi: 10.1093/nar/gkx967

22. Nagpal, S., Haque, M. M., and Mande, S. S. (2016). Vikodak - a modular framework for inferring functional potential of microbial communities from 16S metagenomic datasets. PLoS ONE 11:e0148347. doi: 10.1371/journal.pone.0148347

23. Nagpal, S., Haque, M. M., Singh, R., and Mande, S. S. (2019). iVikodak—A Platform and Standard Workflow for Inferring, Analyzing, Comparing, and Visualizing the Functional Potential of Microbial Communities. Front. Microbiol. 9:3336 doi: 10.3389/fmicb.2018.03336

24. Narayanasamy, S., Jarosz, Y., Muller, E. E. L. A., Heintz-Buschart, Herold, M., Kaysen, A., et al. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol. 17:260. doi: 10.1186/s13059-016-1116-8

25. Plotly Technologies Inc. (2015). Collaborative Data Science. Plotly Technologies Inc. Available online at: https://plot.ly

26. Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188–7196. doi: 10.1093/nar/gkm864

27. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35, 833–844. doi: 10.1038/nbt.3935

28. Reddy, R. M., Mohammed, M. H., and Mande, S. S. (2014). MetaCAA: a clustering-aided methodology for efficient assembly of metagenomic datasets. Genomics 103, 161–168. doi: 10.1016/j.ygeno.2014.02.007

29. Rossen, J. W. A., Friedrich, A. W., and Moran-Gilad, J. (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clin. Microbiol. Infect. 24, 355–360. doi: 10.1016/j.cmi.2017.11.001

30. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. doi: 10.1093/bioinformatics/btr026

31. Tandon, D., Haque, M. M., Gote, Manoj., Jain, Manish., Bhaduri, Anirban., Dubey, A. K. et al. (2019). A prospective randomized, double-blind, placebo-controlled, dose-response relationship study to investigate efficacy of fructo-oligosaccharides (FOS) on human gut microflora. Sci. Rep. 9:5473. doi: doi.org/10.1038/s41598-019-41837-3

32. Tandon, D., Haque, M. M., Saravanan, R., Shaikh, S., Sriram, P., Dubey. A. K., et al. (2018). A snapshot of gut microbiota of an adult urban population from Western region of India. PLoS ONE 13:e0195643. doi: 10.1371/journal.pone.0195643

33. White, R. A., Brown, J., Colby, S., Overall, C. C., Lee, J.-Y., Zukcer, J., et al. (2017). ATLAS (Automatic Tool for Local Assembly Structures) - a comprehensive infrastructure for assembly, annotation, and genomic binning of metagenomic and metatranscriptomic data. PeerJ 5:e2843ve21. doi: 10.7287/peerj.preprints.2843v1

34. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. Nature 486, 222–227. doi