Large-scale databases for genomics enable efficient research-integrated training

David A. Duchêne

Department of Public Health
University of Copenhagen

Introduction

Teaching is increasingly becoming tied with cutting-edge research or industry (Beath et al., 2012). The link between teaching and a highly-competitive workforce has led to a renewed marriage between research and teaching, also fueled by the growth of courses at masters and PhD levels (Clark, 1997; Houghton et al., 2021). The most effective courses involve consistent activities, or active learning, which can be challenging to implement in large cohorts, or research fields that are abstract or costly (Michael, 2006). Publicly available data, however, can greatly facilitate active learning and cutting-edge research alike (Thomas & Mancy, 2004).

The extent to which students get actively involved in research can be examined via diagrams of the research-teaching nexus (Healey, 2005). The adoption of very large cohorts at universities has led to teaching with an emphasis on content, where students are a passive audience (Fig. 1). Minor extensions to lecture-based teaching can be research-oriented, for instance where researchers perform experiments or engage in debates. Students might also engage in guided discussions, laboratory experiments, and computational exercises, but these generally have an indirect connection with ongoing research. Limited involvement in research counters existing evidence that graduate students benefit from a curriculum that aligns research, supervision, collaboration, and teaching (Acar & Tuncdogan, 2019; Feldon et al., 2011; Justice et al., 2009).

Involving students in research more profoundly can be done via groups that conceive and carry out an independent project. Feedback from the instructor is then formative, focusing entirely on the process (Nicol & Macfarlane-Dick, 2006). This type of teaching is known as *research-*

integrated teaching or *inquiry-based learning* (Dostál, 2015). This teaching style is often impractical, but can be implemented in data science contexts, where public databases are routinely used for cutting-edge research.

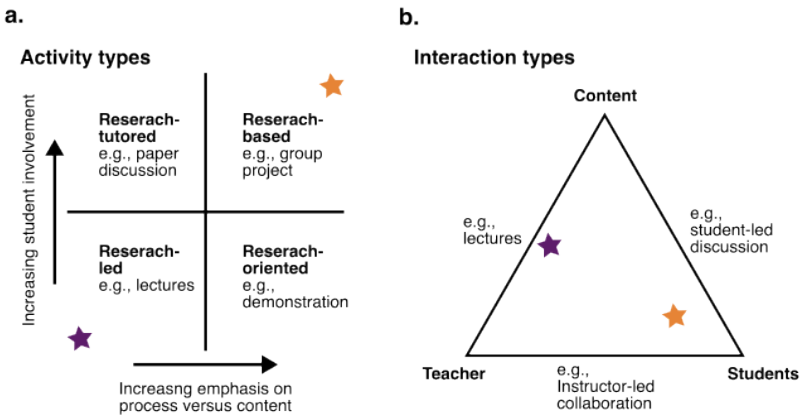

Fig. 1. Models of incorporating research content into teaching, adapted from (Healey, 2005, and Keiding, 2010, *unpublished*), with stars indicating the formats used in the *Molecular Evolution and Phylogenomics* PhD course. (a) Activities can have various degrees of student involvement (*y*-axis) and emphasis on the research process (*x*-axis). (b) The three main types of interactions can differ in strength as represented in a triangle plot. The purple star in indicates lectures delivered in short 30-minute sessions in this course. The orange stars show the research project component of the course.

In both academic and industrial data science, it is routine to carry out short-term projects with some structure but highly flexible content and outcomes. This is often framed as *agile* project development (Alsaqqa et al., 2020). For instance, in the *scrum* framework (Takeuchi & Nonaka, 1986), small cross-functional teams work through the phases of product development, often called *sprints*, passing on responsibilities as seen in rugby. The related CRISP-DM process includes further steps that resemble research, including data preparation, modeling, and reporting (Schröer et al., 2021). When adapted to education, these frameworks can involve small teams with defined tasks and joint steering.

Aiming to advance teaching at the PhD level, I designed and coordinated a new research-integrated course at the University of Copenhagen, with its first instalment held in October 2024. Leveraging publicly available genomic sequence data (e.g., GenBank), I provided a

lightly structured environment for project development where students in small groups had near full autonomy to decide a research question, data set, and analyses in the field of molecular evolution. The intended outcome was a substantive report worthy of submission for publication. Here I focus on the outcomes of this activity, evaluating whether it strengthened or undermined the alignment between course activities and intended learning outcomes.

Methods

A PhD course on *Molecular Evolution and Phylogenomics* was designed to last one week full time. The intended learning outcomes surrounded the preparation and evolutionary analysis of genomic data, and interpretation of results in the field. The fundamental theory behind genomic evolution was delivered via three or four 30-minute lectures daily. On the first day, students were exposed to research talks with state-of the-art findings, and browsed the literature in groups. They also gave one-slide talks on one active question being researched in the field.

Subsequent days involved project development under suggested boundaries of team size and role division (Appendix 1). Groups were planned to have three or four students with the suggested roles of literature searching, data analysis, results presentation, and reporting. Group members were selected at random to minimize any hidden biases in diversified sampling. Two sessions were dedicated to cross-talk between teams to help to tackle major obstacles, and the week ended with presentations on the findings.

Feedback on the course was requested from students twice in the week, and groups were encouraged to develop their projects further towards publication if interested. I also collected the perceptions of students about the style and duration of the course, and on the reasons that groups chose to either pursue further or abandon their projects.

Results and discussion

Twenty-six students enrolled in the *Molecular Evolution and Phylogenomics* PhD course, coming from across faculties and backgrounds in biology, bioinformatics, microbiology, pharmacy,

archaeology, public health, and computer science. Two students dropped out in the last minute, resulting in a project group having two students (Fig. 2a). Students showed high motivation with overwhelmingly positive comments on the course (Appendix 2). Engagement from the instructor with the material and the projects was seen as very positive, consistent with previous findings on the priorities of graduate students when choosing a supervisor (Ives & Rowley, 2005; Ray, 2007). A minority of students wished to extend the course to two weeks (Fig. 2b). Four of the seven groups were strongly encouraged to pursue their project towards publication, and two agreed (Fig. 2c).

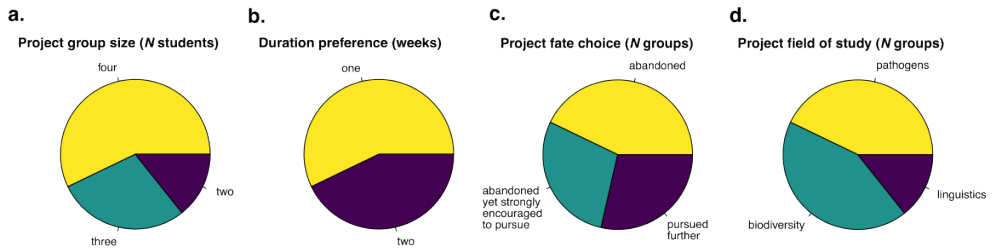

Fig. 2. Quantifiable outcomes of the research projects undertaken in the *Molecular Evolution and Phylogenomics* PhD course.

Students showed high motivation, even choosing to work on their projects at times that were allocated for breaks. Many students brought relevant previous knowledge to their project, possibly enhancing their chances of retaining the content. Students with lower motivation appeared to be those from less-related fields, such as pharmacy and clinicians. This is consistent with evidence that a feeling of identity is closely linked with motivation and wellbeing in doctoral students (Sverdlik et al., 2018). However, many students were motivated by adapting to tasks that fell closer to their expertise, such as literature search instead of coding. Therefore, the format involving project roles improved congruence between the material taught, student background, and course activity, and was likely an important building block towards effective learning (Hounsell & Entwistle, 2005).

The projects that students chose were diverse, including data sets from pathogens, biodiversity, and linguistics (Fig. 2d). The groups that were strongly encouraged to continue towards publication involved data sets of all three types. The field of research and the viability of projects

as publications were also both anecdotally unrelated with student motivation. However, the two groups that chose to pursue their projects further included less than four students. Colleagues that observed the progression of the project commented on the need of smaller groups for additional support, which might have influenced the quality of the project and student motivation near completion.

Groups of four-to-six students are often recommended for effective discussion (Jaques, 2000), but it is likely that four-person groups are more difficult to manage in research-integrated contexts. Future instalments of the course will continue to have the mix of groups of three and four students, aiming to improve our understanding on group dynamics in this context. Similarly, near the end of the course, additional time will be set aside for planning of hypothetical future steps, which might alleviate the overwhelming feeling of having a completed project or the feeling that things were left fully unfinished. This might also motivate more groups to work towards completion after the course.

A risk in the short project setting is that it is overly ambitious. Students might not come to the course with the expectation of a 'bootcamp' or 'hackathon', and stress was cited in comments (Appendix 2) and is a widespread difficulty in PhD education (Mackie & Bates, 2019). The groups that were encouraged to continue but opted to abandon the project cited a lack of free time to explain relinquishing the work. Reducing the expectations from the instructor could improve motivation, but might also reduce the chances of successfully exposing students to the research in the field.

One way to reduce stress is to move part of the lectures component to an online format. In this way, there is more time for students to engage with each other and the instructor. Critically, this would need a reinforced message that the two components of the course, lectures and the project, are closely interlinked and part of one whole body of knowledge. This can be challenging given the very different forms of control in the two formats (Figure 1). Nonetheless, in future installments of the course it will be crucial to clearly formulate a didactical contract with students where we build an understanding of the roles of the two types of activities.

The unique opportunity of this research-integrated approach was that students might build their CV in multiple directions. They likely

retained substantive amounts of material via an intimate relationship with their project. However, the short project might consistently feel as being at an early stage, when unrealistic ambitions tend to peak (Grover, 2007). The course involved no homework to allow students to balance work and their private lives. Nonetheless, the flexible format allowed for ambitious after-hours work, depending on their motivation, likely leading to some disappointment when the desired outcomes were not reached. This topic also relates to the need for a balance between setting a productive pace while avoiding stress. One topic cited by students was the time required and difficulty in identifying an interesting research question to address. One possible solution that might save time is provide group with pre-defined data sets from which several research questions could arise, this way shortening the time before begging work on the project.

Conclusions

A profound form of research integration is possible when teaching evolutionary genomics by leveraging public research databases. A flexible project structure can accommodate a broad diversity of student backgrounds by allowing students to take different 'roles' in project development. Anecdotally, smaller groups are more likely to pursue their projects further, such that larger groups might bring out the stress and other difficulties in PhD progression. Similar courses might benefit from non-random group choice, or pre-defined research questions that can be tackled over such a short period. Two-week projects were largely rejected by students, but might allow for a mix of research-oriented and research-tutored forms of learning (de Jong et al., 2023). Lastly, an initial forum of expected outcomes could lead to a stronger alignment of expectations between instructor and students, leading to a progressive student-centered form of PhD level research-integrated education.

Acknowledgements

The new unit of study and short project format was designed with the help of Lars Ulriksen, Katrine Lindvig, Antton Alberdi, and Mark Khurana.

References

Acar, O. A., & Tuncdogan, A. (2019). Using the inquiry-based learning approach to enhance student innovativeness: a conceptual model. *Teaching in Higher Education*, *24*(7), 895–909.

Alsaqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). Agile software development: Methodologies and trends. *International Journal of Interactive Mobile Technologies (IJIM)*, *14*(11), 246.

Beath, J., Poyago-Theotoky, J., & Ulph, D. (2012). University funding systems: Impact on research and teaching. *Economics The Open-Access Open-Assessment E-Journal*, *6*(1), 1.

Clark, B. R. (1997). The modern integration of research activities with teaching and learning. *The Journal of Higher Education*, *68*(3), 241.

de Jong, T., Lazonder, A. W., Chinn, C. A., Fischer, F., Gobert, J., Hmelo-Silver, C. E., Koedinger, K. R., Krajcik, J. S., Kyza, E. A., Linn, M. C., Pedaste, M., Scheiter, K., & Zacharia, Z. C. (2023). Let's talk evidence – The case for combining inquiry-based and direct instruction. *Educational Research Review*, *39*(100536), 100536.

Dostál, J. (2015). *Inquiry-based instruction. Concept, essence, importance and contribution*. Univerzita Palackého v Olomouci.

Feldon, D. F., Peugh, J., Timmerman, B. E., Maher, M. A., Hurst, M., Strickland, D., Gilmore, J. A., & Stiegelmeyer, C. (2011). Graduate students' teaching experiences improve their methodological research skills. *Science (New York, N.Y.)*, *333*(6045), 1037–1039.

Grover, V. (2007). Successfully navigating the stages of doctoral study. *International Journal of Doctoral Studies*, *2*, 009–021.

Healey, M. (2005). Linking research and teaching to benefit student learning. *Journal of Geography in Higher Education*, *29*(2), 183–201.

Houghton, K. A., Bagranoff, N., & Jubb, C. (2021). The funding of higher education: An empirical examination of the cost of education in business schools. *Abacus*, *57*(4), 780–809.

Hounsell, D., & Entwistle, N. (2005). Enhancing teaching-learning environments in undergraduate courses in electronic engineering: An introduction to the ETL project. *International Journal of Electrical Engineering Education*, *42*(1), 1–7.

Ives, G., & Rowley, G. (2005). Supervisor selection or allocation and continuity of supervision: Ph.D. students' progress and outcomes. *Studies in Higher Education*, *30*(5), 535–555.

Jaques, D. (2000). *Learning in groups: A handbook for improving group work (3rd ed.)* (Vol. 25). Routledge Falmer.

Justice, C., Rice, J., Roy, D., Hudspith, B., & Jenkins, H. (2009). Inquiry-based learning in higher education: administrators' perspectives on integrating inquiry pedagogy into the curriculum. *Higher Education*, *58*(6), 841–855.

Mackie, S. A., & Bates, G. W. (2019). Contribution of the doctoral education environment to PhD candidates' mental health problems: a scoping review. *Higher Education Research & Development*, *38*(3), 565–578.

Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education*, *30*(4), 159–167.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self‐regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218.

Ray, S. (2007). Selecting a doctoral dissertation supervisor: Analytical hierarchy approach to the multiple criteria problem. *International Journal of Doctoral Studies*, *2*, 023–032.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534.

Sverdlik, A., C. Hall, N., McAlpine, L., & Hubbard, K. (2018). The PhD experience: A review of the factors influencing doctoral students'

completion, achievement, and well-being. *International Journal of Doctoral Studies*, *13*, 361–388.

Takeuchi, H., & Nonaka, I. (1986). The new new product development game. *The Journal of Product Innovation Management*, *3*(3), 205–206.

Thomas, R. C., & Mancy, R. (2004, June 28). Use of large databases for group projects at the nexus of teaching and research. *Proceedings of the 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*. ITiCSE04: Innovation and Technology in Computer Science Education, Leeds United Kingdom. https://doi.org/10.1145/1007996.1008039

Appendices

Appendix 1

Mini project description - Molecular Evolution and Phylogenomics PhD Course

The objective is to undertake a research project in groups of three or four students with a series of objectives:
Learn how research is done in molecular evolution and phylogenetics by contributing to a field. Experience the discomfort related with project progression.
Advance collaboration skills.
Build motivation to study the theory and applications of molecular evolution and phylogenetics.
Produce a research manuscript by the end that is submittable to a preprint server and research journal. This can also be a summary of the journey, material learnt from failures.
Receive feedback from the instructor on your progress in learning and research in the field.
Cover any gaps that are not filled in the lecture material of the course.

The project must revisit an existing published phylogenetic data set, based on the premise that previous work only performs a minor subset of possible analyses. In other words, your study will be a re-analysis of existing data.

The team involves three key roles. No need to be experts in each, and it is key to reach out when in need. What matters is the effort put in:
Theory/synthesis/writing: one or two members reading the literature and drafting the manuscript. Aim for references that justify the goal of the project.
Analysis / coding: one member examining the previous methods and executing the most adequate and simple form or reanalysis.
Figures: one member examining methods for data presentation.

The team will meet regularly to discuss progress and support each other, as well as with the instructor and with the broader group of students in the course.

The report is a research article with the following sections, where each section has 3-5 paragraphs in length:
Introduction: the broad question, recent work that produced the data, and your distinct approach.
Methods: Description of the data set and your approach in detail.
Results: Description of the novel results.
Discussion: the difference between past and present approaches and results, and future perspectives based on other topics covered.

As a help for guiding the scope of the study - accounting for the short time available to carry out the project - consider the project as ideally publishable in broad-scope journals that do not focus on impact, but only on scientific validity. These include Royal Society Open Science, Scientific Reports, PeerJ, PLOS ONE, BMC Ecology and Evolution, Journal of Molecular Evolution, Journal of Heredity, Zoologica Scripta, or similar.

Attribution will follow an agreement from the start. All students must have equal contribution and random author order, with the instructor as senior last.

Appendix 2

Responses to the final feedback requested to students for the *Molecular Evolution and Phylogenomics* PhD course at the University of Copenhagen.

What is your opinion on the group assignment and the roles system?	Are you satisfied with the frequency, length, and content of lectures?	Would you prefer the course to be distributed over 2 weeks?	What is your overall opinion on the course?
Quite stressfull, it is a hard task and quite open task. With the time available it may be the best approach for getting a good product, but I'm not sure it is the best process	Content was very good, would maybe be usefull with small gruop discussions with neightbour etc.	yes	Very good introduction to the advanced topics in molacular phylogeny. I liked the lectures more than the project, but good to get some hands on - and I like the idea of not doing copy-paste- evaluate output exercises
I like the approach, and think I've learnt a lot about the model we used. I think its such a great opportunity and time to get into the models in depth, but finding the question and data that we could realistically analyse in a week was tough. Either with time before or the course split 2 days and then a month later 3 days (maybe?) we could put together new datasets tor publication, run lots of models and then come back to discuss this as a group?	Yes, i think some of the more complex topics could have been slowed/simplified a little.	Yes, I think this would allow some of the material to sink in a bit, and help the project ideas etc.	Really great. I like the mini project idea, and core theory. We could benefit from hearing more about each others projects maybe (even dead end ideas, or we couldnt find the data etc. to see what decisions other groups were making, that we might face in our own research).

I think it was fair to create the groups randomly. I found tge role system very helpful to distribute the project responsabilities.	I think lectures were very broad, attempting to cover a lot of material. I personaly struggled to follow, but i also understand it might be a bot challenging given that we only had one week.	If covering the same number if topics, might be valuable to have a bit more time to cover the material.	I found the fourse very interesting, especially thanks to the commitment from David, which was crucial to pish forward the project.
It was great especially since I have not worked with Phylogenetics before and never done coding related to that it was great to learn from group members who already knew how to do things and learn from them. So I took the job of writing.	Yes, loved the structure where it emphasised more on doing a project work which could probably turn into a publication instead of running codes which I would probably not remember what it means later.	Maybe, yes then we could work a lot towards the project and run through some ideas more and maybe we will be sure that this work is something we should publish. Also then people who are weiting and has no experience in coding like me might have the opportunity to learn and contribute more towards analysis.	The coursework designed by David is definetly great. I have no other comments, and the PhD students did a great job in explaining their PhD work and concepts.
I like this group project a lot, I learned a lot throught out the whole brainstorming, role assignment, coding and putting things together	Although the lectures are quite intense, I do learn a lot by being exposed to intense knowledge, so, could be leveled down a little, but this is good either way	I would prefer that, with more background information on the basics : molecular dating, prior, models, and other math stuff.	Great group projects, great lectuer, and very participants from very diverse background which is very cool

I think our group worked very well together. The role system was nice and made the project not seem too stressful. However, in my opinion, it did limit the learning outcomes for the single person - as naturally, people drift towards the role they are already good at - instead of challenging themselves. I do not feel like there was enough time to engage fully in what other people were doing.	Short answer yes! Very much with the frequency and length! Short, consice lectures are nice! I did often get lost in the math heavy parts such as explaining formulas. I come from a very biological and not mathematical background - which got the better of me. Sometimes, stuff like this was either quickly brushed over or explained in "lingo" I do not understand. But I did feel like I always got the gist of what was going on.	No, I think the timing was good. I like intense short courses, and planned my work so that I could only focus on this course.	I think the course was very nice! It was not what I had expected it to be - but I definently learned something. I went in thinking I was going to learn how to build trees using different software, which I think would have been nice. Or even learning how to use basic code for phylo analyses. I think most of my learning outcomes came from the lectures, as well as working within a group and being creative in thinking originally. This is a great outcome!!
it's nice! the assignment is fun and your lack of real requirements makes it very stress-free which is nice. the roles i think are only logical given the time and they work well.	yes! very much so. i like the shift between group project en lectures is very nice, the content was the right level of challenging for me and the length was good	no! not necessarily, i like this one week 'marathon'	super positive. i said previously too, I think your teaching method and energy are super refreshing, the way your personality shines through in your work is very inspiring to me. the content was the right level of challenging, the project was fun and your guidance was exactly what we needed to be able to keep chugging on. i feel super proud to have written an introduction and (almost) a discussion in fields i knew nothing about, about results i only started to understand a day ago. will definitely recommend the course to other people.

Regarding the role system, it worked very well in our group, but there was also great support from roles which were less active in periods, to support the active roles for getting things rolling. **I like the concept of the group assignment, but I think the format is to free to really get a feeling of the course itself.** **I will go out here with an idea to look up some resources, but I do not think that I got the teoretical understanding which I had expected.**	Length, very well with a focus week. Content, I would like more rigid lectures or more strict format. Frequency, I very much like to block out a week in the calender. Instead of having something every Thursday afternoon or similar over a longer period.	No	I am somewhat satisfied with the overall introduction to different topics in molecular evolution and phylogenetics. I do not think that I got the depth of understanding of different subtopics of which I had expected. I can understand why we do not devle into some specific models, but i think one of the more simple models could have been shown some equations from and how things are accutly estimated.
I thought the group assignment was very well structured. I also thought that the distribution of roles worked well. Perhaps a closer collaboration between the coder and figure producer roles and better organized moments of exchange between roles would benefit the pricess. this went well at the beginning of the project but took a dive once things really started rolling (could also have been a fault of our specific group).	I was satisfied with the frequency content and length of the lectures. I would have liked to have seen more about how genomic scale analyses can be optimized an how these differ from smaller scale genetic data analyses. apart from that no comments, great content!	It would reduce the pressure of the group project, but that was also what made it fun and exciting. so my answer would be no, keep the duration as is	I thought it was a great course, well taught, in depth while keeping it accessible for a broad audience. I also enjoyed the format of the preprint project, which was new to me. I learned many new ways of understanding fundamental concepts in phylogenetics as well as expanding my knowledge-base, specifically with regard to molecular evolution.

In our case the group system worked incredibly well! We split the roles a tiny bit differently with seperating the methods/analysis into two parts instead of separating code and figures, but I think that really depends on the project itself. It is a good idea to give these roles in order to make sure that not everyone is supposed to do everything	Yes! :) I like that they were packaged into smaller parts. you focus on one specific topic rather than providing too much information at once	I think it is great to really focus on the project for a whole week. in the beginning it sounded like a lot, but i think limiting it to one week makes it really intense (in a positive way).	I think its the best PhD course that I have taken so far :) I loved the mix between information and actual application.
i think it was very good, it was like a simulation of what collaboration looks like in real life!	the lectures i felt like it was a bit much and detailed. i felt dozing out from time to time because i couldn't understand and when i couldn't understand one thing the rest was seeming a bit harder to focus	yes	even though the lectures sometimes were a bit too harsh on me, overall i loved the course and especially how david was engaged with every single one of us, continuesly giving answers and suggestions was amazing
the mandatory assignment was a good idea. the roles worked fairly well, but I think everyone struggled with the problem ow waiting for results from others (eg you need results to make figures)	I think the content was too broad, I would have preferred a deeper introduction to the basics, eg max likelihood. active engagement in the context of the lectures would have been very useful I think	no	very nice, very good lecturer, but slightly overambitious, I think, on behalf of the students. but for me and - I think - others in the course, there was a constant sense of not totally grasping what is being presented
would be smarter to have two writers one figure and one coder since litterateur position was good but seemed quiet useless at times	yes very nice good lectures and they were very interesting	no would be to long	extremely good learned a lot

I think the roles system really made the project manageable and easier to organize. Towards the end, we didn't stick to the roles so much because we were all just working on things that needed to be done, but I learned a lot from my group mates about how to approach things that I'm not so skilled at. It was also really nice to figure out problems together or get feedback immediately	Yes. I did enjoy it more on the last few days when we were done with lectures by the morning and then just focused on the project, but I think it was easier to focus on the lectures when they were peppered among the groupwork. I think the approaches to lectures were timed well, so to be more frequent when there was less groupwork and less frequent towards the end	no	I really enjoyed it and I learned a lot, not just about phylogenetics but also about how to conduct and execute research project, which I think is really valuable as someone just starting out in the field