PORTFOLIO TAB – MARIE DIECKMANN

Personal presentation: what did I expect to learn?

I didn't know exactly know what to expect from the TAB course. I chose the course as I wanted to learn more about the topics of bioinformatics, and I think it could be an interesting field for me to continue in the future. I wanted to learn how to put the theories related to bioinformatics and theoretical things that I learned previously in my bachelor into practice. At the end of the course I wanted to be able to do basic data analyses and create diagrams to represent data. Also, I wanted to understand what programs, workflows and visualizations are best to use for what to be able to figure out what to apply to a certain problem and in general understand how used workflows work.

Final discussion: What have I learned?

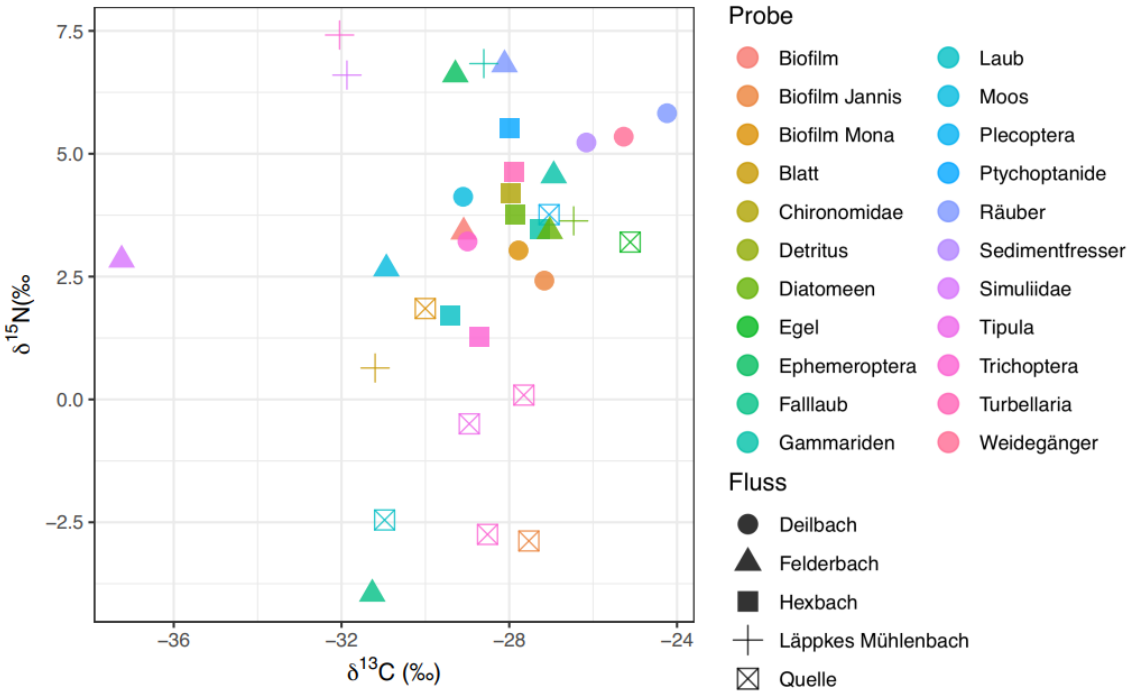
I have learned a lot of things in this course. Starting with the function of Linux and using basic commands to create and delete repositories or processing and transforming data. Because I didn't know anything about Linux before, this helped me a lot to better understand the function of computers, where things are saved and how you can change things in the data directly yourself. In the following practicals we got familiar with ggplot. As I had worked with ggplot before, I kind of knew how it worked. It was good to revise the different datatypes and concepts such as the data-ink-ratio or the components of the grammar of graphics. The second practical especially also helped me to understand how to build a plot layer for layer (without having to google everything and just adjusting it). Also, it helped to develop a sense of which graphs are best to use for a specific dataset. In the third practical we did how to do a Genome-Wide Association Study, which taught me how to prepare and manipulate data for that, how to do an association analysis and to visualize it in R. Additionally I learned more about the genetic background and reason for doing such studies. I can say, that I gained genetical knowledge as well whilst doing this practical, as it was necessary for me to research on those topics to understand what we were doing. Finally in the last practical I learned more about different types of normalization of data, doing statistical test and an enrichment analysis.

In the different conferences I learned how big and diverse the field of bioinformatics and how many different job opportunities there are especially in research. Which I found interesting and what made me more curious about pursuing a career in that field.

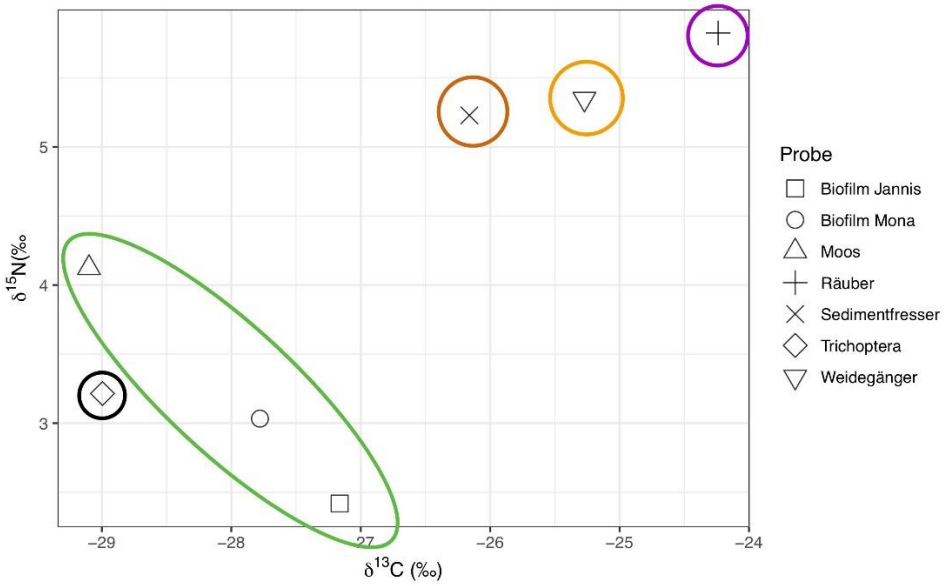
Calaix de sastre

Another thing, I did related to bioinformatics in the beginning of this semester was for a course called "Study of aquatic ecosystems". In that course two groups collected samples of different rivers and streams in our area. Then everyone of us had to evaluate these samples for a certain groups of organisms or a certain type of samples. I chose to work on the field of stable isotopes. In this field the isotopes most used are nitrogen and carbon. The evaluation of these I based on the fact, that there is a lighter isotope (14N for nitrogen, 12C for carbon) and a heavier isotope (15N for nitrogen, 13C for carbon). All the samples were prepared in many steps and then the amount of the different isotopes was measured, and a value was standardised with a formula. Both of the values give us information about the position of an organism in the food chain and therefor also in its ecosystem. The heavier isotopes accumulate, the higher you go in the food chain. Furthermore the values of the heavy 15N is higher in the biomass then in the atmosphere and the concentration of the heavy 13C is higher

in terrestrial habitats and a lower concentration in aquatic ones. So there is a lot of information we can get from these values, but I found that just having these tables with hundreds of values and medians was quite chaotic. So, I wanted to visualize the medians, which I did by using R. With ggplot I first created a plot which shows all the samples from the different rivers in one plot: the different rivers according to different symbols and different samples (type/ group of organisms we sampled according to different colours as you can see in the legend).



To get a better overview I also filtered the data to the different rivers and made plots for them, for which you can see an example here: with just the samples of one river shown and the different types of samples/ groups of organisms according to different symbols. The coloured circles are for the different feeding types (as shredder, filterer, primary producers, predators etc.)



Summary of an attended conference

Clara Inserte: clevR-vis: innovative visualisation techniques for clonal evolution

Topic:

In her presentation Clara was talking about the tool clevR-vis she developed in her master project to improve the visualization of clonal evolution.

Content:

Her career path:

She started by quickly introducing herself. She did her bachelor in Genetics and afterwards her master's in bioinformatics at UAB. During her masters degree she was part of the research group of Antonio Barradilla's. In that time she was creating an interactive application for the functional characterization and priorization of the adaptive genome variants in the human genome. It was her task in the project to implement data from databases from SNPs of the 1000 Genome-Project and additional data with population genomic metrics and to make an interactive tool for the visualization, which is called "PopHumanVar". In that tool you can decide what you need and want to see of the data (description or position of the gene for example).

Now she is doing her PhD in Münster and was developing innovative visualization techniques for clonal evolution. It is a tool called "clevR-vis" which is embedded in R and allows detailed informative visualization for the clonal evolution of cancer genomes. She structured her presentation in 5 parts: clonal evolution, previous tools, ClevR-vis, an example and conclusions.

Scientific Background & Visualization approaches:

At first, she reviewed the topic of clonal evolution. It basically evolves around how tumors evolve. Mostly the mutation of a gene in a normal cell in a normal tissue doesn't have any consequences. But if that mutation happens in certain genes called driver genes it can escape the cell cycle revolution and will follow a path of rapid growth and extension. And with each cell division the cell will continue to acquire mutation. That leads to different cell population with different mutations. These cell populations that share the same mutational profile are called clones. They evolve, resulting in intratumor heterogeneity (ITH).

There are already various computing models to show tumors in different ways: basic trees (with nodes as different variants), fish plots and a coloured blob of cells that represents how the tumor would have physically looked if a sample was taken at the last time point. In the fish plot the clonal evolution is displayed. On the x-axis we see which clones are evolving along time and on the y-axis the clonal prevalences (percentage of each clone at that time point). There is two models for the evolution of tumors: linear and branching. If there is a new driver mutation in the linear evolution, it has a higher fitness and therefor outcompetes the previous ones. A special case of linear evolution is the punctuated evolution where there is a genomic aberration in a short time at early stages of the tumor initiation, but one of them is more predominant. The other type of model is branching. That means that there is multiple clonal lineages. If they have a common ancestor it is called dependent branching and if they have different ancestors they are independent. The extreme case is neutral branching: if there is random mutations occurring over time, but there isn't any changes in the fitness. So if we would take a sample at the last time point there is no predominant clone. For the

reconstruction of clonal evolution three steps are needed. The first step is the Clustering, in which we obtain a mutational profile. The samples are being sequenced and the variance allele frequency is compared. The cancer cell fractions are calculated. The resulting tables contain a number of clones and percentages of those clones. In the following trees are being reconstructed with that data. The last step is the Visualization of the CCF data. Therefor she explained two approaches: tree graph (previously explained) and the fishplot. Again, in the fishplot the time is plotted on the x-axis and the y-axis represents the cancer cell fractions (so how the clones develop over time).

Previous tools:

Clara presented the two previous tools fishplot and timescape, listing their advantages and disadvantages over each other and lacks of both of them, which is why she created her own tool mainly taking fishplot and taking inspiration from timescape. The main disadvantages of fishplot being that it struggles when there is only two or a few timepoints and many clones and not being able to show clonal evolution which is possible with timescape as it assumes there is a gradual development of the clones. The problem with timescape is that it cannot show the clonal evolution of a single clone or a branched independent evolution and is not able to show the tumor size changes.

ClevR-vis:

The objective for her project was to overcome the current visualization short comings, to develop basic, detailed and allele-level plots and designing an interactive interface that everyone would be able to use.

The pipeline of clevR-vis starts with the input of the cancer cell fraction tables which show the cancer cell fraction of each clone at each time point. With that it was already possible to reconstruct the parental relations (from the tree). After running a validity check her tool is creating a seObject to save the necessary information to plot the clonal evolution afterwards. In that object it is also possible to add extra estimated time points to make the visualization better and to be able to see all clones. Additionally, also the estimate effect of therapy can be added. The last step of the visualization is the Visualization. The are three options for that are the Shark Plot, the Dolphine Plot and the Plaice Plot. The Shark Plot only shows the relation between the clones, no time course. There is also the option of an extended shark plot, which shows a little bit of the time course, next to the original basic shark plot. It is kind of like the cell fraction table: each row is a clone (coloured) and each column a time point, but instead of numbers the point have different sizes. By hovering over the points you can see the CCF value. Then there is the dolphin which are kind of designed like the fishplot before. You also have options of changing the design, layout position and adding a legend or even also showing the sharkplot next to it to better see the relation between the clones. The last type of plots in clevR-vis are the plaice plots which are new to show biallelic events. They consist of a top and a bottom part: the top part being the dolphin part shown as a bottom visualization (all clones emerge from the x-axis). The different clones are colored differently. The bottom part mirrors the upper part but is not colored initially. That means that there still is a healthy copy for each gene at all times. In the case of a biallelic event it is possible to colour that part. The used colour is corresponding to the colour of the clone where that function loss of the gene appeared for the first time. With that it is possible to see directly how many cells are missing that gene completely. To know which gene is missing in which clone it is also possible to add annotations. The only disadvantage is that for the first time it is necessary to manually paint

the clones for the biallelic event, because the CCF table doesn't show it it is a biallelic event or not. But on the other hand then you can show the model without having to show any data. Another thing possible with this model is to add therapy effect. If the timepoint of therapy is known, the model can show the tumor size changes and which clones specifically where resistant to therapy and have caused the relapse.

Conclusion

She developed a tool that could do everything what fishplot and timescape could and added additional features like allele-aware visualization which enables the user to identify biallelic events at a glance and visualize the clonal evolution on an allelic level. Then with the new tool it is possible to approximate the tumors development between measured timepoints and also to add a therapy effect between measured time points. Furthermore her tool has a graphical interface which makes it easier to use for everyone.

In the last part Clara was answering questions about her tool and the development of her tool and also explaining a bit about how she got that position at the university of Münster and the advantages and disadvantages of doing a PhD abroad.