Peer-reviewed version available at *Journal of Integrative Bioinformatics* **2019**; doi:10.1515/jib-2019-0005

Integrating molecular biology and bioinformatics education

Boas Pucker[1*], Hanna Marie Schilbert[1], Sina Franziska Schumacher[2]

1 Genetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, Bielefeld, Germany

2 Bielefeld University, Bielefeld, Germany

BP: bpucker@cebitec.uni-bielefeld.de

HMS: hschilbe@cebitec.uni-bielefeld.de

SFS: sina.schumacher@uni-bielefeld.de

ORCID:

BP: 0000-0002-3321-7471

HMS: 0000-0003-0474-7753

SFS: 0000-0003-1471-1287

* Corresponding author: Boas Pucker, bpucker@cebitec.uni-bielefeld.de

Key words: sequencing technologies, NGS, genome research, genome assembly, variant calling, RNA-Seq, transcriptome assembly, bioinformatics, molecular biology, education

Abstract

Combined awareness about the power and limitations of bioinformatics and molecular biology enables advanced research based on high-throughput data. Despite an increasing demand for scientists with a combined background in both fields, the education in dry lab and wet lab is often separated. This work describes an example of integrated education with focus on genomics and transcriptomics. Participants learn computational and molecular biology methods in the same practical course. Peer-review is applied as a teaching method to foster cooperative learning of students with heterogeneous backgrounds. Evaluation results indicate acceptance and appreciation of this approach.

Introduction

There is an increasing demand for life scientists trained in both, molecular biology and bioinformatics [1–3]. Academia and industry are trying to find candidates with a strong combined background. Although there are numerous study programs which are addressing this demand for bioinformaticians [3, 4], single courses at a university are usually focused either on the wet lab or the dry lab site. Frequently, lecturers with a bioinformatics background teach the bioinformatics aspect, while biologists teach the molecular biology part. Probably as a result of this strict separation, many students tend to be substantially more interested in one aspect of their program e.g. focusing on bioinformatics causing a lack of knowledge about biology. Truly combining both aspects in a single course by looking at both sites of an experiment could help to reduce the separation of wet lab and dry lab thinking finally leading to a new awareness [5]. In addition, bioinformatics students as well as life science students can be interested in such a course thus facilitating exchange and cooperative learning between students with different educational backgrounds [6].

Combining substantial knowledge and experience about bioinformatics and biology in a single person, would lead to very powerful and urgently needed scientists [1, 3, 7, 8]. These scientists are not just able to communicate efficiently with scientists form both fields, but are even able to address most challenges on their own [9]. When generating bioinformatics predictions, they can adjust the output to facilitate experimental validation in the wet lab. The awareness of possibilities and limitations of methods in both fields is very important for successful projects. Due to a continuous increase in publicly available data sets, the ability to harness their power efficiently is gaining relevance [9]. There is a broad range of different topics that could be included in a bioinformatics education program [10] thus focus on a certain field seems necessary when discussing best practices.

This work describes the concept and content of two courses, which are committed to integrate molecular biology and bioinformatics education with focus on genomics and transcriptomics. Presented results are the experiences from these courses over the last two years. The intention is to provide an inspiring example to lecturers at universities.

Course concept and content

Complementary courses

This approach to educate students about the web lab and dry lab aspects of genome research was developed over the last three years and resulted in two courses which complement each other. First, a course about bioinformatics methods (‘Applied Genome Research’, https://github.com/bpucker/AppliedGenomeResearch) was substantially enriched with molecular biology content. Second, a molecular biology course was enriched with bioinformatics methods to mirror this concept from the web lab site (‘Molecular Methods in Genome Research’, https://github.com/bpucker/MolecularMethodsInGenomeResearch) (Fig. 1). Both courses were designed to attract bioinformatics students as well as life science students and to increase their

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2018 doi:10.20944/preprints201811.0183.v2

Peer-reviewed version available at *Journal of Integrative Bioinformatics* **2019**; doi:10.1515/jib-2019-0005

engagement with the other field. In addition, exercises in these courses require often knowledge from both fields.

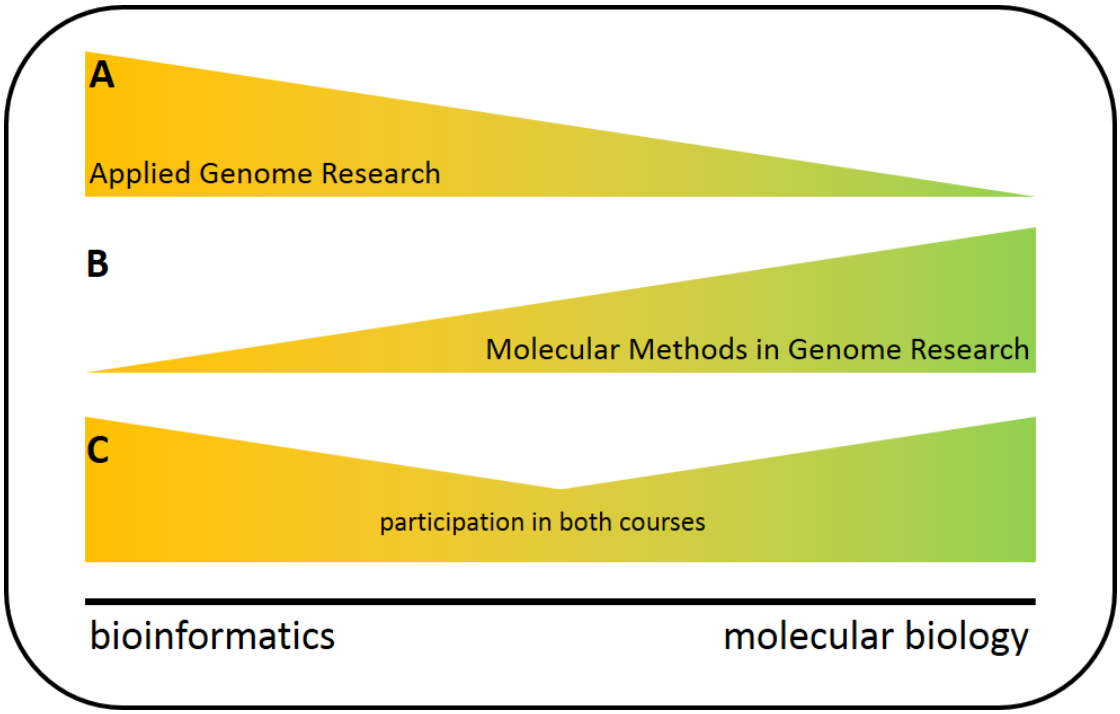


Figure 1: Course content. Complementing design of two courses integrates bioinformatics and molecular biology education. The proportion of bioinformatics content (yellow) and molecular biology content (green) is illustrated for the courses 'Applied Genome Research' (A), 'Molecular Methods in Genome Research' (B), and for the combination of both courses (C).

Applied Genome Research

The content of this course is separated into a genomics section and a transcriptomics section (Fig. 2). There are also three layers involved in this teaching process: general concept/aim, method/tool, and the material/data type. Since some participants have a pure life science background without any prior knowledge in bioinformatics, a short introduction into LINUX was given to achieve familiarity with commands in a terminal.

Starting the genomics section with the biological challenge of isolating DNA (plasmids, BACs, genomic) of sufficient quality and quantity, the introduction provides background knowledge about sequencing technologies and relevant file formats. Next steps were composed to reconstitute a real workflow in

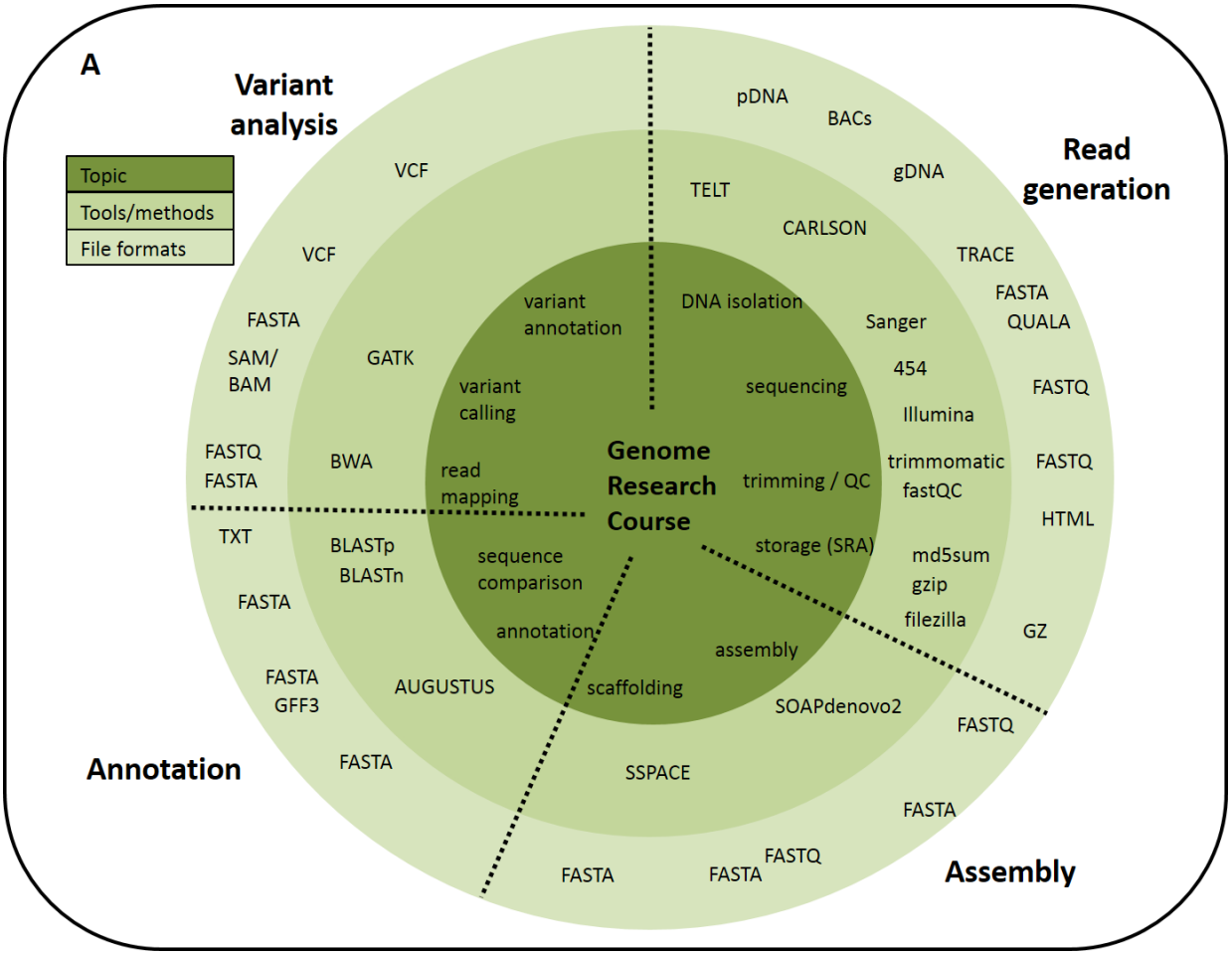
plant genome research [11] including preparation for submission to a standard repository like the Sequence Read Archive [12], trimming of reads via trimmomatic [13], and quality control via fastQC [14].

Since the computation of a plant genome assembly consumes a substantial amount of time and computational resources, the sequencing read data set was reduced to a subset just representing about 3 Mbp of the *Arabidopsis thaliana* Niederzenz-1 (Nd-1) genome sequence [11]. Generating an assembly via SOAPdenovo2 [15] and assessing different ways of scaffolding were the next steps. Exercises and discussion about the performance of different tools and the impact of certain parameters were a central teaching part. AUGUSTUS [16] was applied for structural gene prediction and BLAST [17] was used in supplied Python scripts [11] to transfer functional annotations to the predicted genes. This whole process of genome annotation was accompanied by discussions about the biological interpretation of results, possible pitfalls, alternatives, and next steps.

As high quality reference genome sequences become available, *de novo* assemblies are often replaced by read mappings against an existing reference thus enabling the investigation of populations [18]. Therefore, the next step was the mapping of the above described Nd-1 sequencing reads via BWA MEM [19] against a reference sequence (TAIR10, [20]). Variants were called via GATK [21] and functional implications were predicted using SnpEff [22] and NAVIP (https://github.com/bpucker/NAVIP). The tools applied in this course are not necessarily the best performing ones for a specific step, but overall provide the experience of running a complete genomics workflow. While initially the usage of tools is explained in great detail, students were continuously trained to retrieve usage information from the documentation to facilitate independent application of various bioinformatic tools.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2018 doi:10.20944/preprints201811.0183.v2

Peer-reviewed version available at *Journal of Integrative Bioinformatics* **2019**; doi:10.1515/jib-2019-0005



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2018 doi:10.20944/preprints201811.0183.v2

Peer-reviewed version available at *Journal of Integrative Bioinformatics* **2019**; doi:10.1515/jib-2019-0005

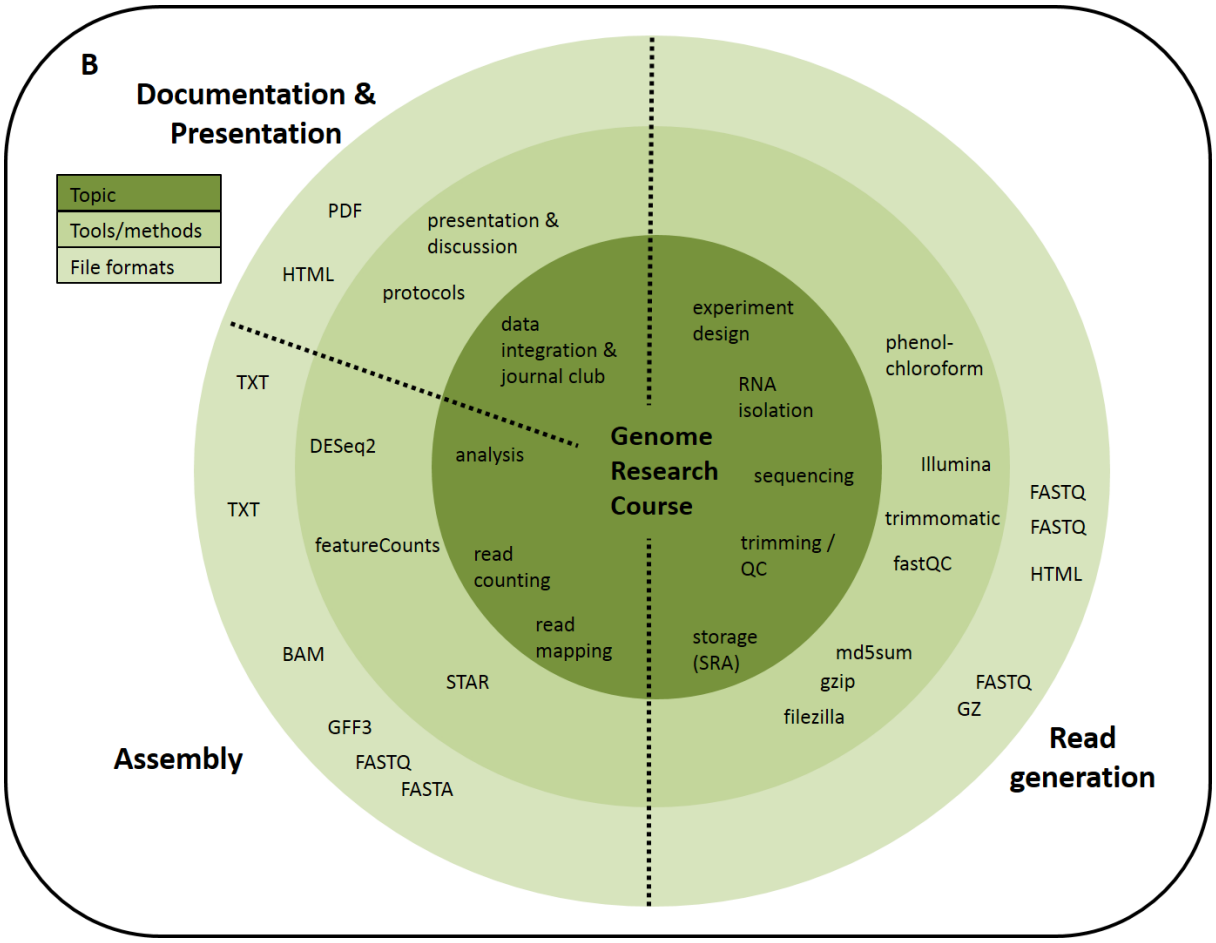


Figure 2: Applied Genome Research course content overview. The content of this course is distributed over two weeks: one genomic (A) and one transcriptomic (B) week. The inner circle contains topics, the middle circle contains methods and tools, and the outer circle contains materials and file formats. Abbreviations in these figures (excluding tool and file format names): plasmid DNA (pDNA), Bacterial Artificial Chromosome (BAC), genomic DNA (gDNA), Sequence Read Archive (SRA).

The transcriptomics part started with an introduction to experiment design and RNA isolation. Differences between DNA and RNA processing were discussed. Since some steps are redundant with the genomics part, it was a helpful repetition. Mapping RNA-Seq reads via STAR [23] and quantifying expression of genes with featureCounts [24] were the first practical steps. To reduce the computational costs associated with the RNA-Seq read mapping, replicates were randomly generated using a customized Python script. Afterwards, DESeq2 [25] was applied for statistical analysis of the results. Different ways to interpret the results were discussed and participants engaged with databases of different model organisms e.g. Araport11 and TAIR10. Besides gene expression analysis, RNA-Seq reads were also used for a transcriptome assembly workflow [26]. Differences between genome and transcriptome assemblies were discussed to identify unique challenges.

Finally, participants demonstrated their enhanced understanding of genomics and transcriptomics in a journal club during the discussion of scientific publications. Each participant gave an approximately 15 minute talk about a recent publication in the field to complete this course. In addition, participants had to write a report about the course topics, applied methods, and results (SupplementaryFile1). The report quality was increased by double blind peer-review thus each participant assessed and commented on two reports [27]. This assessment of reports facilitated a better understanding of the content and improved important skills e.g. providing constructive criticism about a scientific work.

Molecular Methods in Genome Research

This course was about validating bioinformatics findings through wet lab experiments (Fig. 3). Structural variations between *Arabidopsis thaliana* accessions were previously identified [11] and provided as a start off point. Participating students had a background in biology or bioinformatics without prior knowledge about the other field. Students selected appropriate targets and subjected them to bioinformatics tools and approaches to prepare their experiments. For example, participants extracted the surrounding sequence from assemblies, designed oligonucleotides for PCR assays, and validated these via customized Python scripts based on sequence alignments. These initial steps enabled the acquisition of basic Linux skills. Participants became familiar with running scripts on the command line. As all participants worked on different loci, the following molecular biology experiments were unique as well. Moreover, all participants were working on a unique set of *A. thaliana* accessions taken from the Nordborg collection [28]. As a result, all participants were generating new scientific knowledge contributing to the field of Arabidopsis genomics. To bridge the time for ordered oligonucleotides to arrive, some experiments derived from recent genome research projects [11, 29–31] were repeated on different biological material. Therefore, participants were carrying out actual research with unknown outcome. At the same time, it was possible to include positive controls.

The results were documented online in a wiki (SupplementaryFile2) to facilitate cooperative learning by avoiding isolated lab reports. Students were able to directly interact with each others' work by commenting on the wiki pages. Basic knowledge about HTML and wiki code was provided during seminars. Peer-review was applied to enhance the quality of individual wiki pages thus each participant was assessing the wiki pages of two others. The use of a wiki requires some work during setup, but enables the compliance with data protection laws, which might differ between universities and countries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2018 doi:10.20944/preprints201811.0183.v2

Peer-reviewed version available at *Journal of Integrative Bioinformatics* **2019**; doi:10.1515/jib-2019-0005

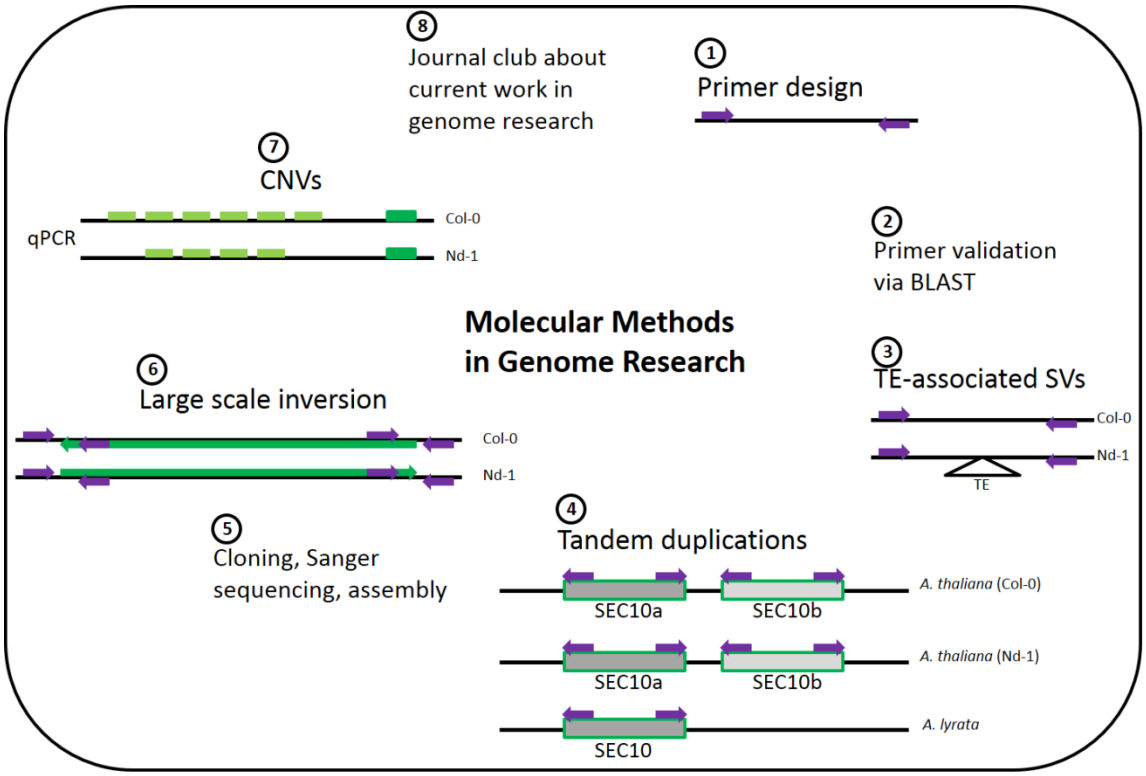


Figure 3: Molecular Methods in Genome Research course content overview. Course content overview displays the interleaved use of bioinformatics and molecular biology.

Lessons learned – evaluation results

Participants were asked to provide feedback about these courses. Some evaluation results of 'Applied Genome Research' were previously described and discussed [27]. Small course sizes (n<10) prevented detailed statistical analyses of these results, but response rates of usually over 50% and repetitions of the courses allowed inference of general trends. All participants would recommend these courses to their fellow students. Usage of peer-review to improve the quality of reports or wiki pages, respectively, was seen as a good approach, but the reviewer qualification was the main concern. Nevertheless, participants stated that they improved several skills like critical reading and providing feedback through this process. In addition, this repetition of the course content was appreciated.

Discussion

The presented courses provide an example for interdisciplinary and innovative teaching methods. Their evaluation indicated participants' satisfaction and a good match with participants' expectations. More detailed evaluation results of two iterations of the 'Applied Genome Research' course were described before with focus on peer-review as a teaching method [27]. In combination with novel insights of more

recent iterations, a more controlled version of this process could further increase the benefit. Currently, a strong heterogeneity in the review quality is a major concern brought up by several participants. Implementing a system in which all reports are evaluated by many peers as it is postulated by many open science movements (reviewed in [32]), could be a solution. Reviewers might be more motivated thus producing better reports when they know that their reports will be published. In addition, errors in reviews could be identified and removed if a large number of peers are inspecting them.

Another important point revealed by the evaluation is the proximity to actual research. Students appeared to be more motivated when working on their own experiments and this was reported before by others [33]. Despite learning valuable skills about experiment design and project management, an extended independence during practical courses could increase the overall interest of students in a subject as well as their self-confidence. However, this comes with higher costs of these innovative experiments. One example is the need for custom oligonucleotides per student as described for the 'Molecular Methods in Genome Research' course. To enable similar courses without external funding, the accumulation of material over years could be the way to go. Some of the materials e.g. oligonucleotides could be used again for following repetitions of a course. Students within one cohort could perform individual experiments, while these experiments are derived from a pool of experiments repeated in every year. In addition, it is feasible that experiments are repeated within one course thus having randomly selected students unknowingly perform the same experiments. This approach enables the validation of results through replicates and can save resources. As all responding students are recommending this course, chances are high to repeat it with success.

Students appreciated the integration of innovative teaching methods. The majority liked the replacement of classical lab reports by digital documentation in a wiki. Although, the application of a wiki as a teaching method is not completely novel [34], it is rarely used in practical courses. It makes students think about displaying their results in an engaging way and connecting them to existing knowledge via links. Learning some HTML basics during the wiki construction is an additional benefit, because students learn the concept of markup languages and the foundation for the development of websites. Finally, the interaction between students with different backgrounds during the peer-review process enables additional exchange and cooperative learning. This provides an opportunity for students to practice science communication very early during their education. They can develop skills that are beneficial and required for future projects in teams.

Although, this example is focused on the combination of bioinformatics with molecular biology, there are other fields in the life sciences, which would benefit from computational methods as well. Therefore, this description is intended to inspire the development of similar courses in other life science fields to facilitate integrated teaching.

Acknowledgements

The authors thank all supporters of this work. Katharina Kemmet and Maximilian Edich supported the 'Molecular Methods in Genome Research' course in the lab. Funding and support for this course was

provided by the Chair of Genetics and Genomics of Plants and through a 'Fellowship for Digital Innovations in Academic Teaching'. The Bioinformatic Resource Facility support team of the CeBiTec provided a wiki and general support. Daniela Holtgräwe provided helpful comments on the manuscript and Nathanael Walker-Hale supported this work by proof-reading.

Authors' contributions

BP wrote the initial draft. BP and HMS revised the manuscript. HMS and SFS wrote SupplementaryFile1 and SupplementaryFile2, respectively. All authors read and approved the manuscript.

References

1. Spotlight on Bioinformatics. NatureJobs. 2016. doi:10.1038/nj0478.

2. Attwood TK, Blackford S, Brazas MD, Davies A, Schneider MV. A global perspective on evolving bioinformatics and data science training needs. Brief Bioinform. doi:https://doi.org/10.1093/bib/bbx100.

3. Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, Gaeta B, et al. Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLOS Comput Biol. 2014;10:e1003496.

4. Ranganathan S. Bioinformatics Education—Perspectives and Challenges. PLoS Comput Biol. 2005;1. doi:https://doi.org/10.1371/journal.pcbi.0010052.

5. Bialek W, Botstein D. Introductory Science and Mathematics Education for 21st-Century Biologists. Science. 2004;303:788–90.

6. Abeln S, Molenaar D, Feenstra KA, Hoefsloot HCJ, Teusink B, Heringa J. Bioinformatics and Systems Biology: bridging the gap between heterogeneous student backgrounds. Brief Bioinform. 2013;14:589–98.

7. Rubinstein A, Chor B. Computational Thinking in Life Science Education. PLOS Comput Biol. 2014;10:e1003897.

8. Goodman AL, Dekhtyar A. Teaching Bioinformatics in Concert. PLOS Comput Biol. 2014;10:e1003896.

9. Via A, Blicher T, Bongcam-Rudloff E, Brazas MD, Brooksbank C, Budd A, et al. Best practices in bioinformatics training for life scientists. Brief Bioinform. 2013;14:528–37.

10. Mulder N, Schwartz R, Brazas MD, Brooksbank C, Gaeta B, Morgan SL, et al. The development and application of bioinformatics core competencies to improve bioinformatics training and education. PLOS Comput Biol. 2018;14:e1005772.

11. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11:e0164321.

12. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011;39 suppl_1:D19–21.

13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30:2114–20.

14. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 14 Dec 2017.

15. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1:18.

16. Hoff KJ, Stanke M. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 2013;41:W123–8.

17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

18. Poland JA, Rife TW. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome. 2012;5:92–102.

19. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013. http://arxiv.org/abs/1303.3997. Accessed 16 Oct 2018.

20. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40 Database issue:D1202–10.

21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

22. 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.

23. 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.

24. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma Oxf Engl. 2014;30:923–30.

25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15. doi:https://doi.org/10.1186/s13059-014-0550-8.

26. Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J, et al. High Quality de Novo Transcriptome Assembly of Croton tiglium. Front Mol Biosci. 2018;5. doi:https://doi.org/10.3389/fmolb.2018.00062.

27. Friedrich A, Pucker B. Peer-review as a teaching method. working Paper. 2018. https://pub.uni-bielefeld.de/record/2920633. Accessed 16 Oct 2018.

28. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, et al. The Pattern of Polymorphism in Arabidopsis thaliana. PLOS Biol. 2005;3:e196.

29. Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, et al. Dissecting a hidden gene duplication: the Arabidopsis thaliana SEC10 locus. PloS One. 2014;9:e94077.

30. Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, et al. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci. 2016;113:E4052–60.

31. Pucker B, Holtgraewe 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. bioRxiv 407627. doi: https://doi.org/10.1101/407627.

32. Tennant JP. The state of the art in peer review. FEMS Microbiol Lett. 2018;365. doi:https://doi.org/10.1093/femsle/fny204.

33. Williams KC, Williams CC. Five key ingredients for improving student motivation. Res High Educ J. https://scholarsarchive.library.albany.edu/cgi/viewcontent.cgi?article=1000&context=math_fac_scholar.

34. Parker K, Chao J. Wiki as a Teaching Tool. Interdiscip J E-Learn Learn Objects. 2007;3:57–72.

Supplementary Material

Supplementary File 1: Report about 'Applied Genome Research' by Hanna Schilbert. This example provides a more detailed impression of the course content and also illustrates how participants perceive it.

Supplementary File 2: Report about 'Molecular Methods in Genome Research' by Sina Franziska Schumacher. This example provides a more detailed impression of the course content and also illustrates how participants perceive it.