EMBnet.journal
Bioinformatics in Action

Gisel et al. (2023) EMBnet.journal 29, e1045
http://dx.doi.org/10.14806/ej.29.0.1045

Reports

EpiCass and CassavaNet4Dev Advanced Bioinformatics Workshop

Andreas Gisel[1,2], Livia Stavolone[3], Temitayo Olagunju[1], Michael Landi[4], Renaud van Damme[4], Adnan Niazi[4], Laurent Falquet[5], Trushar Shah[1], Erik Bongcam-Rudloff[4]✉

[1] International Institute of Tropical Agriculture, Ibadan, Nigeria
[2] Institute for Biomedical Technologies, National Research Council, Bari, Italy
[3] Institute for Sustainable Plant Protection, National Research Council, Bari, Italy
[4] Swedish University of Agricultural Sciences, Uppsala, Sweden
[5] University of Fribourg, Fribourg, Switzerland

Competing interests: AG none; LS none; TO none; ML none; RvD none; AN none; LF none; TS none; EBR none

Abstract

EpiCass and CassavaNet4Dev are collaborative projects funded by the Swedish Research Council between the Swedish University of Agriculture (SLU) and the International Institute of Tropical Agriculture (IITA). The projects aim to investigate the influence of epigenetic changes on agricultural traits such as yield and virus resistance while also providing African students and researchers with advanced bioinformatics training and opportunities to participate in big data analysis events. The first advanced bioinformatics training workshop took place from May 16th to May 18th, 2022, followed by an online mini-symposium titled "Epigenetics and crop improvement" on May 19th. The symposium featured international speakers covering a wide range of topics related to plant epigenetics, cassava viral diseases, and cassava breeding strategies. A new online and on-site teaching model was developed for the three-day workshop to ensure maximum student participation across Western, Eastern, and Southern Africa. Initially planned in Nigeria, Kenya, Ethiopia, Tanzania, and Zambia, the workshop ultimately focused on Nigeria, Kenya, and Ethiopia due to a lack of qualified candidates in the other countries. Each classroom hosted 20 to 25 students, with at least one bioinformatician present for support. The classrooms were connected via video conferencing, whereas teachers located in different places in Africa and Europe joined the video stream to conduct teaching sessions. The workshop was divided into theoretical classes and hands-on sessions, where participants could run data analysis with support from online teachers and local bioinformaticians. To enable participants to run guided, CPU and RAM-intensive data analysis workflows and overcome local computing and internet access restrictions, a system of virtual machines (VMs) hosted in the cloud was developed. The teaching platform provided teaching and exercise materials to support the use of the VMs. Some students could not run heavy data analysis workflows due to unforeseen restrictions in the cloud. Currently, these issues have been solved and in the future all participants will have the opportunity to run the analysis steps independently in the cloud using the protocols hosted on the teaching platform.

Introduction

EpiCass and CassavaNet4Dev are two projects that aim to investigate the epigenetics of cassava (Manihot esculenta) and its influence on agronomic traits. They also aim to disseminate the data and information generated and train African researchers in high-throughput data analysis. EpiCass focuses primarily on research, while CassavaNet4Dev focuses on dissemination, training, and networking within the African plant science and breeding community.

EpiCass is divided into two research lines. The first line examines the DNA methylation profiles of four cassava genotypes and correlates these profiles with the highly variable storage root yield. The second line investigates the DNA methylation profile of two genetically similar cassava genotypes with different resistance against the African Cassava Mosaic Virus. In both cases, the goal is to extract and analyse sequencing data to produce knowledge that can improve these traits or better monitor them during breeding processes. EpiCass also proposes a workshop on next-generation

Article history
Received: 2022
Accepted: 2023
Published: 2023

sequencing (NGS) data analysis for African researchers to gain a basic understanding of sequencing technologies and introduce them to epigenetics and its potential in crop improvement.

On the other hand, CassavaNet4Dev aims to develop a network of African scientists and students interested in learning NGS data analysis and joining a network of plant scientists and breeders. This network plans to include selected participants to collaborate on the data analysis of the EpiCass project, which intends to generate raw data of about 36 GB long-read re-sequencing data, 2.7 TB DNA methylation profiling data, and 130 GB small non-coding RNA data.

The two projects collaborated to organise the first bioinformatics training workshop, which took place from 16-18 May 2022, with a mini-symposium on 19 May providing an overview of epigenetics in plants and possible applications for crop improvement. In the following paragraphs, we describe in more detail the organisation, execution, problems, and reviews of this workshop.

Workshop set-up

With our extensive experience in physical and online bioinformatics training, we have developed a hybrid workshop model combining the benefits of physical interaction with the ability to involve participants and teachers over a larger geographical area, thereby reducing travel and accommodation costs for both participants and organisers. The idea of such a workshop model was developed based on our observation that it is challenging to follow participants with different experience levels in online courses, especially during hands-on sessions. We also wanted to allow scientists and students from various African regions to participate in our workshop without incurring the high costs of long-distance travel.

The EpiCass project covers Nigeria and Kenya in Eastern and Western Africa, respectively. Additionally, the International Institute of Tropical Agriculture (IITA) has essential stations in Tanzania and Zambia in Southern Africa and significant collaborations in Ethiopia, such as the Bio & Emerging Technology Institute (BETin) in Addis Ababa. We planned physical classes of up to 20 participants at IITA stations to ensure a certain infrastructure level, including a good internet connection and a bioinformatician in each classroom to follow the group. However, due to limited funds, only local students or those with their own finances could participate. Fortunately, BETin offered to host the classroom and sponsor the event in Addis Ababa and allowed Ethiopian students from other universities outside Addis Ababa to travel to BETin.

The workshop classrooms were connected through a video conference system (in our case, Zoom) to allow for comments and questions in each classroom, with additional teachers from Africa and Europe joining the live stream to teach and follow the practical hands-on sessions. It's worth mentioning that the three classrooms were in two different time zones, with Ibadan in WAT (UTC +1), Nairobi and Addis Ababa in EAT (UTC+3), and the teacher from Europe in CEST (UTC+2). When the classes started at 8 am in Ibadan, it was 9 am in Europe and 10 am in Nairobi and Addis Ababa. This posed a logistical challenge, especially during the breaks. Another important element of the workshop was the development of documented classes on a teaching platform. In our case, we used the open-access platform Moodle, where each teacher hosted presentations and detailed exercises for the hands-on sessions. This platform enabled participants to preview future classes, follow presentations on their laptops, run hands-on sessions at their own pace, and run hands-on sessions as many times as needed to understand the different data analysis steps. Each participant was registered in Moodle, hosted at SLU in Uppsala, and had full access to the teaching material and communication channels.
The programme for the 3-day workshop introduced participants to NGS data analysis, with one full day dedicated to learning and using the operation system (OS) LINUX. This was necessary because many African scientists know little or nothing about LINUX. The morning was dedicated to introducing LINUX and explaining the most important commands to enable participants to start with NGS data analysis. In the afternoon, each participant was connected to an external LINUX server to run all the exercises and become more familiar with the OS.

The second day introduced the concept of NGS and the first steps in the NGS data analysis workflow, including quality control of raw reads, cleaning raw reads, and mapping them against a reference sequence. In the afternoon, participants were exposed to data analysis running tools for quality control, filtering, and mapping.

On the third day, we gave participants an introduction to epigenetics, the main topic of the EpiCass project, followed by classes on NGS assembly and hands-on sessions in the afternoon.

Selection of participants

To ensure a successful workshop, selecting motivated participants who are engaged in the research work and at a similar level is crucial. As the workshop focuses on

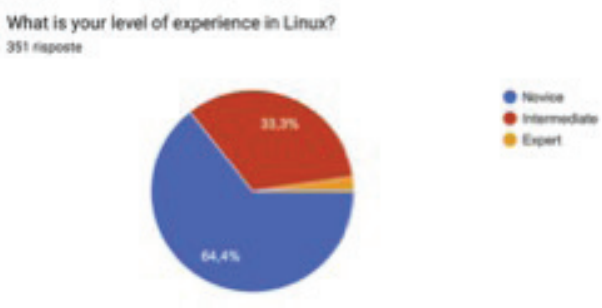

Figure 1. Percentage of LINUX knowledge among the workshop's applicants.

Gisel et al. (2023) EMBnet.journal 29, e1045
http://dx.doi.org/10.14806/ej.29.0.1045

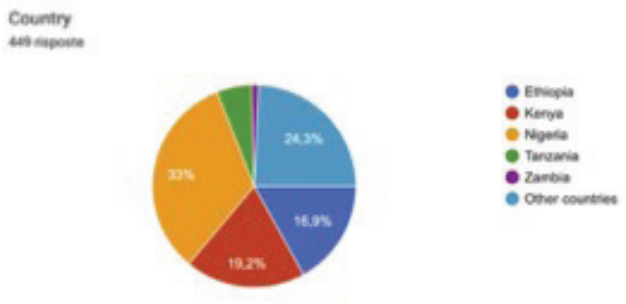

Figure 2. Distribution of the applicants by country

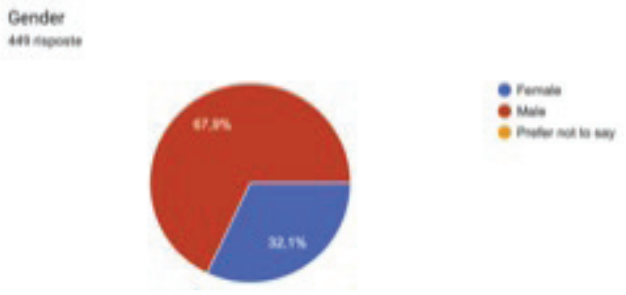

Figure 3. Distribution of gender and level of education of the applicants.

NGS data analysis, we set basic bioinformatic knowledge as a selection criterion for access to the workshop. We developed a Google Form registration module that collects personal information and includes questions about bioinformatics and LINUX to measure the level of the applicants.

The registration form was distributed within the IITA network in Africa and to collaborating universities in Nigeria, Ethiopia, Kenya, Tanzania, and Zambia. We received 449 registrations, mainly from Nigeria, Kenya, and Ethiopia (Figure 2). Interestingly, 25% of the registrations came from countries other than the selected ones, including Egypt, Sudan, Uganda, Tunisia, and countries outside Africa such as India and China. Most of the applicants were male researchers (almost 70%), and half of the applicants were master's students (Figure 3).

As mentioned earlier, we were looking for applicants with basic bioinformatics knowledge. To this aim, we included five questions about basic bioinformatics and five questions to evaluate LINUX experience in the registration form. Each question was worth one point, and a maximum of ten points were possible. Figure 4 shows that most applicants had zero points, indicating no experience in either bioinformatics or LINUX. Only 25% (114) of the applicants had at least one point, which indicates an urgent need for bioinformatics education in Africa. Many applicants reported being involved in analysing biological data without proper training, highlighting the need for bioinformatics training in Africa.

Unfortunately, we did not receive enough applications from Zambia, and the applications from Tanzania were not qualified enough. As a result, we

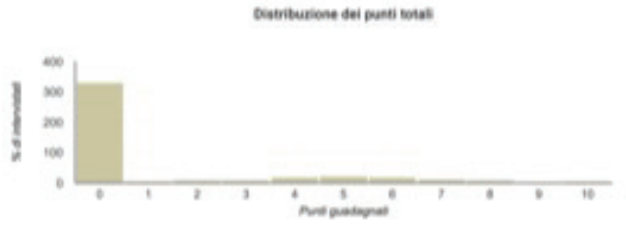

Figure 4. Distribution of the total points of all applicants for the quiz with five questions about basic bioinformatics and five questions about LINUX.

focused our workshop on Nigeria (Ibadan - IITA), Kenya (Nairobi - IITA), and Ethiopia (Addis Ababa – BETin). Each local team selected the final participants based on their bioinformatics and LINUX experience and whether they had a running project requiring data analysis. We invited 21 participants from Ethiopia, 22 from Nigeria, and 24 from Kenya and hosted 29 (4 female and 25 male), 19 (10 female and nine male), and 18 (nine female and nine male), respectively.

As previously mentioned, we had hands-on exercises during the afternoons, and each student had access to a LINUX server to execute the data analysis. In the following paragraph, we will describe how we tackled the challenge of hosting about 60 students for high RAM and CPU data analysis.

Set-up of the computing environment

NGS data analysis, especially NGS assembly, is a computationally intensive task that requires high RAM and CPU usage. Laptops and personal computers are inadequate for such applications, and it has become increasingly difficult to host external participants for such workshops on institutional servers due to security measures. To address these issues, we developed a computing strategy using cloud computing. We have extensive experience with the Ubuntu 18.04 LINUX operating system and have set up an Ubuntu 18.04 virtual machine (VM) with all the necessary software (see Table 1) for the 3-day workshop. We use Amazon Web Services (AWS) via its graphical web-based interface and Command Line Interface (CLI) to host the VM. We support various VM file formats, such as Open Virtualization Archive (OVA), Virtual Machine Disk (VMDK), Virtual Hard Disk (VHD/VHDX), and use the free VM software VirtualBox to save the VM in one of those formats.

When you register for an AWS account, you automatically set up an account with complete access to all AWS services and resources in the account. This account is called the AWS account root user, and it is accessed by signing in with the email address and password that you used to create the account. AWS strongly recommends not using the root account for your everyday tasks. Instead, it is best to create an Identity

Gisel et al. (2023) EMBnet.journal 29, e1045
http://dx.doi.org/10.14806/ej.29.0.1045

Figure 5. Classrooms at the three locations.

and Access Management (IAM) user, to which you grant specific permissions to administer and use services in the AWS account. Therefore, you should never expose the password or access key to collaborators. IAM accesses the AWS services with an account ID (created by AWS), a username, and a password, all created by the root user. It is possible to set up different IAMs for different tasks. For our project, we created an IAM that can create storage buckets, upload data from external resources into these buckets, create Amazon Machine Images (AMI), and start and stop instances. After defining the credentials and permissions for the IAM, you can upload the above-mentioned VM into the Amazon storage bucket with a single CLI command from your terminal. While creating the storage bucket, you must decide in which geographical region to create it. You can choose from US, Canada, Europe, Africa, the Middle East, South America, and Asia. The location of your storage is important since you will have to use the cloud computing in the area where you have your storage. Depending on the location, computing prices and available instance types may vary. After you have uploaded the VM into the storage bucket, you need to import it into the computing environment EC2 (Amazon Elastic Computing Cloud), where it will be converted and stored as an Amazon Machine Image (AMI) from where it can be launched at any time, creating an instance. During the launch, you can select your AMI from "My AMIs," describe the instance, choose the instance type and needed storage size, and define who will have access to the running instance. If you want a public IP accessible to anybody, you can choose "Anywhere" under the "network settings" "Allow SSH traffic from." Otherwise, you can select specific IP addresses to limit access and increase security.

We chose to use the EC2 environment because we needed computing power quickly and wanted the freedom to select specific instance types (*i.e.* number of CPUs and amount of RAM) based on the applications we planned to run. For the first hands-on exercise on the afternoon of the first day, during which we used various LINUX commands, we required only a small amount of computing power. Therefore, we started an instance with 2 CPUs and 4 GiB RAM (using Amazon instance type t2.medium) for each classroom. When an instance is started, it is assigned a temporary public IP address which participants use to connect to the virtual machine using the Secure Shell Protocol (SSH). Each time an instance is started, AWS assigns a new public IP address, which will most likely be different from the previous one. If you need a fixed IP address each time you start the same instance, you will need to set up an elastic IP address, which comes at a cost but might be worthwhile in certain situations. All 66 participants were able to connect to one of the three virtual machines and successfully complete the LINUX exercise.

For the hands-on exercise on the second day, in which we planned to perform quality control, filtering, and mapping of a real NGS dataset, we intended to provide each participant with their own virtual machine with 4 CPUs and 16 GiB RAM to obtain results in a reasonable amount of time. Unfortunately, we encountered a limit that we did not experiment with during our tests and studies of AWS documentation. The available CPUs, without any additional request, are limited to 32, whereas we planned to use 264 CPUs.

Gisel et al. (2023) EMBnet.journal 29, e1045
http://dx.doi.org/10.14806/ej.29.0.1045

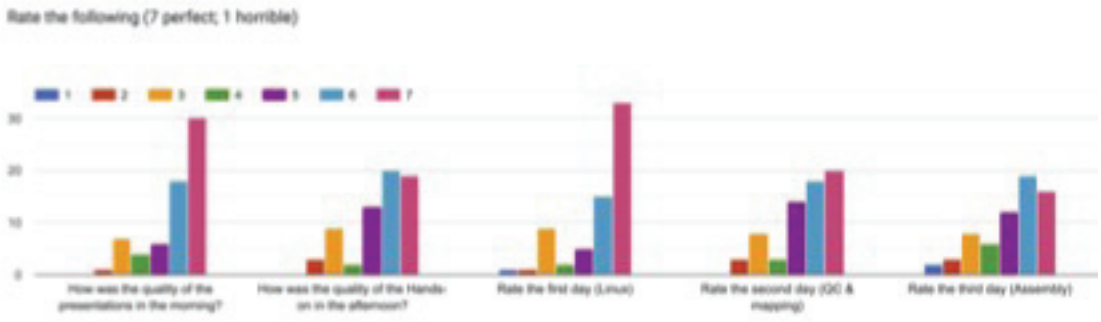

Figure 6. Evaluation of sections of the workshop.

Only in the EC2 dashboard, under "Limits", can you find a CPU limits calculator that shows how much CPU you have available and how much you would need for your task. If you require more CPUs than are assigned, you can request an increase in the actual limit, which we did. However, it took three days for our request to increase the limit from 32 to 264 CPUs, and even then, we only received an increase of 192 CPUs, which was too late for the workshop and not enough for what we had planned. The take-home message is that even when using the EC2 environment and paying a higher price for on-demand computing resources, there are very low limits, and requests for increases can take a long time. Therefore, it is essential to plan well in advance for the needed resources and expect you may not receive the number of requested resources.

With 32 CPUs, unfortunately, we could not offer a reasonable solution for participants to run their own NGS data analysis workflows. We launched three VMs, each with 8 CPUs and 32GiB RAM (Amazon instance type t2.2xlarge), but participants were unable to run their analyses. As a result, local teachers ran the analysis, and participants followed the steps on the screen and had the opportunity to ask questions or request a repetition of some steps. The same scenario occurred during the hands-on exercises on day 3, where we demonstrated how to set up NGS data for an NGS assembly and run the assembly, including some basic quality tests of the assembled sequences.

At the end of the workshop, we requested that participants evaluate the event. Specifically, we wanted to understand their thoughts about the impact of being unable to access their server and follow the data analysis exercises. As shown in Figure 6, the difference between the first day and the following two days reflects the drop in the quality of the hands-on experience due to the inability to use their server. Surprisingly, with the help of local teachers demonstrating the data analysis, we were able to maintain a high level of quality, and participants did not rate this lack as a total failure.

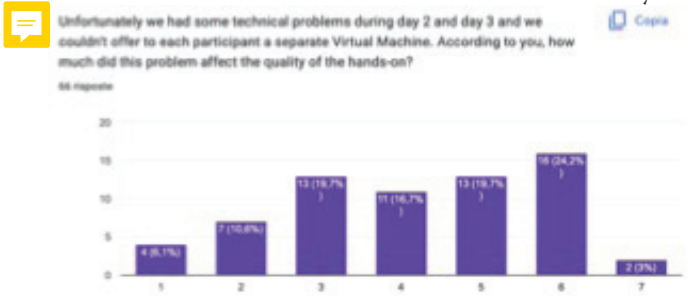

Figure 7. Rating of the influence on the quality of training dur to the lack of access to computing.

Figure 7: Rating of the influence on the quality of training dur to the lack of access to computing.

Mini-symposium

After the three-day training, we closed the workshop with a one-day mini-symposium covering topics related to the EpiCass project, "Epigenetics and Crop Improvement." Once again, participants were present in their classrooms, and speakers connected from Africa, Germany, Switzerland, and the USA. Since we advertised the event as open and free, we also had up to 123 external attendees, mainly from Africa and Europe (Figure 8). We invited speakers to cover topics such as epigenetics in plants, especially in trees, viral diseases in cassava, molecular breeding in cassava, and epigenetics for crop improvement.

Ueli Grossniklaus from the University of Zurich in Switzerland was the first speaker and gave an interesting introduction to plant epigenetics, including an overview of results from his laboratory. He then developed the concept of population epigenetics and performance. Next on the program was Frank Johannes from the Technical University of Munich in Germany, who presented results of epigenetic analysis in trees and demonstrated that trees have different DNA methylation profiles in each branch originating from a different bud, as somatic epimutations accumulate continuously as a function of tree age. Claude Becker from the Ludwig Maximilians University in Munich, Germany, presented his group's work on epigenetics in abiotic and biotic stress, developing the concept that epigenetic variation can be associated with defence capacity.

The next two talks dealt with ongoing activities in cassava research. Stefan Winter from DSMZ (German

Gisel et al. (2023) EMBnet.journal 29, e1045
http://dx.doi.org/10.14806/ej.29.0.1045

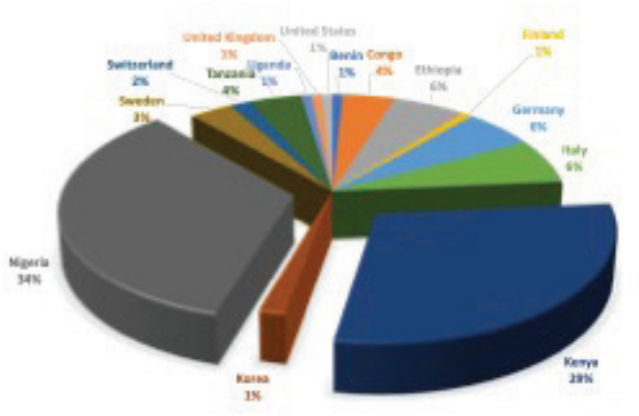

Figure 8. Distribution of countries of external attendees of the mini-symposium.

Collection of Microorganisms and Cell Cultures), who leads the Plant Virus Department in Germany, gave an overview of the viral threats in cassava and the latest research results on Cassava Brown Streak Virus and possible resistance against this virus. This virus is considered the biggest threat to food security in coastal Eastern Africa and around the eastern lakes and is slowly expanding towards Western Africa. Secondly, to provide an overview of advances in cassava breeding, Adenike Ige, a collaborator of Ismail Rabbi from IITA in Ibadan, Nigeria, presented the latest developments in marker-assisted selection and genetic gain they worked on over the last ten years.

The final presentation was given by Chad Niederhuth from Michigan State University in East Lansing, USA. Chad provided an overview of research in the last few years about epigenetics and possible crop improvement applications, mixed with his group's research results. He summarised that epigenetics could have wide-ranging applications in crop improvement, but we need much more research to uncover clear correlations.

The whole workshop, bioinformatics training and mini-symposium was organized by the EpiCass team, including Erik Bongcam-Rudloff, SLU, Uppsala, Sweden; Renaud van Damme SLU, Uppsala, Sweden; Livia Stavolone IPSP-CNR, Bari, Italy; Laurent Falquet, UniFr, Switzerland, Adnan Niazi, SLU, Uppsala, Sweden; Trushar Shah, IITA, Nairobi, Kenya; Michael Landi, IITA, Nairobi, Kenya, Temitayo Olagunju, IITA, Ibadan, Nigeria; and Andreas Gisel, ITB-CNR, Bari, Italy & IITA, Ibadan, Nigeria.

Acknowledgements

This work was supported by the Swedish Research Council (VR) funded projects, EpiCass grant number 2020-04457 and Cassavanet4Dev grant number 2021-05105.