


SOFTWARE

Open Access

The MG-RAST API explorer: an on-ramp for RESTful query composition



Tobias Paczian^{1,2}, William L. Trimble^{1,2}, Wolfgang Gerlach^{1,2}, Travis Harrison^{1,2}, Andreas Wilke^{1,2} and Folker Meyer^{1,2*} 

Abstract

Background: The MG-RAST API provides search capabilities and delivers organism and function data as well as raw or annotated sequence data via the web interface and its RESTful API. For casual users, however, RESTful APIs are hard to learn and work with.

Results: We created the graphical MG-RAST API explorer to help researchers more easily build and export API queries; understand the data abstractions and indices available in MG-RAST; and use the results presented in-browser for exploration, development, and debugging.

Conclusions: The API explorer lowers the barrier to entry for occasional or first-time MG-RAST API users.

Background

Environmental DNA sequence analysis (i.e., metagenomics) is gaining popularity and is frequently used by researchers who are not specialists in genomics or bioinformatics [1]. With expanding reference databases and increasing volumes of raw data, the computational component of environmental sequence analysis is substantial [2]. In metagenomic investigations, the sequence data itself is frequently in the hundreds of gigabytes range, and the datasets often reach the terabyte scale. Searching even a few metagenomes—for example, for all sequences exhibiting similarity with a set of organisms—requires either costly recomputation or storage of index data structures. Researchers handling dozens or hundreds of complex datasets frequently find themselves overwhelmed by the computational requirements of the task at hand.

Although significant advances have been made (see, e.g., [3] or [4]) that help save on computational cost, a considerable amount of compute, storage, and I/O bandwidth is required to analyze metagenomic data. Projects that must analyze many datasets across thousands of

directories and terabytes of data often reach the limit of on-premise or temporary rented public cloud resources. Consequently, remote computation, data management, and indexing of large-scale metagenomic data are a growing community need and will be key features of any future bioinformatics landscape.

Hosted services such as MG-RAST [5], the EBI Metagenomics Portal [5], and the U.S. Department of Energy's JGI IMG/M [6] provide web interfaces to access the data, computational results, and search results. However, these interfaces often are limited to predefined queries, even if the underlying data and indexing support additional query capabilities. Exposing the internal data and indices via an application programming interface (API) helps overcome this limitation, enabling end users potentially to delve more deeply into the data.

In addition, studying larger quantities of datasets is often done best via custom scripts or command line tools. APIs play an essential role by allowing automation. We also note that APIs render the practice of extracting data from web pages (“screen scraping”) obsolete.

* Correspondence: folker@anl.gov

¹Argonne National Laboratory, Lemont, IL, USA

²University of Chicago, Chicago, IL, USA



```

$ curl 'http://api.mg-rast.anl.gov/annotation/sequence/mgm4447943.3?evalue=10&type=organism&source=SwissProt'
sequence id      m5nr id (md5sum)      dna sequence
                semicolon separated list of annotations
mgm4447943.3|GF8803K01B6N5Q|SwissProt  00037faa7c2c1
b088014211234010d0b      CGCTGCTCTATGCTGGCTATAGTCCTTGC
TATCGTAAGGAAGCGGGTGCAGCTGGTAAGTTTAGCCGCGGACTGTTCCGCGT
GCACCAGTTCAATAAGCTGGAAATGTATATTTTCTGTACGCCAGAACAATCAG
CTCAATGCACGAAAAGATCTTGTGATCGAGGAAGAAA [Dictyoglomus
thermophilum (strain ATCC 35947 / DSM 3960 / H-6-12)]

mgm4447943.3|GF8803K01BGX56|SwissProt  00037faa7c2c1
b088014211234010d0b      GACTGGCTTCATGCTGACGAAATCATCGA
CGAGAAAAATCTGCCGCTGCTCTATGCTGGCTATAGCCCTTGTACCCTAAAG
AAGCGGGTGCAGGCTGGCAAGTTTAGCCGCGGACTGTTCCGCGTGCACCAGTTC
AATAAGCTGGAAATGTATATCTTTTGTACACCAGAACAATCAGTCCAGATGCA
CGAAAAAATCTTGTGATCGAGGAAGAAATTTGGCAAACTTAGGCGTGCCGT
ACCATGTGGTGAATATCGCGGCTGGCGACCTCGGTGCGCCAGCAGCAAAAAA
TATGATATCGAGTACTGGTCGCCAGTCGATAACCAGTATCGCGAGTTGACTAG
CTGCTCGAATTGACTGATTTTCAGGCGCGTAATCTCAATATTCGCGTGCGCC
GCAGCGACGGTTCGCTGCAGGACTGCACACGCTTAACGGTACAGGCGGTGAG
CTTGCGCGCTCGCTAGTAGC [Dictyoglomus thermophilum (s
train ATCC 35947 / DSM 3960 / H-6-12)]
mgm4447943.3|GF8803K01C5C07|SwissProt  00124e1437444
cf7c4d4824c1c07d9dd      GCGCATTGAAGACTCCCTTGTGTAACAG
CAGAGGGCTGTGAAATATTGACGACAACATCGAAAGAATTAACGGAGTTGTTT
TAGGAGGAGATTGTGTATGATTACAAGTAATGATTTTCAGACCGGGCGTTACCA
TTGAAATTGACGGTCAGGTTTGGCAGGTTGTGGAATTTTCAGCATGTAAAGCCC
GGCAAAAGGGGCCGCTTTTGTACGTGCTAAAAATTAACCTCGAAACCGGT
GCTGTTGTAGAACGCACATGGAATGCAGGCCAAAAAAGTACAGGAAGGTCGTG
...

```

Fig. 1 URL and the results for the retrieval of SwissProt taxonomy annotations with a cut-off 10^{-10} for dataset mgm4447943.3

In recent years RESTful APIs [7] have become the state of the art, allowing data to be managed and distributed over the internet. These APIs now provide the backbone of commerce and entertainment on the internet, and they have become a necessary tool for the handling of sequencing data and data products.

For example, MG-RAST is designed around a RESTful API [8]. This provides a search engine for datasets, delivers tables of taxa and functional annotations from sequence datasets, allows upload and download of data files, and can deliver sequences with attached annotations. Structured data is delivered in the JSON format; JSON data can be easily converted to tab-delimited tables or other formats used in bioinformatics if needed. While JSON is not particularly user friendly it has emerged as the standard and a myriad of tools exist to assist.

Unfortunately, although RESTful APIs offer more flexibility to access computation and data, they are notoriously difficult to learn [9]. Two factors contribute

to this steep learning curve: the syntax and the abstractions. Syntax refers to the fact that APIs are intended to be consumed primarily by computer programs, not human beings, and require strict adherence to standards and conventions. Abstractions refer to the fact that APIs require some level of understanding of the abstractions used for data storage and indexing. However, when a query against a readily indexed database of metagenomic data can save tens of thousands of dollars of computational (and manpower) costs for reanalyzing metagenomic datasets, learning to work with RESTful APIs becomes an attractive value proposition.

Implementation

The API explorer is implemented as light-weight JavaScript and HTML overlay on top of the MG-RAST API [8] and the MG-RAST infrastructure [10]. The API itself

```

$ curl -X GET "https://api.mg-rast.org/annotation/sequence/mgm4447943.3?source=SEED&type-function&filter=Immunoreactive"
sequence id      m5nr id (md5sum)      dna sequence
                semicolon separated list of annotations
mgm4447943.3|GF8803K01A392B|SEED      568871188d306
0fe85ddbca8035123b1      CTGCGCACCACCTTCCCTCCGCGACGAG
ATGCTGCGCTCCGAACCTATGTTACGTTTATGATGCCACCGTCGGGAAGAAGAA
GGTGAACATCCCCGAAGTGAAAGTTGCCGACGGCGTACTGGCTACCAGCACAC
TGCTCTGTCGACCGCCGCTTCCGCCAATGCCGCTCTGGGAGAGGACAAGTTC
CAGTACACCATCGAGCAAAAGCAGGAAGCACAATCAAGTACCTCATCCAGCA
AGCTCGCATCCGACCCAGCGAATTGCAATCGGCTTCTGTGAAAGACTTTGTGC
AGACTCTCCGAACATCAAAGCCGACGGCAAACGCTTAGAGTTGGGAAATGTG
GAGGTGAGTTCTACGCATCGCCCGACGGCGGTGAGAACTCAACACCAAGTT
GGCACAGAGTCGCGAACGCTTCAGGTAGCTATGTGAAGAGCCAGCTGAAGA
AAATAACCTCGAAGGTGAGGTCAATCCCGTTATACCGCG      [Immu
noreactive 53 kDa antigen PG123]
mgm4447943.3|GF8803K01C4434|SEED      568871188d306
0fe85ddbca8035123b1      GAGGATTTGCGCAAGGCGGCCGAGTTGG
GATCTGACGCCCTTCAAATCGGCTGTTGCCTGCGCGTAATTGCCGCGTGAAT
ATTAAGATTGCCGGCCACGGCGTTATAATTTTTAGCTGCGGCAGCGCTTCTTA
AATAATTTTCAGCGCCTTCGATGTCGCCTTTAGCCAAGGAAAGGAGGGCGTAG
TTGGCGTTTACTTCTGCCGATTGCCTTTTCGCATCGCTGCTTGTGCGAGCCA
ATCTTCGGCAGCATTGAAGTCGGCGCGCTCATAAGCGGCAATGGCCAAATTAT
TGTAGGCTCTGTAGTCGTTAGGGTAGAGGCGGTTGCCGTGCGGTAGATGTTT
TCCTGCTCCGCTTTGACTCTGTCAGCGTAGCATAGTAAAGGAGCTTTCGAT
TGAAAGTTTCGAGGGGTCGGCAGCGTATTGCGCTTGTATTTATCGTCGCTGC
GGCCGATTAGCAGGTAATTTATCGTCAGGCGAGCGGCGGTAGTTCCGG
      [Immunoreactive 53 kDa antigen PG123]

...
Download complete. 14 rows retrieved

```

Fig. 2 URL and the results for the retrieval of SwissProt taxonomy annotations with a cut-off 10–10 for dataset mgm4447943.3

is supported by a complex mixture of databases and object stores.

Like all of MG-RAST, the API explorer is available as open source software on GitHub at <https://github.com/MG-RAST> under a BSD-style license.

Results

To assist researchers in learning the syntax of the MG-RAST RESTful API, the MG-RAST team has developed an API explorer (<https://explorer.mg-rast.org/>). It has one page for search and one for all other API functionality. The explorer allows querying capabilities and displays the results of an API query in-browser. Additionally, it constructs working command line invocations that can be copied and pasted.

The MG-RAST API explorer provides a gentle introduction to the API through a number of simple example queries.

Example 1: Annotated sequence retrieval.

Downloading annotated sequences requires specifying a dataset, a database (one of the databases included in

the union M5nr database [8]), and cut-off thresholds for the similarity table. Indexes for both taxonomy (Fig. 1) and functions (Fig. 2) allow retrieval of annotated sequences with organism name or function labels attached.

Example 2. Query composition. The next example involves a simple query requesting a list of aquatic datasets collected in Chicago. Figure 3 shows the URL and command line representation generated by the API explorer.

Example 3: Dataset properties. Searches for datasets can also include technical properties such as dataset size or type. This example retrieves all shotgun metagenomic datasets using Illumina technology larger than 1 gigabase pair. The resulting URL and command line are given in Fig. 4.

In addition to storing metadata about the datasets, MG-RAST stores the results of the sequence analysis, indexed to allow querying by organism name (such as *Corynebacterium glutamicum*), taxonomic names from the NCBI taxonomy, or protein function

Fig. 3 Using the API explorer to construct a query for aquatic datasets from Chicago. This query was built by selecting “location” from a drop-down menu, entering the value “Chicago,” and selecting “biome” from the same menu and entering the term “aquatic.” All the valid API options are presented as editable fields; the API’s response to the query is shown in the gray box at the bottom of the page labeled “result from API.”

```
# Request from API explorer on one line:
$ curl -g 'https://api.mg-rast.org/search?limit=5&order=created_on&direction=asc&public=yes&seq_meth=illumina&bp_count_raw=[100000000%20TO%20*]&sequence_type=WGS&index=metagenome_index_20180705'
```

```
# Request expanded for ease of reading:
$ curl -g -F 'limit=5' \
  -F 'order=created_on' \
  -F 'direction=asc' \
  -F 'public=yes' \
  -F 'seq_meth=illumina' \
  -F 'bp_count_raw=[100000000 TO *]' \
  -F 'sequence_type=WGS' \
  -F 'index=metagenome_index_20180705' \
  'https://api.mg-rast.org/search'
```

Fig. 4 URL and command line syntax for retrieving a list of Illumina shotgun metagenomes larger than 1 gigabase pair. This query returns a list of matching datasets in JSON format

```

$ curl -X GET "https://api.mg-rast.org/search?limit=5&order=created_on&direction=asc&public=yes&sequence_type=WGS
&index=metagenome_index_20180705&taxonomy=Archaea&taxa_per=25&taxa_level=domain"
{
  "id": "d74ecc29-9256-4ef2-af02-f349a3c4cd9d",
  "status": "submitted",
  "url": "http://api.mg-rast.org/status/d74ecc29-9256-4ef2-af02-f349a3c4cd9d"
}

$ curl 'http://api.mg-rast.org/status/b0a4b919-2c35-42fe-8e22-732794deec7'
{
  "url": "http://api.mg-rast.org/status/d74ecc29-9256-4ef2-af02-f349a3c4cd9d",
  "status": "processing",
  "started": "2019-07-29T12:46:24.363-05:00"
  "parameters": {
    "limit": 5,
    "order": "created_on",
    ...
  }
}

$ # Wait for the server to finish ...
$ curl 'http://api.mg-rast.org/status/d74ecc29-9256-4ef2-af02-f349a3c4cd9d'
{
  "url": "http://api.mg-rast.org/status/d74ecc29-9256-4ef2-af02-f349a3c4cd9d",
  "status": "done",
  "completed": "2019-07-29T12:50:43.553-05:00"
  "started": "2019-07-29T12:46:24.363-05:00"
  "data": {
    "matrix_type": "dense",
    "matrix_element_value": "abundance",
    "data": [
    ...
  ]
}

```

Fig. 5 Searching for datasets with more than 25% Archaea. This complex search request will not return results immediately. In order to avoid client-side timeouts, the queries are turned into asynchronous search requests. The output of the curl command will indicate this and provide a URL for status checking on the complex queries and eventual download. We note that the URL contains a UUID to act as a temporary identifier

labels from the included protein function databases. See [9] for details of the computational pipeline.

Example 4: Dataset content. The MG-RAST RESTful API explorer can also be used to find datasets based on dataset content rather than metadata. In this example we retrieve a list of datasets with a substantial fraction of sequences annotated as Archaea and later aggregate the abundance information at the family level.

The query in Fig. 5 returns datasets with more than 25% of the sequences showing protein similarities to Archaea. We note that the API returns structured data in JSON format and abundance tables in BIOM format [10]. While JSON is not particularly readable, numerous tools are available for converting JSON into other formats, for example, CSV for use in spreadsheets.

The matrix function of the API explorer allows for merging information on taxonomy and function (see Fig. 6). Moreover, the API is not limited to abundance and taxonomy tables; sequences labeled with unstructured, free-text functional annotations from included databases can also be listed or extracted (see Fig. 2).

Example 5: Sequence retrieval. For our next example, the API retrieves sequences, decorated with annotations, that match specified organisms or

functions: the sequences returned by this URL all have similarity to “Immunoreactive proteins” in the SEED database.

Example 6: Metaanalysis-extracting GPS coordinates. With thousands of data sets available in MG-RAST, metaanalyses are becoming more popular. The API supports data extraction and analysis that can be used to explore the coverage of the planet with metagenomic samples.

Conclusion

All the working examples shown here used public data and did not require authentication. However, private datasets can be securely accessed by adding “&auth = MGRKEY” to URLs used in any browser or -H “Authorization: mgrast MGRKEY” to the command line for curl, where MGRKEY is replaced by the text of the MG-RAST authentication key, a password-like string that is available from the user’s upload page.

We also provide a set of client-side Python scripts to assist with standard use cases, allowing, for instance, data upload and download without using the browser and automatically handling authentication and waiting for long-running queries; see the website. <https://github.com/MG-RAST/MG-RAST-Tools>.

The MG-RAST team hopes that the drop-down menus and in-browser troubleshooting environment provided by the MG-RAST RESTful API explorer will help researchers make better use of the metagenomic

```

$ mg-query.py "http://api.mg-rast.org/matrix/organism?value=15&filter_level=function&filter_source=Subsystems&group_level=Family&hide_metadata=0&hit_type=all&id=mgm4447943.3&id=mgm4447192.3&id=mgm4447102.3&result_type=abundance&source=RefSeq" | jq . > Fig6.out
http://api.mg-rast.org/matrix/organism?value=15&filter_level=function&filter_source=Subsystems&group_level=family&hide_metadata=0&hit_type=all&id=mgm4447943.3&id=mgm4447192.3&id=mgm4447102.3&result_type=abundance&source=RefSeq
Making request http://api.mg-rast.org/matrix/organism?value=15&filter_level=function&filter_source=Subsystems&group_level=family&hide_metadata=0&hit_type=all&id=mgm4447943.3&id=mgm4447192.3&id=mgm4447102.3&result_type=abundance&source=RefSeq
Making request http://api.mg-rast.org/status/f56a0e04-9dde-4b73-ab71-57c93f590e71
{
  "data_source": "RefSeq",
  "format_url": "http://biom-format.org",
  "type": "Taxon table",
  "matrix_element_type": "int",
  "id": "mgm4447102.3_mgm4447192.3_mgm4447943.3_organism_family_RefSeq_all_abundance_15_60_15",
  "format": "Biological Observation Matrix 1.0",
  "generated_by": "MG-RAST",
  "url": "http://api.mg-rast.org/matrix/organism?id=mgm4447102.3&id=mgm4447192.3&id=mgm4447943.3&group_level=family&source=RefSeq&hit_type=all&result_type=abundance&value=15&identity=60&length=15",
  "matrix_type": "dense",
  "date": "2019-07-26T20:06:09.184205",
  "shape": [
    355,
    3
  ],
  "source_type": "protein",
  "matrix_element_value": "abundance",
  "rows": [
    {
      "metadata": {
        "hierarchy": {
          "phylum": "Proteobacteria",
          "domain": "Bacteria",
          "order": "Cardiobacteriales",
          "family": "Cardiobacteriaceae",
          "class": "Gammaproteobacteria"
        }
      },
      "id": "Cardiobacteriaceae"
    },
    ...
  ]
}
"data": [
  [
    1122,
    685,
    59
  ]
],

```

Fig. 6 The URL constructed via the API explorer generates an abundance count of SEED subsystem terms for 3 datasets, summarizing abundance at the family level using RefSeq generated taxonomic annotations. The output is again in JSON format; we show the top of the return file using jq to color code it for readability

data and computation already completed and curated at MG-RAST (Fig. 7). The MG-RAST system is open source and is available on github (<https://github.com/MG-RAST>).

Availability and requirements

The API explorer is available as part of MG-RAST.

Project name: MG-RAST

Project home page: <https://github.com/MG-RAST>

Operating system(s): Linux

Programming language: Perl, Python, Go-Lang, HTML5

Other requirements: ElasticSearch, Cassandra, SOLR, MongoDB, SHOCK, AWE

License: BSD type license

Any restrictions to use by non-academics: none.

Abbreviations

API: Application programmer's interface; BIOM: Biological Observation Matrix; I/O: Input and output and movement of computer data; JSON: Javascript object notation; NCB: National Center for Biotechnology Information; URL: Uniform resource locator

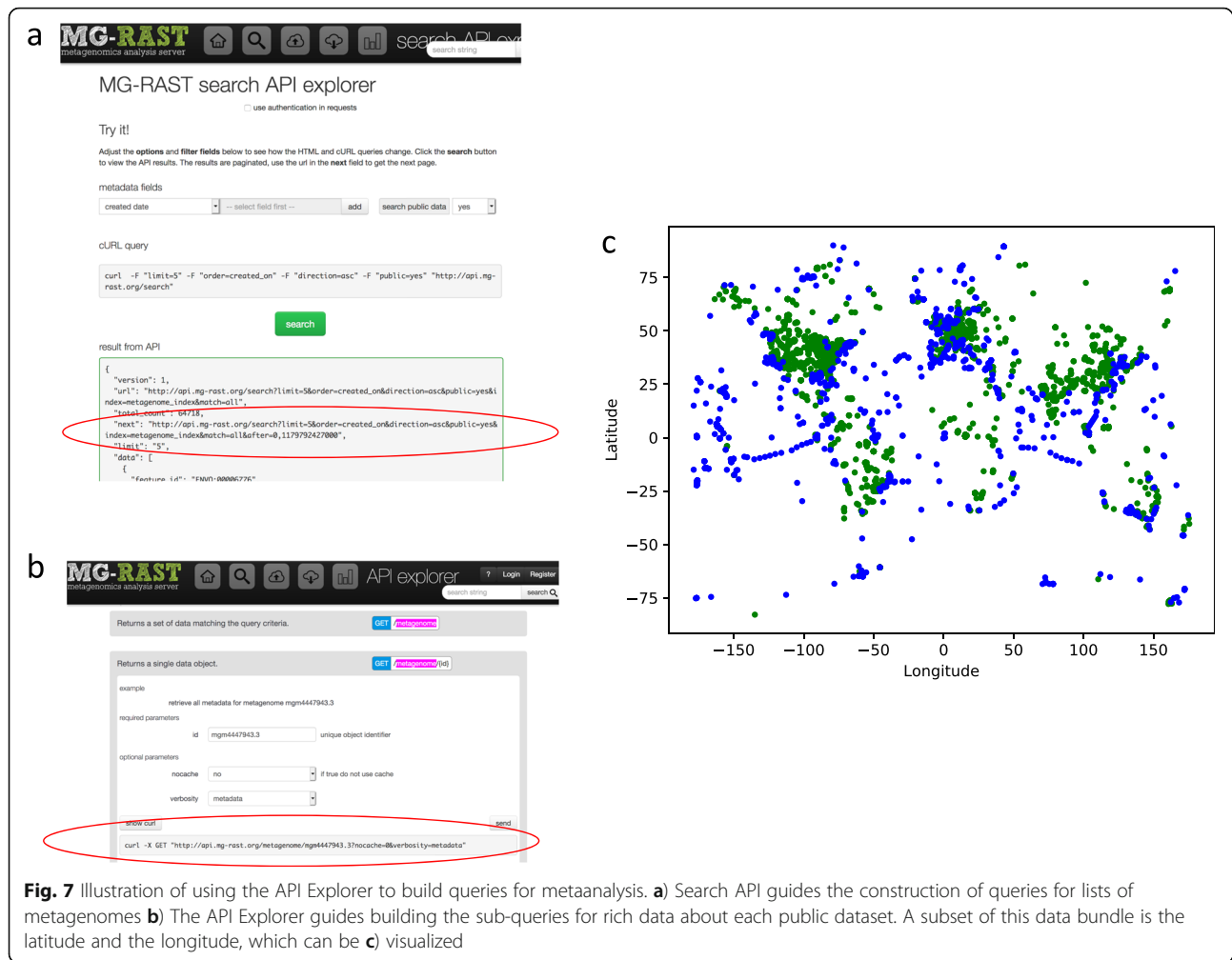


Fig. 7 Illustration of using the API Explorer to build queries for metaanalysis. **a**) Search API guides the construction of queries for lists of metagenomes **b**) The API Explorer guides building the sub-queries for rich data about each public dataset. A subset of this data bundle is the latitude and the longitude, which can be **c**) visualized

Acknowledgments

The authors gratefully acknowledge the help from Gail Pieper with editing this manuscript.

Authors' contributions

WT and FM designed the API explorer. TP, WT, WG, TH, AW and FM contributed equally to the implementation. FM wrote the manuscript. All authors read and approved the final manuscript.

Funding

The work reported in this article was supported in part by a grant from the National Institutes of Health (NIH) grant 1R01AI123037-01. Work on this article was also supported by NSF award 1645609. This material was based upon research supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under contract DE-AC02-06CH11357. The funders had no role in the design or execution of the work.

Availability of data and materials

The source code is available on github under a BSD license. All data used in the examples is publicly available on MG-RAST.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests. Folker Meyer is currently acting as a member of the Editorial Board for BMC Bioinformatics.

Received: 11 December 2018 Accepted: 11 July 2019

Published online: 08 November 2019

References

1. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. *Microbial Inform Exp*. 2012;2(1):3.
2. Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, Arze C, White JR, White O, Fricke WF. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* [electronic resource]. 2011;12:356.
3. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59-60.
4. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35(11):1026-8.
5. Mitchell A, Bucchini F, Cochrane G, Denise H, ten Hoopen P, Fraser M, Pesseat S, Potter S, Scheremetjew M, Sterk P, et al. EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res*. 2016;44(D1):D595-603.
6. Chen IA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Ratner A, Huang J, Andersen E, Huntemann M, et al. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res*. 2017; 45(D1):D507-16.

7. Fielding RT: Architectural styles and the Design of Network-based Software Architectures. University of California, Irvine.; 2000.
8. Meyer F, Bagchi S, Chaterji S, Gerlach W, Grama A, Harrison T, Paczian T, Trimble WL, Wilke A. MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Brief Bioinform*. 2017.
9. Wilke A, Harrison T, Wilkening J, Field D, Glass EM, Kyrpides N, Mavrommatis K, Meyer F: The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics [electronic resource]* 2012, 13:141.
10. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, et al. The biological observation matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience*. 2012;1(1):7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

