2024
PROJECT CATALOG
SciLifeLab Stockholm Summer Intern program

SciLifeLab

LIST OF PROJECTS

1. **Integrating image-based and sequencing-based single-cell transcriptomics data**
   Vicent Pelechano (vicent.pelechano@scilifelab.se), SciLifeLab / Karolinska Institutet

2. **SingleCellGRN - Inferring Gene Regulatory Networks from single cell data**
   Erik Sonnhammer (erik.sonnhammer@scilifelab.se), Stockholm University

3. **Characterization of single-cell DNA replication kinetics in situ**
   Bennie Lemmens (bennie.lemmens@scilifelab.se), Karolinska Institutet

4. **Improving cancer treatment by understanding protein networks**
   Janne Lehtiö (janne.lehtio@scilifelab.se) , Karolinska Institutet

5. **Unsupervised learning techniques for the identification and rapid prediction of animal behaviors**
   Iskra Pollak Dorocic (iskra.pollak@scilifelab.se), Stockholm University

6. **Effects of diet on neuronal gene expression**
   Iskra Pollak Dorocic (iskra.pollak@scilifelab.se), Stockholm University

7. **Elasto-inertial microfluidics for high-throughput particle separation for biomedical applications**
   Aman Russom (aman.russom@scilifelab.se), KTH

8. **Expanding KEGG and Reactome pathways with a network-based approach**
   Erik Sonnhammer (erik.sonnhammer@scilifelab.se), Stockholm University

9. **Integrating and Validating Human Disease Blood Atlas with UK Biobank Plasma Profiles**
   Fredrik Edfors (fredrik.edfors@scilifelab.se), KTH

10. **Evolutionary refinement of DNA nanostructure uptake in cells.**
    Erik Benson (erik.benson@scilifelab.se), Karolinska Institutet

11. **Dynamics of ligand-receptor protein pairs in the blood**
    Maria Pernemalm (maria.pernemalm@ki.se), Karolinska Institutet

12. **Benchmarking taxonomic annotation methods for metabarcoding-based environmental monitoring**
    Anders Andersson (anders.andersson@scilifelab.se), KTH

1.

Integrating image-based and sequencing-based single-cell transcriptomics data

PI: Vicent Pelechano (KI, MTC) – vicent.pelechano@scilifelab.se
Supervisor: Marcel Tarbier (postdoc) – marcel.tarbier@scilifelab.se

Project description:

Single-cell sequencing has revolutionized our understanding of cellular heterogeneity, but since it relies on tissue dissociation it erases all spatial context. It is, however, well understood that gene expression is not just a function of a cell's state but also integrates signals from its spatial context, the so-called micro-environment. So, without spatial information we cannot fully understand molecular mechanisms in complex tissue environments, which in turn limits our understanding of, e.g., cancer recurrence or tissue regeneration.
High-throughput approaches to study spatial patterns of the entire transcriptome do not reach single-cell resolution or are limited to nuclear RNA and lack sensitivity. Therefore, to study the impact of the micro-environment on the single-cell level, one is currently limited to low-throughput imaging approaches, e.g., in situ sequencing (ISS) or sequential fluorescence in situ hybridization (seq-FISH).
We are therefore developing computational approaches to impute spatial context purely based on single-cell RNA sequencing (scRNAseq) data. For this we integrate image-based data (ISS) with scRNAseq data to learn subtle patterns that are indicative for different cellular neighborhoods. Initial analyses show great promise, and this summer will therefore be the perfect time for a student to join our efforts to disentangle the relationship between gene expression and spatial context.
Our lab offers extensive experience in quantitative single-cell analysis as well as data integration. We are looking for a student who has a good understanding of statistics and is experienced in coding with R (experience with omics-technologies comes handy as well). The student will have the opportunity to work independently, receive close supervision as desired or needed, to develop and test their own ideas, and to develop skills in spatial biology, single-cell data analysis and data integration.

- Statistical analysis of single-cell RNA sequencing data
- Statistical analysis of in situ sequencing data
- Data integration
- Advanced statistics / machine learning
- Custom scripting in R

2.

SingleCellGRN - Inferring Gene Regulatory Networks from single cell data

Name/email of PI: Prof. Erik Sonnhammer / erik.sonnhammer@scilifelab.se

Goal: to use single cell expression data from Replogle et al. for perturbation-based gene regulatory network (GRN) inference, and to analyze the quality of the produced GRNs.

Background:
A GRN acts as a descriptive model of how regulator genes affect the transcription of other genes, targets. In recent years the focus of the GRN field has largely shifted towards single cell sequencing data as this data offers a few advantages. The Sonnhammer group has developed a number of algorithms and methods to improve the accuracy of GRN inference, and has applied these to e.g. identify new cancer therapy targets. GRNI in our group builds on the idea that with knowledge of the effect and target of a perturbation, the regulatory system can be reverse engineered using linear models.

The project
The most accurate GRNI methods are based on known perturbations (Secilmis et al., 2022) but this type of experiments have not been done at a large scale with single cells, even though the technology exists. However, recently a large-scale perturbation-based single cell was published by (Replogle et al., 2022) consisting of 8249 genes that were perturbed in 2.5 million cells. While this data opens up the possibility to preprocess it in many ways to optimally handle the incompleteness in each cell's transcriptome, this short project will start by using the form of the data where the single cells with the same perturbation have been 'bulked' together. This also has the advantage that we can compare the raw data to another large-scale study by ENCODE (van Nostrand et al., 2020) which was done in bulk directly for the same cell line K562.

Most of the work can with advantage be done in R or Python, however the inferring GRN step will due to the tools being available there need to be performed in MATLAB.

The steps for the project:

1. Quality control and data parsing:
While we do have the data, the first step will be to check that the quality of the data is high enough to infer networks. Batch effect removal and proper normalization may be necessary. It is also important to verify that the perturbation had the wanted effect by checking the perturbation rank (the fraction of values each experiment has that are knocked down more than the targeted gene). The perturbation rank should be as low as possible and at most a few percent of all expressed genes. The distribution of perturbation ranks should be plotted.

2. Preparing the data and perturbation matrix
Once the data quality is ensured we will need to parse the gene expression into a data matrix of size genes x experiments and create a perturbation design matrix that corresponds to this. The perturbation design matrix is a sparse matrix with a single +/-1 values in each column. The +/-1 value corresponds to a gene that was perturbed in each experiment and as

such describes the experiment and the effect of the perturbation, +1 for over expression -1 for knock out experiments.

3. Infer GRNs
When data and perturbation matrices have been created these should be used in one of the existing GRNI methods. Ten methods are used routinely in the lab, see (Secilmis et al., 2022) and as many as possible should be run. Some of the methods, e.g. LSCO/LSCON, are however not suited for single replicate data, which the pseudobulked data constitutes, and should be avoided. LASSO and Zscore are probably the most relevant methods. One may also use consensus approaches that employ multiple primary inference methods.

4. Compare to ENCODE data
The ENCODE K562 data has 232 genes perturbed. First find out which genes are common to both datasets, and then analyze the correlation between the perturbation-induced expression profiles of each common gene. Use only these common genes for this part. We have inferred GRNs from the ENCODE data that could be used, but it is better to infer GRNs for both single cell and ENCODE data with the same methodology in order to compare them such that the only difference is the data. GRN from either source with the same sparsity will be compared with each other as well as to a reference GRN derived from eCLIP data (also from ENCODE) as validation. The inferred GRNs will be analyzed for various properties such as in/out degree distribution, and for regulatory patterns of interest from a (cancer) biology perspective, which may be validated. This part can be done with several inference methods, at several sparsities, and using several measures of network similarity.

5. Use data from individual cells
The above analysis is done on pseudobulked data, but a 100 times larger dataset is available with gene expression read counts for each cell. It is probably not beneficial to use each cell as a replicate, but we can cluster the single cells into groups to make pseudobulked ‘replicates’ and possibly group them into multiple types. The cells with the same perturbation need to be analyzed for consistency before grouping them, and parameters such as the minimum number of cells need to be established. One might want to allow pseudoreplicates to overlap, i.e. by sampling 50% of the cells in each replicate, or by fuzzy C-means clustering. The replicates shouldn’t be too dependent on each other however. The performance will be measured the same way as in the previous section.

6. Infer a TF-rich GRN
A limitation of the ENCODE data is that it contains almost no transcription factors (TFs), which means it can not be benchmarked using reference GRNs such as TRRUST and RegNetwork since these are based on TF-target links. The Replogle data however contains thousands of perturbed genes, hence we can extract the genes that maximally overlap TRRUST and RegNetwork, infer GRNs for these genes, and then benchmark the results. It would also be possible to do the same thing using ChIP-seq data (also in ENCODE). As before, this part can be done with several inference methods, at several sparsities, and using several measures of network similarity, and using either pseudobulked or single cell data. Another way of benchmarking is to use co-expression from TCGA. A major question to be addressed is how much the knowledge of the perturbation design in the Replogle data adds to the accuracy of the inferred GRNs, which can be studied by randomizing the perturbation design.

3.

Characterization of single-cell DNA replication kinetics *in situ*

Group leader: Bennie Lemmens, bennie.lemmens@ki.se

Supervisor: Bruno Urién González (PhD student), bruno.urien@ki.se

Affiliation: Karolinska Institutet, Medical Biophysics and Biochemistry department

Project description

Eukaryotic cell proliferation requires duplication of nuclear DNA prior to cell division. To ensure the faithful replication of genetic material in a reasonable time, cells orchestrate the simultaneous start of DNA replication in multiple genomic regions. The spatiotemporal control of DNA replication remains poorly understood, due to its complex nature and the lack of proper imaging tools to characterize it. Techniques that allow visualization of individual DNA replication events like DNA fiber assay or Optical Replication Mapping (ORM) lose all spatial information and cell-to-cell heterogeneity. On the other hand, high-content imaging of fixed cells does not have enough resolution to separate the multitude of DNA replication tracks happening inside the nucleus. In our group, we have combined efficient DNA replication labeling with super-resolution microscopy to visualize hundreds of DNA replication events inside the nucleus of individual cells. Now, we aim to develop a computational pipeline that allows us to extract valuable DNA replication features from these images, such as number of DNA replication spots, spatial distribution or DNA synthesis speed.

List of techniques the student will use

- 3D image segmentation (Python)
- High-content imaging analysis (CellProfiler)
- Explore segmented data and extract information (Python, R)

The selected student will work together with Bruno, who has experience in the aforementioned computational softwares.

*** If desired, the student can get involved in the wetlab aspect of the project.

4.

Improving cancer treatment by understanding protein networks

Not all protein variation has a one to one relationship with genetic information, and physical differences between proteins can also result from molecular interactions, modifications like phosphorylation, and other sources of physical and chemical variation. These differences are hard to profile in a systematic way because they are complicated to predict and distinguish, but they have immense functional importance. Our group has developed analytical approaches that resolve and quantify many types of protein variation simultaneously in an unbiased and untargeted way, and which can be used to better understand omics experiments and biological phenotypes. We have also identified applications, notably in identifying drug targets and off-targets, which can improve understanding of clinical observations such as drug repurposing applications. As a summer fellow, the candidate would have the opportunity to contribute to new improvements or applications of these methods, and also to explore complementary strategies to validate findings and enhance our understanding of biological systems.

In addition to the primary project, this internship will provide guidance and support if the candidate is motivated to develop their own hypotheses and independent inquiries, and development of these projects will be supported by our many omics datasets and analytical resources. Our group is integrated within the cancer proteomics research group at SciLifeLab led by Professor Janne Lehtiö, and our shared focus is developing and applying proteomics methods to enable proteome quantification for applications in clinical research. This extends to many interdisciplinary projects, and we are proud to be part of a collaborative, innovative group with diverse expertise spanning methods development, bioinformatics, molecular and cellular biology, statistics, machine learning algorithms, and clinical implementation. As an intern within our group, the candidate would have the opportunity to learn from the work being conducted in other projects, and they will be included in group meetings, journal clubs, and other activities.

Relevant Techniques:
Statistical analysis
R-programming language, Python programming language, Unix or linux computing
Omics data analysis: proteomics, transcriptomics, genetics, high-throughput drug screening

Optional techniques
Experimental prep: mass-spec proteomics, cell culture, drug screening

Supervision plan: A Postdoc and PhD student will collaborate as co-supervisors (Ioannis Siavelis, Isabelle Leo), and will provide training and day-to-day support as needed. As a group leader, Janne Lehtiö will supervise, assuming responsibility for the scientific direction and project guidance.

5.

Project: Unsupervised learning techniques for the identification and rapid prediction of animal behaviors

PI: Iskra Pollak Dorocic (iskra.pollak@scilifelab.se)
Supervisor: Jakub Mlost (jakub.mlost@scilifelab.se)

Objective: The objective of this summer internship project is to explore and implement unsupervised learning techniques for the identification and rapid prediction of animal behaviors. Leveraging advanced machine learning algorithms, the student will contribute to the implementation of a system that can autonomously identify patterns in rodent tracking data, categorize behaviors without labeled examples, and provide model of motion sequence transitions based on observed patterns. The application of this technology is to correlate the behavioral analysis with neural recordings and neural perturbations in animal models in order to decipher the function of neuromodulatory brain circuits.

Key Responsibilities:

1. **Literature Review:** Conduct a thorough review of existing literature and research on unsupervised learning, particularly focusing on techniques relevant to behavior identification.
2. **Data Exploration and Preprocessing:** Work with diverse datasets to understand the characteristics of behavioral data. Implement preprocessing techniques to prepare data for unsupervised learning algorithms.
3. **Algorithm Selection and Implementation:** Explore and experiment with various unsupervised learning algorithms such as clustering (e.g., k-means, hierarchical clustering) and dimensionality reduction techniques (e.g., t-SNE, PCA) to identify patterns and group similar behaviors.
4. **Model Evaluation:** Evaluate the performance of unsupervised learning models using appropriate metrics. Fine-tune parameters and iterate on models to improve accuracy and efficiency.
5. **Motion Sequence Profiling:** Develop models for profiling motion sequences to capture nuanced patterns and transitions.
6. **Documentation:** Document the entire process, including data preprocessing, model development, and evaluation metrics. Create clear and concise documentation that can be used for knowledge transfer and future reference.

Techniques the Student Will Use/Learn:

- *DeepLabCut* - a state-of-the-art animal pose estimation deep learning algorithm
- One of the unsupervised learning pipelines for behavior identification like *MoSeq*, *B-SOID* or *VAME*
- Modelling of motion sequences with an autoregressive hidden Markov model (AR-HMM)
- Dimensionality reduction techniques (t-SNE, PCA).
- Clustering algorithms (k-means, hierarchical clustering).

Expected Outcomes:

1. A comprehensive understanding of unsupervised learning techniques and their application to behavior identification.
2. Implementation of robust unsupervised learning models for efficient prediction of behavior and transitions between motion sequences

6.

Project: Effects of diet on neuronal gene expression

PI: Iskra Pollak Dorocic (iskra.pollak@scilifelab.se)
Supervisor: Charlotta Henningson (charlotta.hennings@scilifelab.se)

Objective:
The objective of this project is to investigate the impact of diet on single-neuron mRNA expression patterns in specific brain regions of mice. The project aims to elucidate whether differences in diet, particularly between a regular diet and a high-fat diet, influences the expression of neuropeptidergic receptors in distinct neuronal populations. By analyzing fluorescent in situ hybridization images of mouse brain sections labeled with various mRNA probes, the student will explore how dietary factors shape neuronal diversity and receptor expression profiles at the single-cell level.

Key Responsibilities:

1. **Literature Review:**

Conduct a review of literature on the effects of diet on neuronal gene expression, neuropeptidergic systems, and relevant methodologies in neuroscience research.
Summarize key findings regarding the influence of diet on neuronal function, Gal receptor expression, and associated behavioral outcomes. Identify gaps in knowledge and potential research directions for investigating diet-induced changes in mRNA expression.

2. **Data Exploration:**

Familiarize oneself with the fluorescent in situ hybridization images of mouse brain sections labeled with specific mRNA probes.
Utilize image analysis software such as Qupath with Cellpose and Fiji to explore spatial distribution, and mRNA expression patterns at single-neuron resolution.
Collaborate with team members to identify regions of interest and neuronal populations for further analysis.

3. **Data Analysis:**

Employ computational techniques to quantify and compare mRNA expression levels in neurons from mice on regular and high-fat diets.
Utilize statistical methods to identify diet-induced differences in neuropeptidergic receptor expression across distinct brain regions and neuronal subtypes.

Techniques:

RNAscope Hiplex Analysis: Utilize RNAscope Hiplex technology to detect and quantify mRNA expression in situ with single-cell resolution.
Qupath with Cellpose: Employ Qupath software with Cellpose plugin for automated cell segmentation and analysis of fluorescent images.
Fiji/ImageJ: Utilize Fiji/ImageJ for image processing and alignment of multiple rounds of imaging

Expected Outcomes:

1. Gain hands-on experience in advanced molecular imaging techniques and computational analysis methodologies.
2. Contribute to uncovering novel insights into the effects of diet on neuronal gene expression and neuropeptidergic signaling pathways.

7.

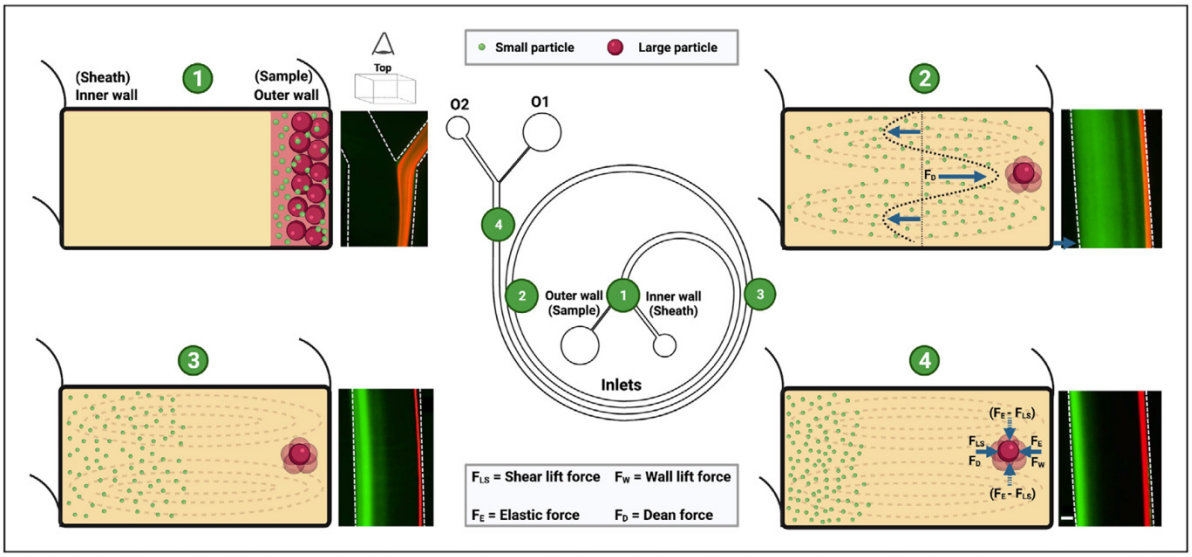
Elasto-inertial microfluidics for high-throughput particle separation for biomedical applications

PI: Aman Russom (aman.russom@scilifelab.se)

Supervisor: Selim Tanriverdi/PhD student (selim.tanriverdi@scilifelab.se)

Microfluidics has been widely used for particle separation, sorting, and manipulation in biomedical applications such as CTC and sepsis research. It is crucial to increase the throughput and improve the particle separation resolution in these applications. Elasto-inertial microfluidics as a passive particle manipulation technique offers high-resolution and high-throughput particle separation. This technique relies on physical forces that arise from microfluidic channel geometry, particle size and fluid rheology. In this project, the student will learn how to fabricate a microfluidic chip and use elasto-inertial microfluidics to separate particles in the range of 1-2 µm for bacteria separation in sepsis studies, and submicron particles for exosome studies. During this experimental work, the student will investigate several microfluidic designs and several particle manipulation parameters, and record images using fluorescence microscope. The outcome of this work will improve our knowledge on high-throughput and high-resolution particle separation for biomedical applications.

The student will learn all the techniques from the direct supervisor, Selim Tanriverdi and independently perform the experiments.



DOI 10.1002/elps.202100140

Techniques that will be used:

- Soft-lithography for microfluidic chip fabrication
- Experimental work using fluorescence microscope
- Image processing via ImageJ

SciLifeLab

8.

Expanding KEGG and Reactome pathways with a network-based approach

Stockholm University (DBB)
Supervisor: Prof. Erik Sonnhammer
Co-supervisor: Ph.D. student Davide Buzzao

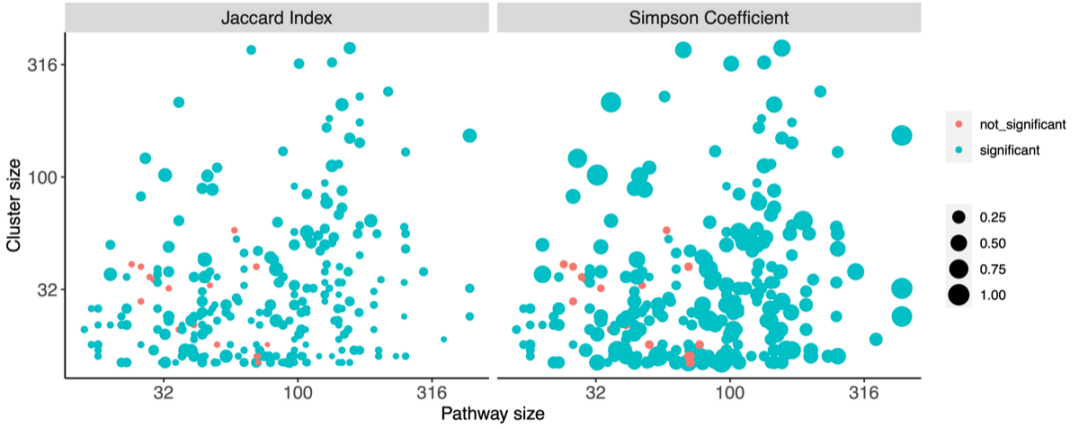
Project description

FunCoup is a web-based network biology resource (Alexeyenko and Sonnhammer 2009; Persson et al. 2021) that is intended to help researchers in identifying and examining functionally coupled genes/proteins. The platform is a useful tool for biologists doing systems biology since it offers a user-friendly interface for interpreting and displaying the results.

A genome-scale biological network like FunCoup could be used to fill the gap of high incompleteness that popular pathway databases such as KEGG (Kanehisa et al. 2014) and Reactome (Croft et al. 2011) suffer from (Gable et al. 2022). As shown in Gable et al., both KEGG and Reactome are (Roussarie et al. 2020) characterized by a low genome coverage and gene annotation biases. With this project we aim at mitigating the incompleteness that plagues most curated functional geneset databases, either by expanding existing and curated pathways with seed-based module detection algorithms (supervised strategy), or by using FunCoup clusters as functional subnetworks with a distinct biological function (unsupervised strategy).

With a supervised clustering approach, KEGG and Reactome pathways may be enhanced and these expanded pathways can then be used for improved pathway analysis. In this project, the FunCoup subnetwork for each KEGG pathway will be extracted and then extended with TOPAS (Buzzao et al. 2022) and MaxLink (Guala, Sjölund, and Sonnhammer 2014). The extent to which new pathway members were added will then be assessed, as well as the quality of the extended pathways by comparing to the Gene Ontology.

With an unsupervised clustering approach, FunCoup clusters represent novel, ab initio-generated pathway-like modules. As FunCoup is a dense network, it first needs preprocessing to find clusters matching the size distribution of KEGG pathways. We can reduce the effect of high-degree genes and accentuate local network structure with algorithms such as SkNN (Roussarie et al. 2020) or NNN (Huttenhower et al. 2007), and then applying clustering algorithms such as Infomap (Rosvall and Bergstrom, 2008). Given the presence of multifunctional proteins in the cells (Becker et al., 2012), we allow overlap between modules in cluster detection. The modules will then be automatically annotated by the predominant member gene functions in other functional databases. Preliminary results support this approach as the generated clusters show significant overlap with KEGG pathways, see Figure 1.



SciLifeLab

8.

Figure 1 - Overlap between human KEGG pathways and the most fitting FunCoup network modules obtained through recursive Infomap runs, following SKNN (k=10) application, with marker size indicating the extent of overlap.

Writing scripts (preferably in R or Python) and results analysis will be required for this project. The student will present to the group and prepare a final report.

Bioinformatics databases and tools to be used

- Online data sources:
  - Networks from FunCoup https://funcoup.org/;
  - Pathways from KEGG https://www.kegg.jp/ , Reactome https://reactome.org/
- Offline programmatic methods:
  - Network visualization with iGraph (python, R), NetworkX (python);
  - Network module analysis with TOPAS https://bitbucket.org/sonnhammergroup/topas/ and MaxLink https://maxlink.sbc.su.se/download/

References

Alexeyenko, Andrey, and Erik L. L. Sonnhammer. 2009. “Global Networks of Functional Coupling in Eukaryotes from Comprehensive Data Integration.” *Genome Research* 19 (6): 1107–16.

Buzzao, Davide, Miguel Castresana-Aguirre, Dimitri Guala, and Erik L. L. Sonnhammer. 2022. “TOPAS, a Network-Based Approach to Detect Disease Modules in a Top-down Fashion.” *NAR Genomics and Bioinformatics*. https://doi.org/10.1093/nargab/lqac093.

Croft, D., G. O’Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, et al. 2011. “Reactome: A Database of Reactions, Pathways and Biological Processes.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq1018.

Gable, Annika L., Damian Szklarczyk, David Lyon, João F. Matias Rodrigues, and Christian von Mering. 2022. “Systematic Assessment of Pathway Databases, Based on a Diverse Collection of User-Submitted Experiments.” *Briefings in Bioinformatics* 23 (5). https://doi.org/10.1093/bib/bbac355.

Guala, Dimitri, Erik Sjölund, and Erik L. L. Sonnhammer. 2014. “MaxLink: Network-Based Prioritization of Genes Tightly Linked to a Disease Seed Set.” *Bioinformatics* 30 (18): 2689–90.

Huttenhower, Curtis, Avi I. Flamholz, Jessica N. Landis, Sauhard Sahi, Chad L. Myers, Kellen L. Olszewski, Matthew A. Hibbs, Nathan O. Siemers, Olga G. Troyanskaya, and Hilary A. Coller. 2007. “Nearest Neighbor Networks: Clustering Expression Data Based on Gene Neighborhoods.” *BMC Bioinformatics* 8 (July): 250.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2014. “Data, Information, Knowledge and Principle: Back to Metabolism in KEGG.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt1076.

Persson, Emma, Miguel Castresana-Aguirre, Davide Buzzao, Dimitri Guala, and Erik L. L. Sonnhammer. 2021. “FunCoup 5: Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-Specificity.” *Journal of Molecular Biology* 433 (11): 166835.

Roussarie, Jean-Pierre, Vicky Yao, Patricia Rodriguez-Rodriguez, Rose Oughtred, Jennifer Rust, Zakary Plautz, Shirin Kasturia, et al. 2020. “Selective Neuronal Vulnerability in Alzheimer’s Disease: A Network-Based Analysis.” *Neuron* 107 (5): 821–35.e12.

9.

Integrating and Validating Human Disease Blood Atlas with UK Biobank Plasma Profiles

Laboratory Principal Investigator (PI): Fredrik Edfors
On-site Supervisor: Maria Bueno Alvez

This project aims to revolutionize the understanding of disease mechanisms by integrating and validating two extensive plasma proteome datasets quantified by Proximity Extension Assay (PEA). One dataset (*in-house*) is from the Human Protein Atlas (20,000 samples), and one is from the UK Biobank (50,000 samples). Integrating these datasets will provide novel insights into early disease signatures, contributing to more accurate diagnosis, effective risk stratification, and improved disease monitoring.

This project stands at the intersection of advanced proteomics and precision medicine, utilizing the Human Disease Blood Atlas and the UK Biobank datasets to refine diagnostic models for various diseases. The integration of these datasets is crucial for validating and enhancing the predictive accuracy of protein signatures for early disease detection. This effort contributes significantly to the field of disease precision medicine, especially in the development of liquid biopsy assays.

A pilot study by Alvez *et al.*, involving 1,477 patients across 12 cancer types demonstrated the feasibility and potential of using machine learning techniques to enhance diagnostic accuracy through protein signature classification. This foundation underscores the project's innovative approach to expanding and validating disease-specific protein signatures on a much larger scale.

Student Objectives

1. Expand this analysis outside the scope of cancer to identify unique protein signatures of specific diseases.
2. To compare protein expressions across various disease groups to understand similarities and differences.
3. To correlate protein profiles with different stages, severities, and other clinical aspects of diseases.

Reference

Álvez, M. B. *et al.* Next generation pan-cancer blood proteome profiling using proximity extension assay. *Nat Commun* **14**, 4308 (2023).

Supervision

The student will be supervised by Ph.D. students and researchers from the Edfors lab, and learn different bioinformatic tools and how these can be applied to PEA data. All in-house data has been generated, and access to the UK-Biobank data is available for 12 cancers. We aim to have access to the full UK-Biobank dataset in March.

No wet lab will be performed.

10.

Name of PI: Erik Benson (bensonlab.se)
Person who will supervise: Anjali Rajwar (postdoc)

Evolutionary refinement of DNA nanostructure uptake in cells.

Over the last decades, we have seen a dramatic increase of potential and realized pharmaceuticals and vaccines based around nucleic acids, typically based around liposome/transfection agents. An alternative mode of delivery can be to fold up the nucleic acid payload to a more compact designed shape through DNA nanotechnology. However, it is currently poorly understood what designs factors influence cell uptake. We are working on a method to produce large libraries of DNA structures that fold from single stranded synthetic 'genomes' and use evolutionary selection to understand what factors control DNA uptake in cell models. We monitor the selection process using nanopore DNA sequencing on the prepared library and the refined structure libraries after round of selection.

A summer project in our group could be experimental, working on DNA nanostructure design assembly, and uptake. Or it could be computational where you could work on the bioinformatics analysis of nanopore sequencing of DNA structures, possibly in combination with high throughput structure prediction by coarse grained molecular dynamics simulations.

Techniques learned (depending on project direction):

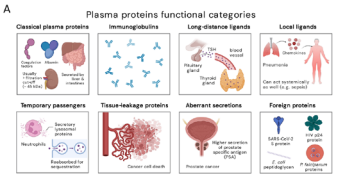
- DNA nanostructure design
- DNA nanostructure synthesis and characterization
- Cellular Uptake studies
- Nanopore sequencing
- Bioinformatics of nanopore sequencing
- Coarse grained MD of DNA (oxDNA)

11.

Group: Maria Pernemalm, PhD
Project Supervisor: Haris Babačić, MD, PhD
Supervisor Email: haris.babacic@ki.se
University: Karolinska Institute
Project Title: *Dynamics of ligand-receptor protein pairs in the blood*

Project Description:

The blood is the only compartment in the human body that acts both as a tissue and a body fluid. As a tissue, the blood contains red blood cells that carry oxygen in the body, white blood cells that maintain the immune response, and platelets that are responsible for coagulation. But as a body fluid, blood's function becomes much more complex in its liquid component, called plasma. Apart from carrying the blood cells, plasma carries a plethora of different molecules that enable communication between the many different cell types in the body. Among these molecules, the plasma proteins act as some of the main executors of bodily functions. The origin and the function of the proteins found in blood plasma can vary to a large degree in a healthy and diseased human body (see Figure A). Among them, protein ligands can act locally, in the tissue where they are produced, or travel through the blood to another tissue where they execute a function. These protein ligands then bind their corresponding receptor, which triggers downstream biological processes in the cell. In this manner, all sorts of biological processes in health and disease are regulated, purposefully or systemically.



The aim of this project is to use computational and statistical methods to explore the dynamics of receptor-ligand pairs in the blood. We will use publicly available proteomics datasets from healthy- and diseased- individuals, to extract quantifications of well-annotated receptor-ligand pairs in the blood. We will use statistical models, including meta-analytical models, to estimate associations between ligands and receptors in the blood. This project will be purely computational, and the student is expected to have some basic knowledge of the statistical software R. Background in Python can be a merit instead. If the student is successful, the project provides a great opportunity for extension into a Master thesis.

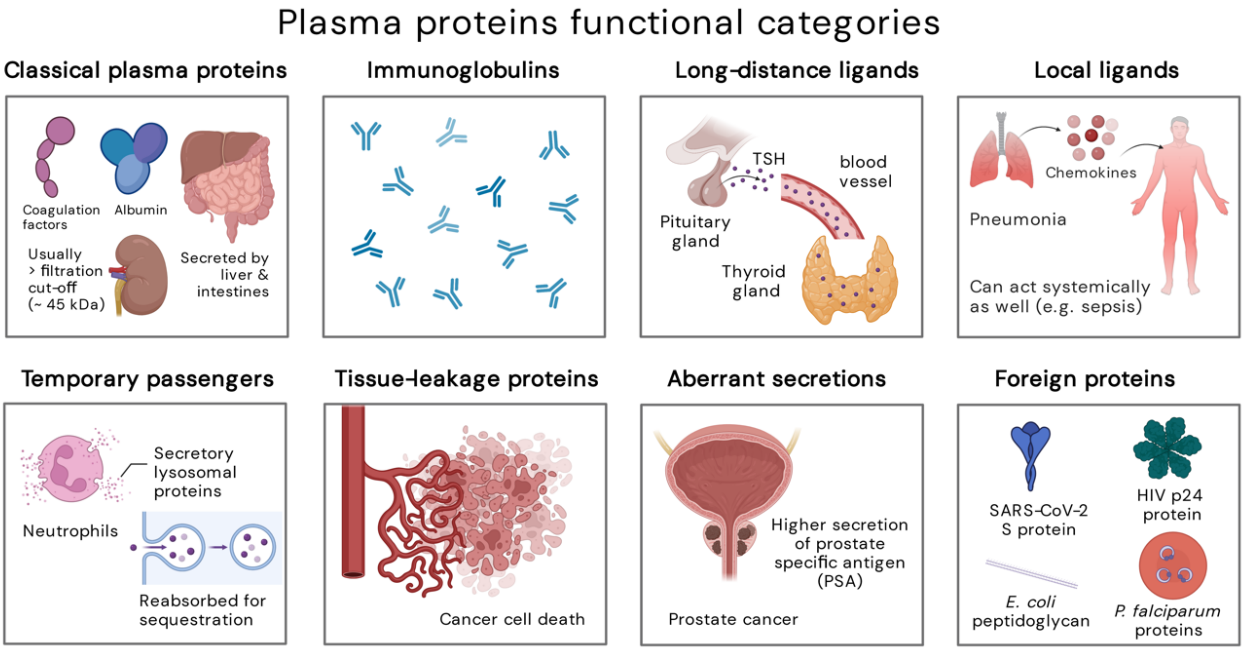
Project Plan:

Week(s)	Tasks
1	Introduction to proteomics and plasma proteomics Introduction to high-dimensional statistics and the statistical software R Introduction to statistical modelling and meta-analysis
2-3	Datasets' curation
4-6	Modelling and meta-analysis

SciLifeLab

11.

A



SciLifeLab

12.

Benchmarking taxonomic annotation methods for metabarcoding-based environmental monitoring

PI
Anders Andersson

Supervisors
Krzysztof Jurdzinski (phd stud), Karin Garefelt (phd stud), Anders Andersson (PI)

Background
The biodiversity crisis is a major challenge to mankind marked by the rapid extinction of species and ecosystems due to human activities like habitat destruction, pollution, and overexploitation of resources. This loss threatens vital ecosystem services and increases the vulnerability of communities dependent on nature. Action is needed to conserve ecosystems, protect endangered species, and promote sustainable practices to mitigate this crisis. Genetic methods are becoming increasingly important for monitoring biodiversity. Specifically, metabarcoding, where taxonomic marker genes are PCR amplified and sequenced from environmental samples is rapidly gaining popularity not only in academia but also in industry and governmental organisations. Metabarcoding can for example be used to monitor community composition of plants (e.g. based on pollen), insects (insect trap samples), fungi and prokaryotes (e.g. soil or water samples). To enable robust taxonomic annotation it is vital that the optimal bioinformatics methods are used and this requires systematic testing of such methods.

Project description
In this project you will conduct in silico benchmarking on a suite of bioinformatics methods for taxonomic annotation of metabarcoding data. The analysis will be conducted using data in reference sequence databases for insects (BOLD), prokaryotes (GTDB), fungi (UNITE), protists (PR2) and possibly other organism groups/databases and simulating metabarcoding data that will be annotated against these databases. A suite of taxonomic annotation tools will be used and compared. The results will have direct implications for how metabarcoding data will be analysed within the Swedish Biodiversity Data Infrastructure (SBDI).

Predicted learning outcome
You will acquire skills in:
Running and interpreting bioinformatic software (dada2, sintax, uchiime, nf-core/ampliseq, vsearch, kraken, and more)
Building Snakemake pipelines
Analysing and plotting data in R and/or Python
Interpreting and presenting scientific data.