Check for updates

SOFTWARE TOOL ARTICLE

DarkQ: continuous genomic monitoring using message queues [version 1; peer review: 2 approved with reservations]

Adrian Viehweger¹, Christian Brandt², Martin Hölzer³

¹Institute of Medical Microbiology and Virology, University Hospital Leipzig, Leipzig, Germany

²Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

³Methodology and Research Infrastructure, MF1 Bioinformatics, Robert Koch Institute, Berlin, Germany

V1 First published: 01 Oct 2021, 10:998
<https://doi.org/10.12688/f1000research.54255.1>

Latest published: 01 Oct 2021, 10:998
<https://doi.org/10.12688/f1000research.54255.1>

Abstract

Newly sequenced genomes are often not noticed by potential stakeholders because submission to public databases is delayed, and search options are limited. However, the discovery of genomes can be vital: in pathogen outbreaks, fast updates are essential to coordinate containment efforts and prevent further spread. Here we introduce DarkQ, a message queue that allows for instant sharing and discovery of genomes.

DarkQ is released under the BSD-2 license at
github.com/phiweger/darkq.

Keywords

outbreak, molecular surveillance, peer-to-peer, pathogen

Open Peer Review

Approval Status ? ?

	1	2
version 1	?	?
01 Oct 2021	view	view

1. **Tessa Pierce-Ward** , University of California, Davis, USA

2. **Rayan Chikhi** , Institut Pasteur, Paris, France

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Adrian Viehweger (adrian.viehweger@googlemail.com)

Author roles: **Viehweger A:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Brandt C:** Conceptualization, Resources, Supervision, Validation, Writing – Review & Editing; **Hölzer M:** Conceptualization, Methodology, Resources, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Viehweger A *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Viehweger A, Brandt C and Hölzer M. **DarkQ: continuous genomic monitoring using message queues [version 1; peer review: 2 approved with reservations]** F1000Research 2021, 10:998 <https://doi.org/10.12688/f1000research.54255.1>

First published: 01 Oct 2021, 10:998 <https://doi.org/10.12688/f1000research.54255.1>

Introduction

Many bioinformatic tasks complement newly sequenced genomes with existing, publicly available ones. For example, when reconstructing a local pathogen outbreak, screening similar genomes can discover related ones from other sampling sites, such as hospitals nearby, and can significantly affect public health responses.^{1,2} However, no tool exists to monitor newly sequenced genomes and automatically identify those of interest to the user. A long delay until they are publicly available explains why many outbreak studies are retrospective and offer limited practical value to the associated outbreak response. Several components are needed: first, a “publisher” needs to be able to send genomic “messages” to a “consumer” using a simple and secure interface. Second, the genomic “message” may require limited space to avoid upload problems or extensive storage infrastructure. Third, a mechanism is needed to route messages only to interested parties, e.g., consumers that search for genomes of a particular species in a specific geography. Lastly, on receiving a relevant message, download of the associated genome should be possible. Several projects currently develop ways to share genomes effectively ([wort](#), [stark](#)). However, to our knowledge, ours is the first end-to-end solution available to users.

Methods

Implementation

DarkQ is implemented using the Nextflow workflow manager to ensure a robust, reproducible, and portable application.³ The user interface of DarkQ is similar to the popular file system service “Dropbox”; the content of a “send” directory is tracked. When a genome is added to it, it is first compressed (“sketched”) using the MinHash algorithm^{4,5} (sourmash, v3.5). The reduction in file size by orders of magnitude allows for efficient transmission. Together with metadata and inferred taxonomy (using sourmash), the genome sketch constitutes a “message” (Figure 1A). The receiving message queue then uses the Advanced Message Queuing Protocol (AMQP)⁶ to route messages (implementation: [RabbitMQ](#), v3.8.9) onto queues, i.e. sequential groups of messages. The original genome is uploaded (“pinned”) to a decentralized, peer-to-peer network (IPFS, v0.7).⁷ Its content-based address is part of the genome message.

The consumer can subscribe to messages using an arbitrary number of filters, so-called “routing keys”. Each routing key is unique and has five properties: name of sender (e.g. “phiweger”), country code (e.g. “DE”), taxon status (“found” or “mystery”), taxon level (either one of superkingdom, phylum, class, order, family, genus, species, and strain) and taxon name at that level (e.g. “Klebsiella” for genus) – these are adapted from and must conform to the Genome Taxonomy

Figure 1. (A) Architecture of DarkQ: compressed genome “messages” (colored arrows) are sent by a publisher (P) onto the message queue. A router (circle) distributes the messages to queues via routing keys (annotated arrows). Consumers (C) can use these keys to receive only a subset of messages and then further filter them with target genomes using MinHash sketches. In parallel, the genomes from the publisher are uploaded on a decentralized peer-to-peer (P2P) network. Once messages pass through the consumer’s filters, they are automatically downloaded from the P2P network. This architecture allows the effective distribution of newly sequenced genomes and enables continuous monitoring, e.g., in outbreak scenarios. (B) Use case simulation: a hospital becomes aware of a local outbreak of an XDR *Klebsiella pneumoniae* (Kp) isolate of subtype (ST, right metadata column) 258 carrying a plasmid-encoded KPC-2 carbapenemase. Using DarkQ, we identified 431 genomes from several countries (leaf colors) from 26 studies (left metadata column) with an average nucleotide identity (ANI) > 99.98% and identical resistance and capsule patterns (not shown). A time-dated phylogeny revealed several non-local isolates, suggesting that the outbreak reached further than previously assumed. An interactive version of the data can be found at microreact.org/project/facEFbDrgwgp9aX97nvpHq. Scale in number of SNVs.

Database (GTDB, release 89).⁸ For example, “phiweger.DE.*.genus.klebsiella” would select all isolate genomes of the stated genus from Germany sent by the author.

Because we can estimate genome similarity using MinHash sketches,⁴ the consumer can quickly filter the received genome messages using target genomes, for example those belonging to a local pathogen outbreak or current research project. If this filter is passed, then the genome is automatically downloaded from the peer-to-peer network using its content hash address, which at the same time locates and validates the downloaded file. If multiple users pin the genome, download speed can increase substantially. A downstream workflow can then be connected to refine these genomes’ analysis further, enabling a complete monitoring system.

Operation

The software can be run on any UNIX-based operating system. Operation of DarkQ requires less than one Gb RAM and a single core. Details of the workflow can be found in the README file.¹²

Use cases

To test DarkQ in a monitoring system, we collected and sent onto DarkQ 9,415 genomes of *Klebsiella pneumoniae*, a pathogen considered an urgent global threat due to extensive antimicrobial drug resistance (CDC, AR threats report, 2019).⁹ We simulated a consumer subscribing to all messages from the *Klebsiella* genus and filtering the received messages using an isolate from a local outbreak at a large tertiary hospital in 2010.¹⁰ 1,461 messages met both routing key and minimum genome similarity criteria of 0.97 at a k-mer size of 51, typically used to estimate the genomic distance at the strain level.¹¹ After downloading the original genomes from the peer-to-peer network, they were further filtered and refined, resulting in a time-dated phylogeny (Figure 1B). The consumer thus received genomes from a total of 26 studies. Two of these studies contained genomes that belonged to the same outbreak clone the consumer used to filter the genomes. Further work is needed to investigate this relationship more thoroughly; however, an initial assessment was already possible by utilizing the mechanics implemented in DarkQ. All methods used in this use case are available elsewhere.¹²

Conclusion

DarkQ allows a user to monitor genomic data with a simple user interface, efficient genome compression, filter-based message routing, and fast download of corresponding genomes using a decentralized peer-to-peer network. The proof-of-concept outlined here scales to thousands of genomes and could be particularly valuable in the context of pathogen outbreaks. However, our approach can also be used to disseminate research more broadly.

Data availability

Underlying data

NCBI BioProject: Context-aware genomic surveillance reveals hidden transmission of a carbapenemase-producing *Klebsiella pneumoniae*. Accession number PRJNA742413. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA742413>.

Software availability

Source code available from: github.com/phiweger/darkq.

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.5503447>.¹³

License: BSD-2 license.

Acknowledgements

We thank Luiz Irber and C. Titus Brown (University of California, Davis) for insightful discussions of the concepts discussed in this article. An earlier version of this article can be found on bioRxiv (<https://doi.org/10.1101/2020.11.12.379560>).

References

1. Grubaugh ND, *et al.*: **Tracking virus outbreaks in the twenty-first century.** *Nat. Microbiol.* 2019; **4**: 10–19.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Armstrong GL, *et al.*: **Pathogen genomics in public health.** *N. Engl. J. Med.* 2019; **381**: 2569–2580.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Di Tommaso P, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat. Biotechnol.* 2017; **35**: 316–319.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Ondov BD, *et al.*: **Mash: Fast genome and metagenome distance estimation using MinHash.** *Genome Biol.* 2016; **17**: 132.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Pierce NT, Irber L, Reiter T, *et al.*: **Large-scale sequence comparisons with sourmash.** *F1000Res.* 2019; **8**: 1006.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. O'Hara J: **Toward a commodity enterprise middleware: Can AMQP enable a new era in messaging middleware? A look inside standards-based messaging with AMQP.** *Queueing Syst.* 2007; **5**: 48–55.
[Publisher Full Text](#)
7. Benet J: **IPFS - content addressed, versioned, P2P file system.** 2014.
8. Parks DH, *et al.*: **A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.** *Nat. Biotechnol.* 2018; **36**: 996–1004.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Wyres KL, Lam MMC, Holt KE: **Population genomics of klebsiella pneumoniae.** *Nat. Rev. Microbiol.* 2020.
[Publisher Full Text](#)
10. Lippmann N, Lübbert C, Kaiser T, *et al.*: **Clinical epidemiology of klebsiella pneumoniae carbapenemases.** *Lancet Infect. Dis.* 2014; **14**: 271–272.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Koslicki D, Falush D: **MetaPalette: A k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation.** *mSystems.* 2016; **1**: e00020–e00016.
[Publisher Full Text](#)
12. Viehweger A, *et al.*: **Context-aware genomic surveillance reveals hidden transmission of a carbapenemase-producing.** *Klebsiella pneumoniae.*
[Reference Source](#)
13. Viehweger A, Hölzer M: **phiweger/darkq: MVP (v0.1).** *Zenodo.* 2021;
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 29 November 2021

<https://doi.org/10.5256/f1000research.57722.r96160>

© 2021 Chikhi R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Ryan Chikhi**

Department of Computational Biology, C3BI USR 3756 IP CNRS, Institut Pasteur, Paris, France

This article presents DarkQ, a proof of concept software for an architecture to perform microbial surveillance. The concept is interesting and original, and appears to be well-engineered. Monitoring microbes is of high interest scientifically. However, as I mentioned, DarkQ is a proof of concept that does not provide a service which end-users can use, it only provides the blueprint for creating such a service.

Some remarks:

1. Please clarify whether the pipeline is designed for monitoring only bacteria, or also supports viruses.
2. The sentence "Using DarkQ, we identified [...]" throws me off, as it is unclear who are the actors in this analysis. It would be beneficial to develop the hospital use case by clearly telling, in this situation, who is the producer and who is the consumer, and whether there is any other third-party involved; i.e. which actor(s) run a DarkQ instance.
3. The text indicates that DarkQ works just like "Dropbox", however Figure 1 looks nothing like Dropbox. Given the originality of the approach, it could be valuable to show the workflow from a user perspective, in addition to (or instead of) the behind-the-hood architecture.
4. I understand the software is distributed as a proof of concept, but as it stands it cannot be used by its intended users (e.g. hospitals), given that there is no central instance. Essentially, to use DarkQ today, one currently needs to act as both the producer and the consumer. This should be noted somewhere in the article.
5. "Operation of DarkQ requires less than one Gb RAM and a single core." looks unlikely, as I copied genomes from "data/test" to "data/send", darkq crashed with the message ".command.sh: line 2: 8734 Killed sourmash lca classify --db gtdb-release89-k31.lca.json.gz -query signature.json > taxonomy.csv" which is out-of-memory error. My config is WSL2 with 4 GB RAM allocated.

6. Is it possible for the pipeline to _miss_ some genomes? I.e. those that are too distantly related to the target genome, or those for which the sourmash signatures are insufficiently similar to the target. Some discussion on this potential limitation would be helpful.
7. Regarding scalability and the sentence "After downloading the original genomes from the peer-to-peer network", is it intended that all newly sequenced genomes worldwide are to be uploaded to IPFS, permanently or for a short period? If the former, it seems that DarkQ will also act as a genome repository.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, data structures

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 08 November 2021

<https://doi.org/10.5256/f1000research.57722.r96159>

© 2021 Pierce-Ward T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tessa Pierce-Ward

Department of Population Health and Reproduction, University of California, Davis, CA, USA

The authors present “DarkQ,” a peer-to-peer network and workflow for pathogen genome monitoring. The COVID-19 pandemic has underlined the need and utility for quickly sharing pathogen genomes across both hospitals and researchers. The workflow and file sharing protocol described here could fill that need, and enable real-time genome sharing.

Questions and concerns:

- Clarification on (& for) intended users:
 - Intended users: hospitals, researchers, etc.?
 - Expectations for skills needed by intended users? Installation and use currently requires familiarity with command-line, and may require familiarity with conda, Nextflow, and IPFS for troubleshooting.
 - IPFS installation may require sudo access, and is not always straightforward. Is sudo access likely to be available for relevant hospital users? Is there a way to access the data (e.g. drop a single genome and find results) without needing this level of access?
- Please address potential security concerns & comment on **patient privacy**:
 - Does the workflow intended to be continuously run (or run, e.g. daily in an automated fashion?) in order to enable the asynchronous genome download? What security concerns might this present for users?
 - While similarity searches are conducted with the sourmash MinHash sketches, the whole genome is uploaded, stored, and downloaded based on similar queries, right? This may present security concerns.
 - Are there protocols in place to prevent accidental (or purposeful) sharing of private patient data? E.g. are only microbial and viral genomes uploaded? Is human data automatically excluded? What about metadata?
- Taxonomic classification method:
 - I would encourage the authors to explore the sourmash tax function, introduced in sourmash v4.2. A *sourmash gather* -> *sourmash tax* workflow is now recommended over *sourmash lca* methods for taxonomic assignment. Note that the rs202 version of GTDB database is also now available [here](#).
 - What compression (scaling) is used for sketching (1000)? How might scaling affect small genome (e.g. viral) pathogen similarity detection?
- A comparison with existing real-time pathogen tracking (e.g. Nextstrain) could be very helpful. Would sharing protocols from DarkQ be combined with nextstrain analysis and visualization workflows?
- Is there any integration of new published data (not uploaded by message queue)? E.g. what additional databases, if any, are searched, and would those genomes regularly become available for comparison? If not, does the utility of DarkQ depend on large-scale uptake and use?

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: I am one of the developers of sourmash, which is used within the workflow presented here.

Reviewer Expertise: Bioinformatics, MinHash sketching.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research