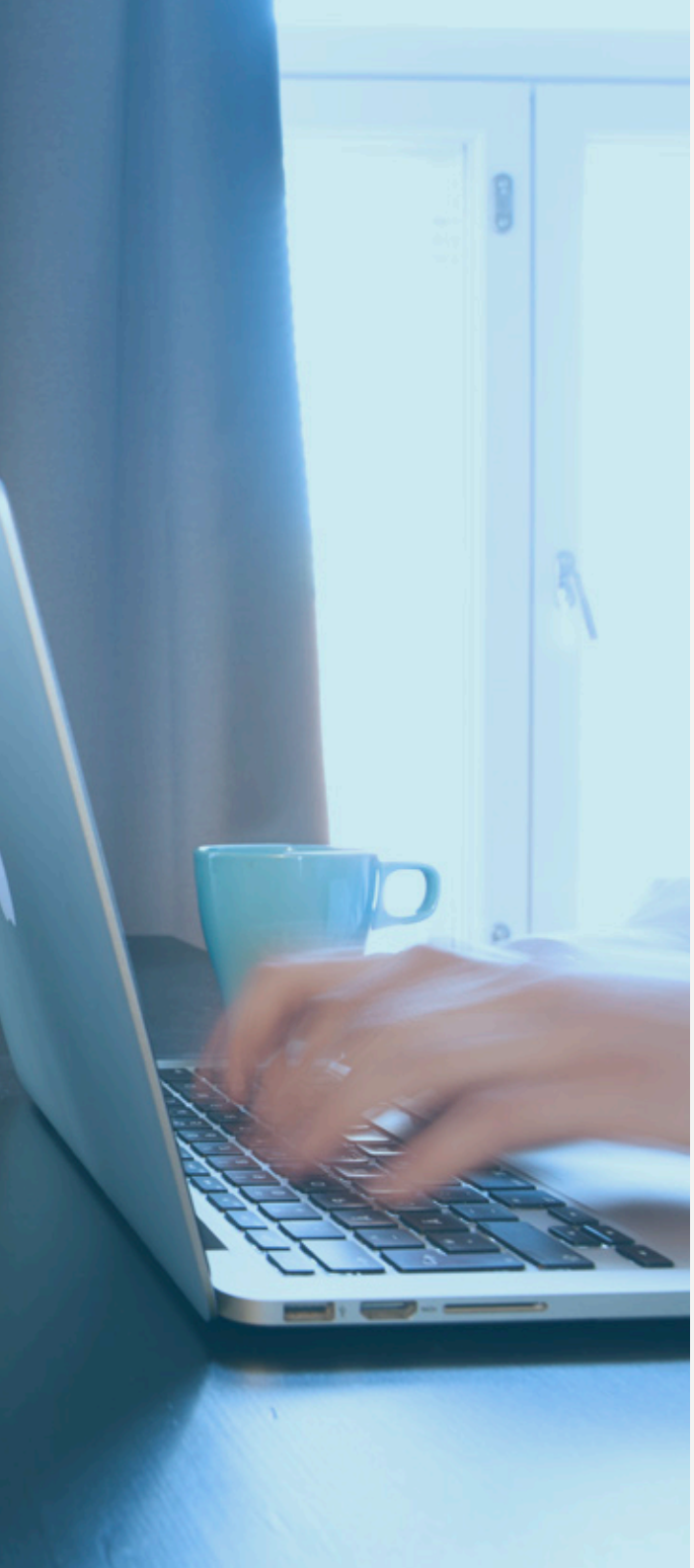
BIOINFORMATICS BUYER'S GUIDE

The Research Manager's Guide to Outsourcing Bioinformatics



CONTENTS

WHO IS THIS GUIDE INTENDED FOR?

The advent of massively parallel experimental assays for the investigation of thousands to millions of molecular events has transformed biological research in recent years. Next-generation sequencing, microarray and mass spectrometry-based methods, among others, enable principal investigators to boost the research output of their groups by many folds — but only if they can embrace new methods for data analysis as well. One could say that molecular biology has taken a huge leap toward becoming a computational science, akin to the transformation that has happened in astronomy and physics.

Despite the apparent learning curve for adaptation to new technologies and skill-sets, this development is exciting; the methodologies of high-dimensional statistics, systems theory and machine learning, first developed for the needs of other data-intensive fields, now lend themselves to biological discovery. Not only do they enable the parallel study of biological entities in remarkable quantities, but also a qualitative transformation to studying systems, or the behavior emerging from the interactions of those very entities.

The pace of this change manifests in a higher-than-ever demand for skilled bioinformaticians who unfortunately are few and far between. This, in turn, has led to many research groups in life sciences experiencing a bottleneck in the analysis and interpretation of large data sets. However, data analysis does not have to be difficult or take seemingly forever to complete. The key to success is a combination of proper planning, efficient project management and, crucially, fluent communication between the wet lab and dry lab. The involvement of an external bioinformatics team to support your research is fundamental in scalable molecular biology research of today.

This guide is written for research managers who are looking for a long-term solution to their recurring bioinformatics support needs. The writers have based this guide on their experiences in discussing and working with a wide variety of research groups since microarrays became mainstream technology. We have gathered our views on the most important questions to ask oneself and a prospective bioinformatics collaborator when planning the omics aspect of a study.

Our goal is to help you identify where you might benefit most from bioinformatics specialists, which parts of the workflow to outsource, and how to rationally compare and select a suitable bioinformatics partner, such as a commercial bioinformatics team or a university core laboratory. However, we believe you will find this guide helpful even in cases where you are considering hiring a bioinformatician or collaborating with an academic bioinformatics group as well, since the scope of the conducted work is the same regardless of your chosen solution.

WHAT DO I NEED?

As a research manager, you are the expert in coming up with novel hypotheses in your own field, validating them using the appropriate experimental methodology, and publishing your findings. Occasionally, experimental methods may include high-throughput measurements, like next-generation sequencing. If this is the case, you may find that you need to acquire external bioinformatics expertise for the project. Often this is easier said than done, especially if it is important for you to remain in control of the study.

In this chapter, we help you dissect your research project in order to identify where and how you benefit the most from bioinformatics expertise and highlight the main sources of costs in each phase. Having a clear vision on these issues will help you save money, time, mitigate potential risks, as well as increase the scientific impact of your findings.

The total cost of all the work necessary to publish an omics study is naturally a sum of many parts. This makes it especially hard to estimate the required resources for a grant application, for example, or to allocate money from an existing grant effectively.

In any research project, a large portion of the expenses are labor costs. Typically, personnel costs are fixed in the budget for a given year, but behind the scenes you will have a lot of influence on the volume of results you can expect for that money. Having your people work on what they do best will give you the most value for a given budget. In order to make optimal use of your research funding, it is helpful to think strategically about which tasks you want your team to focus on, and which parts to outsource. Personal motivation also plays a major role here; some trained biologists are highly motivated to putting time and effort into learning programming and statistics, and eventually becoming biologist-bioinformatician hybrids. However, most are not, or simply do not have enough time left over from their many other tasks.

Next, let's take a look at the different phases of a typical omics project — focusing on a study that requires next-generation sequencing. Moreover, we advise on which steps to do yourself, and where and how to use external bioinformatics expertise to optimally support your study.

The 10 steps of an omics research project

1

EXPERIMENTAL DESIGN

2

SAMPLE PREPARATION

3

NGS LIBRARY PREPARATION

4

NEXT-GENERATION SEQUENCING

5

QUALITY CONTROL

6

BASIC ANALYSIS

7

DOWNSTREAM ANALYSIS

8

INTERPRETATION OF RESULTS

9

EXPERIMENTAL VALIDATION

10

PUBLISHING

STEP 1:
EXPERIMENTAL DESIGN

Experimental design can make or break a study. Design details may dictate not only the cost and data fidelity, but also the types of molecular events that can be inferred from the data (e.g. coding mutations vs. regulatory mutations, or gene expression vs. gene isoform expression). Therefore, to ensure sufficient statistical power from a given budget, we suggest planning the measurements with the team responsible for analyzing the resulting data. Bioinformaticians can help you fix critical design parameters, such as the selection of sequencing kits, required sequencing depth, and defining sufficient numbers of replicates and control samples. It is deceptively easy to overlook the importance of this step, but careful design always pays off later. Well-designed experiments lead to an optimal combination of statistical power and reusability of the data at the lowest cost possible.

STEP 2:
SAMPLE PREPARATION

This step refers to all the work required in order to acquire the biological samples to be sequenced. Sometimes it involves collecting and diagnosing primary tumors, other times it will be conducting experiments on cell lines. Whether the biological material you work with consists of tissue samples, cell lines or blood, it is likely that sample preparation will be one of the most expensive parts of the study, being very labor-intensive and time-consuming. Most work in this phase, before extracting biomolecules to be measured, lies within the core capabilities of a molecular biology lab, and as such is not the most suitable phase for outsourcing.

Extracting the DNA, RNA or other molecular fractions from your samples is the first part of the process that is often outsourced to a measurement service provider. Usually this is a relatively inexpensive service, so you may consider weighing up the time it takes to do this in-house when making the decision. However, not all service providers are able to conduct the extraction from more ‘exotic’ or low-quality samples — RNA extraction from FFPE samples, for instance, might not be offered.

STEP 3:
NGS LIBRARY PREPARATION

Sequencing library preparation refers to the set of biochemical modifications necessary in order to be able to sequence the molecular fractions in which you are interested. Sequencing libraries are usually prepared by the sequencing service provider, but you may also opt to do this yourself with a relevant kit. Keep in mind that not all library prep kits are compatible with all sequencing platforms, so it makes sense to select the sequencing technology first before ordering the kits. Like the extraction step that came before, this service is provided by sequencing service providers at an affordable cost.

STEP 4:
NEXT-GENERATION SEQUENCING

On the one hand, buying sequencing experiments is easy: generating raw NGS data can be outsourced to practically any sequencing service provider around the world. On the other hand, understanding and comparing quotations from different providers may prove difficult due to differences in measurement platforms, library preparation kits, sequencing chemistries and other vendor recommendations. Commercial library prep kits, standard sequencing platforms and protocols ensure that the technical quality rarely varies between different providers for more common experiments such as mRNA or exome sequencing. However, there are differences in turnaround times and prices. Therefore, it is worthwhile to request quotations from a few providers and then seek assistance in comparing them in order to find a balance between the parameters that are important to you and saving your finances for the steps that follow.

STEP 5:
QUALITY CONTROL

The first priority following data generation is to make sure all the samples can be safely included in the downstream statistical analysis. The quality control (QC) analysis can be divided into two parts: 1) technical QC, where the success of the sequencing experiment is assessed with quantitative quality metrics, and 2) downstream QC, where the success of sample preparation (e.g., differentiation experiments) and validity of the sample labeling is ensured.

Computational quality control for the raw sequencing data is a rather straightforward process with plenty of tools readily available. However, since the interpretation of the quality control metrics requires expertise, it makes sense to review the metrics with a bioinformatician experienced in working with similar data. Technical QC is also sometimes provided by the sequencing service provider. Note that some providers provide quantitative QC data without consultation, and you may therefore have to draw the appropriate conclusions with the help of a bioinformatician.

After ensuring the technical quality, you want to make sure that the samples that are expected to look similar or different to each other, actually do. Biological replicates should look similar, outliers should be flagged and inspected. Possible issues in sample preparation or discrepancies in their labeling are often found at this stage. It is tempting to move on to the more interesting analyses without properly addressing each quality issue. However, removing faulty samples from every statistical analysis and figure, or changing sample labels afterwards is extremely cumbersome.

STEP 6:
BASIC ANALYSIS

After the QC analysis, bioinformaticians will typically make the first analysis of the data with a computational analysis pipeline. These pieces of software are pre-designed to match a specific data type, rather than a specific study design, and usually produce only intermediary results, such as annotated lists of molecules, genetic variants, or loci. Therefore, you will still need the help of a bioinformatician in order to dig out and visualize the biological insight from your processed data.

Sometimes you will be able to purchase basic bioinformatics analyses with your sequencing experiments. These analysis offers are usually quite affordable, since the analysis pipelines are pre-designed and are not customizable to accommodate all possible requests. It may or may not be useful to run the first stages of analyses with such pipelines, depending on your project, but the most important thing to remember is that these results will most likely not cover all the requirements related to publishing a paper.

STEP 7:
DOWNSTREAM ANALYSIS

Downstream analysis is the stage where you take the intermediary results from the basic pipelines and design a more bespoke analysis workflow in order to answer your research questions. For example, if you have multiple types of data in your study — say, both transcriptomics and proteomics — then basic analyses are first run separately for each data type, and the results are then integrated to give you a new, deeper view into the biological system being studied. This may require identifying, testing and comparing software and then stitching these together into a tailored downstream analysis pipeline. In other cases, this may mean creating a mathematical model of the system, or using the molecules or variants identified in the previous stage in order to predict biologically or clinically relevant variables by means of machine learning. Unlike the basic analyses, these will usually need to be implemented specifically to match your project.

You will want to have a bioinformatics team at your disposal for this part in particular. If your team has experience with experimental design and the resources to run basic analysis pipelines, it is possible that this is the only phase in which you will significantly benefit from the support provided by an external bioinformatics team. This is the part of study where you need to prepare for salary-level expenses for a month or two, but rarely more if you have access to a skilled team. You will be able to reduce your costs significantly and save time at this point by selecting a bioinformatics team with significant expertise in the required types of analyses.

STEP 8:
INTERPRETATION OF RESULTS

Sometimes the entire analysis workflow up to this point had been designed to answer one or two key research questions in which case the interpretation is usually straightforward. However, many omics studies are explorative in nature, yielding a range of new hypotheses as a result. Since you are likely to end up with more hypotheses than you can validate, we suggest consulting the bioinformatics team responsible for the analysis with a view to shortlisting the most promising candidates.

Selecting the best candidates for experimental validation in such studies requires a solid understanding of both the underlying biology and the statistical methodology. Consulting a seasoned bioinformatician, even briefly, will be invaluable in avoiding costly validation experiments for findings that may later turn out to be false positives.

STEP 9:
EXPERIMENTAL VALIDATION

High-impact journals will require experimental validation of findings resulting from bioinformatic analyses using an orthogonal, and typically low-throughput, method. As a rule of thumb, generating testable hypotheses from large data is significantly easier and faster than validating them. Therefore, it makes sense to allocate resources towards validating the findings. If you happen to be running a molecular biology laboratory, this is likely to be the part where it pays off to do some experimentation on your own with the help of equipment and reagent vendors.

STEP 10:
PUBLISHING

When the computational analyses are done, you will likely need simple, yet informative visualizations that highlight the most interesting findings. Here you would benefit from access to a bioinformatician with an acute eye for detail. Every bioinformatician can create figures, such as heat maps and box plots, but not everyone can make them stand out. Importantly, remember to always request full technical documentation of the computational methods used in order to help you when writing the methods section of your paper. Proper documentation will also enable another bioinformaticians to continue the analysis at a later stage if necessary.

Once the data has been analyzed and the key findings validated, it is time to publish the results. Sometimes it is hard to decide on every figure, panel, table and piece of text to include in the manuscript. You still need to have support for the re-drawing of some of the figures generated earlier in order to emphasize the aspects that you wanted to highlight in the paper. Additionally, it always pays off to ensure that the bioinformatics team is available to support when writing the methods section content and to answer possible reviewer comments.

WHICH SERVICE PROVIDER IS BEST FOR ME?

There are a number of aspects to consider when comparing different bioinformatics services and selecting the most suitable one. Here we have listed the most important questions to ask, but it is ultimately up to you to decide how to weigh the answers — for example, is a longer turnaround acceptable if communication is smoother? The decision may be a combination of various factors, but it is helpful to address each of these aspects in a systematic manner.

WHAT IS THE PROVIDER'S REPUTATION?

Before partnering up with a bioinformatics team, check their credentials. How long have they been in business? How large is the team? Do they mention their bioinformaticians on the website? Who else are they working for? You might want to contact a couple of their current customers and hear about their experience with the service. An experienced team will be proud to let you know who their customers are. Similarly, most researchers are happy to share their experiences of a service with their peers. Besides customer references, take a look at the publications of the team members. Ideally, the team has experience working with data and sample types that are relevant to your research, preferably even developing analysis tools, not just applying them.

IS COMMUNICATION EASY?

No matter how skilled and efficient the data analysts, you have to be able to understand the results and discuss them with the service provider. Communication — or lack thereof — is often a stumbling block in cross-disciplinary collaborations of this type. In fact, it is probably the most common problem, and therefore your first priority after checking the company's basic credentials should be to ensure that you can have a good level of working communication.

If you are a biologist, make sure that your contact person has a background in wet-lab biology or at least significant experience in working with biologists. It also pays to request a meeting or teleconference to ensure there's no language barrier — either in natural language or professional jargon.

IS THE SERVICE PROVIDER ABLE TO SUGGEST ANALYSES THAT I NEED?

A good bioinformatician can execute the analyses that you ask for. An excellent bioinformatician will independently suggest analyses that will help you get the most out of your data. While discussing a potential project with a service provider, expect to be asked what your fundamental research question is and what you expect to get out of the analysis. Understanding the rationale of the experiments and the ultimate research goals will allow for an analysis that is tailored to your needs and delivers what you need in order to understand and communicate the results.

WILL I BENEFIT FROM THE EXPERTISE OF AN ENTIRE TEAM?

Successful bioinformatics projects require expertise in multiple fields, such as statistics, computer science and molecular biology. It is unlikely that a single bioinformatician, however experienced, will have all the skills necessary. An ideal team will include professionals from the entire spectrum, from the computational realm to biology. In a rapidly developing field such as bioinformatics, efficiently coordinated teamwork of a diverse group of professionals is simply the only way.

Another benefit of a team-based service is continuity. If the project rests on the shoulders of a single bioinformatician, the risk of discontinued service might be just too high to consider. It is notoriously difficult for a new bioinformatician to carry on with work started but not completed by another analyst, unless the work is exceptionally well documented.

HOW ARE THE PROJECTS MANAGED?

Ask a prospective bioinformatics partner how they run their projects. Is there a dedicated project manager to ensure that all required resources are planned, allocated and used according to the schedule? Are there systems in place to document, report and assure the quality of all the work that is done? A one-person bioinformatics partner with multiple collaborators or customers will be more risky in terms of sticking to an agreed schedule.

IS THE ANALYSIS PIPELINE TAILORED TO MY NEEDS?

Blind data processing with black-box pipelines is fast and easy. However, rarely — if ever — does one size fit all in high-throughput data analysis. If you are planning to report the findings in a peer-reviewed journal, make sure that each phase of the analysis workflow can be adapted to your needs. From quality control of the raw data to publication-ready visualizations, there are alternative tools and approaches for almost every step; some are better suited for specific organisms, measurement platforms and research questions than others. No one wants to have analyses repeated because of shortcomings that only become apparent upon review of the manuscript.

HOW TRANSPARENT IS THE ANALYSIS METHODOLOGY?

Most new analysis tools in bioinformatics are developed by academic bioinformaticians, published in peer-reviewed journals and made available as open-source software (typically R packages). Thus, a bioinformatician taking advantage of the most recent analysis methods should be able to provide detailed descriptions of the tools and refer to the papers in which they were published.

This is not to say that commercial software should never be used. Some of the best tools are developed by companies and require a paid license. If the underlying algorithms are reviewed by the scientific community and the tool is widely used, it may well be as good, or even better, than openly available alternatives. Using proprietary software that lacks references in quality journals, on the other hand, should raise a red flag (and the question of whether the company promotes their own product at the expense of scientific rigor and transparency).

To summarize, expect a first-class data analysis team to prefer open, published methodology and to use commercial tools only when they are demonstrably better than the open alternatives.

WHAT IF THERE ARE DATA QUALITY ISSUES?

Ask a bioinformatics provider how they control data quality. Different sample types, sample preparation kits and measurement platforms come with their own potential quality issues. A thorough, adaptable QC process can identify them, and potentially even pinpoint the source of the problem.

It is also worth asking how quality issues are dealt with. Bad data happens, and an established team should have experience of routinely addressing a whole range of data-quality problems. In the case of bad data quality, you should have a choice of whether to continue the analysis with the bad samples discarded, to pause the project until you have had the failed measurements re-run or, if possible, to address the quality issues with suitable error/bias correction algorithms.

A reputable team will be open about data quality and let you know if they deem the quality too low to warrant any meaningful analysis, even if it means less business for them.

WHAT IS THE TURNAROUND TIME?

Unexpected things will happen when analyzing data. A quality issue has to be addressed, files have to be wrangled into a different format or an analysis tool may not work correctly. This does not mean that a bioinformatics project should take forever. An experienced team should be able to give you an accurate estimate of turnaround time allowing for small unexpected delays.

For a bioinformatics project involving tailored analyses and visualizations, a good, realistic turnaround time would be around one month. During this time, you should be able to review intermediary results, discuss them with the service provider and internally with your team/collaborators, and mutually agree with the bioinformatics provider on the final, tailored deliverables. For projects involving multiple data types or result-delivery iterations, you can expect a longer turnaround time. On the other hand, for simpler analyses — say, pathway analysis of a list of differentially expressed genes — the turnaround time could be just a few days.

WHAT IS THE COST BASED ON?

In the same way as any kind of project work, bioinformatics projects are typically provisioned with either a fixed price for deliverables or a fixed hourly rate. Both frameworks have their pros and cons. If the project is fixed in terms of cost per deliverable, the obvious upside is a better predictability of the total cost. However, more work will be needed in the beginning of the project in order to define the scientific content in detail and you may need to forgo the possibility of changing the contents during the project.

Conversely, if the service is provided in a more flexible, time-based manner, there won't be a need to fix every detail prior to starting, and you can alter the original plan depending on the intermediary results. In such a case, the total cost and schedule are naturally subject to change, reflecting the completed, rather than planned work. With time-based projects, the key is to agree on the total amount of working hours your chosen bioinformatics team will provide you with, and for

how long, in order to make sure that the work will be completed in line with your schedule.

WILL I BE LEFT TO INTERPRET THE RESULTS ON MY OWN?

An acid test for any good bioinformatics service is whether the result delivery involves the project manager explaining and discussing the outcome with you in order to ensure that you understand the results, can draw biological conclusions from them, and actually use the deliverables in a publication or further research. No-one wants to receive a cryptic, automatically generated report with no possibility of contacting the person responsible for the analysis. Reviewing and discussing the results is the most fruitful part of a bioinformatics project; any service that doesn't offer the opportunity to take time to digest and interpret the results together with a bioinformatician is of limited value.

WHAT IF I NEED THE RESULT FILES IN A DIFFERENT FORMAT?

This is a very simple, yet crucial question. The minimum level of flexibility would involve delivering the figures and tables in a format that you can use and edit with your preferred tools. Furthermore, being able to polish the figures for direct use in an academic publication is always the preferable outcome.

WILL I GET COMPLETE METHOD DESCRIPTIONS?

Another simple yet important question to ask. Any bioinformatician should be able to describe the analysis workflow at the level of detail and clarity required for academic publishing. You do not want to end up writing method descriptions for computational analyses that someone else has done.

DOES THE PROVIDER REQUIRE AUTHORSHIP?

Authorship in an ensuing publication should not be a given condition in paid data analysis service. As the principal investigator, it is up to you to decide who has made a contribution significant enough to warrant authorship. That said, you may well want to ask the service provider's bioinformatician to co-author if their work is central to the paper. Alternatively, for less significant yet helpful contributions, the acknowledgements section of a paper might be more suitable.

WHAT IF THE REVIEWER HAS BIOINFORMATICS-RELATED QUESTIONS?

Remember that you may need to be in contact with the bioinformatics provider after the active analysis phase. Make sure that they will be available to elaborate and justify the chosen methodology in the case where a reviewer has bioinformatics-specific questions.

THANK YOU FOR YOUR INTEREST

Thank you for taking the time to read our thoughts on finding the best bioinformatics partner for your research. We hope that this guide has been helpful in identifying the hallmarks of an excellent service provider and prioritizing prospective partners according to your needs. If you feel better equipped to select and partner up with a bioinformatics team, we have succeeded. If you feel we have missed any important aspects of outsourcing bioinformatics, do let us know. This guide is an ongoing project, and we are always grateful for any suggestions on improvement — just send us an email!

Prof. Matti Nykter

matti.nykter@geneviatechnologies.com

Matti leads the Laboratory of Computational Biology at Tampere University in Finland, and is the Chief Scientific Officer of Genevia Technologies, a bioinformatics company he founded in 2011. He is a computer scientist who turned early on to bioinformatics allured by the possibilities of applying machine learning and signal processing approaches to biological data. His mission is to develop new computational methods for interpreting and integrating omics data and disseminating them to both academic and industrial life science research.

Antti Ylipää

antti.ylipaa@geneviatechnologies.com

A bioinformatician-turned-entrepreneur with over 14 years in the field, Antti has lead the development of Genevia Technologies as its CEO from an early university spinoff to a bioinformatics service provider of choice of top universities and life science companies across Europe and the United States.

Thomas Liuksiala

thomas.liuksiala@geneviatechnologies.com

Thomas helps scientists in identifying how NGS and bioinformatics can transform their research and how to access the expertise required to adopt these new methods. Eight years in academic and industrial bioinformatics and discussions with hundreds of research leaders about their bioinformatics needs has provided him with a uniquely broad view of the challenges and opportunities brought by omics technologies across the biological sciences.

GeneVia
Technologies

www.geneviatechnologies.com