



Utrecht
Bioinformatics
Center

The Power of Collaboration

The 12th UBC Symposium

Friday 2 October 2026

Hogeschool Utrecht
(University of Applied Sciences Utrecht)
Padualaan 101 | 3584 CH Utrecht

[Programme](#)

[Keynote presentations](#)

[Pitch Talks](#)

[Poster Presentations](#)

[Index of presenters](#)

[Practical Information](#)

Part of Utrecht Science Week 2026

UTRECHT SCIENCE WEEK



Programme

08:30 - 09:00	Registration	Registration desk
09:00 - 09:15	Welcome by the UBC Executive Board & HU	Lecture hall 2.404
09:15 - 09:45	Keynote lecture by Saskia Haitjema (UMC Utrecht) <i>'Teamwork: how to create value in digital health'</i>	Lecture hall 2.404
09:45 - 10:15	Keynote lecture by Halima Mouhib (Vrije Universiteit Amsterdam) <i>'Towards Artificial Olfaction'</i>	Lecture hall 2.404
10:15 - 10:45	Coffee Break	HU plein
10:45 - 12:00	Pitch Talks: "Genomics Insights & Applications" <i>Chair: Myrthe Jager (Princess Máxima Center)</i>	Lecture hall 2.404
Anna Spiering (UMC Utrecht) – X chromosomal gene expression is linked to an inflammatory plaque phenotype in women with severe atherosclerosis Roula Farag (Princess Máxima Center) – The Landscape of Chromosomal Instability in Pediatric Solid Tumors Jade Raffle (UMC Utrecht) – From Genotype to Electrophysiology: AI-Derived ECG Latent Factors Reveal 48 Novel Loci in 76,000 UK Biobank Genomes Paul Hop (UMC Utrecht) – Rare variant analysis at scale: from exomes to actionable biology in ALS and Parkinson's disease Joanna von Berg (UMC Utrecht) – Direct haplotyping of parents enables untargeted non-invasive prenatal diagnostics Lianne Suurenbroek (Princess Máxima Center) – From germline variant to pediatric cancer: detecting DNA repair deficiency through somatic footprints		
12:00 - 13:00	Lunch	HU plein
13:00 - 13:30	Keynote lecture by Justin J.J. van der Hooft (Wageningen University & Research) <i>'Connecting the Dots: How Collaboration leads to Improved Spectral Organization and Metabolite Annotation with SpecReBoot and MS2LDA'</i>	Lecture hall 2.404
13:30 - 14:00	Keynote lecture by Jasmijn Baaijens (TU Delft) <i>'Less is more: reference selection for taxonomic profiling'</i>	Lecture hall 2.404
14:00 - 15:00	Poster Session – including competition for the UBC Poster Prize	Poster area
15:00 - 15:30	Coffee Break	HU plein
15:30 - 16:45	Pitch Talks: "Computational Methods & AI in Bioinformatics" <i>Chair: Robin van der Weide (UMC Utrecht)</i>	Lecture hall 2.404
Joske Ubels (Princess Máxima Center) – A bulk gene expression foundation model for clinical cancer prediction Eva van Zelm (Utrecht University) – Modeling meets transcriptomics: understanding how plants forage for nutrients Frank Okyere (Utrecht University) – DeepPheno: A Deep Learning Framework for linking Hyperspectral Imaging and SNP Genotypes in Lettuce Grigory Reznikov (Utrecht University) – Bioinformatics uncovers pregnancy-associated glycosylation patterns in plasma Brigitta Varga (University of Amsterdam) – pip install refs-mcc: Out-of-the-Box Feature Selection for High-Dimensional, Low-Sample-Size Omics Data Chris van Oevelen & Berend Wilkens (Hogeschool Utrecht) – Project based collaborations at the Hogeschool Utrecht		
16:45 - 17:00	Closing Ceremony & Awards	Lecture hall 2.404
17:00	Drinks & Networking	HU plein

Keynote Presentations

Teamwork: how to create value in digital health.

Many digital health solutions that are developed never reach clinical practice. Collaborating in multidisciplinary teams is an important prerequisite for implementation success. Based on experience from local, national and international projects, I will discuss the necessary members of multidisciplinary teams, their roles and their much needed interaction in innovation pipelines to create value in digital health.

Saskia Haitjema (UMC Utrecht)



Saskia Haitjema is an associate professor at the Central Diagnostic Laboratory and head of the Utrecht Patient Oriented Database (UPOD). Her research focuses on improving diagnostics and personalizing healthcare through data analytics, using routine hematology characteristics and large datasets.

She studied Medicine and Linguistics at Utrecht University and completed her PhD in Experimental Cardiology in 2017. Haitjema leads a vibrant data science community, supervises PhD candidates, and participates in multiple European research projects. She is also actively involved in education and strategic initiatives to advance digital health and AI in healthcare.

Towards Artificial Olfaction

In contrast to hearing and vision, the exact function of the olfactory sense remains largely unknown. This impacts many research areas ranging from fragrance chemistry to odor perception, as well as sensor design in artificial noses. Recent advances in AI technologies started to pave the way towards digitizing odors for structure-based odor predictions and broadening our understanding of structure odor relationships. However, there are still many open ends and questions before the sense of smell is successfully digitized, and new knowledge can be transferred to other applications such as bio-mimetic sensor design.

One problem that is currently holding back the advancement in the field of artificial noses is our fragmentary understanding of the molecular detection mechanisms underlying the olfactory sense. During the talk, I will give an overview of the advances and challenges in the field and show different projects that our team is working on at the VU Bioinformatics group of the Vrije Universiteit Amsterdam. Here, one of our main objectives is to unravel and to quantify the uptake mechanisms of odorants and other volatile organic molecules through binding proteins for applications in bio-mimetic sensor units. In future, due to the increasing number of available data on molecular structures and olfactory properties, machine learning and more data-driven approaches will be necessary to address new computational challenges and thus, in the long term, allow us to move towards the digitalization of the olfactory sense.

Halima Mouhib (Vrije Universiteit Amsterdam)



Halima Mouhib is an associate professor and co-head of the Bioinformatics group in the Computer Science department at VU Amsterdam. Her research focuses on understanding complex biological phenomena, such as olfaction and taste, using machine learning and multi-scale computational techniques.

She obtained her PhD at RWTH Aachen University, specializing in the characterization of odorants and volatile organic compounds. She later expanded her expertise into structural bioinformatics and computational biophysics, and was awarded the Descartes-Huygens Prize in 2021 for her work on odorant absorption.

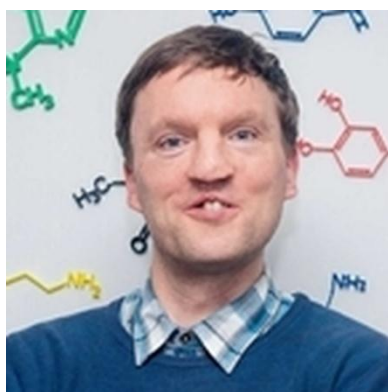
Connecting the Dots: How Collaboration leads to Improved Spectral Organization and Metabolite Annotation with SpecReBoot and MS2LDA

Nature has always been a rich resource of diverse chemistry. Fortunately, in recent years, technological advances in the omics fields have increased our capacity in measuring them. For example, modern mass spectrometers generate information-rich chemical profiles used for untargeted metabolomics studies. Such studies are typical examples of exploratory data analysis, as one does not a priori know what one will discover in the data. However, the current information-rich metabolomics profiles leave researchers with the daunting task to separate valuable signals from everything else.

In this talk, I will highlight recent advances in networking and machine learning-based strategies to organize and annotate metabolite features. These include the work on the SpecReBoot framework that boosts library matching by offering query-focused match scores, and enhances molecular networking organization by trimming unwanted spectral links and focusing on more reliable connections. Another key framework I will showcase is MS2LDA unsupervised substructure discovery to complement spectral library-based annotation by substructure annotations. Both frameworks originate from cross-discipline collaborations to connect the dots in a meaningful manner.

I will finish the talk with highlighting the important role of Open Science and Software in stimulating community-based science. Altogether, I expect that the presented progress in computational metabolomics tools will empower metabolomics researchers navigating their increasingly complex datasets to find and annotate relevant and novel chemistry in nature.

Justin J.J. van der Hooft (Wageningen University & Research)



Justin J.J. van der Hooft is an Assistant Professor in Computational Metabolomics in the Bioinformatics Group at Wageningen University & Research, NL, and an author of over 100 peer-reviewed articles in the metabolomics field. He obtained his PhD (2012) in Systematic Metabolite Annotation and Identification at Wageningen, followed by a postdoctoral period in Glasgow, UK, before returning to Wageningen, where his team develops computational metabolomics strategies to decompose mass spectral data into structure and substructure information.

Recently developed tools and frameworks from his group include SpecReBoot for confidence-aware molecular networking, MS2Query for analogue search, FERMO for prioritizing metabolite features, and NPLinker for multi-omics natural products discovery. Since 2022, he has also been a Visiting Professor in Johannesburg. Justin is a strong advocate of Open Science and community-driven research, contributing to matchms, NPLinker, MIBiG, and the Paired Omics Data Platform.

Less is more: reference selection for taxonomic profiling

As genomic reference databases continue to grow, the question of how we represent biological diversity is becoming increasingly important. More reference genomes do not necessarily translate into better metagenomic analyses: redundancy can reduce computational efficiency, while closely related genomes and lineages can blur the boundaries between taxa and distort abundance estimates. So how do we select references that capture diversity while remaining computationally tractable and biologically informative?

I will explore this question through metagenomic profiling, using benchmarking experiments to understand the impact of reference selection on profiling accuracy. In particular, we ask when and why selecting a subset of references can improve performance. This motivates ReSeT, an optimization-based approach that jointly considers coverage within taxa and separation between taxa, providing a principled alternative to conventional dereplication. More broadly, this work points toward a shift from simply accumulating reference sequences to deliberately designing computational representations of biological diversity.

Jasmijn Baaijens (TU Delft)



Jasmijn Baaijens is a bioinformatician and assistant professor at TU Delft with a focus on pathogen genomics. She develops algorithms and software to leverage sequencing data for reconstructing microbial genomes, enabling the characterization of pathogen diversity in patient or community samples.

With a background in mathematics, she earned her PhD in viral genomics at CWI and expanded her expertise in bacterial genomics during a postdoc at Harvard Medical School. She is also a member of the European Virus Bioinformatics Center and advocates for gender equality in academia through mentorship and outreach initiatives.

Pitch Talks

Genomics Insights & Applications

10:45 – 12:00

Chair: Myrthe Jager (Princess Máxima Center)

Title	Speaker
<i>X chromosomal gene expression is linked to an inflammatory plaque phenotype in women with severe atherosclerosis</i>	Anna Spiering (UMC Utrecht)
<i>The Landscape of Chromosomal Instability in Pediatric Solid Tumors</i>	Roula Farag (Princess Máxima Center)
<i>From Genotype to Electrophysiology: AI-Derived ECG Latent Factors Reveal 48 Novel Loci in 76,000 UK Biobank Genomes</i>	Jade Raffle (UMC Utrecht)
<i>Rare variant analysis at scale: from exomes to actionable biology in ALS and Parkinson's disease</i>	Paul Hop (UMC Utrecht)
<i>Direct haplotyping of parents enables untargeted non-invasive prenatal diagnostics</i>	Joanna von Berg (UMC Utrecht)
<i>From germline variant to pediatric cancer: detecting DNA repair deficiency through somatic footprints</i>	Lianne Suurenbroek (Princess Máxima Center)

[**X chromosomal gene expression is linked to an inflammatory plaque phenotype in women with severe atherosclerosis**](#)

Anna Spiering (UMC Utrecht)

Background: Sex differences in atherosclerosis are well established, with women generally exhibiting more fibrous plaques and men more inflammatory plaques. While these differences are largely attributed to sex hormones, the contribution of sex chromosome-related mechanisms remains poorly understood. We investigated female-biased X-linked gene expression and its association with sex-specific molecular phenotypes in human and mouse atherosclerosis.

Methods: Human carotid plaques from the AtheroExpress biobank were analyzed using histology, bulk RNA sequencing (153 women, 468 men), single-cell RNA sequencing (20 women, 26 men), and plaque proteomics (51 women). Female-biased X-linked genes (FDR<0.01) were used to generate a female-biased X-linked gene score, which was associated with clinical, histological, and molecular plaque characteristics. Vascular bulk RNA sequencing from the STARNET biobank (150 women, 365 men) was used to assess tissue specificity and associations with coronary atherosclerotic burden. The Four Core Genotypes mouse model was used to investigate the effects of sex chromosomes and gonadal sex using single-nucleus RNA sequencing.

Results: We identified 146 genes with higher expression in female versus male carotid plaques, of which 94% were X-linked. Among women, low X-linked gene expression was associated with greater plaque lipid content and an inflammatory macrophage gene program, including reduced expression of *PLTP*, *SPP1*, and *MMP9* (all adjusted $P<0.001$). The female-biased X-linked gene signature showed greater overlap in atherosclerotic aortic tissue than healthy mammary artery tissue and was not associated with coronary disease burden. In mice, X-linked gene expression was associated with the XX genotype in macrophages ($P=0.004$) and correlated with the X-escape gene *KDM5C* ($P<0.001$). Deletion of *Kdm5c* altered expression of a subset of X-linked genes in cultured macrophages.

Conclusion The X chromosome is an important determinant of inflammatory plaque characteristics in women with advanced atherosclerosis. This is mechanistically connected to *KDM5C*, supporting a role for sex chromosome biology in determining plaque vulnerability.

[Back to programme](#)

The Landscape of Chromosomal Instability in Pediatric Solid Tumors

Roula Farag (Princess Máxima Center)

Chromosomal instability (CIN) is characterized by ongoing structural and numerical chromosomal alterations. A recent framework has enabled the deconvolution of the distinct mutational processes underlying CIN via copy number signatures. Seventeen CIN signatures were identified across adult cancers from The Cancer Genome Atlas (TCGA) cohort, but whether these same CIN processes operate in pediatric cancers, and to what extent, remains largely unexplored. This limits insights into the underlying mechanisms and therapeutic potential of CIN in pediatric cancers.

We applied this CIN signature detection framework to a cohort of 312 pediatric solid tumors. CIN was detected in approximately 60% of the tumors and CIN signature extraction identified nine of the 17 TCGA CIN signatures. The most active CIN signatures in pediatric CIN tumors are characterized by chromosome missegregation, impaired homologous recombination (IHR) and IHR in combination with replication stress. Mutational analysis revealed four recurrently mutated genes (*COPS3*, *ALKBH5*, *UBB* and *TP53*) associated with significantly increased activity in the IHR with replication stress CIN signature ($q < 0.02$). In addition, pediatric CIN tumors with both *TP53* mutations and whole-genome duplication showed further elevated activity in CIN signatures associated with IHR and IHR with replication stress. Finally, at least one complex genomic rearrangement (CGR) was identified in 67% of the pediatric CIN tumors and these CGR CIN tumors showed the highest activity in CIN signatures linked to IHR, replication stress and amplifications.

CIN is prevalent in pediatric solid tumors and defined by a combination of CIN-associated gene mutations, whole genome duplication events and complex genomic rearrangements that collectively increase CIN activity linked to IHR with replication stress. Our findings capture the distinct landscape of CIN in pediatric solid tumors and reveal CIN signatures with actionable targets for potential future therapeutic intervention.

From Genotype to Electrophysiology: AI-Derived ECG Latent Factors Reveal 48 Novel Loci in 76,000 UK Biobank Genomes

Jade Raffle (UMC Utrecht)

Cardiovascular disease remains the leading global cause of mortality, yet links between genetic variation, cardiac electrical activity and structural remodelling remain poorly understood. Genome-wide association studies of conventional ECG intervals have identified more than 400 loci, however single-trait interval measures capture only part of the electrophysiological information in a 12-lead ECG. Advances in deep learning enable the full electrophysiological signal to be decomposed into quantitative, genetically informative latent factors (ECG-LFs).

We performed univariate GWAS of 32 ECG-LFs, extracted using the FactorECG model, in 76,244 UK Biobank participants with ~30x whole-genome sequencing, using linear mixed models (REGENIE) adjusted for age, sex, height, heart rate, and population structure. Genome-wide significant hits ($p < 5 \times 10^{-8}$) were LD-clumped (PLINK2, $r^2 < 0.05$, 500kb) and classified as novel against the GWAS Catalog. Multi-trait MTAG analysis across all 32 LFs was used to recover shared-pathway signal underpowered in the univariate models.

Univariate LF-GWAS identified 351 associated loci, including 48 not previously reported for ECG traits, MTAG identified a further 27 novel loci. The strongest novel signal, LF1 on chromosome 5, showed opposing effects on QTc versus PQ/QRS, exposing a conduction-repolarisation trade-off undetectable by single-trait GWAS of conventional ECG intervals. These results demonstrate that AI-derived ECG latent factors capture genetically distinct electrophysiological signal beyond conventional intervals and lay the groundwork for identifying causal variants and protein mediators through multi-omic integration.

Rare variant analysis at scale: from exomes to actionable biology in ALS and Parkinson's disease

Paul Hop (UMC Utrecht)

Rare variants of low-to-moderate penetrance are thought to dominate genetic risk in ALS and Parkinson's disease. To identify such variants, we uniformly processed exome sequencing data from over 20 cohorts and analysed them with rare variant software we developed. In ALS (17,919 cases and 200,703 controls) this led to the discovery of several risk genes. Among them, *ARPP21* stood out with two high-effect variants, segregation in families and a distinctive, rapidly progressive clinical presentation. In familial Parkinson's disease (2,184 index cases and 69,775 controls), analyses converged on *RAB32* p.S71R (OR = 65.5), which segregated in three families and, in functional work, was shown to increase LRRK2 kinase activity. I will show how layering association, segregation, clinical and functional evidence moves a statistical association towards tangible biology.

Direct haplotyping of parents enables untargeted non-invasive prenatal diagnostics

Joanna von Berg (UMC Utrecht)

In the Netherlands, every year more than a thousand risk couples carrying a severe monogenic disease opt for genetic testing before or during pregnancy, being either pre-implantation genetic testing (PGT) or prenatal diagnostics (PND). Both trajectories are highly demanding and burdensome for parents.

Here we will present how untargeted ONT long-read DNA sequencing as well as Illumina TruPath DNA sequencing provides a generic method to haplotype risk couples opting for PND, independent of the disease gene that they carry. We also show how this haplotyping information can subsequently be used to diagnose the fetus based on the analysis of cell-free DNA collected from the pregnant mother. Currently we are using this novel diagnostic all-in-one methodology in a prospective accuracy/validation study with inclusions from the UMCs of Utrecht and Amsterdam, with the aim to subsequently implement and offer MG-NIPD in the Netherlands.

From germline variant to pediatric cancer: detecting DNA repair deficiency through somatic footprints

Lianne Suurenbroek (Princess Máxima Center)

In $\pm 10\%$ of children with cancer, a germline pathogenic variant in a cancer predisposing gene can be identified. Early detection of cancer predisposition syndromes is essential to offer timely surveillance, prevention and alternative treatment options. However, patients with DNA repair syndromes, including constitutional mismatch repair deficiency and Fanconi anemia, don't always present with clear symptoms and are therefore hard to diagnose based on clinical presentation. In this study, we have analyzed patterns in the genomes of 1955 pediatric tumors to detect DNA repair defects. To detect microsatellite instability (MSI), which is a marker for mismatch repair deficiency, we have applied MSMuTect4. Additionally, we extracted mutational signatures, and used insertion/deletion signature 6 (ID6) as a marker for homologous recombination deficiency (HRD). We found that both MSI and ID6 can be used to detect patients with a DNA repair syndrome. In contrast with adult cancer, MSI and HRD are rare in sporadic cases of pediatric cancer.

Computational Methods & AI in Bioinformatics

15:30 – 16:45

Chair: Robin van der Weide (UMC Utrecht)

Title	Speaker
A bulk gene expression foundation model for clinical cancer prediction	Joske Ubels (Princess Máxima Center)
Modeling meets transcriptomics: understanding how plants forage for nutrients	Eva van Zelm (Utrecht University)
DeepPheno: A Deep Learning Framework for linking Hyperspectral Imaging and SNP Genotypes in Lettuce	Frank Okyere (Utrecht University)
Bioinformatics uncovers pregnancy-associated glycosylation patterns in plasma	Grigory Reznikov (Utrecht University)
pip install refs-mcc: Out-of-the-Box Feature Selection for High-Dimensional, Low-Sample-Size Omics Data	Brigitta Varga (University of Amsterdam)
Project based collaborations at the Hogeschool Utrecht	Chris van Oevelen & Berend Wilkens (Hogeschool Utrecht)

A bulk gene expression foundation model for clinical cancer prediction

Joske Ubels (Princess Máxima Center)

High-throughput gene expression measurements have long held promise to truly personalize cancer treatment. However, predicting clinical outcomes from bulk transcriptomic data remains a major challenge, with most classifiers failing to reach the clinic. A major hurdle for these models is their failure to generalize in new populations. Foundation models based on large-scale pretraining have revolutionized the vision and language fields, yet have failed to yield similar breakthroughs in the gene expression space. We demonstrate the standard approaches to leverage pretrained models for downstream (clinical) tasks - such as mean or attention pooling - consistently degrade predictive performance and are mostly outperformed by simple baselines like logistic regression on raw gene expression. To overcome this major bottleneck, we introduce a robust, pretraining architecture using a leJEPa-inspired loss function, paired with a novel supervised transfer learning framework utilizing the bilinear matrix factorization of embeddings.

We show we can predict survival in a variety of cancer types, across different ancestries, treatment regimens, and measurement platforms. By successfully preserving structured information in pretrained embeddings, our approach provides a definitive path forward for building clinically impactful transcriptomic models in rare cancer settings.

Modeling meets transcriptomics: understanding how plants forage for nutrients

Eva van Zelm (Utrecht University)

Where soil nitrate is unevenly distributed, plants can specifically invest in root growth in nitrate-rich patches. This preferential root foraging requires integration of local, long-distance and systemic signals to coordinate and achieve asymmetric root growth. Despite insight into key molecular players, it remains unknown how they collectively drive root growth decisions. We combine a nitrate-signaling-based computational model with time-resolved transcriptomics to understand how known signals are integrated and to identify missing regulatory links. Interestingly, most gene expression dynamics can be classified as being part of either local, long-distance or systemic signaling, with individual genes switching roles as internal plant status develops. Using WGCNA analysis, we aim to identify novel transcriptional regulators that allow genes to mediate these signaling programs. Mismatches between model predictions and observed expression dynamics reveal regulatory insights that directly drive model improvement. Together, this approach aims to decipher how plants use their molecular networks to compute root foraging decisions.

[Back to programme](#)

DeepPheno: A Deep Learning Framework for linking Hyperspectral Imaging and SNP Genotypes in Lettuce

Frank Okyere (Utrecht University)

While whole-genome sequencing captures millions of single nucleotide polymorphisms (SNPs) and hyperspectral imaging (HSI) enables non-destructive plant phenotyping, integrating these modalities to link genotype to phenotype remains challenging due to their high dimensionality and non-linearity. This study presents DeepPheno, a deep learning framework that predicts SNP genotypes from HSI data, using model predictability as a proxy for genotype-phenotype association. HSI data were acquired from 194 lettuce genotypes under field conditions. HSI data patches (20×20 pixels × 224 spectral bands) were used to train a hybrid CNN to predict the variant of a specific SNP. The framework was validated on SNPs with known phenotypic effects (anthocyanin, leaf serration, pale pigmentation), achieving high predictive performance (AUC ranging from 0.806 to 0.935), whereas models trained on randomly shuffled labels performed at chance (mean AUC ≈ 0.51). Extending the workflow to 50 randomly selected putatively neutral SNPs, most yielded low predictability, but two showed high performance (AUC > 0.76), suggesting uncharacterized genotype-phenotype links. Explainable AI, including SHAP and GRAD-CAM, identified relevant spectral and spatial features driving these predictions, particularly the green and red-edge wavelengths associated with pigment dynamics and leaf structure. These results establish a framework for understanding complex genotype-phenotype interactions in plants and extracting these links from HSI data without predefining the exact trait values.. These results demonstrate the potential of DeepPheno as a hypothesis-generating and candidate-screening framework for identifying genotype-associated spectral signatures and provide a foundation for further validation of image-based genotype-phenotype relationships in lettuce.

Keywords: Convolutional Neural Networks, Hyperspectral Imaging, High-Throughput Phenotyping, Plant Breeding, SNP Discovery, Lettuce, GWAS

Bioinformatics uncovers pregnancy-associated glycosylation patterns in plasma

Grigory Reznikov (Utrecht University)

A typical glycoproteomics experiment produces highly multidimensional data, with glycan distributions varying across donors, cohorts, proteins, and individual glycosylation sites. We analyzed plasma glycoproteomics data from pregnant women, cord blood, and non-pregnant controls using distribution-based distances to identify differences between cohorts. We found recurring patterns across glycosites, with pregnancy generally associated with larger glycans and cord blood with smaller ones. However, the magnitude of these changes, as well as specific glycan abundances, were strongly protein- and even site-specific. These results suggest that pregnancy-associated glycosylation cannot be explained by a single global mechanism, and understanding the origins and functional consequences of this site-specific regulation is a natural next step.

pip install refs-mcc: Out-of-the-Box Feature Selection for High-Dimensional, Low-Sample-Size Omics Data

Brigitta Varga (University of Amsterdam)

Introducing refs-mcc, a newly released Python library and Command-Line Interface (CLI) that makes the validated, scalable and reliable Matthews Correlation Coefficient–Recursive Ensemble Feature Selection (MCC-REFS) algorithm easily accessible to the bioinformatics community. Engineered specifically for high dimensional, low-sample-size (HDLSS) data, MCC-REFS provides out-of-the-box, deterministic feature selection that requires no hyperparameter tuning or predefined feature counts. In rigorous benchmarking against widely used feature selection approaches, including Boruta, SHAP, SelectKBest, GRACES, and Deep Neural Pursuit (DNP), MCC-REFS consistently achieved higher or comparable performance while selecting fewer features. Available on PyPI (pip install refs-mcc), refs-mcc provides bioinformaticians with a fast, transparent, and easy to use solution for reproducible biomarker discovery.

Project based collaborations at the Hogeschool Utrecht

Chris van Oevelen & Berend Wilkens (Hogeschool Utrecht)

1. ALS genomic data retrieval with natural language (chris.vanoevelen@hu.nl; ILC)

The research group of Kevin Kenna (<https://research.umcutrecht.nl/research-groups/neurogenetics/>) coordinate Project MinE (<https://projectmine.com/>). This project started in 2015 and is an international collaboration between 27 centers in 21 countries to sequence the genome of ALS patients to map genetic variants. The genetic information is stored in VCF format and is accessible in an SQL database. To retrieve information from this database a profound knowledge of SQL querying and VCF format is required. Here we present a collaboration between the Kenna group and the Institute of Life Science at the Hogeschool Utrecht to make the SQL database accessible using natural language (SQL queries i.e. Text-to-SQL or NL2SQL). Students have tested different LLM's and configurations to test benchmark questions including non-answerable questions. The overall score of the application is about 80% correct answers and 99% hallucination free.

2. Project based education as a base for collaboration (Berend.wilkens@hu.nl; ICT)

In the second part of this pitch we would like to introduce the ICT department to the UBC community. Their project based education provides excellent opportunities for collaborations with a focus on Artificial Intelligence, Cyber Security & Cloud and Software Development.

Poster Presentations

Poster session: 14:00 – 15:00, including the competition for the UBC Poster Prizes.

The posters are spread over four rooms on the 2nd floor: 2.303, 2.307, 2.311 and 2.319 (see the [map](#)). A sign at each door lists the poster numbers inside. Posters on related topics have nearby numbers. Click a poster number to jump to its abstract, or look someone up in the [index of presenters](#).

#	Presenter	Position	Affiliation	Title
1	Christopher Gaenssler	BiBC master student	Utrecht University	Investigating Ambiguous Rhabdomyosarcoma Samples Using Multi-omics Data
2	Victoria M. Cruz de Paz	Bioinformatician	Prinses Máxima Center	The Dutch childhood cancer genome project: Characterizing the landscape of early tumor driver events
3	Vasileios Papageorgiou	BiBC master student	Utrecht University	Calling out genetically engineered oncogenes at the single cell level for rare pediatric cancers.
4	Naresh de Bruijn	BiBC master student	Utrecht University	Identification of Pre- and Post-Infusion Biomarkers Associated with Relapse in Paediatric BCP-ALL Patients Undergoing CD19-Directed CAR T-Cell Therapy
5	Elizaveta Kulaeva	PhD candidate	Prinses Máxima Center	Somatic variant detection using an individual-specific pangenome reference: a study of a pediatric acute myeloid leukaemia patient
6	Nick Bos	BiBC master student	Utrecht University	Characterizing Cell-Free DNA fragmentation in health and pregnancy
7	Fatima Danawar	BiBC master student	Utrecht University	Air Pollution Exposure and Atherosclerotic Plaque Characteristics in the Athero-Express Biobank
8	Varshini Ganesan Selvi	BiBC master student	Utrecht University	Microscopy Video-Based Contraction Tracking for Force Curve Generation and Machine Learning Disease Prediction Using Cuore Platform Force Profiles
9	Gijs van Haaften	Senior researcher	UMC Utrecht	De novo assembly of the SMN locus reveals mechanisms of SMN1 loss in spinal muscular atrophy
10	Jamie Verkaart	BiBC master student	Utrecht University	Dissecting the Regulatory Landscape of Glucocorticoid Receptor Enhancers
11	Annemiek van de Steeg	BiBC master student	Utrecht University	Investigating changes in microglia-neuron communication in amyotrophic lateral sclerosis.
12	Emma van der Meij	BiBC master student	Utrecht University	Developmental origins of a human enriched neuron type
13	Gwen van der Sluis	PhD candidate	UMC Utrecht	Optimalisation and validation of genome phasing using long-read sequencing to enable monogenic Non-Invasive Prenatal Diagnostics.
14	Zita Rátkai	BiBC master student	Utrecht University	How the Absence of Retinal Input Shapes Cell Type Diversity in the Midbrain
15	Arianna Dato	BiBC master student	Utrecht University	CCA identifies a lipid axis of cortical thickness in youth and young adults, with PUFAs and ceramides at opposite ends
16	Feng Cao	BiBC master student	Utrecht University	HuiduRep: A Robust Self-Supervised Framework for Learning Neural Representations from Extracellular Recordings
17	Kahina Baouche	PhD candidate	Utrecht University	Uncovering genetic architecture of regional white matter hyperintensities using a non-parametric multivariate Genome Wide Association Study framework

[Back to programme](#)

#	Presenter	Position	Affiliation	Title
18	Griffin Kutler Dodd	BiBC master student	Utrecht University	High-dimensional HIV-1 quasispecies modeling guides escape-proof antibody design
19	Nikolas Hernandez	BiBC master student	Utrecht University	Mining Shotgun Proteomics for Antibody Gold
20	Jahaciel Villaverde Noguez	BiBC master student	Utrecht University	Ab-Bench-it a benchmarking tool for Ab-Ag cofolding methods
21	Shervin Mehryar	PhD candidate	Maastricht University	Concurrent Clinical Predictions using Electronic Health Care records and Biomedical Ontologies
22	Atman Shah	BiBC master student	Utrecht University	Structural Characterisation and De Novo Design of Protein Phosphorylation Site Pockets
23	Brigitta Varga	Student	University of Amsterdam	pip install refs-mcc: Out-of-the-Box Feature Selection for High-Dimensional, Low-Sample-Size Omics Data
24	Zhonghao Xue	BiBC master student	Utrecht University	Natural variation in co-expression clusters linked to environmental conditions in Arabidopsis
25	Herm Westerveld	BiBC master student	Utrecht University	From Images to Genomes: Decoding Plant Genetics through Automated Phenotyping
26	Fernando Sancho Gómez	PhD candidate	Utrecht University	Sinsei: A command-line toolkit for exploring single-cell (epi)genomics
27	Pim van Helvoort	BiBC master student	Utrecht University	Hyperspectral Imaging for Detection of Tipburn in Lettuce
28	Raphaëlle Versini	Postdoc	Utrecht University	Integrating the MARTINI2 Coarse-Grained Force Field into HADDOCK3 for Faster Modelling of Large Biomolecular Complexes
29	Manuel Leeuwerik	BiBC master student	Utrecht University	snakemake-MLSA-ANI: An automated workflow for multilocus sequence analysis and reproducible phylogenetic inference
30	Ise Maarsen	BiBC master student	Utrecht University	Uncovering the evolution of marine-derived Streptomyces species
31	Apostolia Bachsevanoglou Dimitriadou	Student	Utrecht University	Itaconate-related Post Translational Modifications
32	Laurens de Mooij	BiBC master student	Utrecht University	Gene Clustering in Bacteria and Archaea
33	Kaixin Hu	Postdoc	Utrecht University	SpreadARG: Metagenomic deep learning to support dissemination risk assessment of antimicrobial resistance genes
34	Reyanne Heijerman	BiBC master student	Utrecht University	Lifestyles of bacteriophages in the Griftpark
35	Lira Van Dinther	BiBC master student	Utrecht University	Virality of phages affects evolution of phage lysogeny decisionmaking
36	Pieter Hart	PhD candidate	Utrecht University	Conserved, highly organised genome architecture in Lactobacillaceae
37	Birk van Strien	BiBC master student	Utrecht University	Conflict resolution as an explanation for altruism
38	Annemiek Waajen	Postdoc	Utrecht University	Unraveling and quantifying the impacts of viruses on greenhouse gas emissions from agricultural peat soils
39	Simon García Moreno	BiBC master student	Utrecht University	The spatial configuration of habitat fragmentation: effects on connectivity and the diversity of differently dispersing trees

Poster abstracts

1 Investigating Ambiguous Rhabdomyosarcoma Samples Using Multi-omics Data

Christopher Gaenssler · BiBC master student · Utrecht University

Rhabdomyosarcoma (RMS) is the most common pediatric soft-tissue cancer comprising two major subtypes, embryonal RMS and alveolar RMS, which have different molecular characteristics, clinical behavior, and prognosis. Accurate subtype classification is important for diagnosis and clinical management. Classification is usually fairly straightforward with the subtypes differing by structural variants, copy number variations (CN), RNA expression, somatic mutations, methylation profiles and histology. However, several diagnostically ambiguous cases have been identified in which histology and methylation-based classification indicate alveolar RMS, while CN and RNA-sequencing evidence support embryonal RMS.

This project will investigate the molecular basis of this discordance and determine why these tumors resemble alveolar RMS by some modalities, despite other molecular evidence suggesting embryonal RMS. Six ambiguous RMS cases will be investigated alongside a larger cohort of conventionally classified alveolar RMS and embryonal RMS samples.

Analyses will integrate DNA methylation profiles, gene expression data from RNA sequencing, CNV patterns, and other relevant genomic alterations. By assessing whether the ambiguous cases consistently align with either subtype across molecular data types, the study will seek patterns or biological mechanisms that may explain their conflicting classification. This project aims to improve understanding of biological variability within RMS and help improve diagnostic classification models.

Data acquisition and analysis are ongoing. At the time of abstract submission, methylation data are available for reference cohorts and ambiguous cases. Other planned molecular data types are pending. The poster will present the study rationale, analytical approach, working hypotheses, and preliminary observations available by the symposium.

2 The Dutch childhood cancer genome project: Characterizing the landscape of early tumor driver events

Victoria M. Cruz de Paz · Bioinformatician · Prinses Máxima Center

Despite improved cure rates, cancer remains the leading cause of disease-related deaths among children in high-income countries. As childhood cancers are rare, concerted efforts are needed to identify relevant genomic alterations. The centralized Princess Máxima Center for Pediatric Oncology enables uniform data generation across The Netherlands. Whole-genome sequencing (WGS) and RNA-sequencing are routinely performed, yielding a homogeneous representative national dataset of 1,727 primary tumors, covering 42 hematological, 91 solid and 40 brain tumor entities.

Candidate driver events were derived from single nucleotide variants (SNVs), copy number alterations (CNAs), structural variants (SVs) and gene-fusions using paired tumor-normal WGS. Enrichment testing of SNVs, CNAs and SV breakpoints within genes gave per-variant type p-values, combined into harmonic mean p-values. Gene-fusions were called on RNA seq and sub-selected for high-confidence events with WGS-supporting SV calls. We identified 827 candidate driver genes, with ~98% of tumor samples containing an alteration in at least one. 30% of these genes show contributions from multiple mutation types, highlighting the value of integrating variant types for tumor driver identification. About a third of candidate drivers appear primary group specific. Of candidate drivers affected by an SV breakpoint, ~67% could be linked to a gene-fusion, ~22% of which involve a known oncogene. Preliminary functional analysis confirms cases of amplification-driven oncogene overexpression.

This is the first comprehensive landscape of tumor driver events in a nationally representative cohort of primary pediatric cancers. Functional analyses and analyses of early driver events will be finalized shortly, enhancing understanding of tumor biology and informing future clinical decision-making.

3 Calling out genetically engineered oncogenes at the single cell level for rare pediatric cancers.

Vasileios Papageorgiou · BiBC master student · Utrecht University

Medulloblastoma is the most prevalent pediatric malignant brain tumor, originating in the cerebellar region and accounting for approximately 60% of childhood intracranial embryonal tumors. While historically divided into four main molecular subgroups—WNT, SHH, Group 3, and Group 4—recent transcriptomic and methylation profiling has further refined these classifications. This project focuses on the Sonic Hedgehog (SHH) subgroup, which constitutes 25% of all cases and predominantly affects infants and young adults. SHH medulloblastoma is primarily driven by dysregulation in the SHH developmental pathway, frequently due to loss-of-function mutations. Furthermore, the disease is subclassified and exacerbated by alterations in the TP53 gene, alongside concurrent amplifications of oncogenes such as MYCN and GLI2. To overcome the limitations of traditional 2D cell cultures, which fail to capture true tumor complexity, this study utilizes 3D genetically engineered tumor organoids (GEBTOs). By introducing targeted genetic modifications—specifically the overexpression of MYCN and GLI2 combined with p53 deficiency—these organoids generate ectopic outgrowths that accurately mimic the neoplastic growth of SHH medulloblastoma. This provides a precise in vitro platform for disease characterization and the investigation of potential drug treatments. Historically, tumor profiling relied on bulk RNA-seq, which obscures vital intra-tumoral heterogeneity. To gain deeper insights into tumor biology, stem-like progenitors, and neoplastic populations, modern approaches utilize single-cell (scRNA-seq) and single-nuclei (snRNA-seq) RNA sequencing. Ultimately, this project aims to address current methodological challenges in accurately segmenting transformed from non-transformed cell populations at the single-cell level, paving the way for more targeted therapies.

4 Identification of Pre- and Post-Infusion Biomarkers Associated with Relapse in Paediatric BCP-ALL Patients Undergoing CD19-Directed CAR T-Cell Therapy

Naresh de Bruijn · BiBC master student · Utrecht University

B-cell precursor acute lymphoblastic leukaemia (BCP-ALL) is the most common paediatric malignancy, and while CD19-directed CAR T-cell therapy has transformed outcomes for patients with relapsed or refractory (r/r) disease, many still relapse or develop severe toxicity such as cytokine release syndrome (CRS) and immune effector cells-associated neurotoxicity syndrome (ICANS). Tools do exist to help identify these high-risk patients, such as the CAR-HEMATOTOX (HT) score. However, this and similar tools were developed almost exclusively in adult cohorts and rely on a single biomarker modality captured at a single timepoint.

In a retrospective cohort of 50 paediatric patients with BCP-ALL treated with CAR-T cell therapy at the Princess Máxima Centre, longitudinal features were derived from three data modalities; complete blood count (CBC), biochemistry, and haematological flow cytometry, across pre- and post-infusion time windows. Each modality was modelled independently via XGBoost Cox regression with leave-one-out cross-validation, producing one risk score per patient per modality. These scores were combined via a late-fusion CoxPH model.

Preliminary single-modality results suggest meaningful predictive signal in routinely collected CBC and flowcytometry data. A multi-modal, late-fusion approach combining these modalities may improve prediction of relapse and toxicity after CAR T-cell therapy in paediatric BCP-ALL beyond what a single modality of the adult-derived HT-score can achieve. Findings that will be confirmed once fusion-model validation is complete.

5 Somatic variant detection using an individual-specific pangenome reference: a study of a pediatric acute myeloid leukaemia patient

Elizaveta Kulaeva · PhD candidate · Prinses Máxima Center

Current clinical pipelines using linear reference genomes cause reads carrying mutations to be misaligned or lost, which is particularly problematic for low-coverage assays like single-cell whole-genome sequencing (scWGS), where technical noise can interfere with true somatic events. To overcome this, we applied a haplotype-resolved pangenome reference to a pediatric acute myeloid leukemia (pAML) case and assessed the clinical utility of this approach.

We constructed an individual-specific pangenome from long-read Nanopore sequencing of healthy haematopoietic stem and progenitor cells, incorporating two haplotype-resolved assemblies with Minigraph-Cactus. This added 36.3 Mb of novel sequence, 42% of which represented variation absent from the standard hg38 reference. We compared somatic variant calling across bulk, clonally expanded, and single-cell WGS data of healthy and cancerous cells. Reference bias in pangenome alignments was found to be significantly reduced compared to linear alignments. The personalized pangenome reduced reads with mates mapped to different chromosomes up to 8,000-fold which allowed more accurate split-read event detection and remapping of previously ambiguous alignments.

It also rescued SNVs previously filtered out by mapping quality thresholds and eliminated up to 80% of short indels in low-complexity regions. Subsequent driver analysis revealed that frameshift insertion artifacts in scWGS data were largely removed, and the medically relevant and computationally challenging PRSS1 locus was substantially disentangled due to reduced misalignment. Pangenome-based alignments also resulted in calling more SVs across both healthy and cancerous samples compared to hg38-based alignments, with the most pronounced differences in the number of deletions and duplications.

Our findings provide systematic evidence that personalized pangenome improves somatic variant detection in cancer genomics with the most benefit for scWGS data and complex genomic regions previously misaligned due to linear reference bias.

6 Characterizing Cell-Free DNA fragmentation in health and pregnancy

Nick Bos · BiBC master student · Utrecht University

Cell-free DNA (cfDNA) consists of highly fragmented DNA released into the bloodstream by dying cells. Its fragmentation is a non-random process, with the hematopoietic system as the major contributor to circulating cfDNA. During pregnancy, trophoblasts from the placenta also contribute cfDNA, providing a proxy for the fetal genome.

The first aim of this project is to catalogue characteristics of cfDNA in a cohort of healthy non-pregnant women using open access data. These characterizations support the overarching goal of identifying fragment differences that can help enrich for fetal cfDNA in maternal plasma during pregnancy.

The clearest characteristic of cfDNA is that most fragments have a length of around 166bp with a periodicity of 14bp under 151bp. This is consistent with literature and the length of a single nucleosome unit. Additionally, there is an overrepresentation of C-end motifs suggesting an association with CpG sites. It was found that around 8% of fragments end within a 10bp window of a CpG site. On average, higher methylation is associated with increased cleavage frequency near the Cytosine of a CpG site.

Upcoming work will use CySeq on samples taken during pregnancy and will focus on inferring maternal or fetal origin of each cfDNA fragment by incorporating parental genotype information. CySeq is specifically designed for liquid biopsy, provides increased accuracy and retains the true 3' end, which is often lost when using other platforms.

After inferring the origin of each cfDNA fragment into separate datasets, we expect to find differences in characteristics. On average, fetal cfDNA is shorter and methylation in the placenta is lower, suggesting that there might be a difference in whether fragment ends at a CpG site and therefore perhaps a difference in end-motif frequencies.

Identifying these differences can help to enrich for Fetal Fraction and support non-invasive prenatal testing for genetic diseases.

7 Air Pollution Exposure and Atherosclerotic Plaque Characteristics in the Athero-Express Biobank

Fatima Danawar · BiBC master student · Utrecht University

Air pollution is an important risk factor for cardiovascular disease, but the biological mechanisms linking air pollution to atherosclerosis are not fully understood. This study investigates whether long-term exposure to air pollution is associated with the composition and vulnerability of atherosclerotic plaques. Clinical and plaque data from patients in the Athero-Express Biobank were linked to annual air pollution estimates from RIVM. We investigated exposure to

PM2.5, PM10, nitrogen dioxide (NO₂), and elemental carbon (EC) in relation to multiple plaque characteristics, including inflammatory, lipid, fibrous, and vulnerability-related features. Regression analyses were performed while accounting for relevant patient characteristics and surgery year, and the Benjamini–Hochberg method is used to correct for multiple testing. Measures of urbanization are also explored. By studying different features of atherosclerotic plaques, this project aims to better understand how air pollution and the living environment may influence plaque composition and vulnerability, providing further insight into the pathways linking environmental exposures to cardiovascular disease.

8 Microscopy Video-Based Contraction Tracking for Force Curve Generation and Machine Learning Disease Prediction Using Cuore Platform Force Profiles

Varshini Ganesan Selvi · BiBC master student · Utrecht University

3D tissue engineered skeletal muscles (3D-TESMs) provide a physiologically relevant model for studying muscle function and disease phenotypes. Quantifying their contractile behaviour is essential for characterising disease associated changes, developing reliable functional readouts and possible drug and treatment effects. This thesis investigated two complementary computational aspects of 3D-TESM contractility analysis: automated extraction of contractile force data from microscopy videos and machine learning based classification of disease conditions using force profiles obtained directly from the Cuore platform.

First, microscopy videos of 3D-TESMs were analysed to investigate the feasibility of tracking muscle contraction at the micropillar interface and converting the resulting displacement into force-time profiles. Manual annotation and measurement using TrackMate provided the most reliable results among the approaches investigated, but required substantial user intervention and was time consuming. Optical flow based tracking was subsequently explored as a more automated alternative. Although this approach reduced manual intervention, variations in imaging quality, pixel density, and contraction dynamics resulted in substantial noise in the derived force profiles, limiting the reliability of automated force estimation. These findings highlight the need for more robust tracking approaches, potentially incorporating deep learning or other advanced image analysis techniques.

Second, experimentally measured force-time profiles obtained directly from the Cuore platform, which uses optical fibre sensing to quantify tissue contraction and converts these measurements into force profiles, were used for machine learning based disease classification. Baseline models, random forest models using extracted contractile features, and 1D CNN's using raw and temporally binned force profiles were evaluated. Group aware validation was employed to minimise data leakage between measurements from the same experimental well, while model interpretation approaches were explored to identify informative characteristics of the force profiles.

Together, this work evaluates both the challenges of extracting contractile force from microscopy data and the potential of force profiles for automated muscular disease classification.

9 De novo assembly of the SMN locus reveals mechanisms of SMN1 loss in spinal muscular atrophy

Gijs van Haften · Senior researcher · UMC Utrecht

The SMN locus on chromosome 5 is one of the most complex segmental duplications in the human genome and is the primary causal locus for spinal muscular atrophy (SMA), a severe neuromuscular disorder caused by loss of SMN1. The paralog SMN2 in the same locus can act as a modifier for disease severity; a high SMN2 CN is usually associated with milder SMA phenotypes. Despite this, SMN2 CN alone does not fully explain the SMA phenotypes, particularly among patients carrying three SMN2 copies. These findings suggest that additional genetic factors, such as the structure of the locus itself, could contribute to disease variability. However, the precise structural mechanisms leading to SMN1 loss have not been identified. De novo assembly of long-read sequencing enabled nucleotide resolution maps of the SMN locus of twelve SMA patients. These 24 assemblies allows for the identification of various previously uncharacterized SMN1 deletion breakpoints – including genes SERF1A, SMN1 and partially NAIP – one of which is a recurrent SMA-causing deletion observed in nine alleles. Three other alleles show different structural variation varying in size and orientation. All rearrangement breakpoints consistently occur within NAIP (pseudo)genes and/or GUSB

pseudogenes. The remaining alleles confirm the SMN1-SMN2 gene conversion debate but still show structural differences among alleles. Together, these results demonstrate how high-resolution, assembly-based genetics can uncover the mechanistic recombination events that delete SMN1, reveal recurrent population-shaped founder alleles, and define the structural genetic diversity underlying SMA. This work enables the identification of pathogenic structural variation in complex loci that have remained inaccessible to conventional methods.

10 Dissecting the Regulatory Landscape of Glucocorticoid Receptor Enhancers

Jamie Verkaart · BiBC master student · Utrecht University

Enhancers are crucial for the spatiotemporal gene expression patterns that ultimately result in the emergence of different cell types. Although enhancers have been extensively characterized based on their epigenomic properties, how their distinct regulatory features translate into gene expression at individual loci remains incompletely understood. Here, we employ the glucocorticoid receptor (GR), a nuclear receptor and transcription factor that is known to primarily bind at enhancers, to create an overview of the enhancer-gene landscape of its target genes and characterize their functional properties. We show that these enhancers can be segregated into seven discrete enhancer states based on their epigenomic features, with each state additionally exhibiting differential enrichment for regulatory annotations as well as unique gene-level distribution patterns. Moreover, the enhancer states were also characterized by different motif profiles that could additionally be linked to distinct mechanisms regulating gene expression, as knockout of the corresponding transcription factors had increasingly profound effects on the expression of genes associated with higher proportions of motif-containing enhancers. These findings lay the groundwork for a locus-specific overview of enhancer-mediated regulation of GR target genes and provide a framework for understanding how regulatory elements contribute to GR-mediated transcription.

11 Investigating changes in microglia-neuron communication in amyotrophic lateral sclerosis.

Annemiek van de Steeg · BiBC master student · Utrecht University

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease with dying motor neurons causing a progressive fatal paralysis. Microglia are a likely contributor to motor neuron death: They are a major source of differentially expressed genes (DEG) in ALS spinal cords; They are involved in ALS associated processes, such as neuroinflammation and excitotoxicity and ALS associated mutations, such as C9orf72 HRE, impair microglial function.

We use snRNA-sequencing data from post-mortem spinal cords of 4 controls and 20 patients to investigate the hypothesis that gene expression changes in ALS microglia induce ALS associated changes in neurons through altered microglia-neuron communication. We performed a pseudobulk DEG analysis, finding significant differential expression in homeostatic and disease associated microglia of ligands whose receptors are expressed by various neuronal subtypes. By combining the DEG results with known transcription factor regulons, we were able to identify transcription factors whose activity was significantly changed with ALS. We then used protein-protein interactions to look for potential signaling cascades connecting altered receptor activation with changes in transcription factor activity.

We were able to find potential signaling cascades connecting changes in microglial ligand expression with changes in neuronal transcription factor activity in each of the 7 neuronal subtypes investigated. The predicted signaling cascades contained 2 sections shared between neuronal subtypes, but were otherwise quite diverse, despite originating from a small number of receptors. Although still requiring experimental validation, these potential signaling cascades provide a starting point to further explore how changes in microglia-neuron communication may contribute to ALS progression.

12 Developmental origins of a human enriched neuron type

Emma van der Meij · BiBC master student · Utrecht University

Mouse models are commonly used to study brain development and function, but it is important to recognise which aspects of development can be directly compared between mice and humans, and which cannot. Previous analyses

comparing adult mouse and human brain cell types have identified a neuronal population in the midbrain that appears to be highly enriched in humans compared to mice. These cells were named Midbrain Derived Inhibitory (MDI) neurons. In this project we are looking at possible developmental origins for these MDI neurons to hopefully support the development of in vitro models and contribute to a better understanding of midbrain biology.

13 Optimisation and validation of genome phasing using long-read sequencing to enable monogenic Non-Invasive Prenatal Diagnostics.

Gwen van der Sluis · PhD candidate · UMC Utrecht

Monogenic genetic diseases can result in severe health effects. Knowing the disease state of the fetus enables parents to make reproductive choices or allows them to prepare for an affected child. Current diagnostics rely on invasive methods that increase the risk of miscarriage. This project aims to replace the current invasive diagnostic approaches with Non-Invasive Prenatal Diagnostics (NIPD), a blood test for both parents, enabling an earlier and safer diagnosis.

To determine the carrier status of the unborn child using our NIPD approach, we need to accurately reconstruct the haplotypes from the gDNA of the parents and use these within our cfDNA based relative haplotype dosage analysis (RHDO) statistical approach. In this study, we aim to determine and optimize the quality of haplotype phasing of Oxford Nanopore long-read sequencing. The HG002 Genome In A Bottle (GIAB) sample and trio data are used to validate our phasing quality. Trying out filtering steps on the data such as quality threshold and repeat regions, had a positive effect on the phasing quality. Going from 3923 haploblocks of which 480 contained errors before, to 4390 haploblock of which 148 contained phasing errors after filtering. In conclusion, our current filtering result in more/smaller haploblocks, but with higher quality.

Future analysis will focus on further optimization to become a base for the RHDO analysis and enable an accurate NIPD result

14 How the Absence of Retinal Input Shapes Cell Type Diversity in the Midbrain

Zita Rátkai · BiBC master student · Utrecht University

The superior colliculus (SC) is the principal subcortical target of retinal ganglion cell axons in the mouse and undergoes extensive activity-dependent refinement during the first two postnatal weeks, yet the role of retinal input in shaping the transcriptional identity of individual SC neuron subtypes remains poorly understood. Using single-cell RNA sequencing in the *Pitx3* mutant mouse, a model of complete and permanent retinal input absence from birth, this study profiled 5,866 glutamatergic neurons from postnatal day 8 superior colliculus and identified 18 transcriptionally distinct clusters annotated against the Allen Brain Cell Atlas. Differential expression analysis between wild-type (seeing) and *Pitx3* mutant (blind) animals revealed a graded, cell-type-specific transcriptional response ranging from complete silence in three clusters to 362 robust differentially expressed genes in cluster 16 (SCig Tfp2b Chrb3 Glut) after downsampling validation, with clusters belonging to the same annotated supertype showing divergent levels of responsiveness. A small core of genes, most prominently *Fabp7* and *Gm10076*, was differentially expressed across the majority of clusters with consistent directionality, indicating a shared global component alongside population-specific responses. Gene Ontology enrichment analysis revealed two opposing transcriptional programmes: genes requiring retinal input for their expression were enriched for synaptic transmission, glutamate receptor signalling, and synapse organisation, while genes elevated in blind animals were enriched for ribosomal translation and mitochondrial oxidative phosphorylation. This organisation mirrors the experience-dependent transcriptional structure previously described in the visual cortex and suggests that retinal input actively maintains a synaptic molecular identity in SC glutamatergic neurons that is replaced by a metabolic maintenance state in its absence, extending the principle of input-dependent transcriptional regulation to the subcortical visual pathway at single-cell resolution.

15 CCA identifies a lipid axis of cortical thickness in youth and young adults, with PUFAs and ceramides at opposite ends

Arianna Dato · BiBC master student · Utrecht University

Lipids are the major biomolecules of the brain, with critical roles in synaptogenesis, signalling, and myelination. However, their spatial distribution and relationship to cortical thickness remain sparsely explored. We investigated the patterns between the postmortem Lipidome Atlas of the Adult Human Brain and two developmental cohorts – YOUTH (9- and 12-year-olds) and the Human Connectome Project Young Adults (HCP-YA) – using canonical correlation analysis (CCA).

Across cohorts, we identified a consistent canonical axis separating two main groups: polyunsaturated fatty acids (PUFAs) and phospholipids at one end, and sphingomyelins, ceramides and triglycerides at the other. The PUFA pole was associated with thinner posterior regions, while the ceramide pole was associated with thicker frontal regions. We propose that this axis may reflect a signature of healthy cortical maturation, a possibility that future studies should address directly.

16 HuiduRep: A Robust Self-Supervised Framework for Learning Neural Representations from Extracellular Recordings

Feng Cao · BiBC master student · Utrecht University

Extracellular recordings are transient voltage fluctuations in the vicinity of neurons, serving as a fundamental modality in neuroscience for decoding brain activity at single-neuron resolution. Spike sorting, the process of attributing each detected spike to its corresponding neuron, is a pivotal step in brain sensing pipelines. However, it remains challenging under low signal-to-noise ratio (SNR), electrode drift, and cross-session variability. In this paper, we propose HuiduRep, a robust self-supervised representation learning framework that extracts discriminative and generalizable features from extracellular recordings. By integrating contrastive learning with a denoising autoencoder, HuiduRep learns latent representations robust to noise and drift. With HuiduRep, we develop a spike sorting pipeline that clusters spike representations without ground truth labels. Experiments on hybrid and real-world datasets demonstrate that HuiduRep achieves strong robustness. Furthermore, the pipeline significantly outperforms state-of-the-art tools such as KiloSort4 and MountainSort5 on accuracy and precision on diverse datasets. These findings demonstrate the potential of self-supervised spike representation learning as a foundational tool for robust and generalizable processing of extracellular recordings.

17 Uncovering genetic architecture of regional white matter hyperintensities using a non-parametric multivariate Genome Wide Association Study framework

Kahina Baouche · PhD candidate · Utrecht University

White Matter Hyperintensities (WMH) are hallmark magnetic resonance imaging (MRI) markers of cerebral small vessel disease and early predictors of cognitive decline and Alzheimer's disease (AD). Traditional genome-wide association studies (GWAS) typically apply univariate, parametric models that assume normality and may miss subtle or spatially distributed genetic signals. We introduce a non-parametric, multivariate GWAS framework, the Multivariate Asymptotic Non-Parametric Test of Association (MANTA), to jointly analyze region-specific WMH phenotypes without requiring distributional assumptions. Our study included 1,388 cognitively unimpaired participants from the ALFA (Alzheimer's and Families) cohort (mean age 55.71 years, 61.31% women). WMH volumes were regionally segmented (periventricular, deep, and juxtacortical regions) using a Bayesian inference algorithm applied to T1-weighted and T2-FLAIR MRI. Genetic data were obtained via the Illumina Neurochip array and imputed using the HRC r1.1 panel.

Using MANTA, we tested multivariate associations between genotype dosages and the three correlated WMH measures, adjusting for age, sex, education, total intracranial volume, and the first five genetic principal components. We identified five genome-wide significant loci threshold ($p < 1 \times 10^{-8}$), including novel signals undetected by standard univariate GWAS. To validate the added value of our method, we conducted a parallel MANOVA analysis on quantile-

normalized phenotypes, which failed to detect several of the MANTA, significant loci, underscoring the robustness and power of our non-parametric approach. Results were also compared with conventional univariate GWAS and parametric multivariate models, and sex-stratified GWAS were performed. Functional annotation and pathway analyses were conducted using MAGMA.

Gene-set enrichment of top hits pointed to neurodevelopmental and membrane organization pathways, consistent with known WMH biology and further supporting the biological validity of the findings. Sex-stratified analyses suggested additional sex-specific associations MANTA provides a statistically rigorous alternative to conventional GWAS, particularly suited for multivariate phenotypes with non-Gaussian distributions, and offers a scalable framework for imaging-genetics applications in aging and dementia research.

18 High-dimensional HIV-1 quasispecies modeling guides escape-proof antibody design

Griffin Kutler Dodd · BiBC master student · Utrecht University

Rapidly evolving viruses form diverse quasispecies that enable escape from immune responses and treatments. For example, HIV-1 can rebound within weeks of broadly neutralizing antibody (bNAb) treatment through the outgrowth of high-fitness escape mutants in the quasispecies or the evolution of new escape variants. Most existing models of viral dynamics consider only a small number of viral variants and either assume arbitrary mutant fitness distributions or require extensive fitting to sparse clinical data. Here, we develop a high-dimensional HIV-1 quasispecies model and parameterize this using *in silico* binding affinity predictions that captures the dynamics of millions of viral strains. Without fitting to experimental data, the model qualitatively reproduces viral rebound following bNAb treatment. We use the model to develop a quasispecies-based framework for antibody optimization and identify antibodies predicted to effectively suppress viremia. Together, our results demonstrate that integrating mechanistic genotype-phenotype maps with high-dimensional quasispecies models provides unprecedented insights into viral evolution.

19 Mining Shotgun Proteomics for Antibody Gold

Nikolas Hernandez · BiBC master student · Utrecht University

Immunoglobulins constitute roughly 20% of plasma protein by mass, yet comprehensive characterisation of the circulating antibody repertoire remains an unsolved analytical and computational problem. Shotgun proteomics can read repertoire composition at the amino acid level, but current methods cannot link peptides within a single chain or across heavy–light chain pairs, leaving antibody coupling inaccessible. We are developing a computational approach that infers chain pairing indirectly, by identifying germline-resolved immunoglobulin peptides across large numbers of donors and testing whether peptide abundances co-vary in patterns consistent with clonal co-expression.

We report progress on pipeline construction and data scoping. A custom germline database was assembled from IMGT reference sequences with framework and CDR annotation, and searched together with the full human proteome to prevent misassignment. Searches were configured under tryptic models. The pipeline was validated against a public monoclonal antibody spike-in benchmark, independently reproducing published quantification results, and replicate stability was characterised by coefficient of variation, correlation and peptide identification overlap.

Ongoing work applies the pipeline to COVID-19 plasma and serum cohorts, where antigen-driven clonal expansion, documented convergent variable-gene pairings and longitudinal seroconversion provide both stronger signal and independent validation for inferred chain pairings.

20 Ab-Bench-it a benchmarking tool for Ab-Ag cofolding methods

Jahaciél Villaverde Nogues · BiBC master student · Utrecht University

Co-folding tools are usually compared on a single accuracy number. That number cannot say whether a method generalised or memorised, how it fails when it fails, what the accuracy costs in stereochemistry, or whether the good

model can be found at all. We present Ab-bench-it, a configuration-driven benchmark that measures all four in one run, and apply it to five co-folding configurations – AlphaFold3, AlphaFold3 without an MSA, ESMFold2 (fast and MSA modes) and Boltz-2 – over 96 antibody–antigen targets unseen by every tool, forming 59 interface clusters, and 239,609 scored models.

Sampling is not the bottleneck: 72–86% of clusters contain a model at DockQ ≥ 0.23 , although per-model success is far lower (AlphaFold3 45% with an MSA, 27% without). Failures are not epitope failures – every incorrect best model still places the antibody on the right epitope and misses the pose, with CDR-H3 RMSD of 0.73–1.01 Å (Boltz-2 1.74 Å). Accuracy is partly recall: structural similarity to each tool's own pre-cutoff training set explains 10–15% of DockQ rank variance and the least training-like tertile loses 0.19–0.28 DockQ, while sequence similarity explains nothing. Accuracy also carries a stereochemical cost: the most accurate configuration, ESMFold2 with an MSA, has a MolProbity score of 3.21, driven by collapsed aromatic side chains rather than a worse interface.

The limiting step is ranking. On 38 AlphaFold3 ensembles of 500 models each, four scoring functions order models within a target only weakly (median ρ 0.14–0.24) and none outperforms another. Selecting the top K by score stalls at 71–76% success, whereas taking one model per binding mode reaches the oracle's 82% by K = 20.

Ab-bench-it is open source and runs any subset of these analyses from a single YAML file.

21 Concurrent Clinical Predictions using Electronic Health Care records and Biomedical Ontologies

Shervin Mehryar · PhD candidate · Maastricht University

The growing availability of electronic health records (EHR) presents increasing opportunities for data-driven applications in healthcare. These records, however, are represented using disparate coding systems and often suffer from fragmentation and structural inconsistencies. Such data fragmentation can lead to substantial performance gaps across different predictive tasks. In this work, a framework is proposed in order to mitigate these artefacts and enable multi-task predictions. To address performance gaps, we introduce a concurrent loss that guides the learning process over fragmented entities. To address structural misalignment, we propose a neural message-passing algorithm that enables the integration of ontological information. Experiments conducted on cardiovascular disease (CVD) patients demonstrate that the proposed system overcomes prediction gaps across multiple clinical tasks, including laboratory tests, diagnoses, medications, and mortality prediction, performed concurrently. The results highlight the potential of ontology-aligned representation learning for high-precision decision-making in healthcare systems.

22 Structural Characterisation and De Novo Design of Protein Phosphorylation Site Pockets

Atman Shah · BiBC master student · Utrecht University

Phosphorylation is the most widespread reversible modification in eukaryotic signalling. Its biological effect depends on recognition of the phosphorylated residue by reader proteins. These readers are spread across multitudes of different protein folds, and their binding pockets have mostly been described family by family rather than as a single structural class. We set out to test whether phosphosite pockets share a common structural class across different protein folds, and whether that architecture is specific enough to be scored and reproduced by de novo design. Here we characterise 341 buried phosphosite pockets containing phosphoserine, phosphothreonine or phosphotyrosine using three independent structural descriptions; discontinuous geometric motif search, graph-based topology combined with Rosetta energetics, and atomic interaction fingerprinting of the phosphate group. Geometric search of the AlphaFold database recovered known phosphorylation machinery, including receptor and non-receptor tyrosine phosphatases. A five-feature Gaussian mixture model trained on these pockets reached 97% leave-one-out recall and applied to the yeast proteome, retrieved a protein kinase A orthologue and ubiquitin. Unsupervised clustering of atomic interaction fingerprints separated the three modification types with a homogeneity of 0.981, while sequence embeddings from a protein language model failed to separate them under any tested setting, suggesting that the key differences come from structural information rather than the sequence. We then designed binders against consensus phosphosites by hallucination using iterative design sequencing and compared the resulting synthetic pockets to the natural set. Designs converge on direct electrostatic coordination, contacting the phosphate with basic residues far

more often than natural pockets do, which instead favour polar side chains. De novo design therefore recovers the general principle of gripping the phosphate electrostatically, but not the coordination chemistry seen around phosphosites in natural structures.

23 **pip install refs-mcc: Out-of-the-Box Feature Selection for High-Dimensional, Low-Sample-Size Omics Data**

Brigitta Varga · Student · University of Amsterdam

Introducing refs-mcc, a newly released Python library and Command-Line Interface (CLI) that makes the validated, scalable and reliable Matthews Correlation Coefficient–Recursive Ensemble Feature Selection (MCC-REFS) algorithm easily accessible to the bioinformatics community. Engineered specifically for high dimensional, low-sample-size (HDLSS) data, MCC-REFS provides out-of-the-box, deterministic feature selection that requires no hyperparameter tuning or predefined feature counts. In rigorous benchmarking against widely used feature selection approaches, including Boruta, SHAP, SelectKBest, GRACES, and Deep Neural Pursuit (DNP), MCC-REFS consistently achieved higher or comparable performance while selecting fewer features. Available on PyPI (`pip install refs-mcc`), refs-mcc provides bioinformaticians with a fast, transparent, and easy to use solution for reproducible biomarker discovery.

24 **Natural variation in co-expression clusters linked to environmental conditions in Arabidopsis**

Zhonghao Xue · BiBC master student · Utrecht University

The core idea of my project is to explore that whether natural variation in constitutive gene expression and the underlying hard coded architecture reflect the environmental conditions of accession origin.

25 **From Images to Genomes: Decoding Plant Genetics through Automated Phenotyping**

Herm Westerveld · BiBC master student · Utrecht University

Automated phenotyping can reveal the genetic variation underlying phenotypic traits. Here, we developed an automated image-based framework to quantify phenotypic plant features. The framework combines automated image processing with extraction of numerical phenotypic traits. Beyond conventional whole-leaf traits, the framework captures spatial variation within leaves and complex margin features such as leaf lobing and serration. By integrating phenotypic traits with genetic data, we employed the framework to investigate genetic factors underlying plant growth, color, shape, and architecture of individual leaves across 194 *Lactuca sativa* accessions. We further investigated these accessions under three environmental conditions (control, far-red light, and stress) to assess how interactions between genetics, leaf development, and environmental context shape phenotypic variation.

26 **Sincei: A command-line toolkit for exploring single-cell (epi)genomics**

Fernando Sancho Gómez · PhD candidate · Utrecht University

Emerging single-cell sequencing protocols allow researchers to study multiple layers of epigenetic regulation while resolving tissue heterogeneity. However, despite the rising popularity of such single-cell epigenomic assays, a lack of user-friendly computational tools for flexible quality control and genome-wide data exploration hinders their broad adoption. We introduce the Single-Cell Informatics (sincei) toolkit, a command-line interface for the exploration of data from a wide range of single-cell (epi)genomics protocols directly from aligned reads stored in BAM format. Sincei provides tools for preprocessing, feature identification, signal aggregation, dimensionality reduction, clustering and visualization. Results are stored in the AnnData format for seamless compatibility with a wide variety of tools. Additionally, sincei includes a Python API that enables advanced use cases, such as generating cell embeddings using

Generalized Principal Component Analysis (GLM-PCA). We show that sincei resolves cellular heterogeneity and improves interpretation of genomic regions across species, tissues and protocols.

27 Hyperspectral Imaging for Detection of Tipburn in Lettuce

Pim van Helvoort · BiBC master student · Utrecht University

Tipburn is a physiological disorder that can cause large losses in field-grown lettuce. Currently, early detection of tipburn is destructive, as it requires the lettuce to be cut open for visual inspection. In this project, we investigate whether tipburn can instead be predicted from raw hyperspectral images collected in the field. From each plot image, 150 patches of 30x30px are taken and grouped into a bag. These are processed by a 3D-2D convolutional neural network (CNN) and classified using a multiple-instance learning (MIL) model. An advantage of this approach is that the model also provides attention weights for individual patches, which can give an indication of which areas of the plot contributed most to the prediction.

The model achieved a ROC-AUC of 0.88 on the held-out data. To check whether this performance could be explained by chance, we compared it with a 20-run shuffled label null distribution, resulting in a one-sided p-value of 0.048. The attention weights are generally concentrated around the center of the lettuce. The model's confidence is not equally reliable for both classes. Highly confident healthy predictions tend to be reliable, while highly confident tipburn predictions are less reliable. We also found that standard classification metrics such as accuracy and F1-score can be unstable due to the class balance in the dataset. Further work is needed to determine whether the model is learning tipburn-related features or effects related to lettuce subtype, as well as to evaluate how stable the results are between different runs.

28 Integrating the MARTINI2 Coarse-Grained Force Field into HADDOCK3 for Faster Modelling of Large Biomolecular Complexes

Raphaëlle Versini · Postdoc · Utrecht University

The integration of coarse-grained (CG) approaches into docking workflows offers a powerful strategy for modelling large biomolecular assemblies with reduced computational costs. We present here the implementation of the MARTINI2 coarse-grained force field into the HADDOCK3 integrative modelling platform. This development enables the use of the CG representations and parameters within HADDOCK3 for efficient sampling and scoring of large protein-protein complexes. The implementation takes advantage of the modular and flexible architecture of HADDOCK3, allowing a seamless combination of MARTINI2 representation with the various modules. Conversion from and to all-atom models is integrated into the coarse-grained modelling workflow. The performance of the protocol is first assessed on protein-protein and protein-DNA benchmarks and then illustrated on a few representative large-scale systems, demonstrating a significant reduction in computational costs while maintaining biologically relevant accuracy. HADDOCK3 is freely available from <https://github.com/haddocking/haddock3>.

29 snakemake-MLSA-ANI: An automated workflow for multilocus sequence analysis and reproducible phylogenetic inference

Manuel Leeuwerik · BiBC master student · Utrecht University

The increasing availability of whole-genome sequencing data has expanded opportunities for fungal species characterization, genotyping, and phylogenetic analysis. However, species differentiation in many fungal taxa cannot often be reliably resolved using a single genetic marker due to limited specificity. Consequently, multilocus sequence analysis (MLSA) is widely applied in clinically and taxonomically relevant fungi, including *Aspergillus*, *Cryptococcus*, *Fusarium*, and *Penicillium*. Despite its widespread use, manual MLSA workflows can be time-intensive and are difficult to standardize. We therefore aimed to develop a reproducible open-source workflow for automated MLSA-based phylogenetic analysis.

A reproducible workflow was developed using Snakemake for workflow management and Conda for dependency management and shareability. Users provide genome assemblies together with loci of interest. Homologous target loci are automatically identified from genome assemblies through mapping and extraction approaches. Retrieved loci are then processed through locus-specific sequence alignment, concatenation, and phylogenetic inference. Additionally, the workflow incorporates Average Nucleotide Identity (ANI) analysis and automated locus-recovery reporting to support quality assessment. Publicly available genome assemblies may be integrated for reference inclusion and comparative analysis.

This open-source workflow, `snakemake-MLSA-ANI`, enables reproducible, user-friendly MLSA-based phylogenetic reconstruction directly from genome assemblies, reducing manual processing requirements. Potential applications include species confirmation, comparative phylogenetics, and taxonomic studies in fungal groups where single-marker approaches provide insufficient resolution.

The workflow is available through the Westerdijk Fungal Biodiversity Institute GitHub repository: <https://github.com/westerdijk-wm/snakemake-mlsa-ani>. Allowing users to run automated MLSA-based phylogenetic analysis for their fungus of interest.

30 Uncovering the evolution of marine-derived *Streptomyces* species

Ise Maarsen · BiBC master student · Utrecht University

The bacterial genus *Streptomyces* is a highly important group in medicine, being responsible for over 66% of antimicrobials used today, and is generally regarded as soil-dwelling. However, numerous species have been isolated from different habitats, including aquatic and highly saline environments. Many unique and medically relevant compounds have been identified from marine-derived *Streptomyces* species. Therefore, understanding this group could aid the ongoing search for new antimicrobial compounds. This project aims to identify a split between terrestrial and marine *Streptomyces* species within their phylogeny. Additionally, the genomic differences between terrestrial and marine *Streptomyces* species are investigated. Preliminary results show a clear clustering of marine-derived species within the *Streptomyces* phylogeny, as well as synteny between environment and melanin production. These patterns suggest marine-derived *Streptomyces* may represent a genomically distinct lineage with traits relevant to future antimicrobial discovery.

31 Itaconate-related Post Translational Modifications

Apostolia Bachsevanoglou Dimitriadou · Student · Utrecht University

Itaconate is an immunometabolite produced in activated macrophages by the mitochondrial enzyme Immune-Responsive Gene 1 (IRG1). Itaconate links metabolic rewiring to inflammatory and stress-response pathways, in part through covalent modification of proteins. Despite growing evidence that itaconate-derived post-translational modifications (PTMs) contribute to macrophage function, their protein targets and global prevalence remain poorly defined. In this study, we aim to systematically map itaconate-related PTMs in macrophage proteomics data and improve their computational detection. We analysed a publicly available data-dependent acquisition (DDA) dataset of LPS-stimulated mouse bone marrow-derived macrophages from wild-type (WT) and IRG1-knockout (IRG1-KO) conditions. The data were analysed using the MSFragger search engine in a closed-search framework to enable identification of the mass shift associated with itaconation. Surprisingly, itaconation events were also identified in IRG1-KO sample. This observation suggests the presence of an isobaric modification or false-positive assignments and highlights the need for additional features to distinguish true itaconation. Using matched itaconated and non-itaconated peptide pairs, we first investigated whether itaconation causes consistent changes in peptide fragmentation patterns or generates diagnostic ions. Neither feature appeared to provide a robust diagnostic signature of itaconation. Next, we assessed the retention time (RT) of our peptide pairs. Interestingly, a characteristic positive RT shift was observed for modified peptides in WT, but not in IRG1-KO. We then selected the population with this RT shift and subjected them to pathway enrichment analysis. Of note, enriched processes were associated with known aspects of itaconate biology, such as oxidative stress and mitochondrial function. These findings suggest that retention-time behavior can provide an additional feature for increasing confidence in the identification of itaconated peptides.

Ultimately, this work provides a systematic framework for detecting itaconation and advances our understanding of how metabolic reprogramming shapes macrophage signalling through covalent protein modification.

32 Gene Clustering in Bacteria and Archaea

Laurens de Mooij · BiBC master student · Utrecht University

In Bacteria and Archaea, co-regulation of genes is often achieved by polycistronic encoding, where a single operon codes for two or more different proteins. Eukaryotes, in contrast, are typified by monocistronically encoded genes. While much is known about gene regulation in Bacteria, the mechanisms for co-regulation of functionally related genes in Archaea are less understood. Existing research suggests functional gene groups may be clustered together within the genome, i.e. encoded monocistronically but on nearby sections of the chromosome. This hypothesis, however, requires appropriate testing.

In this study, which is still in progress, we aim to investigate gene clustering across various bacterial and archaeal lineages, in order to better understand the fundamental evolutionary processes driving gene regulation. Genes from a large number of genomes will be identified and annotated across selected lineages, so that genome-architectural properties of co-regulated genes can be quantified. We will detect operons using public tools, and develop methods to identify gene clustering beyond polycistronic encoding. Comparing the obtained results to the phylogenetic tree could then reveal potential evolutionary transitions from one regime to another. Additionally, if clusterings are distinct enough, it may be possible to use this information to provide additional functional insight to the many genes of unknown function in these domains, providing a potential new framework for gene annotation in Archaea. All in all, this project will provide a new view on the evolution of gene co-regulatory dynamics across domains.

33 SpreadARG: Metagenomic deep learning to support dissemination risk assessment of antimicrobial resistance genes

Kaixin Hu · Postdoc · Utrecht University

Genomic surveillance has become a powerful tool for identifying antimicrobial resistance determinants that predict antimicrobial susceptibility. However, antimicrobial resistance genes (ARGs) vary in their dissemination potential, and most studies emphasize mobile genetic elements, with limited integration of broader genomic context to assess ARG dissemination risk. Moreover, existing approaches typically assign static risk values to individual ARGs rather than providing sample-specific risk scores. To investigate genomic drivers of resistance spread across the One Health continuum, we leveraged metagenomic data to develop a deep learning model that scores ARG dissemination risk per sample.

We annotated 1228 metagenome samples for ARGs and other coding regions, identifying 462 ARGs with a total of 33,381 occurrences. For feature construction, we built a pseudo-pan-genome graph connecting all samples carrying a specific ARG, with each node representing a coding region. To generate training labels, we quantified the ARG dissemination risk across two dimensions: the number of species associated with ARG-containing contigs and the number of countries from which ARG-containing samples were collected. For the model architecture, we employed Graph Attention Networks (GATs) to aggregate information across nodes in the graphs and fed the resulting ARG-node embeddings into a multilayer perceptron (MLP) to predict the two-dimensional risk score.

Our work developed an aggregative scoring scheme that provides a per-sample risk metric of ARG dissemination threat. We further pinpointed the predictive genomic context associated with ARG spread by identifying a compact subgraph structure in the learned GAT model. Given a new metagenomics or genomics sample, we are developing a pipeline that performs assembly, predicts the species, annotates ARGs, and subsequently generates a dissemination risk score for each ARG, together with the genomic contexts that contribute to their dissemination.

34 Lifestyles of bacteriophages in the Griftpark

Reyanne Heijerman · BiBC master student · Utrecht University

Nowadays a recreational park, the Griftpark once gave home to a gas factory. This caused severe hydrocarbon pollution in the groundwater. To clean this up, a waste water treatment plant was implemented, in which the local “Griftpark bacteria” metabolized the hydrocarbons. One way in which these bacteria might have acquired this metabolic activity, is by the presence of auxiliary metabolic genes. These genes might be horizontally transferred by bacteriophages. Therefore, this poster discusses the different lifestyles of the bacteriophages in the Griftpark metagenome.

35 Virality of phages affects evolution of phage lysogeny decisionmaking

Lira Van Dinther · BiBC master student · Utrecht University

We created an agent based model of bacteria and bacteriophages, in which the bacteriophages could sense the growth rates of their hosts, and evolve their response to different levels of host growth. We observed that a strategy wherein phages lyse their host when it is starving, but remain lysogenic when the host is growing seemed to be the evolutionary stable strategy. However, if virality of the phage is high, it could struggle to find this strategy due to it needing to cross an evolutionary valley before reaching this strategy.

36 Conserved, highly organised genome architecture in Lactobacillaceae

Pieter Hart · PhD candidate · Utrecht University

In fast growing bacteria, essential and highly expressed genes are typically enriched near the origin of replication, where gene dosage is highest during rapid growth. We find the opposite in Lactobacillaceae: across 180 complete genomes, translation and replication/repair genes are consistently biased towards the replication terminus. Other functional categories show the reverse preference with transport and metabolism genes sitting closer to the origin. Gene age follows a separate gradient that runs independently of functional category, with evolutionarily old genes concentrated towards the terminus and young genes towards the origin. Both patterns are found across the entire Lactobacillaceae family, in genomes differing widely in size, lifestyle and predicted growth rate. Together these observations point to a genome architecture that is organised along an origin–terminus axis and raises the question of what keeps that organisation in place. Identifying the processes behind it is the aim of the next part of the project.

37 Conflict resolution as an explanation for altruism

Birk van Strien · BiBC master student · Utrecht University

Within the field of evolution, the topic of altruism has been a heavily debated topic. How can altruism exist if it is by definition unfavorable for the individual? Two possible explanations are group selection and enforcement. For this project I combined a group selection model with an epistasis model to test whether groupbselection and conflict resolution are possible explanations for why social groups are not taken over by cheaters. The results show that by increasing intergenomic epistasis, more conflicts get resolved in favour of the group and therefore cheating is reduced. Furthermore, if individuals increase their intragenomic epistasis to counter the groups influence, conflicts get resolved in favour of neither. In this case, cheating is still minimal. Therefore, conflict resolution can explain the a reduction of cheating. Furthermore, when decreasing group selection pressure, results are inconclusive. In some cases cheating is increased while other times it is reduced. In short, this unique model shows that altruism can be explained by enforcement of the group, with group selection possibly supporting the altruists.

38 Unraveling and quantifying the impacts of viruses on greenhouse gas emissions from agricultural peat soils

Annemiek Waajen · Postdoc · Utrecht University

Viruses are the most abundant biological entities on Earth. 97% of viruses are found in soils, and the vast majority of these are bacteriophages, viruses that infect bacteria. Despite their substantial abundance, soil bacteriophages are severely understudied. Recent studies on marine bacteriophages show that 25% of the microbially-fixed carbon is being dissolved each day through bacteriophage-induced bacterial cell lysis. This demonstrates the substantial impact that bacteriophages may have on their environment. In soils however, these mechanisms remain elusive.

Soils contain vast amounts of soil organic carbon (SOC), especially peatlands, which contain 21% of the global SOC while only covering 3% of the terrestrial surface. Soil bacteria can degrade SOC into greenhouse gases (GHGs) such as carbon dioxide (CO₂), methane (CH₄) and nitrous oxide (N₂O). Bacteriophages can influence the microbial community composition by bacteriophage-induced bacterial cell lysis, which releases nutrients and organic material that can be utilized as substrate by surviving microorganisms.

This research investigates the effect of viruses on greenhouse gas emissions from agricultural peat soils. We investigate the bacterial and viral biodiversity in a Dutch agricultural peatland, as well as after laboratory treatments with increased and decreased virus content. Multi-omics datasets are used to elucidate the virus-host-SOC-GHG interactions in these soils.

The emission of GHGs from peatlands has implications for climate change, as well as for the productivity of agricultural land, since SOC retention enhances crop productivity. By understanding the impact of viruses in peatland soils, agricultural practices could be altered to reduce SOC loss due to viruses, and conservation practices could be altered to reduce climate change impacts of these soils.

39 The spatial configuration of habitat fragmentation: effects on connectivity and the diversity of differently dispersing trees

Simon García Moreno · BiBC master student · Utrecht University

Habitat loss is one of the main causes for declines in biodiversity and ecological connectivity (henceforth connectivity). Connectivity describes the extent to which a habitat's spatial structure allows for a species to move among said habitat. Habitat loss decreases species' connectivity, making their populations more (genetically) isolated and thus more vulnerable. To counter the negative effects of habitat loss many attempts have been made at restoring lost habitat. Unfortunately, such attempts are often unsuccessful, in part because knowledge on restoration ecology is limited. How to measure connectivity has been a topic of discussion, and the effect of habitat's spatial configuration on restoration success is not well known. Here we used a near-neutral, spatially semi-explicit individual based model to simulate Amazon tree dynamics in a restoring habitat. In the model, different tree species, varying in their dispersal capacity, disperse through the landscape. Habitat was iteratively removed and subsequently restored using different spatial configurations (clustered, fractal and random). Using this model, we will address questions such as whether certain configurations restore diversity quicker, and how these configurations affect connectivity for different species. Our preliminary results show that clustered restoration generally leads to a quicker increase in species diversity, supporting more kinds of dispersal distances at earlier time steps. When there is little habitat, short distance dispersers outperform long distance dispersers because they are less likely to disperse into non-habitat. As habitat increases, there are more unbroken stretches of habitat for long distance dispersers to disperse in without landing in uninhabitable patches. Clustered restoration creates these conditions earlier in the simulation, as these configurations are more similar to pristine habitat conditions than patchy configurations are. We hope to still gain more insights into the effect of habitat configuration on species' connectivity.

[Back to programme](#)

[Back to programme](#)

Index of presenters

Keynote speakers, pitch speakers and poster presenters, alphabetically by surname. Click a reference to jump to the abstract.

B

Baaijens, Jasmijn	Keynote
Bachsevanoglou Dimitriadou, Apostolia	Poster 31
Baouche, Kahina	Poster 17
Berg, Joanna von	Pitch talk (10:45)
Bos, Nick	Poster 6
Bruijn, Naresh de	Poster 4

C

Cao, Feng	Poster 16
Cruz de Paz, Victoria M.	Poster 2

D

Danawar, Fatima	Poster 7
Dato, Arianna	Poster 15
Dinther, Lira Van	Poster 35

F

Farag, Roula	Pitch talk (10:45)
--------------	------------------------------------

G

Gaenssler, Christopher	Poster 1
Ganesan Selvi, Varshini	Poster 8
García Moreno, Simon	Poster 39

H

Haaften, Gijs van	Poster 9
Haitjema, Saskia	Keynote
Hart, Pieter	Poster 36
Heijerman, Reyanne	Poster 34
Helvoort, Pim van	Poster 27
Hernandez, Nikolas	Poster 19
Hooft, Justin J.J. van der	Keynote
Hop, Paul	Pitch talk (10:45)
Hu, Kaixin	Poster 33

K

Kulaeva, Elizaveta	Poster 5
Kutler Dodd, Griffin	Poster 18

L

Leeuwerik, Manuel	Poster 29
-------------------	---------------------------

M

Maarsen, Ise	Poster 30
Mehryar, Shervin	Poster 21
Meij, Emma van der	Poster 12
Mooij, Laurens de	Poster 32
Mouhib, Halima	Keynote

[Back to programme](#)

O

Oevelen, Chris van
Okyere, Frank

[Pitch talk \(15:30\)](#)
[Pitch talk \(15:30\)](#)

P

Papageorgiou, Vasileios

[Poster 3](#)

R

Raffle, Jade
Rátkai, Zita
Reznikov, Grigory

[Pitch talk \(10:45\)](#)
[Poster 14](#)
[Pitch talk \(15:30\)](#)

S

Sancho Gómez, Fernando
Shah, Atman
Sluis, Gwen van der
Spiering, Anna
Steeg, Annemiek van de
Strien, Birk van
Suurenbroek, Lianne

[Poster 26](#)
[Poster 22](#)
[Poster 13](#)
[Pitch talk \(10:45\)](#)
[Poster 11](#)
[Poster 37](#)
[Pitch talk \(10:45\)](#)

U

Ubels, Joske

[Pitch talk \(15:30\)](#)

V

Varga, Brigitta
Verkaart, Jamie
Versini, Raphaëlle
Villaverde Nogues, Jahaciel

[Pitch talk \(15:30\)](#), [Poster 23](#)
[Poster 10](#)
[Poster 28](#)
[Poster 20](#)

W

Waajen, Annemiek
Westerveld, Herm
Wilkens, Berend

[Poster 38](#)
[Poster 25](#)
[Pitch talk \(15:30\)](#)

X

Xue, Zhonghao

[Poster 24](#)

Z

Zelm, Eva van

[Pitch talk \(15:30\)](#)

[Back to programme](#)

Practical Information

How to get to the venue

The symposium will take place at Hogeschool Utrecht, [Padualaan 101, 3584 CH Utrecht](#).

Public transport: The venue is easily accessible by public transport. From Utrecht Central Station, trams 20, 21 and 22 take you to the Padualaan stop, which is a 1 minute walk to the venue.

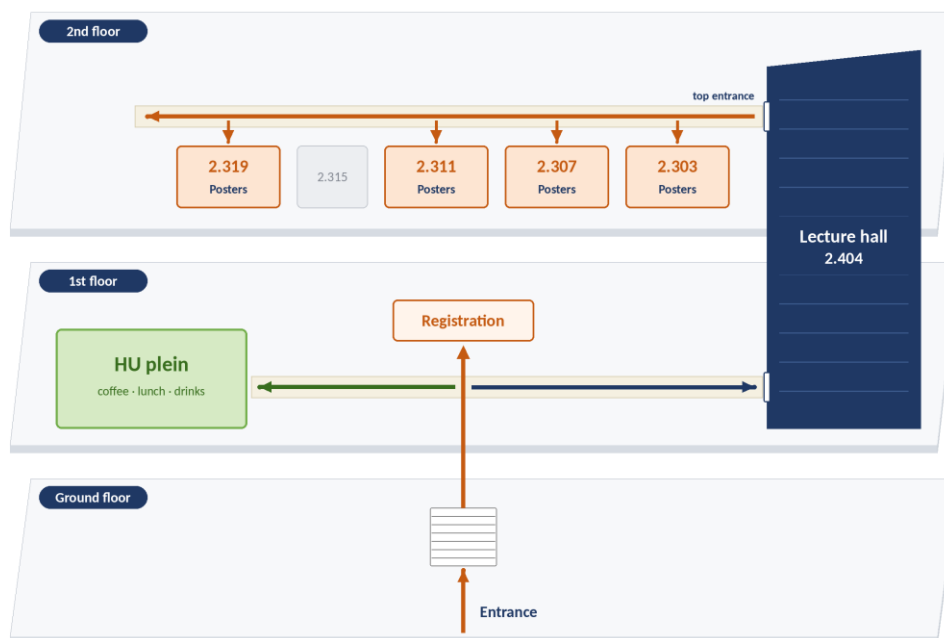
Car: If you're coming by car, there is a P+R facility nearby. More information about accessibility and travel to Utrecht Science Park can be found [here](#).

Bike: The venue is easily accessible by bike, and you can park your bike nearby.

Locations & directions

Location	What happens there
Registration desk	1st floor, at the top of the stairs from the entrance
HU plein	1st floor, to the left of the registration desk: coffee breaks, lunch, drinks & networking
Lecture hall 2.404	To the right of the registration desk (a few steps up): welcome, keynote lectures, pitch talks, closing ceremony & awards
Poster rooms	2nd floor: 2.303, 2.307, 2.311 and 2.319. Easiest route: leave the lecture hall via the top entrance and follow the corridor

Finding your way • Padualaan 101



Cloakroom

There is an unattended cloakroom next to the HU plein where you can leave your coat. Please keep valuables with you.

Photos

Photos will be taken at the symposium. If you would like to not be photographed, please let us know at the registration desk or send an e-mail to ubc@uu.nl.

[Back to programme](#)

Contact

Utrecht Bioinformatics Center — ubc@uu.nl

Utrecht Science Week

The UBC Symposium is proud to be part of Utrecht Science Week, a yearly celebration of science, innovation, and collaboration at Utrecht Science Park. This year marks the 6th edition, taking place from 25 September to 4 October 2026.

