Communication

Data Literacy in Genome Research

Katharina Wolff , Ronja Friedhoff , Friederieke Schwarzer , Boas Pucker *

Posted Date: 4 September 2023

doi: 10.20944/preprints202309.0079.v1

Keywords: computational biology, genomics, sequencing, data literacy, bioinformatics, education

Communication

Data Literacy in Genome Research

Katharina Wolff [1], Ronja Friedhoff [1], Friderieke Schwarzer [1] and Boas Pucker [1,*]

[1] Plant Biotechnology and Bioinformatics, Institute of Plant Biology & BRICS, TU Braunschweig, Braunschweig, Germany

* Correspondence: authors: Boas Pucker, b.pucker@tu-braunschweig.de.

Keywords: computational biology; genomics; sequencing; data literacy; bioinformatics; education

Summary

With an ever increasing amount of research data available, it becomes constantly more important to possess data literacy skills to benefit from this valuable resource. An integrative course was developed to teach students the fundamentals of data literacy through an engaging genome sequencing project. Each cohort of students performed planning of the experiment, DNA extraction, nanopore sequencing, genome sequence assembly, prediction of genes in the assembled sequence, and assignment of functional annotation terms to predicted genes. Students learned how to communicate science through writing a protocol in the form of a scientific paper, providing comments during a peer-review process, and presenting their findings as part of an international symposium. Many students enjoyed the opportunity to own a project and to work towards a meaningful objective.

Introduction

We live in a world of data with collections and databases growing in size and complexity at an ever increasing pace. Examples are the sequence databases European Nucleotide Archive [1], GenBank [2], and SwissProt [3] that all showed an exponential growth in recent years [4]. Large data sets pose a valuable resource that could be harnessed for scientific discoveries, economic endeavors, and the good of society. Examples are genome sequences of crop wild relatives that can be utilized to identify pathogen resistance genes for introduction into high yield cultivars through breeding [5,6]. However, a lack of knowledge about the methods for exploring data sets and the interpretation of data is an obstacle to gaining the aforementioned benefits. There are also some challenges concerning the access to sensitive data types e.g. in biomedical research [7]. Education about data management and data literacy is largely restricted to dedicated data science study programmes. However, we see a necessity to equip students of other subjects and especially in the life sciences with the necessary skills. This need was also recently summarized in the form of grand challenges in bioinformatics education [8]. Furthermore, it is important to teach data literary and data management skills in close connection to the scientific discipline and not as an isolated subject. The combined teaching in bioinformatics and molecular biology [9] is an example for successful interdisciplinary education. Previous publications also reported success with transdisciplinary approaches used for diverse cohorts [10,11] and the particular importance of a practical methodology and problem-based learning approaches [12].

Genomics is a research field that deals with inherently large data sets. Many researchers in the field are committed to open data and are sharing the sequencing data, genome sequences, and related data sets through well established databases. Examples of important plant genome databases are Phytozome [13], PLAZA [14], banana genome hub [15], Sol Genomics Network [16], and the rice genome hub [17]. While databases like ENA and GenBank are universal, other databases like the banana or rice genome hub support specific communities through focus on a set of closely related species. This free exchange of data allows large scale studies and provides the basis for benchmarking

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2023 doi:10.20944/preprints202309.0079.v1

studies [4,18]. Sequencing of genomes has produced 'big data' for over 20 years [19–21] and there are plans for future activities to systematically study the genomes across the full taxonomic diversity of plants [22,23]. While initial genome sequencing projects were conducted by large international consortia [24], nowadays individual research groups can complete genome sequencing projects [23]. The onset of nanopore sequencing resulted in a democratization of sequencing that enables even students to engage in genomics.

Given the availability of data sets and the accessibility of cutting-edge sequencing technologies, we developed a data literacy course with a focus on genomics. Owing to our background in plant biology, we focus on plant genomics, but the concepts and methods are also applicable to research projects investigating other organisms.

Course content

The objective of this six-week course was to provide students with theoretical knowledge and hands-on experience of an entire genome sequencing project from the planning of the project to the reporting of final results (Figure 1). Theoretical background about genome sequencing strategies, sequencing technologies, and data analysis methods was provided through a lecture in the first week. Following open education principles, the materials of the lecture are freely available through GitHub [25]. The lecture enables the students to plan their own sequencing project. The precise objectives of the different cohorts differed and were subject to availability of plant material and students' preferences. Courses running in the winter were relying on plants that were grown specifically for this purpose. Many cohorts investigated plant pigmentation of some kind, because a visible phenotype is advantageous in a teaching setting.

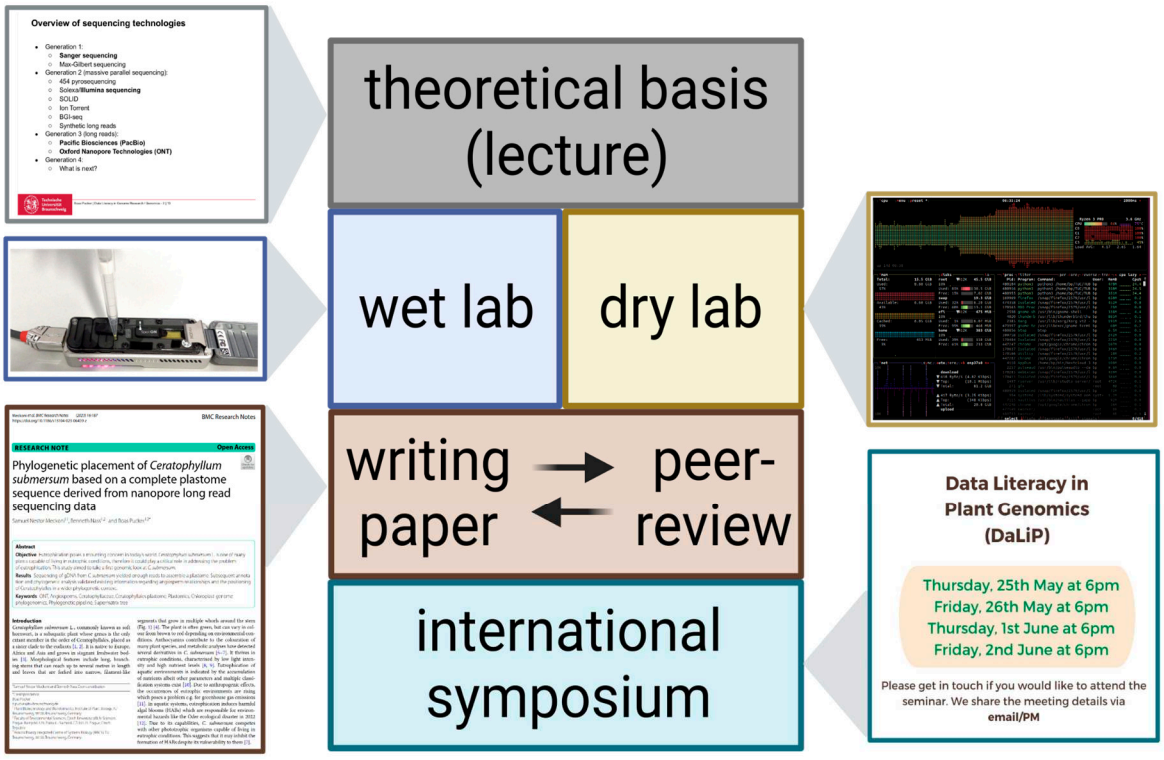

Figure 1. Overall structure of the module comprising lecture, practical parts, paper writing, and oral presentation of the results. Credits for taking a MinION picture and design of the symposium announcement to Melina Nowak. The screenshot belongs to a related open access publication [26].

The practical course part can be separated into two weeks of wet lab work and three weeks of data analysis (Figure 2). The students estimated the genome size by integrating public information about the species and closely related species. Next, the amount of required sequencing data was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2023 doi:10.20944/preprints202309.0079.v1

calculated and served as the basis for the planning of the sequencing experiment. DNA extraction was usually performed based on a universal CTAB protocol [27], but students had the possibility to explore other options. A NanoDrop measurement served as the first quality check. Only DNA containing samples were taken forward to a quality assessment via agarose gel. Students were able to roughly evaluate the fragment size and to determine the amount of RNA contamination on the gel. A precise DNA quantification was performed via Qubit measurement with a dsDNA broad range kit. The whole DNA extraction process was repeated several times to allow students to gain some routine and to see if the quality of the outcome improves.

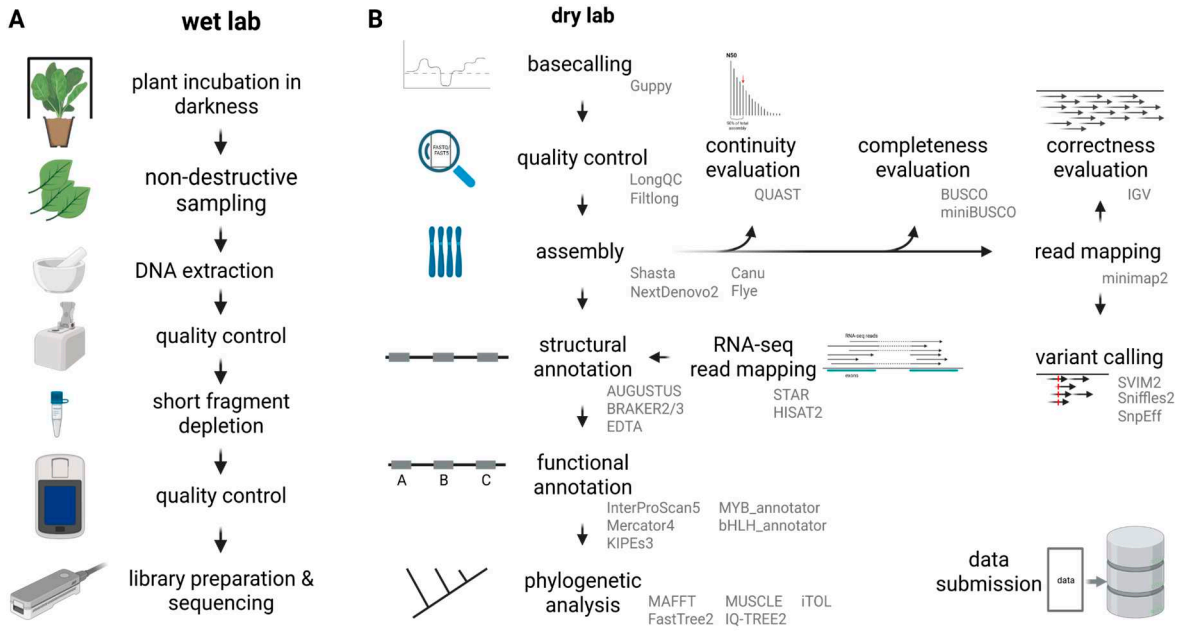

Figure 2. Graphical summary of practical course content. (A) Steps of the sequencing workflow performed in the wet lab. (B) Data analysis steps performed with bioinformatics tools. Names of selected tools are given for the different data analysis and processing steps.

DNA samples passing through all these iterative filters were subjected to removal of short DNA fragments with the Short Read Eliminator kit (Circulomics/PacBio). During the process students learnt to make decisions about the next steps based on quality check results. Finally, students picked samples that were suitable for library preparation and nanopore sequencing. Students were provided with the SQK-LSK109 and SQK-LSK110 protocols for library preparation. Another Qubit measurement was performed prior to loading flow cells to ensure that the process was completed successfully. Students prepared R9.4.1 flow cells for sequencing and loaded their successfully prepared libraries. Sequencing was performed for about 16 hours in most cases. The course was designed in a way that multiple students share a flow cell. One student started a sequencing run on a flow cell on the first day. The next student performed a flow cell washing on the next day and prepared and loaded a new library on the flow cell. Following this strategy, up to four students were able to use one flow cell.

During the first cohorts, the conversion of raw nanopore sequencing data into actual sequences (basecalling) was performed with Guppy (ONT) using a graphic card in the de.NBI cloud. After upgrading the local computational resources, real time basecalling was performed to allow students the monitoring of even more parameters during the sequencing experiments. Next, students assessed the quality of the generated data sets. LongQC [28] and Filtlong [29] were often applied for this step. Afterwards, the reads were subjected to a genome sequence assembly with different tools. Students were encouraged to identify suitable tools for this step. Frequently, Shasta [30] was deployed due to its short run time and low computational costs. Cohorts working on plant species with smaller genomes also tried other assemblers like Canu [31], NextDenovo2 [32], or Flye [33]. Different tools

 doi:10.20944/preprints202309.0079.v1

were deployed to evaluate the quality of the resulting assemblies with respect to the three 'C's: continuity, completeness, and correctness. While some cohorts relied on basic Python scripts for counting contig numbers and N50 calculation (FASTQ_stats3_graphical_v0.2.py, [26]), others applied QUAST [34]. BUSCO [35,36] and compleasm [37] were frequent choices for the completeness assessment. Individual students tried to evaluate the assembly correctness through coverage analyses based on long read mapping with minimap2 [38] and alignment inspection with Integrative Genomics Viewer (IGV) [39,40].

While the assembly of plant genome sequences is transitioning into a routine task, the structural and in particular the functional annotation are becoming the new bottlenecks [23]. Assembled genome sequences were structurally annotated with AUGUSTUS [41,42], BRAKER2/3 [43,44], or EDTA [45]. Since AUGUSTUS and BRAKER focus on protein coding genes, EDTA was applied to annotate transposable elements that comprise a substantial amount of the genome and consequently account for a substantial amount of the genome sequence. AUGUSTUS can perform a gene prediction without the integration of additional hints. BRAKER is based on AUGUSTUS and permits the use of RNA-seq data sets or sequences of a close relative as gene prediction hints. The mappings of RNA-seq reads for the generation of hints for the gene prediction process were performed with STAR [46,47] or HISAT2 [48]. Students had the opportunity to explore other gene prediction tools as well.

The functional annotation was largely based on the identification of orthologs in *A. thaliana* and a following transfer of the TAIR annotation terms. BLASTp [49,50] was applied for the identification of reciprocal best BLAST hits (RBHs) which can serve as indicators for ortholog connections [51]. A general annotation of all predicted peptide sequences was also performed based on InterProScan5 [52] and Mercator4 [53,54]. Additional annotation steps were performed based on the specific research question in the respective cohort. Projects focussing on flower pigmentation performed an in-depth annotation of the flavonoid biosynthesis via KIPEs [55]. Additionally, students analyzed the MYB and bHLH transcription factor families, which harbor important transcriptional regulators of the flavonoid biosynthesis, with dedicated tools [56,57].

The relationship of individual candidate sequences was studied based on phylogenetic trees constructed by FastTree v2 [58] or IQ-TREE v2 [59] usually based on MAFFT v7 [60] or MUSCLE v5 [61] alignments. A customized Python script algntrim.py [62] or pxclsq [63] were applied for the alignment trimming i.e. to remove columns with low occupancy. Finally, the phylogenetic trees were visualized in iTOL [64]. Students learned how to root a tree and how to color different elements to highlight specific aspects.

An optional element in several courses was the detection of sequence variants based on a mapping of reads against a reference genome sequence. Depending on the particular research question in the respective course, this element was included. Students were encouraged to identify tools for the mapping of long reads. A frequently selected tool was minimap2 [38]. Conversion of SAM to BAM file format and other BAM processing steps were conducted with samtools [65]. SVIM2 [66] and Sniffles2 [67] were applied for the identification of sequence variants between the long reads and the reference genome sequence. The functional consequences of the detected sequence variants were predicted with SnpEff [68].

Students wrote their report about the entire course content in the form of a scientific paper. Background about the respective project and the fundamental research question were presented in the introduction. All methods applied by the students were presented in the method section. With respect to reproducibility, students were encouraged to include all necessary details like tool version number and parameters used in their analyses. This method section covers the wet lab and dry lab part of the course. The results section showcases the sequencing output and the findings of the following data analysis. The discussion section allowed the student to demonstrate their newly acquired skills by interpreting their large data sets. This requires an integration of various analysis results to answer a biological question. A peer-review process was exercised to assess and further improve the report quality and give the students a first hand experience of the steps involved in the publishing process. If project planning, conduction of all experiments, and data analysis were very successful, students of sequencing courses could submit their results for publication [26].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2023 doi:10.20944/preprints202309.0079.v1

An online symposium with an international audience concluded the practical courses. This was the opportunity for students to demonstrate their ability to present their main findings in a concise presentation and to defend their choice of methods in the following discussion. All students of a cohort worked on the same biological question thus their reports would only differ in the methodological approaches they applied. To avoid strong redundancy throughout an international symposium, participants of different cohorts were mixed.

Teaching strategy

The design of this course considered the integration of numerous teaching strategies that were previously tested, evaluated, and established with addition of ideas that were developed and evaluated over six cohorts of this course (Figure 3). An interdisciplinary project allows the students to connect their biology skills and knowledge with competencies in bioinformatics as this has proven successful before [9]. Students can participate in this course without comprehensive knowledge about various laboratory methods, because the project allows for an individual learning pace. There are also several time slots within come protocols (e.g. 30 minutes centrifugation) that can be utilized to practice skills for the following steps or to go over theoretical concepts of complicated steps. A solid understanding of the wet lab and dry lab part of a project enables the students to communicate effectively with both molecular biologists and bioinformaticians. There are also synergistic effects as students improve in both fields. Although all students performed all tasks on their own to get hands-on experience, there were plenty of opportunities for exchange and collaborative learning. Writing a protocol about the entire course in the form of a scientific research article allows the students to apply all their newly acquired skills.The result was assessed and criticized by peers. This provides the students with suggestions for improvements and also trains their ability to provide constructive comments. Peer-reviews have been successfully used as teaching methods before [69]. Finally, students practiced presentation and communication skills when sharing their work and derived insights with an international audience as part of an online symposium. The students learned how to give a scientific presentation and how to interact with other scientists.

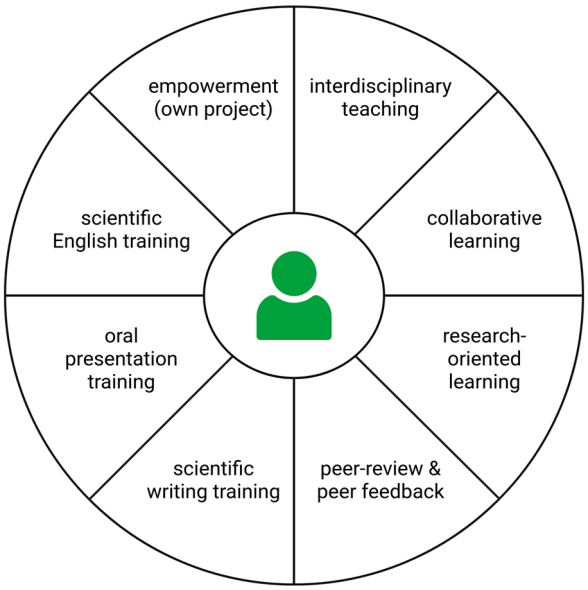

Figure 3: Summary of the didactic elements embedded in the course. (1) Interdisciplinary teaching allows the students to connect elements of different subjects and think about a specific project without borders imposed by different subjects. (2) Collaborative learning has many advantages, one of which is an increased motivation. (3) Research-oriented learning makes it obvious to students why they need to acquire certain skills, which again boosts motivation. It provides also a solid basis for upcoming challenges in a research career. (4) Utilizing peer-review and peer-feedback are successful teaching methods, because students can benefit from providing and from receiving

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2023 doi:10.20944/preprints202309.0079.v1

comments. We utilized these methods throughout the course at various stages. (5), (6), and (7) were all connected to the final presentation of the project in the form of a scientific publication and an oral presentation during an international symposium. (8) Empowering students by allowing them to develop their own project resulted in a motivation boost that we consider crucial for the success of this course.

Lessons learned and recommendations

The described course was offered six times at TU Braunschweig during the period of one year with a total of up to 12 slots available per cohort. After participation in the course, students were asked to provide feedback for further improvements and to highlight strengths of the concept. Here we summarize their comments and derive some recommendations for future courses. Improvements were made between cohorts according to the feedback that students provided after participating in the course. While this ensures an ideal teaching outcome, it prevents a quantitative evaluation. It is also important that each cohort might have different and sometimes conflicting preferences concerning specific aspects. Nevertheless, some overarching points for improvement were identified through observation during the course and evaluation meetings after each course (Figure 4).

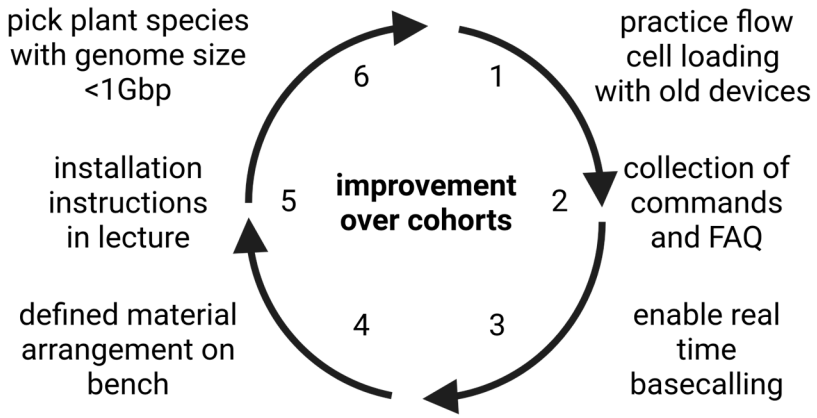

Figure 4. Continuous improvement of the Data Literacy in Genome Research course based on feedback and experience from six consecutive cohorts. (1) Extensively practicing the flow cell loading with old devices avoided the destruction of expensive materials, because the introduction of a single air bubble can already destroy a flow cell. (2) A collection of all fundamental commands enabled all students to start working in a linux environment within the virtual machine more efficiently. A collection of the most frequently appearing errors and questions (FAQ) reduced the need to answer these questions redundantly. (3) Upgrading the computational resources to enable real time basecalling provided the students with additional monitoring options during the sequencing. (4) Defining the arrangement of different reagents and equipment on the bench reduced the chances of students dropping those expensive items. (5) Extending the introduction about software installation reduced frustration when students were facing too complex challenges. (6) Restricting the investigation of plants to genome sizes <1 Gbp helped to avoid too long waiting times when tools are running.

While students appreciated the opportunity to pick their own tools for bioinformatic analyses, the installation of these tools was a substantial challenge. This holds especially true for life science students who usually have never installed any software on a Linux system before. Therefore, a collection of the most important Linux commands was prepared and shared with students (Supplementary File 1). Spontaneous short presentations given by students were encouraged to facilitate exchange between the students who might decide to work with different tools. This provides extended insights into one of the big obstacles to successful bioinformatics research: installation of novel tools. As some students asked for additional instructions, an introduction into software installation was included in the lecture. This includes an extension of the section about

sharing tools to specifically introduce the general structure of a GitHub repository which is one of the most frequently used platforms to share bioinformatic tools. Virtual environments are very important for the contained installation of new tools. Therefore, additional background information about configuring virtual environments was also added to the lecture. The practical part was extended by a short hands-on session about the installation of selected bioinformatics tools to demonstrate frequently used installation strategies. Students also asked for the presentation of more tools for the various applications during the lecture. Consequently, we updated the lecture and presented additional tools for some steps and added the presentation of tools for other steps which were previously presented without any tool recommendations. Additionally, one standard tool was defined for each analysis and students were asked to start their analysis with this tool before looking for alternatives. Nevertheless, we preserve the opportunity that students can identify their own tools and we encourage the students to do so.

Many learning successes are the consequence of errors. However, not every student needs to repeat all possible errors to gain the best learning outcome. Instead, we compiled a collection of frequently asked questions and frequently appearing errors (Supplementary File 1) which helps students to also learn from the errors of others. An example is that students should prepare a document containing all their commands. This needs to be a plain text file (TXT) and not a Word document (DOCX). Learning about these tiny details can help students to avoid frustration. Also, students are encouraged to construct a graphical overview of all tools to indicate how every element contributes to the entire workflow.

To cover one additional aspect of data literacy, we included some strategies about literature search in the lecture. Students are now informed about the most important databases and search strategies. They can put their theoretical knowledge to practice by searching for publications pertinent to the specific project of the cohort. There is allocated time during the practical course part to ensure that all students gain sufficient experience with literature search, which was appreciated by the students. This was crucial when students were asked to identify potential tools for a certain analysis and students appreciated these opportunities to learn how to solve these little data analysis challenges. Experiencing success on a daily basis enables students to feel confident in future analyses and has substantial advantages over a predefined set of commands that all students have to execute. Students liked the opportunity to explore their own ideas, to test alternative tools, and to follow up on their own research questions. As this does not require any additional preparations, we believe that this high degree of flexibility is transferable to many other courses.

Students asked for more discussion of the obtained results in the group. Therefore, we introduced short presentations by all participants about specific results that they obtained. All students were provided with the opportunity to comment on the results of others and also received feedback regarding their own work. This applies to the selected methods, the obtained results, and the interpretation of the results. We extended this and now ask students to summarize the content of each day at the end of the day and also at the beginning of the next day. In a similar way, the content of a week is summarized on Friday before leaving for the weekend. As repetition is an important element of learning, we expect to improve the long term learning outcome through these measures.

Although the specific research question is different for each cohort, some elements remain the same and allow the reuse of teaching materials. We consider different parts of the course as modules that can be combined in different ways. However, it is important to note that the precise run time of tools depends largely on the provided data set which can pose a challenge. In our experience, a genome size above 1 Gbp is not suitable for most analyses in the practical course. Based on a script for the practical part and instructions for the bioinformatics section, the students are challenged to develop their own summarizing documentation (Supplementary File 2).

Students appreciated that they generated their own sequencing datasets and that all following analyses were based on their own datasets. It created a feeling of responsibility and also achievement when completing all required analyses. The motivational boost resulting from the continuous work

 doi:10.20944/preprints202309.0079.v1

on a specific dataset can help to increase the learning outcome. The analysis of complementary data sets by different groups of students was considered an excellent solution that makes the course even more exciting.

Close supervision of up to 12 students by two experienced plant scientists during the DNA extraction parts was described as excellent. The sequencing supervision was also perfect with one experienced scientist supervising up to 3 students at a time. The inclusivity and failure culture in the courses was rated as excellent by many students. Many students learned a lot by making mistakes. The speed and difficulty of the course were constantly adjusted to meet the requirements of the respective cohort. Many students recommended reducing the course capacity to 5-8 students per cohort. That would ensure a close supervision leading to optimal outcome. The calculation that four students would be able to share a flow cell was too optimistic for realistic course settings. While this would work in an ideal situation, some flow cells were destroyed by students, when they accidentally introduced air bubbles into the system. Despite all precautions, it is probably more realistic to assume that on average two students can share a flow cell.

Some students have already applied recently developed large language models like ChatGPT v3 (https://chat.openai.com/auth/login) or the Bing chat via Microsoft Edge (https://news.microsoft.com/the-new-Bing/). These artificial intelligence (AI) tools were rated as very helpful tools when searching for installation instructions. However, there are still cases where these LMMs fail to provide an accurate answer to a specific question. Future AI development will enable students to write their data analysis scripts with AI support and fundamentally change the way we need to teach coding and bioinformatics.

The proportion of one week with 3-4 hours of lectures per day followed by two weeks of wet lab work and three weeks of bioinformatics was rated as very good by most students. Especially the combination of wet lab and dry lab work in the same course was appreciated. Many life science students have not experienced dry lab work before and would not have selected a course on the topic. The combination of wet lab and dry lab can be an elegant way to expose such students to bioinformatics. Many students also liked the opportunity to experience all stages of the project from designing the experiments, performing the data generation and data analysis to the final presentation of their findings in the form of a written and oral report.

Sharing all course materials freely through GitHub [25] and other platforms is important for the students. This enables the students to access everything even after leaving TU Braunschweig which results in loss of access to all internal file exchange systems.

We evaluated the integration of different settings for the bioinformatic analysis part. Flipped classroom and other concepts were tested. However, the students preferred to work on site in a seminar room. We consider this an important finding, because our expectation was that students would benefit from the increased flexibility and the saved time by not commuting to a university building. They explained that peer pressure is important for many of them to actually work on the project. The students suggested an alternative to presence in a seminar room: an online meeting where all students keep their video on.

Conclusion

This ‘Data Literacy in Genome Research’ course was developed as a teaching innovation connecting different subjects and enabling students to develop their own project. Over the course of six weeks, the students gained theoretical background knowledge and hands-on experiences in genomics and computational biology. We identified a number of key elements like the ‘own project’, peer-feedback, and opportunity for individual approaches. Now, we are utilizing our experiences to establish a permanent course with similar content. In line with open education concepts, we make our teaching materials and data sets freely available to the community.

Supplementary Materials**:** Supplementary File 1: Basic commands for working in a virtual machine on linux. Supplementary File 2: Documentation of the entire workflow by a course participant

Authors' contribution: BP acquired funding for this teaching innovation. KW and BP designed the course 'Data Literacy in Genome Research'. KW, RF, and BP supervised the students during their wet lab work. KW supervised the students during the computational biology part. FS participated in the second cohort of the course, provided example documentation, and contributed the students' perspective. BP wrote the manuscript with input from all authors. All authors read the final version of the manuscript and approved its submission.

Acknowledgements: This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). We acknowledge funding from the 'Stiftung Innovation in der Hochschullehre' for our project 'DaLiP' (FRFMM-13/2022) through 'Freiraum2022'. bioRender.com was used to construct some of the figures. We would like to thank all students who participated in the course and agreed to have their anonymised feedback summarized in this publication.

Conflict of interest statement: KW, RF, and FS declare no conflicts of interest. BP is head of the technology transfer center Plant Genomics and Applied Bioinformatics at iTUBS.

References

1. EMBL-EBI. European Nucleotide Archive. 2023. https://www.ebi.ac.uk/ena/browser/home. Accessed 23 Jul 2023.
2. NCBI. GenBank. 2023. https://www.ncbi.nlm.nih.gov/genbank/. Accessed 23 Jul 2023.
3. Coudert E, Gehant S, de Castro E, Pozzato M, Baratin D, Neto T, et al. Annotation of biologically relevant ligands in UniProtKB using ChEBI. Bioinformatics. 2023;39:btac793.
4. Sielemann K, Hafner A, Pucker B. The Reuse of Public Datasets in the Life Sciences: Potential Risks and Rewards. 2020.
5. Zhang H, Mittal N, Leamy LJ, Barazani O, Song B-H. Back into the wild—Apply untapped genetic diversity of wild relatives for crop improvement. Evol Appl. 2017;10:5–24.
6. Capistrano-Gossmann GG, Ries D, Holtgräwe D, Minoche A, Kraft T, Frerichmann SLM, et al. Crop wild relative populations of Beta vulgaris allow direct mapping of agronomically important genes. Nat Commun. 2017;8:15708.
7. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25:37–43.
8. Işık EB, Brazas MD, Schwartz R, Gaeta B, Palagi PM, van Gelder CWG, et al. Grand challenges in bioinformatics education and training. Nat Biotechnol. 2023;41:1171–4.
9. Pucker B, Schilbert HM, Schumacher SF. Integrating Molecular Biology and Bioinformatics Education. J Integr Bioinforma. 2019;16.
10. Dorn M, Ligabue-Braun R, Verli H. Transdisciplinary Approach for Bioinformatics Education in Southern Brazil. Front Educ. 2021;6.
11. Johnston IG, Slater M, Cazier J-B. Interdisciplinary and Transferable Concepts in Bioinformatics Education: Observations and Approaches From a UK MSc Course. Front Educ. 2022;7.
12. Garzón A, Rubio A, Pérez-Pulido AJ. E-learning strategies from a bioinformatics postgraduate programme to improve student engagement and completion rate. Bioinforma Adv. 2022;2:vbac031.
13. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40 Database issue:D1178–86.
14. Van Bel M, Silvestri F, Weitz EM, Kreft L, Botzki A, Coppens F, et al. PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. Nucleic Acids Res. 2022;50:D1468–74.
15. Droc G, Martin G, Guignon V, Summo M, Sempéré G, Durant E, et al. The banana genome hub: a community database for genomics in the Musaceae. Hortic Res. 2022;9:uhac221.
16. Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, et al. The Sol Genomics Network (SGN)--from genotype to phenotype to breeding. Nucleic Acids Res. 2015;43 Database issue:D1036-1041.
17. Rice Genome Hub. Rice Genome Hub. 2023. https://rice-genome-hub.southgreen.fr. Accessed 23 Jul 2023.
18. Schilbert HM, Rempel A, Pucker B. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants. 2020;9:439.
19. Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20 years of plant genome sequencing. Nat Plants. 2021;7:1571–8.
20. Sun Y, Shang L, Zhu Q-H, Fan L, Guo L. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 2022;27:391–401.
21. Kress WJ, Soltis DE, Kersey PJ, Wegrzyn JL, Leebens-Mack JH, Gostel MR, et al. Green plant genomes: What we know in an era of rapidly expanding opportunities. Proc Natl Acad Sci. 2022;119:e2115640118.
22. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, et al. 10KP: A phylodiverse genome sequencing plan. GigaScience. 2018;7:giy013.
23. Pucker B, Irisarri I, Vries J de, Xu B. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quant Plant Biol. 2022;3:e5.

24. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.
25. Pucker B. Data Literacy In Genome Research. GitHub. 2023. https://github.com/bpucker/teaching/tree/master/FRX_DataLiteracyInGenomeResearch. Accessed 20 Jul 2023.
26. Meckoni SN, Nass B, Pucker B. Phylogenetic placement of Ceratophyllum submersum based on a complete plastome sequence derived from nanopore long read sequencing data. BMC Res Notes. 2023;16:187.
27. Siadjeu C, Pucker B, Viehöver P, Albach DC, Weisshaar B. High Contiguity de novo Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long Read Sequencing. Genes. 2020;11:274.
28. Fukasawa Y, Ermini L, Wang H, Carty K, Cheung M-S. LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. G3 GenesGenomesGenetics. 2020;10:1193–6.
29. Wick R. Filtlong. 2023.
30. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38:1044–53.
31. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
32. GrandOmics. NextDenovo. 2023.
33. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.
34. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinforma Oxf Engl. 2013;29:1072–5.
35. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
36. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54.
37. Huang N, Li H. miniBUSCO: a faster and more accurate reimplementation of BUSCO. 2023;:2023.06.03.543588.
38. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
39. Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics. 2023;39:btac830.
40. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol. 2011;29:24–6.
41. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34 suppl_2:W435–9.
42. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
43. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 2021;3:lqaa108.
44. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. 2023;:2023.06.10.544449.
45. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.
46. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
47. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinforma. 2015;51:11.14.1-11.14.19.
48. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
50. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
51. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, et al. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLOS ONE. 2019;14:e0216233.
52. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.

53. Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, Hallab A, et al. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Mol Plant. 2019;12:879–92.
54. Bolger M, Schwacke R, Usadel B. MapMan Visualization of RNA-Seq Data Using Mercator4 Functional Annotations. In: Dobnik D, Gruden K, Ramšak Ž, Coll A, editors. Solanum tuberosum: Methods and Protocols. New York, NY: Springer US; 2021. p. 195–212.
55. Rempel A, Choudhary N, Pucker B. KIPEs3: Automatic annotation of biosynthesis pathways. 2023;:2022.06.30.498365.
56. Pucker B. Automatic identification and annotation of MYB gene family members in plants. BMC Genomics. 2022;23:220.
57. Thoben C, Pucker B. Automatic annotation of the bHLH gene family in plants. 2023;:2023.05.02.539087.
58. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010;5:e9490.
59. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37:1530–4.
60. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.
61. Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun. 2022;13:6968.
62. Pucker B, Iorizzo M. Apiaceae FNS I originated from F3H through tandem gene duplication. PLOS ONE. 2023;18:e0280155.
63. Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for unix. Bioinformatics. 2017;33:1886–8.
64. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49:W293–6.
65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
66. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. 2019;35:2907–15.
67. Smolka M, Paulin LF, Grochowski CM, Mahmoud M, Behera S, Gandhi M, et al. Comprehensive Structural Variant Detection: From Mosaic to Population-Level. 2022;:2022.04.04.487055.
68. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6:80–92.
69. Friedrich A, Pucker B. Peer-review as a teaching method. working Paper. 2018.