UBDS^3 2025 Syllabus

Content of the document

UBDS3 2025 Syllabus
Content of the document
Schedule
Plenary Lecture of Research Highlights
Free / Biological class (week 1-2)
Week 1: Courses
Week 2: Projects

Schedule

Schedule 2025 - public

Plenary Lecture of Research Highlights

Lecturer	Lecture title	Abstract
Michael Love	Estimating gene-to-trait effects from large gene expression studies	The genetic liability for many complex traits appears enriched in non-coding regions of the genome. One mechanism that would result in associations of genetic variation in the non-coding genome with downstream traits is that alleles may affect gene transcription in certain cell types and contexts. Knowledge of genetic regulation of gene expression therefore may help us understand key molecules, pathways, cell types, and time points involved in these complex traits. I will discuss a number of experiments and methods to measure context-specific genetic regulation of expression, including chromatin accessibility and expression quantitative trait locus studies, and massively parallel reporter assays, and how these may be used to understand effects of genes on downstream traits.
Roderic Guigo Sr.	Finalizing the human and mouse gene sets	Accurate and complete gene annotations are indispensable for understanding how genome sequences encode biological functions. For more than twenty years, the GENCODE consortium has developed reference annotations for the human and mouse

		genomes, becoming a foundation for biomedical and genomics communities worldwide. Nevertheless, collections of important yet poorly-understood gene classes like long non-coding RNAs (lncRNAs) remain incomplete and scattered across multiple, uncoordinated catalogs. To address this, GENCODE has undertaken the most comprehensive lncRNA annotation effort to date. This is founded on the manually supervised computational annotation of full-length targeted long-read sequencing, on matched embryonic and adult tissues, of orthologous regions in human and mouse. Altogether 17,931 human genes (140,268 transcripts) and 22,784 mouse genes (136,169 transcripts) have been added to the GENCODE catalog representing a 2-fold and 6-fold growth in transcripts, respectively - the greatest increase in the number of annotated human genes since the sequencing of the human genome. Our targeted design assigned human-mouse orthologs at a rate beyond previous studies, tripling the number of human disease-associated lncRNAs that have mouse orthologs. Novel lncRNA genes consistently exhibit biological signals of functionality, and they greatly enhance the functional interpretability of the human genome. While poorly expressed in bulk RNA-Seq samples, many of them are highly expressed in specific cell populations, maybe even contributing to cell-type determination. The expanded GENCODE lncRNA annotations mark a critical step toward deciphering the human and mouse genomes.
Aleksandra Pękowska	Maturation of chromatin structure in development	CTCF is a transcriptional regulator that helps partition mammalian genomes into topologically associating domains (TAD). CTCF defines the borders of TADs and the anchors of chromatin loops. In our past work we uncovered that TAD boundaries and loops progressively strengthen upon embryonic stem (ES) cell differentiation, underscoring the importance of chromatin topology in ontogeny. Yet, the mechanisms driving this process remain unclear. Here, I will discuss our recent discoveries of a pervasive increase in CTCF-RNA binding protein (RBP) interactions upon ES to neural stem (NS) cell differentiation. While dispensable in ES cells, RBPs reinforce CTCF-anchored chromatin topology in NS cells. We identify Pantr1, a non-coding RNA, as a key facilitator of CTCF-RBP interactions, promoting chromatin maturation. Using acute CTCF degradation, we find that through its insulator function, CTCF helps maintain neuronal gene silencing in NS cells, acting as a barrier to untimely gene activation in development. Our data reveal a fundamental mechanism driving developmentally linked chromatin structural consolidation and the contribution of this process to the control of gene expression in differentiation.
Wolfgang Huber	Precision oncology and systems medicine	Using multiomics, single cell transcriptomics, spatial omics to improve cancer therapies. Many promising, modern drugs rely on the tumor microenvironment (checkpoint inhibitors, bispecific antibodies, CAR T cells…), responses can be spectacular but vary highly between patients. There are reasons to assume that a better understanding of tumor microenvironment will help improve these therapies.
Kateryna Pantiukh	TBA	TBA

Free / Biological class (week 1-2)

Week	Lead	Title	Course description	Language
1	Volodymyr Shvadchak	Structural biology	Introduction to structural biology. Types of interactions in proteins. Why do we need post-translational modifications? PDB and protein structure visualization. Interaction of proteins with ligands and drugs. Homo- and hetero-oligomers and supramolecular assemblies. Protein misfolding. Experimental methods for structure determination.	Ukr
1	Aleksandra Pękowska	Chromatin Biology, Hi-C	Introduction on the 3D organisation of chromatin - why we wish to study it and what information about gene expression we can learn from it. We will focus on Hi-C, an unbiased method for studying chromatin conformation. I will outline the tools that can be used to process Hi-C data and the features that can be extracted from the data. I will discuss one project that we are currently focusing on, revealing how chromatin structure matures during development. In the second series of talks, I will outline the recent progress in our understanding of brain evolution and how genomics can help to unravel the mechanisms of gene expression changes in evolution.	Eng
1	Lisa Korneeva	Next-Generation Sequencing	We will explore the world of NGS together - from different sequencing platforms and their applications to the technology behind it. You will learn the complete sample journey, from lab bench to processed data ready for your analysis.	Eng/Ukr
1	Roderic Guigo	Bioinformatics		Eng
1	Daryna Zavadska	Protistology	Learn a bit about the best organisms in the universe - eukaryotes which are NOT Embryophyta, Metazoa or Eumycota!!!	Eng/Ukr
1 (1 session on Friday only) - available as a replacement of "Bioinformatics" and "Chromatin Biology, Hi-C"	Roman Lysenko	1h session on biodiversity databases and biodiversity preservation	Основи аналізу відкритих баз даних біорізноманіття, використання їх для збору власних польових даних, різні варіанти обробки даних про біорізноманіття в ГІС системах. Приклад використання отриманих даних для створення охоронних зон для збереження біорізноманіття.	
1 (1 session on Friday only) -	Ihor Kendiukhov [in ZOOM]	1h session on large transformer-based models in biology	We will talk about performance and promise of large transformer-based models in biology, their classification, and	

available as a replacement of "Bioinformatics" and "Chromatin Biology, Hi-C"			will focus specifically on single-cell and systems biology models in more detail. We will put the emphasis on the topics of scaling, mechanistic interpretability, and generalization.	
2	Roy Jacobson	Information Theory in Biology	We will study the foundations of information theory and understand how cells have been practicing it already since the beginning of life.	English
2	Fyodor Kondrashov	Writing in English	Scientists write all the time, and the quality of the texts you produce is more important when you are just beginning. How do you write to a faculty member to accept you for an internship? In our informal course, I will show you examples of texts that I regularly receive that are either OK or are really, really bad. I will then ask you to start writing your own letters to real or imaginary professors. We will spend most of the time in class dissecting sentences and structures that you will be writing in class with me providing specific feedback.	Eng
2	Oleh Prylutskyi	Measuring Biological Diversity	Introduction to quantitative assessment of biological diversity: from designing your study in community ecology to biostatistical techniques for data analysis.	Ukr
2	Leonid Shevchenko & Olena Solianyk	Practical course on molecular biology methods	DNA extraction + PCR	Ukr

Week 1: Courses

	Day 1	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Day 6**
Statistics lecture	Visualization and data exploration, grammar of graphics, tidy data (MIL)	Data and randomness (MIL)	Regression	Testing (Stats lecture)	Clustering	Single Cells & Spatial
Statistics lab	R: Importing data, data types, data wrangling, data exploration (e.g. working with chromosomal data) labs/exploratory	R: simulation, simple inference, bootstrapping and permutation (re-use chromosomal data?) labs/randomness	ANOVA, t-test Linear regression and Metrics	Non-linear ML models (ML lecture) (Guillem)	R-Python interoperability: read a dataframe in R, do a simple wrangling & exploratory plot, call a ML method in Python, then use ggplot2 to visualize the ML result	Single-cell data and plotting

ML lecture	Data manipulations (Guillem)	Dim reduction PCA, tSNE, UMAP	Logistic regression		Neural network (ML lecture) (Guillem)	Neural network (Transformers)
ML lab	Pandas + numpy + matplotlib /seaborn(Sviat)	Advanced dim reduction (Roderic)	logistic regression and metrics (Artem)	Social activity	sklearn, pytorch (Oleh)	LLM pipelines(classifiers using using promt)(Guiilem)

Week 2: Projects

Lead	**Project title** (Please adjust according to your preference)	**Project description and learning objectives**	**Expected programming skills for enrollment (**E.g.: basic R, visualization... **)**	**Spoken Language**	**Course materials link (if you feel comfortable sharing)**
1. Daryna Yakymenko	Microscopy Image Analysis using Machine Learning based algorithms	This project focuses on the computational analysis of high resolution confocal microscopy 3D imaging data of cell nuclei with fluorescently labelled replication sites, histones γH2AX and 53BP1 proteins (involved in the repair of double-strand DNA breaks; data source: Rybak P. *et al. Low level phosphorylation of histone H2AX on serine 139 (γH2AX) is not associated with DNA double-strand breaks*. Oncotarget. 2016; 7: 49574-49587). The primary goal is to explore spatial patterns of H2AX phosphorylation using segmentation models from the BioImage Model Zoo and to perform downstream analysis using basic machine learning and statistical techniques. You will learn how to: • apply a pre-trained deep learning model on confocal microscopy imaging data for objects segmentation, • extract features describing objects (fluorescent signal foci) and preprocess obtained numerical data in Python, • analyse and visualise results using statistical methods and machine learning techniques to compare signal distributions and investigate potential clustering or classification based on image-derived metrics. The project will integrate principles of image analysis, machine	Python (at least Basic/Intermediate) Basic knowledge of mathematics and statistics is welcome	Eng/Ukr	TBA

		learning, and quantitative data interpretation, providing a practical pipeline from raw imaging data to meaningful biological insights.			
2. Daryna Zavadska	Protist time!	Piles of diverse data were obtained from diverse eukaryotic microorganisms (Cryptophytes, Kinetoplastids, Heteroloboseans, Rhodophytes and other branches across the eukaryotic tree of life) by protistologists who are going to supervise this project. This data awaits to be analysed by you. You are free to choose the quest(s) - de novo assembly, decontamination and annotation of transcriptomes/genomes from novel protist species, phylogenetic/phylogenomic tree building or metagenomic data analysis - and together we'll go through all the adventures of "bioinformatics for non-model organisms" edition. Available subproject 1 – identification and characterisation of protists' endosymbionts from single-cell data. In this project we will look at putative endosymbionts in protists' single-cell metagenome, aiming to obtain their genomes and answer the questions about their evolution. Topics such as metagenomic analysis, prokaryotic genome assembly and annotation will be covered, with focus on the analysis of a real-life dataset.	R, bash (for subproject 1 - basic Linux terminal/bash)	Eng/Ukr	No.
3. Florian Pflug	Evolution of DNA replication programs	Before dividing, cells copy their genome so that both offspring cells end up with a full copy. To ensure this, sophisticated "replication programs" have evolved. We learn how these replication programs are organized in different species, and will compare the programs of different species of yeast in detail by using published replication timing data. If the time suffices, we will study the link between replication programs and gene expression. You will learn how to work with genomes, genome annotations, and replication time data.		Eng	
4. Helene Tonnele & Luisa Santus	Who's eating whose poop? Learn how to write a pipeline in Nextflow and find out.	**TL;DR** The main focus of the course will be on learning how to code in Nextflow. You will learn how to create Nextflow modules and how to use them to build a complete Nextflow pipeline. We will use allo-coprophagy detection as our application: you will learn why it matters, understand the analysis steps, and analyze some results from real data. **A bit more in-depth (more in the course material link!) :** Bioinformatics analyses typically involve several steps, from raw data processing to generating results. Pipelines help you organize and automate these steps so you don't have to run each one manually. This makes your analyses faster, more reliable, and easier to repeat or share with colleagues. They also make it simple to process large numbers of samples at once, which is becoming essential in modern research. During this course, you	Any programming language (best if R or Python): Nextflow is its own language, so we will learn its syntax together and do not expect any previous Nextflow knowledge. To be able to follow the project nicely, you would need to be able to handle a terminal and have some previous knowledge/experience of any programming language.	Eng	Course material

		will learn how to use Nextflow, one of the most widely used tools to write pipelines. And now you may be wondering, but then who eats whose poop? Well, you will find out at the end of the course. In fact, we will build a Nextflow pipeline to detect allo-coprophagy in rats. The bacteria in our gut (called the gut microbiome) are very important for our health. In lab animals like rats, one rat can eat the poop of another rat in the same cage (cage mate). This is called allo-coprophagy. When this happens, bacteria from one rat can move to the other rat's gut. This can change the types of bacteria living in their gut. To measure how often one rat eats another rat's poop, we use reads from deep shotgun sequencing of the rat cecum (a specific part of the gut). We look at all the DNA found in the cecum. This DNA comes from both the bacteria, the rat itself and what the rat ingested, i.e. the poop of the cage mate. If one rat eats another rat's poop, we expect to find some DNA from the first rat inside the gut of the second rat (coming from the cells that are expelled in the poop). We can tell which DNA belongs to which rat by looking at the genetic differences between the DNA found in the gut and the genotypes sequenced from the spleen of each rat. We will learn Nextflow, write a pipeline from raw reads to results for the allo-coprophagy detection study. We will use it on real data, and we will understand and discuss the results together.			
5. Ian Kouzel	Making Sense of RNA-seq Data	We will analyse publicly available RNA-seq datasets, starting from gene count data, and work toward gaining biological insights. Several real-world datasets will be provided, and we will perform exploratory and differential gene expression analysis followed by gene ontology (GO) and pathway analysis. You will learn how to import and explore the data in R/RStudio using the tidyverse, analyse RNA-seq count data with the DESeq2 package, and become familiar with the capabilities of Bioconductor. You will turn your analysis into a reproducible R Markdown report.	R	Eng	
6. Illia Savchenko & Tetiana Holdanova	Molecular Dynamics Simulations of Peptide–Protein Binding with OpenMM	We'll start from the theory behind Molecular Dynamics and go through all the main concepts with underlying physics. The theory will be accompanied by hands-on experience in OpenMM (I'm choosing OpenMM over the other MD engines as it allows more control over the process, and it is easier to learn and understand MD with this MD engine), with practical examples for each introduced concept. We'll learn to prepare and run MD and analyze its results (with emphasis on estimation of binding between biomolecules). Depending on the group's motivation / skills / time available, we can cover one or more enhanced topics (such as Alchemical Calculations or Metadynamics). The project itself would be to calculate the binding energy between the chosen protein-peptide system using the acquired	Python	Ukr/ Eng (up to students)	tba

		knowledge and analyze the results. The enhanced version of the project includes suggesting mutations to the peptide and assessing the mutant's binding. There are two main learning objectives: - To understand core concepts and physics behind the MD, as this understanding will make the future experience with MD much easier - To get an understanding of the possibilities of MD applications for the "Drug Design"- like tasks, and to get some practical experience in this			
7. Iryna Poplevichiva & Bohdana Hurieva	Explore Protein Evolution and Design of Lipid Transfer Proteins	In this project, you'll explore the structure evolution of the START (StAR-related lipid transfer) domain family. Using AlphaFold-predicted 3D structures, you'll investigate how this α/β helix-grip fold—specialized for binding lipids, sterols, and other small hydrophobic molecules—has diversified across bacteria, plants, and animals. By performing structural alignments and building phylogenetic trees, you'll trace how changes in the START fold led to new ligand-binding specificities and functional innovations in different lineages. In the second part of the project, you'll get hands-on experience with de novo protein design. Using tools like ProteinMPNN and RFdiffusion, you'll redesign selected proteins based on what you've learned. You'll then evaluate your designs with AlphaFold2/3, and, if time allows, explore their potential interactions using PLIP. **During this project, you will learn how to:** ● Use AlphaFold and PyMOL to explore protein structures ● Align sequences and structures to build evolutionary trees ● Design new protein variants with deep learning tools ● Analyze protein-ligand interactions computationally	Python, bash	Eng/Ukr	
8. Joao Henrique Diniz Brandao Gervasio	Antibody evolution	Antibody Repertoire is the set of all antibodies present in an individual at a given time. This project aims to understand how persisting lineages of antibodies evolve to better recognise antigens. Persisting lineages are those antibodies that are present both before and after an immunization event like a vaccination or being infected by a pathogen. We will look and them, their phylogenetic tree topology and make sense to what changed so they can protect the body	Basic bash, Python	Eng	
9. Julia Ruehle	How Cells Make Decisions: A Single-Cell View of Gene Regulation	Each cell in our body contains the same genome, yet a muscle cell expresses a completely different set of genes than a neuron. In this course, we will explore how single-cell RNA sequencing helps us understand how individual cells read and interpret the	basic R	Eng	

		genome to decide what type of cell to become. We'll examine how this process is influenced by short regulatory DNA sequences called enhancers, and how their activity can be measured at single-cell resolution. You will learn how to analyze and visualize single-cell RNA data using R, how to identify different cell populations based on gene expression patterns, and how to combine gene expression with enhancer activity data to uncover how cells make decisions.			
10. Kateryna Pantiukh	Metagenomics: Human gut microbiome in health and diseases	In this practical project, students will explore the composition and diversity of the human gut microbiome and how it varies across different phenotypic traits, such as age and sex. Using publicly available data from the Curated Microbiome Data project, which aggregates microbiome profiles from 68 real-world studies—students will perform taxonomic analyses, identify key microbial features, and examine factors driving variation in microbial communities. The project offers hands-on experience in microbiome data processing, statistical analysis, and biological interpretation using Python.	Python	Eng/Ukr	
11. Lada Isakova	tRNA and proteins in sequence space	We will have two subprojects - the tRNA one and the protein one. The tRNA team will study the evolution of tRNA gene and across the tree of life with an emphasis on the size of the occupied sequence space and spatial distribution of sequences in it. The protein team will work with experimental data from an enzyme mutagenesis study trying to figure out how many possible functional sequences of this protein can exist. We will compare the mutagenesis data from more and less epistatic parts of this protein to see if and how epistasis affects sequence span and distribution in sequence space. We will work with large sequence alignments, perform sequence comparison, and apply computational and mathematical methods to study the sequence space.	Python, bash	Eng/Ukr	
12. Matthias Meyer-Bender	AI under the microscope - Predicting protein localization with transfer learning	Highly multiplexed immunofluorescence imaging is a technique that enables the characterization of tissues at single-cell resolution by detecting the localization of many proteins in a spatial context. This produces complex image data that require multiple steps of quality control and image processing, which can be significantly improved using deep learning approaches. For example, if we can automatically detect whether a protein is localized to the cell membrane or nucleus, we can assess whether an antibody has bound specifically to its intended target or not — a key part of validating the staining quality. In this project, we take a representation learning approach by applying pretrained vision transformers such as DINOv2 to	Python	Eng	Course Material (WIP)

		imaging data. We will use transfer learning to evaluate how well these zero-shot embeddings capture biologically meaningful features, and explore fine-tuning the model on our data to improve performance for downstream tasks.			
13. Max Ticó & Iseult Leahy	Tracing the Evolution of Selenoproteins in Metazoans	In this project, students will explore the unique biology of selenoproteins. Selenoproteins are proteins which incorporate selenocysteine, the 21st amino acid in response to a TGA stop codon. Selenocysteine is similar in structure to the cysteine amino acid but contains sulphur rather than selenium. This provides them with a catalytic advantage in redox regulation and protection against oxidative damage. Selenoproteins play crucial roles in cellular function, yet their diversity remains vastly underexplored. Focusing on vertebrates such as fishes and amphibians, students will investigate how selenoprotein families have been gained, lost, or duplicated across evolution, including in species with whole-genome duplications like salmon. Through hands-on work, students will learn to process real genomic data, run similarity searches and genome annotations, and use specialized tools like Selenoprofiles, SECISsearch, and Secmarker. The course also introduces practical skills in bash scripting, Python visualization, and utilising docker images. By the end, students will have annotated selenoproteins in the genome and traced their evolutionary history across vertebrate species — gaining both technical expertise and insight into an intriguing molecular story.	Python, bash	Eng	Course Material (WIP)
14. Oleh Prylutskyi	Community ecology based on environmental DNA data	Environmental DNA offers a rapid, unbiased, and (moderately) low-cost approach for assessing species composition at a given site. However, interpreting large metabarcoding datasets might not be easy and requires not only substantial bioinformatics skills but also a deep understanding of the diversity and ecology of the focal group. In this project, we will dive into a real-world case of soil fungi data from Ukraine to explore both opportunities and pitfalls of such kind of data and find out what use we can make of it to answer the basic ecological questions.	basic R	Ukr	