



METHOD ARTICLE

Organizing and running bioinformatics hackathons within Africa: The H3ABioNet cloud computing experience [version 1; peer review: 3 approved with reservations]

Azza E. Ahmed ^{1,2*}, Phelelani T. Mpangase^{3*}, Sumir Panji⁴, Shakuntala Baichoo ⁵, Yassine Souilmi ⁶, Faisal M. Fadlelmola ¹, Mustafa Alghali¹, Shaun Aron³, Hocine Bendou ⁷, Eugene De Beste⁷, Mamana Mbiyavanga⁴, Oussema Souiai ⁸, Long Yi⁷, Jennie Zermeno⁹, Don Armstrong⁹, Brian D. O'Connor¹⁰, Liudmila Sergeevna Mainzer^{9,11}, Michael R. Crusoe ¹², Ayton Meintjes ⁴, Peter Van Heusden ⁷, Gerrit Botha⁴, Fourie Joubert¹³, C. Victor Jongeneel ⁹, Scott Hazelhurst^{3,14}, Nicola Mulder ⁴

¹Centre for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum, Khartoum, Sudan

²Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Khartoum, Khartoum, Sudan

³Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa

⁴Computational Biology Division, Integrative Medical Biosciences, University of Cape Town, Cape Town, South Africa

⁵Department of Digital Technologies, University of Mauritius, Reunion, Mauritius

⁶Australian Centre for Ancient DNA, University of Adelaide, Adelaide, Australia

⁷South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa

⁸Institut Pasteur De Tunis and Institut Supérieur des Technologies Médicales de Tunis, University Tunis Al Manar, Tunis, Tunisia

⁹Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

¹⁰Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA

¹¹National Centre for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, IL, USA

¹²Common Workflow Project, Vilnius, Lithuania

¹³Centre for Bioinformatics and Computational Biology, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa

¹⁴School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

* Equal contributors

v1 First published: 18 Apr 2018, 1:9 (<https://doi.org/10.12688/aasopenres.12847.1>)

Latest published: 07 Aug 2019, 1:9 (<https://doi.org/10.12688/aasopenres.12847.2>)

Abstract

The need for portable and reproducible genomics analysis pipelines is growing globally as well as in Africa, especially with the growth of collaborative projects like the Human Health and Heredity in Africa Consortium (H3Africa). The Pan-African H3Africa Bioinformatics Network (H3ABioNet) recognized the need for portable, reproducible pipelines adapted to heterogeneous compute environments, and for the nurturing of technical expertise in workflow languages and containerization technologies. To address this need, in 2016 H3ABioNet arranged its first

Open Peer Review

Reviewer Status 

Invited Reviewers

Cloud Computing and Reproducible Workflows Hackathon, with the purpose of building key genomics analysis pipelines able to run on heterogeneous computing environments and meeting the needs of H3Africa research projects. This paper describes the preparations for this hackathon and reflects upon the lessons learned about its impact on building the technical and scientific expertise of African researchers. The workflows developed were made publicly available in GitHub repositories and deposited as container images on quay.io.

Keywords

Bioinformatics, hackathon, workflow, reproducible, pipeline, capacity building



This article is included in the [African Society of Human Genetics](#) gateway.

REVISED

version 2

published
07 Aug 2019

version 1

published
18 Apr 2018

1

2

3

?




report

?

report

?

report

- 1 **Steffen Möller** , Rostock University Medical Center, Rostock, Germany
- 2 **Juan Ruiz-Alzola**, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
- 3 **C. Titus Brown** , University of California, Davis, Davis, USA
- Rayna Harris** , University of California, Davis, Davis, USA

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

Corresponding author: Nicola Mulder (nicola.mulder@uct.ac.za)

Author roles: **Ahmed AE:** Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Mpangase PT:** Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Panji S:** Conceptualization, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Baichoo S:** Software, Supervision, Validation, Writing – Review & Editing; **Souilmi Y:** Software, Supervision, Validation, Writing – Review & Editing; **Fadlilmola FM:** Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Alghali M:** Software, Validation, Writing – Review & Editing; **Aron S:** Software, Validation, Writing – Review & Editing; **Bendou H:** Software, Validation, Writing – Review & Editing; **De Beste E:** Software, Validation, Writing – Review & Editing; **Mbiyavanga M:** Software, Validation, Writing – Review & Editing; **Souiai O:** Software, Validation, Writing – Review & Editing; **Yi L:** Software, Validation, Writing – Review & Editing; **Zermeno J:** Software, Validation, Writing – Review & Editing; **Armstrong D:** Software, Validation, Writing – Review & Editing; **O'Connor BD:** Software, Validation, Writing – Review & Editing; **Mainzer LS:** Conceptualization, Project Administration, Resources, Supervision, Writing – Review & Editing; **Crusoe MR:** Software, Validation, Writing – Review & Editing; **Meintjes A:** Conceptualization, Software, Validation, Writing – Review & Editing; **Van Heusden P:** Conceptualization, Software, Validation, Writing – Review & Editing; **Botha G:** Conceptualization, Software, Validation, Writing – Review & Editing; **Joubert F:** Conceptualization, Project Administration, Resources, Software, Supervision, Writing – Review & Editing; **Jongeneel CV:** Conceptualization, Project Administration, Resources, Supervision, Writing – Review & Editing; **Hazelhurst S:** Conceptualization, Project Administration, Resources, Supervision, Writing – Review & Editing; **Mulder N:** Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: MRC in his role as CWL Community Engineer, has had his salary supported in the past by grants from Seven Bridges Genomics, Inc to its employers. The other authors declare that they have no competing interests.

Grant information: H3ABioNet is supported by the National Institutes of Health Common Fund [U41HG006941]. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) programme of the African Academy of Science (AAS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Ahmed AE *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ahmed AE, Mpangase PT, Panji S *et al.* **Organizing and running bioinformatics hackathons within Africa: The H3ABioNet cloud computing experience [version 1; peer review: 3 approved with reservations]** AAS Open Research 2018, 1:9 (<https://doi.org/10.12688/aasopenres.12847.1>)

First published: 18 Apr 2018, 1:9 (<https://doi.org/10.12688/aasopenres.12847.1>)

Introduction

As an inherently interdisciplinary science, bioinformatics depends upon complementary expertise from biomedical scientists, statisticians and computer scientists¹. This opportunity for collaborative projects also creates a need for avenues to exchange knowledge¹. Hackathons, along with codefests and sprints, are emerging as an efficient mean for driving successful projects². They can be in the form of science hackathons that aim to derive research plans and scientific write up³, community-driven software development⁴, and data hackathons or datathons⁵. In addition to the scientific and technical outcomes, these intensive and focused activities offer necessary skills development and networking opportunities to young and early career scientists.

On the African continent, there is generally limited access to such events. However, with the growing capacity for Africans to generate genomic data, the need to analyze these data locally by African scientists, is also growing. H3ABioNet⁶, the Bioinformatics Network within the H3Africa initiative⁷, has invested in capacity building via different approaches⁸. The H3ABioNet Cloud Computing hackathon was a natural extension of the network's efforts in developing **Standard Operating Procedures (SOPs)** via its Network Accreditation Task Force (NATF)⁹; aimed at building and assessing capacity in genomic analysis. This also follows other efforts by the H3ABioNet Infrastructure Working Group (ISWG) towards setting up infrastructure at various H3ABioNet Nodes at the hardware, software, networking, and personnel level. The H3ABioNet Cloud Computing hackathon, therefore, provided an excellent opportunity to

assess the computational skills capacity development of the network through training, learning and adoption of novel technologies (**Figure 1**). These technologies included workflow languages for reproducible science, containerization of software, and creation of computational products that can be used in heterogeneous computing environments encountered by African and international scientists in the form of standalone servers, cloud allocations and High Performance Computing (HPC) resources.

In this paper, we discuss the organization of the H3ABioNet Cloud Computing hackathon, the interactions between the participants, and the lessons learnt. A paper describing the technical aspects of the pipelines developed has been in preparation (unpublished paper- Baichoo *et al.* 2018¹), whereas the code and pipelines themselves have been made publicly available via **H3ABioNet's GitHub page** in the following repositories: (**h3agatk**, **h3abionet16S**, **h3agwas** and **chipimputation**) as well as container images hosted on **Quay.io**.

¹Full author list for unpublished study: Shakuntala Baichoo, Yassine Souilmi, Sumir Panji, Gerrit Botha, Ayton Meintjes, Scott Hazelhurst, Hocine Bendou, Eugene de Beste, Phelani T. Mpangase, Oussema Souiai, Mustafa Alghali, Long Yi, Brian D. O'Connor, Michael Crusoe, Don Armstrong, Shaun Aron, Fourie Joubert, Azza E. Ahmed, Mamana Mbiyavanga, Peter van Heusden, Lerato E. Magosi, Jennie Zerneno, Liudmila Sergeevna Mainzer, Faisal M. Fadlelmola, C. Victor Jongeneel and Nicola Mulder

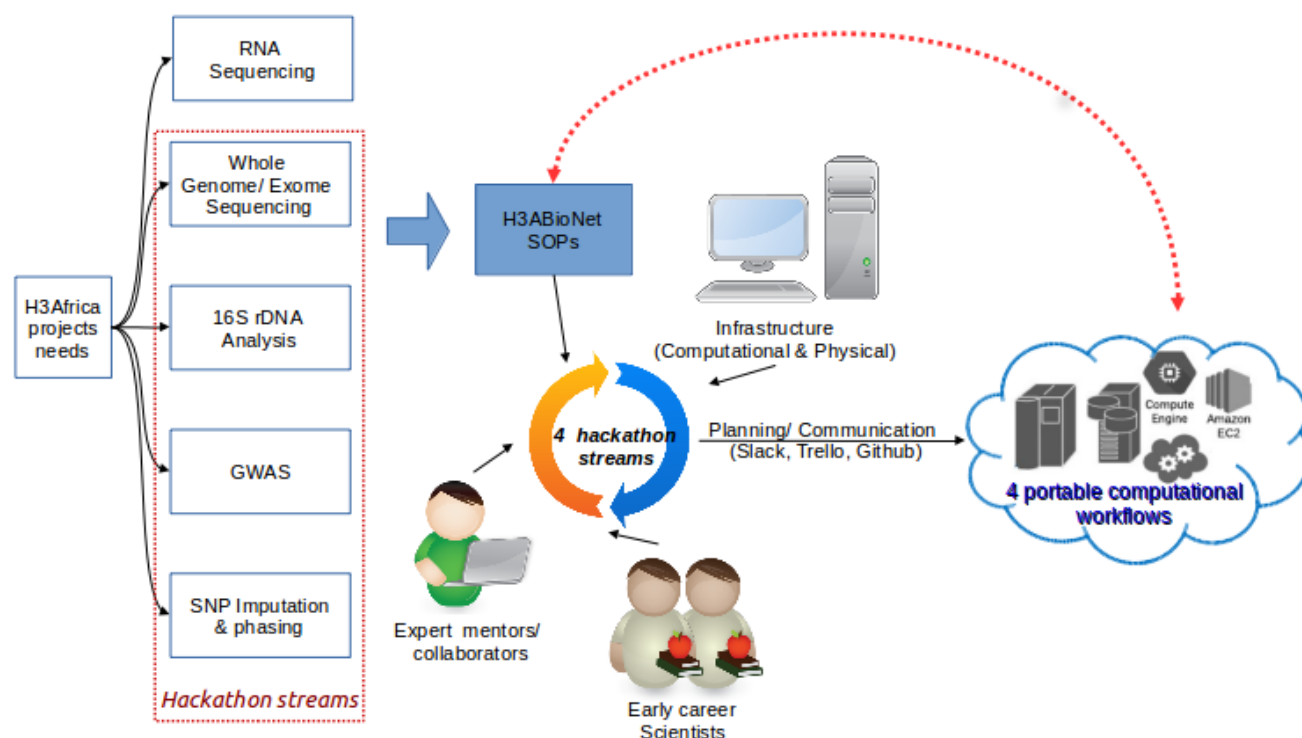


Figure 1. Planning and execution of the H3ABioNet Cloud computing hackathon. (SOPs: Standard Operating Procedures).

H3ABioNet Cloud Computing Hackathon Activities

Prior to the H3ABioNet Cloud Computing hackathon, H3ABioNet, via its Infrastructure Working Group (ISWG), formed a Cloud Computing task force to investigate cloud computing technologies, to familiarize H3ABioNet members with current cloud implementations and gauge their suitability for H3Africa data analyses. The H3ABioNet Cloud Computing hackathon was one of the first deliverables of this task force, with the specific objective to test and implement the use of four analysis workflows that can be ported on multiple compute platforms. Figure 1 shows this hackathon within the broader H3Africa context and provides a broad overview of the planning and execution of this activity, with details in the following subsections.

Pre-hackathon preparations

The computational pipelines put forward for development during the H3ABioNet Cloud Computing hackathon were identified based on the data being generated by different H3Africa projects and the SOPs used for the H3ABioNet Node Accreditation exercises. Reproducibility and portability were also identified as key features for the workflows, due to the heterogeneous computational platforms available in Africa. H3ABioNet Nodes that used or helped develop current H3ABioNet workflows and SOPs were part of the planning team, as well as other nodes that had technically strong scientists who were willing to extend their skills.

In the course of planning for the H3ABioNet Cloud Computing hackathon, two technical areas were identified where additional expertise was required. These were containerization technology such as Docker, and the writing of genomic pipelines in popularly used workflow languages and newly emerging community-standards like Nextflow¹⁰ and the Common Workflow Language (CWL)¹¹, respectively. While expertise for Nextflow already existed within the network, two collaborators from outside of Africa were interested to join the project given their expertise in running genomic pipelines in cloud environments, containerization of code¹² and developing CWL¹¹. They subsequently joined the planning and participated in the hackathon.

The H3ABioNet Cloud Computing hackathon was announced on the internal H3ABioNet consortium mailing list as a call for interested applicants and in some cases, individuals were invited based on their specific expertise. Most of the participants selected were early career scientists with strong computational skills, an understanding of genomic pipelines and willingness to work in teams. The pipelines for the Cloud Hackathon were divided into four “streams”: 1) Stream A: variant calling from whole genome sequencing (WGS) and whole exome sequencing (WES) data (<https://github.com/h3abionet/h3agatk>), 2) Stream B: 16S rDNA Diversity Analysis (<https://github.com/h3abionet/h3abionet16S>), 3) Stream C: Genome Wide-association studies (Illumina array data) (<https://github.com/h3abionet/h3agwas>) and 4) Stream D: SNP Imputation and phasing using different reference panels (<https://github.com/h3abionet/chipimputation>). Successful applicants were given a choice to select a project stream based on their skills and interest or if unsure, assigned to a specific stream. Streams A and B decided to use CWL for their pipeline development, whereas Streams C and D opted to use Nextflow Language, due to their prior experience using Nextflow.

Vital in setting up the teams for each of the streams was that each team had a balanced composition. This included bioinformaticians with strong computational skills to create the Docker containers and implement the workflow languages, knowledge in the specific genomic analyses and computational tools required, and strong system administration skills to assist with the installation of numerous software components. We also included bioinformaticians with experience in running the workflows or components of the workflows, and software developers who could assist with creating Docker containers, troubleshoot and implement workflow languages.

To maximize the learning experience, upon selection, participants were given prerequisite tutorials and materials (Github, Nextflow, CWL, Docker and the SOPs) to go through. Communication and planning infrastructure in the form of Slack channels and Trello boards were created beforehand with all the participants added in order to allow them to brainstorm and share ideas with team members before the hackathon began (Table 1).

Table 1. Communication channels used for the hackathon.

Channel	Link	Purpose
Mailing list	-	Group wide announcements and communications
Mconf	https://mconf.sanren.ac.za/	Online meetings
Slack	https://slack.com/	Inner group discussions and chat
Trello	https://trello.com/	Plan goals and activities, and track progress
GitHub	https://github.com/	Code repository and version control

Fortnightly planning meetings were held starting from 3 months in advance in order for hackathon participants to get involved in planning their proposed tools and to get to know one another and develop a working rapport before the start of the hackathon.

The hackathon ran in August 2016 and was hosted at the University of Pretoria Bioinformatics and Computational Biology Unit in South Africa. The choice of the hackathon venue was based on availability of Unix/Linux desktop machines with the facility for sudo/root access enabling participants to install software and deploy Docker containers for testing. Besides the local machines, participants also had access to cloud computing platforms such as [Azure](#) and [Amazon](#), [Nebula](#) (made available by the National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign), and the [African Research Cloud](#) (through a collaboration with the University of Cape Town eResearch initiative). After the hackathon, more testing was also done on [EGI Federated Cloud](#) resources (through an agreement contract with the University of Khartoum).

Hackathon week activities

The initial day of the H3ABioNet Cloud Computing hackathon was dedicated to introductions, expectations by the participants and practical tutorials covering the use of CWL, Nextflow and creation of Docker containers to ensure all participants had the same basic level of knowledge. The teams had a breakout session where overall milestones for the streams during the hackathon week were refined, tasks were identified and assigned to team members and Trello boards updated with the specific tasks. Each stream reported back on their progress and overall work plan for the coming hackathon days. For the remaining days of the hackathon, participants were split into their respective streams to work on developing and containerizing their pipelines as well as creating the related documentation. To ensure a successful hackathon with concrete outcomes, the streams spent the first 30 minutes of each hackathon day reviewing their prior progress and updating their Trello boards and reporting to the group what they will be working on. At the end of the day each stream provided a progress report to the whole group on what they had achieved, what they struggled with and what they will be working on. The start and end of day reporting proved useful as it allowed groups that had encountered and solved an issue to share the implemented solution with another stream, and for different streams to work together to solve any shared issues encountered, thus speeding up the development of the pipelines. Area experts and collaborators would switch between the streams to provide necessary technical expertise.

Communication during the hackathon was facilitated by Slack integration with Trello (for tasks managements and progress tracking) and code developed was pushed to GitHub (for live code integration). [Table 1](#) lists the various communication media used during the hackathon. Some groups also

utilized Google docs for documenting their progress prior to migrating documentation into github README files. Remote participation in the hackathon was facilitated through the MConf conference system. One stream had a participant with very strong coding skills working remotely from the US; who managed to make progress on the corresponding workflow when the other group members were not working due to the big time difference between the USA and South Africa (SA). This ensured continuous development on the workflow when the team in SA would clock off and provide a to-do list which was accomplished by the participant from the USA. Noticeable during the hackathon was the team spirit created and the increasingly later end time for the days with most days ending at 8:30 pm as participants continued working after the different streams provided their daily reports. All participants wished for an extra day or two to complete their pipelines.

Post-hackathon feedback and actions

After the week-long hackathon at the University of Pretoria, members of each stream continued working on their respective pipelines communicating via Slack and Trello. Meetings were held over [MConf](#) every two weeks to report on the progress of each pipeline. Upon completion, each group handed their pipeline to other groups to test on different platforms to avoid any bias in implementation and improve the documentation and hence facilitate the ease of use of the four pipelines developed.

Discussion

The H3ABioNet Cloud Computing Hackathon was aimed at producing portable, cloud-deployable Docker containers for a variety of bioinformatics workflows including variant calling, 16S rDNA diversity analysis, quality control, genotype calling, and imputation and phasing for genome-wide association studies. Dockerization provides a method to package and manage software, tools and workflows within a portable environment/container, similar to virtualization but with a smaller computing overhead compared to virtualization. Docker containers can easily be developed and deployed on computing environments (especially cloud based infrastructure) and can be used by a variety of groups to ensure reproducible analysis using the same tools, software versions, and workflows used.

The novelty of the H3ABioNet Cloud Computing Hackathon was that all the participants selected were involved in the latter stages of the planning and the setting of some of the outcomes for the hackathon. Critical recommendations during the hackathon planning meetings were that the resulting Docker containers and pipelines developed should be compatible with heterogeneous African research compute environments with portability and good documentation being key. This is especially important considering the fact that access to Cloud computing environments within Africa is still in its infancy. Hence, it was decided that development and testing of the pipelines should occur on a single machine, with the ability to be ported to a cluster or an HPC environment, and ultimately tested and deployed on cloud-based platforms (Amazon, Microsoft

Azure, EGI FedCloud, IBM Bluemix, and the new African Research Cloud initiative).

Lessons learnt and concluding remarks

The opportunity to link people physically and focus solely on one project has been highly effective in providing the main outline and proof of concept outputs. However, once people were back home, continuing the tasks has been a challenge. Clearly defining the roles and commitment of all the participants in the papers reporting the results should encourage them to complete the work, and increase their accountability.

The communication and management tools used for this hackathon were important as these tools facilitated interaction between and across team members and enabled the participants to continue to work in a structured manner once back at their respective institutions, despite time zones differences.

The H3ABioNet Cloud Computing Hackathon has been an important milestone for the Network as it brought together people with various skills to work on focused projects. It signalled the shift from capacity building to utilizing the capacity developed in order to tackle problems specific to the heterogeneous African computing environments, as defined and implemented by the mostly African participants.

As software packages and computing environments evolve with varying build cycles and new bioinformatics tools become available, we envision that hackathons to keep these pipelines current, adopt new technology implementations such as Singularity, and develop new workflows such as for RNA-Seq analysis will occur. The pipelines developed during the H3ABioNet Cloud Computing hackathon will be used for training and data analyses for intermediate level bioinformatics workshops, and for scientific collaborations requiring bioinformatics expertise for data analysis such as with the H3Africa genotyping chip and GWAS analyses. Future H3ABioNet hackathons would also provide an opportunity to utilize the skills of trained bioinformaticians at intermediate and advanced levels, who would not otherwise attend bioinformatics training

workshops, to come together to derive practical solutions that are of benefit to the African and wider scientific community.

Data and software availability

All data underlying the results are available as part of the article and no additional source data are required.

The four pipelines are available publicly via H3ABioNet's github organization page <https://github.com/h3abionet> in the following repositories: (h3agwas, chipimputation, h3agatk and h3abionet16S) as well as container images on quay.io at https://quay.io/organization/h3abionet_org.

All code is available under MIT license, except for the h3agtk pipeline which is available under Apache license 2.0

Competing interests

MRC in his role as CWL Community Engineer, has had his salary supported in the past by grants from Seven Bridges Genomics, Inc to its employers. The other authors declare that they have no competing interests.

Grant information

H3ABioNet is supported by the National Institutes of Health Common Fund [U41HG006941]. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) programme of the African Academy of Science (AAS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We acknowledge the advice and help from Ananyo Choudhury from Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa.

References

- Yanai I, Chmielecki E: **Computational biologists: moving to the driver's seat.** *Genome Biol.* 2017; **18**(1): 223.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Möller S, Afgan E, Banck M, *et al.*: **Community-driven development for computational biology at Sprints, Hackathons and Codefests.** *BMC Bioinformatics.* 2014; **15** Suppl 14: S7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Groen D, Calderhead B: **Science hackathons for developing interdisciplinary research and collaborations.** *eLife.* 2015; **4**: e09944.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Crusoe MR, Brown CT: **Channeling Community Contributions to Scientific Software: A Sprint Experience.** *J Open Res Softw.* 2016; **4**(1): pii: e27.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aboab J, Celi LA, Charlton P, *et al.*: **A "datathon" model to support cross-disciplinary collaboration.** *Sci Transl Med.* 2016; **8**(333): 333ps8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mulder NJ, Adebisi E, Alami R, *et al.*: **H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa.** *Genome Res.* 2016; **26**(2): 271–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- H3Africa Consortium, Rotimi C, Abayomi A, *et al.*: **Research capacity. Enabling**

- the genomic revolution in Africa. *Science*. 2014; **344**(6190): 1346–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Aron S, Gurwitz K, Panji S, *et al.*: **H3abionet: developing sustainable bioinformatics capacity in africa**. *EMBNet j*. 2017; **23**: e886.
[Publisher Full Text](#)
9. Jongeneel CV, Achinike-Oduaran O, Adebiyi E, *et al.*: **Assessing computational genomics skills: Our experience in the H3ABioNet African bioinformatics network**. *PLoS Comput Biol*. 2017; **13**(6): e1005419.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows**. *Nat Biotechnol*. 2017; **35**(4): 316–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
11. Peter A, Crusoe MR, Nebojša T, *et al.*: **Common Workflow Language, v1.0**. 2016.
[Publisher Full Text](#)
12. O'Connor BD, Yuen D, Chung V, *et al.*: **The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows [version 1; referees: 2 approved]**. *F1000Res*. 2017; **6**: 52.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 22 June 2018

<https://doi.org/10.21956/aasopenres.13913.r26312>

© 2018 Brown C et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



C. Titus Brown

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

Rayna Harris

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

Summary and impression

The article describes the organization and execution of a hackathon in Africa with the goal of producing cloud-deployable Docker containers for four bioinformatic workflows.

The article begins with a brief history of H3ABioNet and the need for bioinformatics infrastructure, training, and community in Africa.

Then the authors describe their pre-hackathon preparations, the weeks of activities during the hackathon, and their lessons learnt. No quantitative assessment of the hackathon was provided, but the authors do provide links to the bioinformatic workflows used.

Hackathons are an increasingly popular way of building tools and community for research and education. This article nicely describes why a hackathon was needed in Africa and how the authors went about organizing and executing their first hackathon. The article contains some useful suggestions for what bioinformatic or social tools can be used to facilitate research and communication. I think it is important to communicate this type of activity to the broader scientific community, however, I have a few concerns.

Major issue

My main concern is that this paper was submitted as a "method article". My understanding is that methods articles should describe new empirical or computational methods that are described in sufficient detail such that they can be reproduced. While this article does describe the authors' strategy for organizing and running a hackathon, this article is written more as a retrospective piece on the event rather than a recipe for running a hackathon. I am not convinced that it should be published as a method paper; unfortunately, there doesn't appear to be a more suitable platform within the AAS journal.

<https://aasopenresearch.org/for-authors/article-guidelines/method-articles>

Minor issues

1. The figure legend does include a title, but it does not include a description of the key points nor does it explain the meaning of the arrows. According to the AAS guidelines, "the legend should be sufficiently detailed so that it can stand alone from the main text". Additionally, it is not clear from first glance that "4 portable computational workflows" is the goal of the hackathon. This could be made more clear.
2. Table 1 provides a nice overview of communication channels. Can you elaborate and add what tools you used for sharing documents (e.g notes, slides, pdf)?
3. On page 4, the authors state: "Vital in setting up the teams...". Does this paragraph refer to the expertise of the learners/participants or to the people who are leading the stream or both?
4. On page 5, the first sentence of the discussion provides the first clear statement of the aim of the hackathon (in my opinion). The goal is mentioned in the abstract and intro, but it isn't as clear. It is in this paragraph that I realized that you had working pipelines, but they were not "dockerized". After reading this, the figure made a lot more sense. I recommend revising the abstract and intro to make it clear what the starting point (5 bioinformatic workflows not in the cloud) and the endpoints (4 bioinformatic workflows in the cloud).
5. The paragraph on "Post-hackathon feedback and actions" seems incomplete or perhaps is mislabelled. This paragraph describes communication and work that extended past the hackathon, but it does not describe any assessment or feedback mechanisms that were used. Also, what happened after groups traded platforms? Was the documentation improved or was this simply the goal?
6. The article jumps from "Introduction" to "Discussion" without providing a clear a description of some of the results or outcomes. Is there a reason why some statistics regarding participation or progress toward the goal are not reported? If there is a reason why demographic information cannot be provided, that is fine, but a report about how much progress was made toward the goal of dockerization would be useful.
7. While I appreciate that you provided a link to the GitHub repo and Docker container, I highly recommend getting a DOI for these repositories. The AAS data guidelines page provides a list of providers. Some of your GitHub repositories already have versions, so it should be fairly easy to import these into Zenodo for a DOI. <https://aasopenresearch.org/for-authors/data-guidelines>.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No source data required

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

Partly

Competing Interests: No competing interests were disclosed.

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

Reviewer Report 02 May 2018

<https://doi.org/10.21956/aasopenres.13913.r26376>

© 2018 Ruiz-Alzola J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Juan Ruiz-Alzola

Department of Signals and Communications, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

I have overall enjoyed reading this paper. The implementation of the hackathon is well explained, and provides good hints for others to replicate both in Africa and in the rest of the World. In particular, a large array of ICT-based collaborative tools has been applied whose use is very interesting everywhere. The motivation is also very relevant, as hands-on knowledge sharing is important to build distributed strong scientific communities that can develop new solutions and keep collaboration active within wide multi-site networks. This is particularly so in the case of bioinformatics, a profoundly interdisciplinary and complex field, key to modern knowledge-based economy, health and welfare, where Africa could leverage important opportunities as far as some hurdles are properly managed. It is in this framework where I see the main contribution of this paper, and I must point out the relevance of the collaboration among African and rest of the World institutions and researchers, in addition to the internal African collaboration. In my view all of this is essential to unlock all the potential, and achieve not only scientific success but a wider interaction of great value for economic development. I'm not getting into any technical detail of the pipelines since, as reported in the paper, they will be presented elsewhere.

Nevertheless I have some constructive criticisms that I'd like to share:

1. I've pointed out that the rationale is only partly explained essentially because some boundary conditions should be explained: how many nodes and researchers are participating, what their access to Internet is like in their sites, what difficulties they're experiencing in their home institutions to access to knowledge and to interact with other international colleagues, etc. Africa is large with very different situations across countries and within each country. It'd be good that the reader could understand better the everyday situation that motivates an action like this, or a wider program such as H3ABioNet. It'd also be good to explain a bit longer what H3A and H3ABioNet are, and how they become important programs for the scientific development of Africa.

2. The description of the method is technically sound, but I have some concerns about how realistic it is to expect such sophisticated ICT-based collaboration, difficult anywhere, considering the existing difficulties for high bandwidth connections in many areas of Africa, including many Universities in main cities. Our group is currently involved in a cooperation with research groups in four African countries. Internet access, good enough for sharing medical datasets, for example, or just to hold videoconferences, is something that we cannot take for granted. Again, this is also related to (1) and boundary conditions, but it is my experience that dealing with this sort of issues is key for success and, at least, to replicate the experience elsewhere.
3. The conclusions are somehow conditioned by what I've already pointed out.

In summary, I think this is a very interesting and motivating paper that could benefit of a better explanation of the boundary conditions, and of how the specific difficulties have been overcome, so that others can replicate the experience in the same or in different fields across the continent.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

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

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical imaging, medical technology for sustainable development, cooperation for development

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 18 April 2018

<https://doi.org/10.21956/aasopenres.13913.r26315>

© 2018 Möller S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Steffen Möller

Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany

The paper describes an in-person meeting in Africa to further develop workflows for genomic analyses and distribute that skill throughout the continent. As a Northern European I can only remotely assess the difficulties of computational biology and bioinformatics services in Africa in local areas. I can see how important such Hackathons are to edutRain research groups with difficulties to attend international conferences. And, with some experience in attending and organizing such events, I was very curious about what may be different in Africa. After all, in Europe there are communities having difficulties to commute, too. The European Union has extra funds for graduate students from Eastern European countries for instance. But there are also highly talented high school students who are under age or do not have the funds or scientific contacts to travel/be trained. So, if you have developed principles for African talents to overcome such obstacles to learn about such Hackathons and prepare for them, then I would be very much interested to see what we can adopt over here.

From what I then read in the article, there was not so much that the authors did differently, except that they seem to have done it particularly well. Table 1 describes a whole bunch of communication channels when typically it is a mere Wiki or co-editing site orchestrating the participants. Still, every attendee had to physically attend, except that the event was in Africa describing an African set of resources and there was external expertise flown in. I tend to think that here the event fell short of what could have been (and is often) done to invite remote participation. Also, one could videograph training material to support the further distribution of these Africa-specific analysis skills.

I am also a bit critical about the dominance of online resources that demand good internet connections as in the download of gigabytes of data: Docker. Here I wish more would be done to promote offline services. But I am biased, the authors NM and MRC know me as a contributor to Debian Med, MRC being a Debian contributor himself, which basically means that participants could take a DVD home and perform analyses of their samples without Internet access involving as many local-to-them machines as they like. For the workflows one doesn't need most of the complicated bits for which one needs computational expertise the article is describing. And that may have helped the post-Hackathon drop in participation, while, of course every event sees that drop and that is a main reason to have such a dedicated time to jointly develop our research environments in the first place.

Concerning the scientific results, I understand that there is a separate paper prepared. Still, can you say as much as if there are scientific papers out there that already employed the pipelines for their analysis? I mean, from the time before the Hackathon? This would emphasize that you are indeed redistributing a very current set of skills with practical acceptance in the community.

There is something else to it all. Hackathons form a social network of trust. And you need trust in locally well-described samples. The genetic diversity of Africa is a gem, but one needs to be aware of the demands for population stratification. And because of the harsh environment conditions, one can expect considerable batch effects on samples taken, for which the emphasis on equally collected control

samples is important. You can certainly read about it all, but it will help to hear in person about that study that was ruined because cases and controls were kept separate and one box had the dry ice evaporated early – factor variation and confounding technical parameters cannot be communicated enough. Well-established workflows allow participants to analyse their local data independently, have extra parameters matching local concerns, with matching local controls, which will all improve the quality of the pan-African study at large.

So, I do not think that anything performed for the organisation of this Hackathon was specific to Africa. In the contrary, there should have been some teleconferencing to it. The “initial day to get everyone up to speed” reminds me a bit more of a Summer School than a Hackathon, for which it is not uncommon that most participants already know each other for some time and would not need that day. Could you possibly elude a bit more on how the work performed was structured? And how was the external expertise intermingled with your local needs for Africa?

I would want to retitle the paper towards something like “Hackathon on workflows for genomics held in Africa”. The H3ABioNet workflows are nicely accepted everywhere, I tend to think, so, why not have an H3ABioNet workflow Hackathon in Europe with some Africans flying in? Clarifying that a bit more may help the paper.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

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

Partly

Competing Interests: No competing interests were disclosed.

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.

Author Response 21 Apr 2018

Victor Jongeneel, University of Illinois at Urbana-Champaign, Champaign, USA

Thanks very much to Steffen Möller for his pointed comments. It is certainly true that this hackathon did not differ in any major way from similar events held in high-income countries, and

did not incorporate any features specific to the African context. Its primary value was to expose African scientists to the practical aspects of community development of computer code, and to try to create a community around the maintenance of a set of workflows that implement methods that are useful to the H3Africa research community and beyond.

There are aspects of the work that may not have come across in the paper. For one thing, the workflows implementing haplotyping, imputation, and GWAS analysis were based on work done in the framework of the H3Africa AWIGen project, and are in production for the analysis of data generated by this project. Similarly, the workflow for variant calling from WGS data was used in the analysis of 350 African genomes that has led to the design of a novel genotyping chip optimized for African populations, and the 16S rDNA sequence analysis was derived from work done to analyze bacterial populations present in leg ulcers of sickle cell patients in Nigeria. Therefore, all of the code developed during the hackathon was solidly anchored in existing genomic analysis projects in Africa.

Secondly, the workflows developed in the hackathon serve as practical implementations of Standard Operating Procedures for the H3Africa Accreditation Exercises, which are used to evaluate the capacity of African research groups to analyze complex genomic datasets being produced by its research projects (see Jongeneel et al, PLoS Comput Biol, [PMC5453403](#)). Success in taking one of the exercises is considered a landmark for African groups who are preparing to step into the existing gap between data production and data analysis, where the analysis is typically undertaken by First World groups.

It is true that the authors, including myself, could have done a better job at explaining how the hackathon and its products are anchored in the H3Africa research ecosystem. I hope that the above clarifies this.

As a final remark, while highly Internet-dependent tools were used extensively during the hackathon, to my knowledge none required a very high bandwidth. At least two of the participants attended remotely from North America, and were able to contribute substantially in part because of their time zone differences and asynchronous contributions.

Competing Interests: No competing interests were disclosed.

Reader Comment 04 May 2018

Steffen Möller,

Thank you for your constructive reply to my initial comments. This addresses all my reservations and I am looking forward for a revised upload. SM

Competing Interests: No competing interests were disclosed.