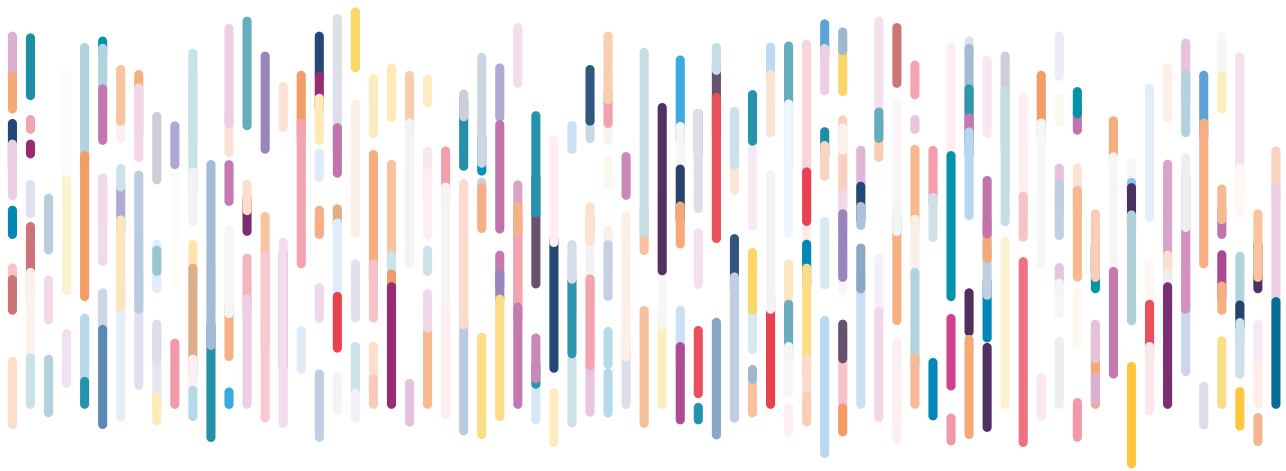
2025 Gene Forum

The 24th Annual International

9–10 September 2025 Tartu, Estonia



UNIVERSITY OF TARTU
Institute of Genomics

Eesti Geenikeskus
Estonian Genome Foundation

Funded by
the European Union

POSTER PRESENTERS

EVOLUTION AND POPULATION GENETICS		
1	The Underlying Evolutionary Dynamics of the Human Immune Response	Rutvi Rajpara
2	The conundrum of nc886 – methylation pattern with clear association to ancestral origins, but no genetic determinants?	Saara Marttila
3	Sperm competition intensifies purifying selection on testis-specific genes in primates	Vasili Pankratov
4	Application of spherical variational autoencoders to human population genetic structure	Elia Tiso
5	Signatures of selection and adaptation in an Arctic Circle population	Rodrigo Flores
GENETICS OF HUMAN TRAITS AND DISEASES		
6	GWAS reveals importance of vaginal epithelium associated genes in case of recurrent vaginitis	Evelin Mutli, Reet Mändar
7	Music style preferences and well-being: A genetic perspective	Anastasiia Bratchenko
8	Genetic architecture of the Big Five personality traits and their associations with health behaviours	Kerli Ilves
9	Integrating brain structure and function for the neurobiology and genetics of language	J.S. Amelink
10	Genetic non-additivity reveals variants in complex traits plasticity	Ralph Porneso
11	Association between plausible genetic factors and weight loss from GLP1-RA and bariatric surgery	Uku Vainik
12	GWAS Meta-Analysis of ADHD Symptoms in Adults Reveals Differential Genetic Architecture of Inattention and Hyperactivity-Impulsivity Domains	Triinu Varvas
13	Coding variant analysis based on 5,522 ICD-10-based disease phenotypes in the Estonian biobank	Kanwal Batool
14	Recessive variants in the DARS2 gene as a novel cause of axonal Charcot-Marie-Tooth disease	Siiri Sarv
15	Interpreting artificial neural networks to detect genome-wide association signals for complex traits	Burak Yelmen
16	Leveraging primary health care data from Estonian Biobank to discover novel genetic associations	Anastasiia Alekseienko
17	**Characterisation of rare copy number variation effects on circulating metabolic biomarkers**	**Maarja Jõeloo**
18	Text-based approach for detecting cases of ADEs from EHRs of participants of the Estonian Biobank	Dage Särg

MICROBIOME		
19	Exploring the Estonian Gut Microbiota: Composition, Structure, and Research Potentia	Hindrek Teder
20	**Metagenome-assembled genomes from a population-based cohort uncover novel gut species and strain diversity, revealing prevalent disease associations**	**Kateryna Pantiukh**
21	A Bioconductor workflow for microbiome-based survival analysis	Sneha Das
22	Probabilistic alternatives for microbiome research in R/Bioconductor	Rasmus Hindström
23	Predictive markers of future depression in the gut microbiome	Annabel Klemets
24	Modeling geographic variation of human microbiome	João Paulo Cassucci dos Santos
25	Global antimicrobial resistance patterns in human gut metagenomes are structured along socio-economic gradients	Mahkameh Salehi
26	Biome-specific genome catalogues reveal functional potential of shallow sequencing	Dattatray Mongad
REPRODUCTIVE GENETICS		
27	Leveraging genome-wide association study summary statistics to identify possible drug targets for uterine fibroids	Lisette Haug
28	Maternal Mood Disorders Are Associated With Increased Proportion of Placental Immune Cell Types Driving Gene Expression Differences	Soile Hytti
29	**Multi-ancestry, trans-generational GWAS meta-analysis of gestational diabetes and glycaemic traits during pregnancy reveals limited evidence of pregnancy-specific genetic effects**	**Valentina Rukins**
30	Genetic risk factors of vasomotor symptoms	Mia Golob
RISK & PREDICTION		
31	Multi-omic risk prediction scores for rheumatoid arthritis in the Estonian Biobank	Galadriel Velázquez
32	No evidence for a causal effect of thyroid function and disease on risk for attention-deficit/hyperactivity disorder: A two-sample Mendelian randomization study	Triinu Peters
33	PREV-GEN – PREVentative outreach with GENetic testing	Kalle Pärn
34	**Association between meat intake and subclinical atherosclerosis in the population-based Swedish CardioPulmonary BioImage Study**	**Getachew Arage**

35	Integrating Polygenic Risk Scores into OMOP Common Data Model using the Estonian Biobank: A Preliminary Study of AI-Driven Cardiovascular Disease Prediction in Atherosclerosis	Djeane Debora Onthoni
36	Identifying predictors of weight loss from a diverse set of biological, behavioural, and psychological factors	Birgit Malken
37	Predicting blood metabolite profiles using multi-target neural networks	Merve Nur Güler
38	CVDLINK: Federated AI for Advancing Predictive Models of Cardiovascular Disease	Urmo Võsa
PERSPECTIVE ON ADVANCED GENOMIC TECHNOLOGIES		
39	EASIGEN-DS: Designing a distributed Research Infrastructure on Advanced Genomics Technologies	Mireia Vaca-Dempere
40	The Genome of Europe: A Reference Dataset of Genomes	Andres Metspalu, Merit Kreitsberg
EVOLUTION AND POPULATION GENETICS		
41	Genetic origins of the Kiritimati population from central-eastern Micronesia	Kai Tätte

1

The Underlying Evolutionary Dynamics of the Human Immune Response

Rutvi Rajpara[1], Karlis Pleiko[2], Tambet Teesalu[2], Monika Karmin[1], Irene Gallego Romero[1,3], Michael Dannemann[1]

[1]Institue of Genomics, University of Tartu, Estonia
[2]Institute of Biomedicine and Translational Medicine
[3]Human Genomics and Evolution, St Vincent's Institute of Medical Research, Fitzroy, Australia

E-mail address of presenting author: rutvi.rajpara@ut.ee

Background and study question: Anatomically Modern Humans have been exposed to a wide variety of pathogens throughout their existence. Being able to adapt to them was a matter of life and death. Genetic variants that helped to fight against pathogens remained in the human gene pool. Today, we still carry some of these genetic modifications in our DNA that impact human health. However, linking specific pathogen outbreaks to their impact on shaping modern human genomes is challenging. This research delves into the evolutionary dynamics of the human immune response, focusing on how past episodes of natural selection and admixture with archaic humans, such as Neanderthals and Denisovans, have shaped the genetic composition of present-day human populations.

Methods: To achieve this, we use genomic and immune-related datasets to identify loci associated with pathogen-specific immune responses and investigate their evolutionary pathways. By reconstructing allele frequency changes, examining signatures of selection, we aim to understand how pathogens shaped these genetic adaptations. Furthermore, we investigate past admixture events and the contribution of archaic introgressed alleles to immune response loci to determine their role in pathogen-specific adaptations. The integration of genomic data, GWAS, eQTL data, and computational tools helps identify associations between evolutionary processes and immune response phenotypes.

Main results: The findings of this research will yield novel insights into the relationship between pathogens and disease-associated variants in present-day individuals. They will highlight how historical evolutionary events shape human health and disease susceptibility.

Wider implications: These insights may inform innovative therapeutic strategies by leveraging the legacy of evolutionary adaptations for medical advancements. Moreover, understanding the evolutionary context of immune variation can also help future studies in personalised medicine and infectious disease risk prediction.

2

THE CONUNDRUM OF *NC886* – METHYLATION PATTERN WITH CLEAR ASSOCIATION TO ANCESTRAL ORIGINS, BUT NO GENETIC DETERMINANTS?

Saara Marttila[1], Sonja Rajic[1], Olli Raitakari[2], Terho Lehtimäki[3], Emma Raitoharju[1]

[1]Molecular Epidemiology (MOLE), Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[2]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland
[3]Department of Clinical Chemistry, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

E-mail address of presenting author: saara.marttila@tuni.fi

Background: *nc886* (VTRNA2-1) locus displays a maternal polymorphic imprinting pattern. The locus has gained interest as a potential mediator of the DOHaD hypothesis; *nc886* imprinting status of an individual is associated with maternal phenotypes but also in later life with their own cardiometabolic health. At population level, the proportion of individuals with an imprinted *nc886* locus shows clear association to ancestral origins (Figure 1), implying a genetic component in the establishment of *nc886* imprinting status.

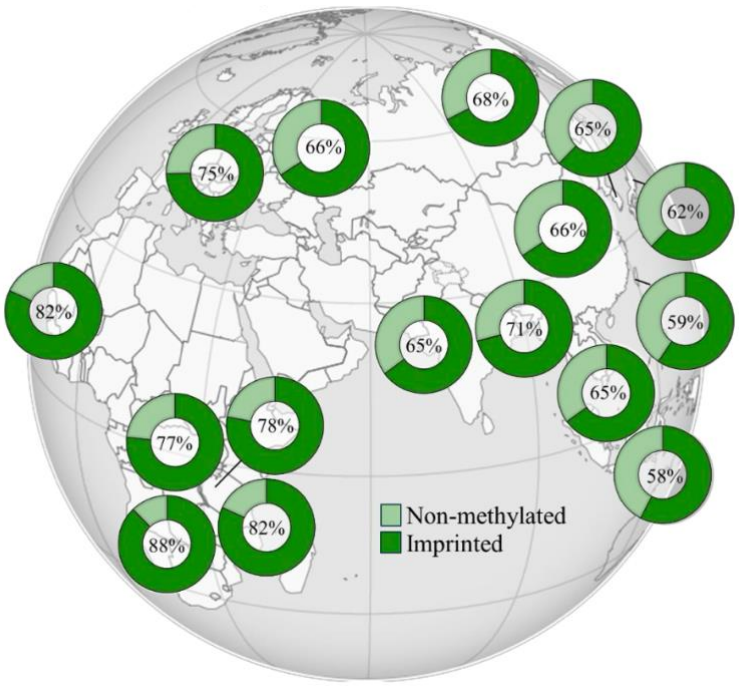


Figure 1. In populations with African, European and Asian ancestral origins, the proportions of individuals with an imprinted *nc886* locus are ~80%, ~75% and ~65%, respectively.

Study question: Our aim was to identify the genetic determinants of *nc886* imprinting status.

Methods: Potential inheritance patterns of *nc886* methylation status were analysed in three-generational family data (183 families) in the Young Finns Study. Genetic determinants of *nc886* methylation status were analysed (GWAS, n>50 000) as part of the GoDMC2 consortium and through a maternal GWAS (n=1000).

Main results: Family data did not reveal obvious patterns of *nc886* methylation status inheritance. So far, we have not been able to identify genetic variants to be associated with *nc886* methylation status.

Limitations: Number of individuals was limited in the family analysis and in the maternal GWAS.

Wider implications: Currently, we can't explain the robust population level pattern of *nc886* methylation, as we can't identify any genetic determinants. We hope future collaborative efforts would help to solve this conundrum. As an example of polymorphic imprinting and a potential mediator of DOHaD, understanding *nc886* could offer keys to understand these wider biological phenomena.

3

SPERM COMPETITION INTENSIFIES PURIFYING SELECTION ON TESTIS-SPECIFIC GENES IN PRIMATES

Vasili Pankratov, Bjarke Meyer Pedersen, Mengjun Wu, Juraj Bergman, Mikkel Heide Schierup

Aarhus University, Department of Molecular Biology and Genetics, Bioinformatic Research Centre

E-mail address of presenting author: vasilipankratov@gmail.com

Background: Mating systems with one female mating with different males within a relatively short time lead to sperm competition. In turn, sperm competition is thought to impose strong selective pressures on male reproductive biology traits, including relative testis size and efficiency of spermatogenesis.

Study question: We test whether more intense sperm results in stronger purifying selection acting on testis-specific genes in primates.

Methods: To address this, we use the ratio of nucleotide diversity at nonsynonymous and synonymous sites in protein-coding genes ($\pi N/\pi S$) as a measure of purifying selection. We estimate this value in the concatenated sequence of 821 genes with testis-specific expression, and compare it across 136 primate species with different levels of sperm competition.

Main results: We show that $\pi N/\pi S$ in testis-specific genes is lower (resulting from more intense purifying selection) in species with higher levels of sperm competition (p-value=9.8×10^{-5}).

Limitations: One limitation of our work is that due to low levels of genetic polymorphism and small sample size for many primate species, we cannot estimate $\pi N/\pi S$ for individual genes and had to concatenate the target genes. We plan to detect the subset of genes driving the signal by counting synonymous and non-synonymous polymorphisms across species grouped based on levels of sperm competition.

Wider implications: Our results show that species with low sperm competition are prone to accumulating (slightly) deleterious mutations affecting spermatogenesis. Given that humans are a low sperm competition species, identifying genes with relaxed selection under low sperm competition will allow us to better understand the genetics of male fertility.

4

APPLICATION OF SPHERICAL VARIATIONAL AUTOENCODERS TO HUMAN POPULATION GENETIC STRUCTURE

Elia Tiso[1], Alessandro Raveane[2], Michela Carlotta Massi[2], Massimo Mezzavilla[4]

[1]University of Tartu
[2]Human Technopole
[4]University of Padua

E-mail address of presenting author: elia.tiso@ut.ee

Background: The analysis of human genetic variation makes it possible to trace the evolutionary past of populations and investigate its influence on complex traits and phenotypes. A key challenge in GWAS is that, unless population structure is properly detected or accounted for, it may introduce confounding, leading to spurious associations of genetic variants with traits.

Study question: Traditional methods starts with a dimensionality reduction step, such as PCA — which has many advantages over more complex techniques, but remains limited in representation and visualization if not combined with other tools —. Such tools need to improve *interpretability* to better assess or correct for population structure.

Methods: Drawing inspiration from the link between genetics and geography, we propose a tool based on variational autoencoders, to process PCA output and obtain a visual representation on a sphere (S^2) that combines and condenses the information of the top PCs given in input.

Main results: Our main example analyses the European genetic variation on a model trained with worldwide genetic data. We combined multiple PCs and translated our ideas for what represents a *good visualization*, obtaining interesting results, capable of correctly address genetic insights. These visualizations are much more powerful than the simple 2D or 3D PCs plots.

Limitations: We consider a flexible yet simple architecture which can be studied further. We may also refine our metrics and test population-specific models to see how it works on a smaller scale. We applied this model to PCA outputs, which inherently bring some errors or at least limit the captured information.

Wider implications: If properly tuned, it could be a starting step for a clustering analysis (detection of population structure) also in ancient individuals. It could also be implemented in PRS to increase trasferability to under-represented populations.

5

SIGNATURES OF SELECTION AND ADAPTATION IN AN ARCTIC CIRCLE POPULATION

Rodrigo Flores[1], Liisa Loog[2], Anders Eriksson[1]

[1]Center for Genomics, Evolution and Medicine, Institute of Genomics, University of Tartu
[2]Estonian Genome Center, Institute of Genomics, University of Tartu

E-mail address of presenting author: rodrigo.flores@ut.ee

Background: The Northern Finland Birth Cohort (NFBC) is a population-based prospective cohort of ~12,000 genotyped individuals from Northern Finland, of which ~5,000 reside within the Arctic Circle. Importantly, phenotypic and physiological data—including metabolite levels, BMI, diagnoses, and lifestyle questionnaires—have been collected since the cohort's establishment in the 1960s and 1980s.

Study question: This unique dataset enables the study of genomic signatures of adaptation to Arctic-specific environmental pressures such as cold climate, low UV exposure, and region-specific diets. We aim to identify such regions and link candidate adaptive variants to relevant available phenotypes.

Methods: We performed exploratory principal component analysis (PCA) of NFBC genotype data alongside external datasets. Signatures of selection were investigated using the population branch statistic (PBS), for example comparing the Arctic subset to Saami and a non-Arctic outgroup. Extended haplotype homozygosity (EHH) scans to detect recent selection events were also employed.

Main results: PCA revealed a subtle but detectable genetic differentiation of the Arctic subset along the north–south European genetic cline, indicating a distinctive background within the NFBC. Selection scan results are preliminary and currently under interpretation.

Limitations: The genetic differentiation between Arctic and non-Arctic NFBC participants is modest, potentially diluting adaptation signals. Additionally, recent migration history into the region may limit the time frame for strong selection to leave detectable genomic footprints.

Wider implications: This work will contribute to understanding how human populations adapt to extreme environments, with potential insights into physiological, metabolic, and genetic mechanisms relevant to health in cold, low-light, and specialized-diet settings.

GWAS REVEALS IMPORTANCE OF VAGINAL EPITHELIUM ASSOCIATED GENES IN CASE OF RECURRENT VAGINITIS

Evelin Mutli[1], **Reet Mändar**[1], Kairi Koort[2], Andres Salumets[1,3,4], Estonian Biobank Research Team[5], Triin Laisk[1]

[1]University of Tartu, Tartu, Estonia
[2]Tallinn University, Tallinn, Estonia
[3]Celvia CC AS, Tartu, Estonia
[4]Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden
[5]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia

E-mail address of presenting author: reet.mandar@ut.ee; evelin_mutli@hotmail.com

Background: Recurrent vaginitis is a leading reason for visiting a gynaecologist, with bacterial vaginosis (BV) and vulvovaginal candidiasis (VVC) being the most common diagnoses. Reasons and mechanisms behind their recurrent nature are poorly understood.

Study question: Which genetic risk factors for recurrent vaginitis can be revealed using large-scale GWAS study?

Methods: The study included 6,870 cases (at least two episodes of vaginitis) and 5,945 controls (no vaginitis episodes) from the Estonian Biobank. GWAS approach included single marker and gene-based analyses, followed by functional annotation of associated variants and candidate gene mapping.

Main results: The gene-based association analysis identified KRT6A gene that is a member of the keratin protein family. Comparison of our results with previous studies provided support for LBP (associated with immune response to vaginal bacteria) and PRKCH genes (possible role in keratinocyte differentiation and susceptibility to candidiasis). Thus, both the results of the gene-based analysis and the testing of candidate genes/variants point to the unexplored role of the epithelium in the development of vaginitis. Although the vaginal epithelium is categorised as non-keratinised, it still uses keratin proteins to create a durable and complete structure.

Limitations: The background data of the subjects may not be fully complete, especially in older women. These preliminary findings described above need independent replication.

Wider implications: This study is the first highlighting a potential role of the vaginal epithelium in recurrent vaginitis.

Acknowledgements: Enterprise Estonia (grants No. EU48695), Estonian Research Council (grant No. IUT34-19, TEM-TA28 and PRG1076), Horizon 2020 innovation grant (ERIN, grant no. EU952516), Estonian Ministry of Education and Research (grant No. KOGU-HUMB).

MUSIC STYLE PREFERENCES AND WELL-BEING: A GENETIC PERSPECTIVE

Anastasiia Bratchenko[1,2], Penghao Xia[1,3], Dorret I. Boomsma[4], Miriam A. Mosing[1,3,5,6], Fredrik Ullén[1,3], Laura W. Wesseldijk[1,3,7]

[1]Department of Neuroscience, Karolinska Institutet, Sweden
[2]Institute of Genomics, University of Tartu, Estonia
[3]Department of Cognitive Neuropsychology, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany
[4]Department of Complex Trait Genetics, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[5]Melbourne School of Psychological Sciences, Faculty of Medicine, Dentistry, and Health Sciences, University of Melbourne, Australia
[6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden
[7]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, the Netherlands

E-mail address of presenting author: anastasiia.bratchenko98@gmail.com

Background: The association between music and well-being is complex, spanning emotional, mental, physical, and social domains. However, it remains unclear whether preferences for specific music styles are related to well-being and to what extent genetic factors influence these associations.

Study question: We aimed to investigate whether music style preferences are associated with well-being and whether genetic factors underlie these associations.

Methods: Data were collected from 8,879 adult monozygotic and dizygotic twins from the Swedish Twin Registry, including a genotyped subset of 3,764 individuals. Well-being was assessed using the WHO-10 questionnaire, and participants rated their preferences for 19 music styles. Sex, age and levels of education were included in the models as covariates. Associations between music preferences and well-being were examined using regression models accounting for familial clustering, and co-twin control analyses in monozygotic pairs discordant for music preferences. We also tested whether polygenic indices for well-being were associated with music preferences.

Main results: Higher well-being was associated with stronger preferences for pop, gospel, and Swedish dance band music, while a preference for indie music was associated with lower well-being. The co-twin control analyses suggested that these associations were influenced by shared genetic/family environmental factors, providing no support for a causal relationship. Polygenic indices for well-being did not predict the music preferences.

Limitations and wider implementations: Given the cultural context—most participants were around 50 years old and favored music styles unique to Sweden—generalizability

may be limited. Future research should explore genetic and environmental factors influencing music preferences to clarify their underlying biology.

Note

This abstract is based on work previously published in Personality and Individual Differences (Bratchenko et al., 2025).

References

Bratchenko, A., Xia, P., Boomsma, D. I., Mosing, M. A., Ullén, F., & Wesseldijk, L. W. (2025). Music style preferences and well-being: A genetic perspective. Personality and Individual Differences, 241, 113162. https://doi.org/10.1016/j.paid.2025.113162

GENETIC ARCHITECTURE OF THE BIG FIVE PERSONALITY TRAITS AND THEIR ASSOCIATIONS WITH HEALTH BEHAVIOURS

Kerli Ilves

Institute of Genomics, University of Tartu

E-mail address of presenting author: kerli.ilves@ut.ee

Background: Personality traits have replicable associations to a range of life outcomes. The NEO Personality Inventory has long been used to measure these individual differences across five domains of personality. Uncovering the genetic basis of personality can improve our understanding of many clinical and behavioural outcomes. However, past attempts in detecting the genetic architecture of personality have been underpowered, beyond Neuroticism.

Study question: We aimed to explore the genetic architecture of the Big Five and its genetic correlates with other socially relevant behavioural phenotypes.

Methods: We performed genome-wide association meta-analysis across 46 cohorts (N~611K to 1.14 million) and published a new well-powered GWAS on personality. Analyses included population-based and within-family (N~51K) GWAS, polygenic predictions, genetic correlations and Mendelian randomisation.

Main results: We identified 1257 lead SNPs, including 823 novel variants. Common variants explained 4.8%-9.3% of the trait variance. Genetic effects were consistent across geography, reporter (self vs other), age, and instrument, and spousal assortment was minimal. We quantified the widespread relevance of personality genetics to several socially relevant behaviours; for example, Extraversion increased COVID infection risk, Conscientiousness reduced likelihood of smoking initiation, and Neuroticism decreased liability for longevity. Furthermore, we found bidirectional effects, suggesting health behaviours that may shape personality development.

Limitations: Future work should further triangulate on the MR-based results with additional robust methods to strengthen the causal inference. Granular personality data would enable more sophisticated downstream analyses.

Wider implications: Our findings along open-access statistics support further research into the dynamic interplay between personality and life experiences, and personality-informed health risk predictions.

INTEGRATING BRAIN STRUCTURE AND FUNCTION FOR THE NEUROBIOLOGY AND GENETICS OF LANGUAGE

J.S. Amelink[1], S. Soheili-Nezhad[1], G. Alagöz[1], A. Llera[2,3], M-Y. Wang[1], K. V. Haak[3,4], S. E. Fisher[1,3], C. F. Beckmann[2,3] and C. Francks[1,2,3]

[1]Language & Genetics department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]Department of Medical Neuroscience. Radboudumc, Nijmegen, The Netherlands
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands
[4]Department of Cognitive Science and Artificial Intelligence, School of Humanity and Digital Sciences, Tilburg University, Tilburg, The Netherlands

E-mail address of presenting author: jitse.amelink@mpi.nl

Background: Brain structure and function have largely been studied separately in relation to the neurobiology and genetics of language. This distinction is partly artificial, as structural and functional organization are intertwined.

Study question: Which aspects of structural brain variance are associated with functional language connectivity, and what are the genetic and behavioural correlates of the shared patterns?

Methods: We used linked independent component analysis to integrate language network functional connectivity with brain volumetric and white matter structure in 32,677 UK-Biobank participants, followed by analysis of behavioural and genetic correlates of the derived multimodal imaging components.

Main results: Stronger functional connectivity between brain language areas was associated with expansion of parts of the cerebellum and motor cortex, together with smaller ventricles and sensory parietal and occipital areas. The multimodal language components mediated an association between picture vocabulary level and the polygenic influence on reading ability. We report 18 genomic loci associated with these components, corroborating in particular the role of PLCE1 and INPP5A in brain language connectivity. SNP-heritability estimates for multimodal components were roughly 25-30%, and there was significant enrichment of heritability in primate-conserved genomic loci, and foetal brain human-gained enhancer elements.

Limitations: Our dataset is limited to a relatively healthy, white middle-aged dataset with a brief functional scan, which may limit the generalization of findings. Language-related cognitive tests are also limited in this dataset.

Wider implications: This study revealed that structural correlates of functional language network connectivity extend well beyond previously defined language areas of the brain, and highlight the value of multimodal brain phenotyping for human neurogenetic discovery.

GENETIC NON-ADDITIVITY REVEALS VARIANTS IN COMPLEX TRAITS PLASTICITY

Ralph Porneso[1], Alexandra Havdahl[2], Espen Moen Eilertsen[1], Eivind Ystrøm[1]

[1]PROMENTA Research Center, Department of Psychology, University of Oslo, Oslo, Norway
[2]PyschGen Centre for Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, Oslo, Norway

E-mail address of presenting author: ralphp@uio.no

Background: The contribution of non-additive genetic variation in human traits is nil for dominance, small for epistasis and, theory suggests, minimal for gene-environment interaction. Evidence from behavioral genetics indicates genetic interactions play a role in cognitive traits, and recent genome-wide association studies (GWAS) on height and BMI have identified single nucleotide polymorphisms (SNPs) involved in interactions. This study is a first step to uncover the molecular mechanisms behind gene-gene and gene-environment interactions and, in part, bridge the gap in the missing heritability problem.

Study question: Are there SNPs associated with complex traits plasticity?

Methods: We used a non-linear mixed effects framework to concurrently estimate SNPs' mean (i.e. additive) and within individual variability (i.e. non-additive) effects on height and BMI at 6, 8 and 12 months (N=~45,000) and Math and Reading scores at grades 5, 8 and 9 (N=~65,000) from the Norwegian Mother Father and Child cohort (MoBa). We compared our models against a newly developed pipeline, trajGWAS, which restricts discovery to variants with linear within subject (WS) variance effects. Our approach relaxes this restriction. In addition, we investigated whether time-varying effects overlap with non-additive effects.

Main results: We found putative "plasticity" variants with linear and non-linear effects. We provide within-individual variability summary statistics for height, BMI, Math and Reading. We found SNP time-varying effects explain ~1% of non-additive effects on average but found no overlap.

Limitations: At present, our pipeline runs only using multiple cores and takes a considerable amount of compute hours.

ASSOCIATION BETWEEN PLAUSIBLE GENETIC FACTORS AND WEIGHT LOSS FROM GLP1-RA AND BARIATRIC SURGERY

Uku Vainik[1] and The genomics of GLP1-RA consortium

[1]University of Tartu & McGill University

E-mail address of presenting author: uku.vainik@ut.ee

Background: Obesity is a major public health challenge. Glucagon-like peptide-1 receptor agonists (GLP1-RA) and bariatric surgery (BS) are effective weight loss interventions; however, the genetic factors influencing treatment response remain largely unexplored. Moreover, most previous studies have focused on race and ethnicity rather than genetic ancestry.

Study question: We tested genetic predictors of weight loss after GLP1-RA & BS.

Methods: We analyzed electronic health records of 10,960 individuals from 9 multiancestry biobank studies across 6 countries. As predictors, we used 15 genetic variants either related to GLP1R or weight change. We further tested polygenic scores for body mass index (BMI) and Type 2 Diabetes.

Main results: Between 6 and 12 months, GLP1-RA users had an average weight change of −3.93% or −6.00%, depending on outcome definition, with modest ancestry-based differences. BS patients experienced −21.17% weight change between 6 and 48 months. We found no significant associations between GLP1-RA-induced weight loss and polygenic scores for body mass index or type 2 diabetes, nor with any genetic variants. A higher BMI polygenic score was modestly linked to lower weight loss after BS (+0.7% per s.d., $P = 1.24 \times 10^{-4}$).

Limitations: Our study was underpowered for a genome-wide analysis; we could not accurately estimate treatment adherence; we used ancestry labels based on major continental ancestry groups; we had no control group so hard to distinguish intentional from unintentional weight loss.

Wider implications: Our findings suggest known genetic factors have limited impact on GLP1-RA effectiveness with respect to weight change and confirm treatment efficacy across ancestry groups.

GWAS Meta-Analysis of ADHD Symptoms in Adults Reveals Differential Genetic Architecture of Inattention and Hyperactivity-Impulsivity Domains

Triinu Varvas[1], Elis Haan[1], Laura Hegemann[2,4], Natàlia Pujol-Gualdo[3], Sofiya Babok[1], Kristi Krebs[1], Katri Pärna[1], Kadri Kõiv[1], Siim Kurvits[1], Triin Laisk[1], Alexandra Karoline Saasen Havdahl[2], Elizabeth C. Corfield[2,4,5], Martje Bos[6], Melissa Vos[6], Harold Snieder[7], Isabell Brikell[7,8,9], Henrik Larsson[7,10], Estonian Biobank Research Team[1], Catharina Hartman[6], Helga Ask[2,12], Kelli Lehto[1]

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Estonia
[2]PsychGen Centre for Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, Oslo, Norway
[3]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Catalonia
[4]Psychiatric Genetic Epidemiology Group, Research Department, Lovisenberg Diaconal Hospital
[5]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[6]University of Groningen, University Medical Center Groningen, Department of Psychiatry, Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), Groningen, the Netherlands
[7]University of Groningen, University Medical Center Groningen, Department of Epidemiology, Unit of Genetic Epidemiology and Bioinformatics, Groningen, the Netherlands
[8]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[9]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway
[10]Department of Biomedicine, Aarhus University, Aarhus, Denmark
[11]School of Medical Sciences, Örebro University, Örebro, Sweden
[12]PROMENTA Research Center, Department of Psychology, University of Oslo, Oslo, Norway

E-mail address of presenting author: triinu.varvas@ut.ee

Background: Attention-deficit hyperactivity disorder (ADHD) diagnoses in adults are increasing worldwide. However, there is limited knowledge about the genetics underlying ADHD symptoms in adulthood.

Study question: What is the genetic basis of self-reported ADHD symptoms in adults?

Methods: Six-item Adult ADHD Self-Report Scale (ASRS v1.1) was used in the Estonian Biobank (EstBB, n=77,726) and the Norwegian Mother, Father and Child Cohort Study (MoBa, n=64,229) parent generation. GWAS meta-analyses (N=141,505) of ADHD

inattention (INA), and hyperactivity/impulsivity (HYP/IMP) domain score were performed. The meta-analyses were followed by downstream analyses, including polygenic score (PGS) prediction in the Lifelines cohort (n=19,857) and PheWAS in the EstBB subsample (n=125,290).

Main results: We identified five independent genetic loci for INA (*HTR1D*, *KCNN2*, *H4C8*, *FNBP4*, *CADM1*), and two for HYP/IMP domain (*ARNTL/BMAL1*, *FOXP2*). SNP heritability (h^2_{SNP}) was higher for INA (h^2_{SNP}=0.09, s.e.=0.005) compared to HYP/IMP (h^2_{SNP}=0.05, s.e.=0.005). The genetic correlation (r_g) between the domains was moderate (r_g=0.35, 95% CI [0.26, 0.43]). Tissue set enrichment was shown in 13 brain regions for the INA, and in two for the HYP/IMP domain. The PGS analyses showed significant associations with respective phenotypes in the Lifelines (PGS_{INA}: β=0.09, 95% CI [0.07, 0.10]; $PGS_{HYP/IMP}$: β=0.07, 95% CI [0.06, 0.09]). R_g-s with 55 related traits revealed significant differences between the two symptom domains. PheWAS findings showed 37 somatic and psychiatric ICD-10 diagnoses associated with $PGS_{HYP/IMP}$, while five were detected for PGS_{INA}.

Limitations: HYP/IMP items primarily capture hyperactivity rather than impulsivity.

Wider implications: These findings point to the importance of understanding the distinct mechanisms underlying ADHD symptoms across the lifespan.

CODING VARIANT ANALYSIS BASED ON 5,522 ICD-10-BASED DISEASE PHENOTYPES IN THE ESTONIAN BIOBANK

Kanwal Batool[1], Erik Abner[1], Anastasiia Alekseienko[1], Klavs Jermakovs[1,2], Hele Haapaniemi[5], Satu Strausz[5,6], Estonian Biobank Research Team[1], FinnGen[5], Health Informatics Research Team[4], Lehte Türk[1], Anu Reigo[1], Teele Palumaa[1,3], Hanna M. Ollila[5,6,7,8], Jaanika Kronberg[1], Urmo Võsa[1], Tõnu Esko[1], Priit Palta[1]

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Estonia
[2]Yusuf Hamied Department of Chemistry, University of Cambridge, UK
[3]Department of Ophthalmology, Emory University, Atlanta, USA
[4]Institute of Computer Science, University of Tartu
[5]Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland
[6]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[7]Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
[8]Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

E-mail address of presenting author: kanwal.batool@ut.ee

Background and study question: Genome-wide association studies (GWAS) have identified numerous variants underlying complex traits, yet population-specific studies, especially for low-frequency and rare variants, remain limited. We examined how the low-frequency coding variants in EstBB are linked to underlying biology with potential clinical relevance.

Methods: Using data from 212,955 Estonian Biobank (EstBB) participants, we conducted a multi-trait GWAS to investigate coding variant associations. Genome-wide significant variants were intersected with gnomAD v4.1.0 exonic variants. We annotated variants, calculated allele frequency (AF) enrichment relative to global and non-Finnish European populations. Clinically relevant variants were identified via DrugBank and ACMG SF v3.2.

Main results: Overall, we identified 3,540 genome-wide significant coding variant associations, of which 109 variants ranging from low-frequency to ultra-rare, were enriched in EstBB. For example, the protective variant *SCN11A*:p.Gly1736Val associated with reduced migraine risk (OR=0.71, $p=2.74\times10^{-9}$, AF=0.001). This neuron-specific voltage-gated sodium ion channel participates in pain signalling and the variant may alter ion channel intracellular localization, implicating it as a promising therapeutic target. Conversely, the *GOT1*:p.Gln208Glu variant is enriched ~10 fold more in the Estonian

population and increases chronic liver disease risk up to threefold (OR=3.02, $p=3.6\times10^{-11}$, AF=0.0046).

Limitations: Population specific low frequency variants are difficult to replicate in other biobanks and depend on functional studies for validation.

Wider implications: However, these results demonstrate analysis of the population-specific rare variants can reveal unique genetic risk factors, enabling more precise disease prediction and targeted clinical strategies.

RECESSIVE VARIANTS IN THE *DARS2* GENE AS A NOVEL CAUSE OF AXONAL CHARCOT-MARIE-TOOTH DISEASE

Siiri Sarv[1,*], Berta Estévez-Arias[2,3,*], Nathalie Bonello Palot[4,5], Laura Carrera-García[2,6], Carlos Ortez[2,6,7], Jesica Expósito-Escudero[2,6], Delia Yubero[7,8], Jordi Muchart[9], Emilien Delmont[10], Eve Õiglane-Shlik[11], Teele Meren[12], Sanna Puusepp[13], Ülle Murumets[12], Gajja S Salomons[14], Bjarne Udd[15,16,17], Andrés Nascimiento[2,6,7], Katrin Õunap[1,12], Janet Hoenicka[2,7], Francesc Palau[2,7,18], Daniel Natera-de Benito[2,6,7]

[1]Institute of Clinical Medicine, University of Tartu, Tartu, Estonia
[2]Neuromuscular Unit, Department of Neurology, Hospital Sant Joan de Déu, Barcelona, Spain
[3]Laboratory of Neurogenetics and Molecular Medicine, Center for Genomic Sciences in Medicine, Institut de Recerca Sant Joan de Déu, Barcelona, Spain
[4]Marseille Medical Genetics, Aix-Marseille University–Inserm UMR 1251, Marseille, France
[5]Genetic Department, Timone Hospital, APHM, Marseille, France
[6]Applied Research in Neuromuscular Diseases, Institut de Recerca Sant Joan de Déu, Barcelona, Spain
[7]Center for Biomedical Research Network on Rare Diseases (CIBERER), ISCIII, Madrid, Spain
[8]Department of Genetic and Molecular Medicine-IPER, Hospital Sant Joan de Déu and Institut de Recerca Sant Joan de Déu, Barcelona, Spain
[9]Department of Radiology, Hospital Sant Joan de Déu, Barcelona, Spain
[10]Reference Center for Neuromuscular Disorders and ALS, APHM, CHU La Timone, Filnemus, ERN Neuro-NMD, Marseille, France
[11]Children's Clinic, Tartu University Hospital, Tartu, Estonia
[12]Genetics and Personalized Medicine Clinic, Tartu University Hospital, Tartu, Estonia
[13]Department of Pathology, Tartu University Hospital, Tartu, Estonia
[14]Department Laboratory Medicine, Laboratory Genetic Metabolic Diseases, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[15]Folkhälsan Research Center, Helsinki, Finland
[16]Medicum, University of Helsinki, Helsinki, Finland
[17]Tampere Neuromuscular Center, Tampere University Hospital, Tampere, Finland
[18]Division of Pediatrics, Faculty of Medicine and Health Sciences, Universitat de Barcelona (UB), Barcelona, Spain

E-mail address of presenting author: siiri.sarv@ut.ee

Background: Pathogenic variants in the aspartyl-tRNA synthetase 2 (*DARS2*) gene are known to cause leukoencephalopathy with brainstem and spinal cord involvement and lactate elevation (LBSL, MIM#611105).

Study question: Are *DARS2* recessively inherited variants the main cause of axonal Charcot-Marie-Tooth (CMT) phenotype in five individuals from three unrelated families?

Methods: We retrospectively collected clinical information of all individuals and performed functional studies on three affected persons.

Main results: All individuals showed progressive distal muscle weakness starting in childhood, with absent lower limb deep tendon reflexes. Electrophysiological studies indicated an axonal motor and sensory neuropathy. Two pairs of adult siblings had normal cognitive function, while one patient had moderate intellectual disability. Brain MRI scans were normal, and whole exome sequencing revealed compound heterozygous variants in the *DARS2* gene (c.713C>T; p.Ser238Phe and c.1006C>T; p.Arg336Cys in family 1; c.74del; p.Ile25Thrfs*38 and c.713C>T; p.Ser238Phe in family 2; c.492+2T>C and c.1508C>T p.Pro503Leu in family 3). Functional analysis showed reduced DARS2 protein activity (25% for c.1508C>T) and altered mitochondrial localization with disrupted mitochondrial network morphology.

Limitations: Our limitation is the size of the study group, since we only have five patients. And as with other transferases, how DARS2 causes axonal CMT remains unclear, but it suggests a non-canonical function of ARS proteins in peripheral nerve.

Wider implications: We present a new phenotype of mostly pure axonal CMT in five individuals linked to recessively inherited *DARS2* variants. This highlights the role of mitochondrial aspartyl-tRNA synthetase in peripheral nerve function and potential therapeutic targets for CMT.

Funding: PRG2040, 2024 FI-1 00075, CP22/00141.

15

INTERPRETING ARTIFICIAL NEURAL NETWORKS TO DETECT GENOME-WIDE ASSOCIATION SIGNALS FOR COMPLEX TRAITS

Burak Yelmen[1], Maris Alver[1], Merve Nur Güler[1], Estonian Biobank Research Team[1], Flora Jay[2], Lili Milani[1]

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[2]CNRS, INRIA, LISN, Paris-Saclay University, Orsay, France

E-mail address of presenting author: burak.yelmen@ut.ee

Background: Investigating the genetic architecture of complex diseases is challenging. Although GWAS have identified thousands of variants for multiple complex traits, conventional statistical approaches can be limited by simplified assumptions such as linearity and lack of epistasis in models.

Study question: Deep learning can potentially overcome these limitations with holistic modelling of genomic data and detect novel variants.

Methods: We first trained artificial neural networks to predict complex traits using both simulated and real genotype-phenotype datasets. We extracted feature importance scores via different post hoc interpretability methods to identify potentially associated loci (PAL) for the target phenotype and devised an approach for obtaining p-values for the detected PAL.

Main results: Simulations with various parameters demonstrated that associated loci can be detected with good precision using strict selection criteria. By applying our approach to the schizophrenia cohort in the Estonian Biobank, we detected multiple loci associated with this highly polygenic and heritable disorder. There was significant concordance between PAL and loci previously associated with schizophrenia and bipolar disorder, with enrichment analyses of genes within the identified PAL predominantly highlighting terms related to brain morphology and function.

Limitations: Main limitations of the method stem from difficulties in training large neural networks (e.g., computational resources, stochasticity) and lack of generalized methodologies for obtaining statistical significance for findings.

Wider implications: With advancements in model optimization and uncertainty quantification, neural networks have the potential to enhance the identification of genomic loci associated with complex diseases, offering a more comprehensive approach for GWAS and serving as initial screening tools for subsequent functional studies.

LEVERAGING PRIMARY HEALTH CARE DATA FROM ESTONIAN BIOBANK TO DISCOVER NOVEL GENETIC ASSOCIATIONS

Anastasiia Alekseienko[1,*], Erik Abner[1,*], Kanwal Batool[1], Klavs Jermakovs[1,2], Hele Haapaniemi[5], Satu Strausz[5,6], Estonian Biobank Research Team[1], FinnGen[5], Health Informatics Research Team[4], Lehte Türk[1], Anu Reigo[1], Teele Palumaa[1,3], Hanna M. Ollila[5,6,7,8], Jaanika Kronberg[1], Urmo Võsa[1], Tõnu Esko[1,#], Priit Palta[1,#]

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
[2]Yusuf Hamied Department of Chemistry, University of Cambridge, UK
[3]Department of Ophthalmology, Emory University, Atlanta, USA
[4]Institute of Computer Science, University of Tartu
[5]Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland
[6]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[7]Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
[8]Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

E-mail address of presenting author: anastasiia.alekseienko@ut.ee

Background: Genome-wide association studies (GWAS) have significantly advanced the understanding of genetic mechanisms underlying complex human diseases and traits by systematically identifying genetic variants linked to diverse phenotypic traits across diverse populations. Large-scale analyses that combine multiple phenotypes are especially valuable, as they can reveal shared genetic architectures and patterns of comorbidity, refining disease classification and risk prediction.

Study question: Overall, this research aims to enhance our understanding of the genetic determinants of human traits and diseases.

Methods: The presented work includes performing GWAS and applying several post-GWAS steps to the genetic data from 206,159 participants from the Estonian Biobank. The workflow included analysing genetic and phenotypic correlations between the phenotypes, estimating SNP heritability, and identifying novel variants, all done in a high-throughput manner. Our study includes 5,522 phenotypes defined using the ICD-10 disease classification coding system: 4,884 non-sex-specific, 546 female-specific, and 62 male-specific disease phenotypes.

Main results: By analysing EstBB genotype data, altogether 18,977,777 SNV and indel variants (including common, low-frequency and rare variants), our analyses revealed 813

unique genome-wide significant loci, including 176 putatively novel locus-phenotype associations.

Limitations: One limitation of this GWAS-based approach is its reliance on ICD-10-coded disease phenotypes, which may substantially reduce statistical power despite cost- and time-efficiency. Additionally, the results from such analyses need to be evaluated critically, especially when phenotype definitions are heterogeneous and exhibit highly skewed sex-specific distributions.

Wider implications: Our study highlights the potential of population-based biobanks in discovering common and rare genetic associations, contributing to identifying novel disease-associated loci and expanding the catalogue of human disease-related genetic variants.

CHARACTERISATION OF RARE COPY NUMBER VARIATION EFFECTS ON CIRCULATING METABOLIC BIOMARKERS

Maarja Jõeloo[1], Adriaan van der Graaf[2,3], Nele Taba[1], Chiara Auwerx[2,3], Reedik Mägi[1], Urmo Võsa[1], Kaur Alasoo[4], Zoltán Kutalik[2,3], Lili Milani[1]

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[4]Institute of Computer Science, University of Tartu, Tartu, Estonia

E-mail address of presenting author: maarja.joeloo@ut.ee

Background: Metabolites are small molecules in the bloodstream that regulate physiological processes and offer insights into disease mechanisms. They are valuable for genetic studies seeking accessible biomarkers and drug targets. While single nucleotide variant (SNV) genome-wide association studies (GWAS) have examined metabolite levels, copy number variants (CNVs)—despite their potential for having highly deleterious effects—have so far been excluded.

Study question: Here, we present the results from the first ever biobank-scale CNV GWAS for circulating metabolic markers and examine the implications of metabolite-associated CNVs for drug response.

Methods: We conducted a GWAS meta-analysis on microarray-based CNV events for 249 metabolic traits using 189,167 Estonian Biobank (EstBB) and 187,250 UK Biobank participants. We mapped the results to 340 genes involved in drug absorption, distribution, metabolism and excretion (ADME).

Main results: We detected 499 genome-wide significant ($p < 1.65 \times 10^{-6}$) associations comprising 39 unique CNV loci. The frequencies of the associated CNVs in EstBB varied between 0.01% and 5.04%. Five detected CNVs intersected an ADME gene, including the 22q11.21 duplication overlapping *COMT* enzyme associated with decreased levels of omega-3 fatty acids ($p < 2.67 \times 10^{-12}$) and the 10q26.3 CNV overlapping *CYP2E1* associated with acetone ($p < 4.82 \times 10^{-49}$).

Limitations: A key limitation of our study is that microarray-based CNV calls capture only a limited portion of the CNV landscape. Moreover, differences in microarray platforms can obscure associations and impede the effectiveness of meta-analysis.

Wider implications: Our analyses provide the first insights into how CNVs shape the human metabolome, offering a valuable resource for understanding the genetic basis of metabolic biomarkers and drug efficacy beyond small-variant studies.

TEXT-BASED APPROACH FOR DETECTING CASES OF ADES FROM EHRS OF PARTICIPANTS OF THE ESTONIAN BIOBANK

Dage Särg[1,2], Kairit Sirts[2], Kristi Krebs[1], Markus Tamm[4], Alise Metsküla[4], Marek Oja[2], Sven Laur[2], Jaak Vilo[2,3], Lili Milani[1]

[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[2]Institute of Computer Science, University of Tartu, Tartu, Estonia
[3]STACC, Tartu, Estonia
[4]Faculty of Medicine, University of Tartu, Tartu, Estonia

E-mail address of presenting author: dage.sarg@ut.ee

Background: Adverse Drug Events (ADEs) are a major public health concern. Some ADEs are linked to genetic variation, making their detection essential for pharmacogenetic research. However, ADE information in Electronic Health Records (EHRs) is embedded in unstructured text, requiring large-scale manual review.

Study question: This study develops computational methods for detecting ADEs from EHRs of participants of the Estonian Biobank. It focuses on creating manually annotated datasets and improving ADE detection efficiency by combining rule-based and machine learning (ML) approaches.

Methods: To detect potential ADE mentions, we employed a lexicon-based approach to extract text snippets containing both a drug name and a symptom. We developed a rule-based and an ML-based system to prefilter the extracted text snippets. Both systems were applied before manual annotation, and we assessed their impact on the annotation process.

Main results: Prefiltering reduced annotation workload up to 24-fold compared to no filtering, and ML-based filtering outperformed rule-based filtering, requiring only 1.3–1.5 snippets per positive case. We produced annotated datasets for antidepressants (520 patient–drug pairs) and antipsychotics (1,329 pairs). Pharmacogenetic validation revealed significant genotype–ADE associations for Escitalopram, Sertraline, and Quetiapine.

Limitations: Manual annotation remains necessary, and the study focused on two drug classes. Prefiltering may exclude positive examples, and the symptom lexicon cannot capture unknown side effects. Large Language Models (LLMs) could improve performance, but privacy constraints prevent their use with sensitive clinical data, and their effectiveness on Estonian text remains underexplored.

Wider implications: This study demonstrates that a hybrid computational approach can enable scalable creation of high-quality ADE datasets, supporting pharmacogenetic discoveries.

Exploring the Estonian Gut Microbiota: Composition, Structure, and Research Potential

Susan Pihelgas[1,2], Kristel Ehala-Aleksejev[1], Aili Kallastu[1], **Hindrek Teder**[1], Jekaterina Kazantseva[1]

[1]TFTAK, Tallinn, Estonia
[2]School of Natural Sciences and Health, Tallinn University, Tallinn, Estonia

E-mail address of presenting author: hindrek.teder@tftak.eu

Background: Understanding how the human gut microbiota varies across populations and life stages is essential, given its links to health and diet. However, region-specific data, such as from Estonia, remain limited.

Study question: What is the taxonomic composition and enterotype structure of the Estonian gut microbiota, and how can microbiota profiling support dietary intervention research?

Methods: Approximately 1,800 Estonian gut microbiota samples were collected and characterized by 16S rRNA sequencing, alongside anthropometric, dietary, and disease-related metadata. Community structure was assessed via principal coordinate analysis, hierarchical clustering, and diversity metrics to describe the community structure. Differential abundance analysis and multivariate methods were used to detect enterotype-specific patterns and their associations with metadata variables. As a case study, a three-week intervention involving daily consumption of fermented vegetables was conducted in a subset of participants to illustrate how gut microbiota profiling can support nutritional studies and inform evidence-based conclusions.

Main results: Our analysis revealed the most dominant bacterial genera in the Estonian population and identified distinct enterotypes that may serve as a foundation for future studies on disease and dietary associations. Dietary intervention showed that consumption of fermented vegetables increased the abundance of butyrate-producing and anti-inflammatory bacterial species. Improvements in phase angle suggest enhanced metabolic health.

Limitations: 16S rRNA sequencing provides limited functional and strain-level resolution compared to shotgun metagenomics. Self-reported metadata might introduce potential bias.

Wider implications: Our large-scale characterisation of the Estonian gut microbiota lays a foundation for future studies and demonstrates how microbiota profiling can inform evidence-based dietary strategies.

20

METAGENOME-ASSEMBLED GENOMES FROM A POPULATION-BASED COHORT UNCOVER NOVEL GUT SPECIES AND STRAIN DIVERSITY, REVEALING PREVALENT DISEASE ASSOCIATIONS

Kateryna Pantiukh, Kertu Liis Krigul, Oliver Aasmets, Elin Org

Institute of Genomics, Estonian Genome Centre, University of Tartu, Tartu, Estonia

E-mail address of presenting author: pantiukh@ut.ee

Background: Metagenomic profiling has advanced our understanding of microbe–host interactions. However, widely used read-based methods are limited by incomplete reference databases and insufficient resolution of strain-level variation.

Study question: We present a scalable, genome-resolved framework that incorporates population-specific metagenome-assembled genomes (MAGs) to identify novel species, strain diversity, and disease associations.

Methods: Using 1,878 deeply sequenced shotgun metagenomic samples from the Estonian Microbiome Cohort, we reconstructed MAGs via a multi-binning approach and performed genome-resolved analysis . Associations with 33 prevalent diseases were tested using linear regression models.

Main results: We recovered 84,762 MAGs representing 2,257 species, including 353 (15.6%) previously uncharacterized species reaching up to 30% relative abundance in some individuals. An expanded MAG reference enabled profiling of 2,509 EstMB individuals and revealed associations with 25 diseases, 8 of which involved newly identified species. To assess within-species diversity, we developed the Strain Richness Index (SRI), a novel MAG-based metric. Based on SRI, we prioritized *Odoribacter splanchnicus*, a prevalent species with low strain heterogeneity, enabling strain-level analysis. Two dominant strains, N1 and N2, showed distinct gene repertoires and divergent disease associations. Notably, strain N1 was negatively associated with gastritis, duodenitis, and hypertensive heart disease—associations undetected at the species level.

Limitations: MAG quality may limit resolution, which could be improved through long-read sequencing and targeted cultivation.

Wider implications: This study expands the gut reference landscape, demonstrates the utility of population-specific MAGs for uncovering novel microbial diversity, and highlights the importance of genome-resolced approches for revealing strain-level associations that remain obscured at higher taxonomic levels.

21

A BIOCONDUCTOR WORKFLOW FOR MICROBIOME-BASED SURVIVAL ANALYSIS

Sneha Das[1], Geraldson Muluh[1], Tuomas Borman[1], Teemu Niiranen[2,3,4] , Aki Havulinna[1,4,5], Leo Lahti[1]

[1]Department of Computing, University of Turku
[2]Division of Medicine, University of Turku
[3]Department of Internal Medicine, Turku University Hospital and University of Turku
[4]Department of Public Health, Finnish Institute for Health and Welfare
[5]Institute for Molecular Medicine Finland (FIMM), Finland

E-mail address of presenting author: snedas@utu.fi

Background: There is growing interest in exploring how the human microbiome influences human health. This is no longer limited to cross-sectional association studies alone. With the increasing availability of cohorts with considerable follow-up times, researchers are now also turning to time-to-event, or survival analysis, studies to examine the link between the taxonomic, functional, and other features of microbiome, and the occurrence of significant events- typically the onset of certain diseases or mortality.

Current limitations and study question: The accumulation of well phenotyped microbiome cohorts with a substantial follow-up time profiling has made it easier to analyze the association between microbes and future health status. Despite the growing interest, the current studies remain limited by methodological challenges. The complex and compositional microbiome data sets necessitate specialized statistical approaches. There is a lack of comprehensive methods and limited availability of R packages to support microbiome-based survival analyses, and benchmarking studies comparing the available methods in microbiome data sets.

Methods, main results and wider implications: We introduce a Bioconductor workflow for microbiome-based survival analysis using the (Tree)SummarizedExperiment framework and the mia package. This can support future survival analysis studies in future cohort studies.

22

PROBABILISTIC ALTERNATIVES FOR MICROBIOME RESEARCH IN R/BIOCONDUCTOR

Rasmus Hindström & Leo Lahti

Department of computing, University of Turku, Turku, Finland

E-mail address of presenting author: rasmus.hindstrom@utu.fi

Background: Microbiome and -omics data have features that require special considerations for statistical inference and testing. Probabilistic methods offer promising benefits over commonly used classical methods. Warranting research into the applicability of probabilistic methods to raise awareness of their use and offered benefits.

Study question: We compare and benchmark alternative probabilistic methods against common classical univariate group comparison methods used in microbiome research.

Methods: In the context of two-sample and multiple-group comparisons, we compare Bayesian estimation to the classical methods of ANOVA, Kruskal-Wallis, and their post-hoc procedures pairwise t-tests, Tukey's HSD (honestly significant difference), and Dunn's test. Performance is evaluated in terms of inference clarity, handling of multiple comparisons and interpretability.

Main results: The probabilistic approach produces consistent results with classical methods. However, it offers key advantages in more intuitive interpretation and inherent correction for multiple comparisons due to partial pooling.

Limitations: Probabilistic methods are hindered by higher computational demands, unfamiliarity among practitioners, and the need for more thorough model diagnostics.

Wider implications: Raising awareness of probabilistic methods as a suitable alternative, and incorporating them into the well established R/Bioconductor ecosystem can bring the unique benefits they offer into the hands of a wider community of researchers in the field of microbiome research and beyond.

PREDICTIVE MARKERS OF FUTURE DEPRESSION IN THE GUT MICROBIOME

Annabel Klemets[1], Kateryna Pantiukh[1], Kadri Kõiv[2], Reidar Andreson[3], Kelli Lehto[2], Oliver Aasmets[1] & Elin Org[1]

[1]Microbiome research group, Institute of Genomics, University of Tartu, Estonia
[2]Neuropsychiatric genomics research group, Institute of Genomics, University of Tartu, Estonia
[3]Bioinformatics research group, Institute of Molecular and Cell Biology, University of Tartu, Estonia

E-mail address of presenting author: annabel.klemets@ut.ee

Background: Major depressive disorder (MDD) is a prevalent mental health condition that severely impacts the general life quality[1]. Several studies have reported alterations in the gut microbiome composition related to MDD[2-6]. However, most studies so far have been conducted on individuals with already active disease. As depression can lead to major lifestyle changes and is paired with drug use known to alter the microbiome composition, confounding can be a significant issue, and detecting MDD-specific changes in the microbiome remains a challenge.

Study question: We aim to identify MDD-specific alterations in the gut microbiome by analyzing incident MDD cases using the prospective Estonian Microbiome Cohort (EstMB) (n=2181)[9].

Methods: Electronic health records and microbiome shotgun sequencing data were used to associate gut microbiome with prevalent and incident depression. A subset of 325 EstMB participants with two-timepoint microbiome measurements was used to further pinpoint MDD-induced microbiome changes. Linear and Cox regression models adjusted for covariates were used to analyze associations between prevalent and incident depression and CLR-transformed microbial species, respectively.

Main results: We identified 13 bacterial species predictive of MDD onset (FDR<0.1). However, we found no bacterial species with consistent associations in prospective and retrospective analysis, indicating significantly different microbial signatures in active and future disease.

Limitations: The main limitations of this study include a short follow-up time (4.5 years), and a small sample size in the follow-up cohort (n=325).

Wider implications: The results of our study warrant further investigation into using prospective data to accurately predict onset of MDD and identify disease-specific effects.

References

1. Cryan, J. F. et al. The Microbiota-Gut-Brain Axis. Physiological Reviews 99, 1877–2013 (2019).
2. Brushett, S. et al. Gut feelings: the relations between depression, anxiety, psychotropic drugs and the gut microbiome. Gut Microbes 15, 2281360 (2023).
3. Valles-Colomer, M. et al. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol 4, 623–632 (2019).
4. Radjabzadeh, D. et al. Gut microbiome-wide association study of depressive symptoms. Nat Commun 13, 7128 (2022).
5. Kurokawa, S. et al. Fecal Microbial and Metabolomic Change during treatment course for depression: An Observational Study. Journal of Psychiatric Research 140, 45–52 (2021).
6. Xie, Z. et al. Integrated multi-omics analysis reveals gut microbiota dysbiosis and systemic disturbance in major depressive disorder. Psychiatry Research 334, 115804 (2024).
7. Aasmets, O. et al. Long-term consequences of drug usage on the gut microbiome. 2024.07.17.24310548 Preprint at https://doi.org/10.1101/2024.07.17.24310548 (2024).
8. Michaelis, L., Berg, L. & Maier, L. Confounder or Confederate? The Interactions Between Drugs and the Gut Microbiome in Psychiatric and Neurological Diseases. Biological Psychiatry 95, 361–369 (2024).
9. Aasmets, O., Krigul, K. L., Lüll, K., Metspalu, A. & Org, E. Gut metagenome associations with extensive digital health data in a volunteer-based Estonian microbiome cohort. Nature Communications 13, 1–11 (2022).

24

MODELING GEOGRAPHIC VARIATION OF HUMAN MICROBIOME

João Paulo Cassucci dos Santos, Leo Lahti, Aki Havulinna

UTU

E-mail address of presenting author: N/A

Background: The human microbiome varies within and across geographical regions. However, modeling spatial variation is challenging due to data sparsity and confounding factors.

Study question: How to better model the geographical variations of the human microbiome?

Methods: We use probabilistic analysis with the conditional autoregression (CAR) method to analyse population-level geographical variation in human microbiome. Since such methods are computationally costly, we used a scalable approach-the integrated nested Laplacian approximation (INLA) to perform Bayesian inference. The INLA method is supported by a community of R developers, and we are working towards Bioconductor-compatible methodology for geospatial metagenomics.

Main results: We evaluate the methodology with simulated and real data and discuss current challenges in geospatial analysis of microbiome variation.

Limitations: The methods used are only available in the R language currently.

Wider implications: We will develop a workflow for the analysis of metagenomics together with geographical data that can be applied by future researchers using the Bioconductor package environment.

25

GLOBAL ANTIMICROBIAL RESISTANCE PATTERNS IN HUMAN GUT METAGENOMES ARE STRUCTURED ALONG SOCIO-ECONOMIC GRADIENTS

Mahkameh Salehi[1], Katariina Pärnänen[2], Shivang Bhanushali[1], Ville Laitinen[1], Eetu Tammi[1], Aura Raulo[1], Guilhelm Sommeria-Klein[3], Peter Collignon[4], John J. Beggs[4], Johan Bengtsson-Palme[5], Leo Lahti[1]

[1]Department of Computing, University of Turku, Finland
[2]Department of Microbiology, University of Helsinki, Finland
[3]Université de Montpellier, France
[4]Australian National University, Canberra, Australia
[5]Chalmers University of Technology, Sweden

E-mail address of presenting author: mahksa@utu.fi

Background: Antimicrobial resistance (AMR) in the human gut is a critical but often overlooked part of the global AMR crisis. While clinical surveillance tracks resistance in pathogens, it misses resistance genes in the wider gut microbiome, and their ecological and socio-economic drivers.

Study question: How do socio-economic conditions, geography, and demographic factors shape the human gut resistome globally, and how is microbial composition linked to antibiotic resistance gene (ARG) abundance?

Methods: We analyzed over 60,000 human gut metagenomes from public repositories, representing more than 30 countries. ARG abundance and diversity were quantified and linked to standardized country-level data on infrastructure, governance, education, GDP, and antibiotic use. We used distance-based models to evaluate resistome similarity across geographic and socio-economic gradients. Subgroup analyses were performed on more than 14,000 samples with gender and age metadata.

Main results: ARG load was more strongly associated with socio-economic conditions than with antibiotic use. Countries with similar infrastructure or governance tended to have more similar ARG profiles, regardless of geography. In high-income countries, women had 9% higher ARG load and greater ARG diversity than men, with differences emerging in adulthood. In low- and middle-income countries, these gender differences were reversed or not statistically significant. ARG levels followed a consistent U-shaped trend, with peaks in infancy and old age.

Limitations: Data coverage was uneven across regions, and not all samples had complete metadata.

Wider implications: Patterns of global gut resistome reflect global inequalities more than just antibiotic exposure. Large-scale metagenomic data can help identify hidden demographic and geographic risks and guide strategies to control AMR.

BIOME-SPECIFIC GENOME CATALOGUES REVEAL FUNCTIONAL POTENTIAL OF SHALLOW SEQUENCING

Matti Ruuskanena[1,*], Alejandra Escobar-Zepedab[2,*], Martin Beracochea[2], Jennifer Lu[2], **Dattatray Mongad**[1], Lorna Richardson[2], Robert D. Finn[2], Leo Lahti[1]

[*]Equal contribution
[1]Department of Computing, University of Turku, Vesilinnantie 5, 20500, Finland
[2]EMBL-EBI, Microbiome Informatics Team, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

E-mail address of presenting author: dattatray.mongad@utu.fi

Background: The use of 16S rRNA metabarcoding is limited by known taxonomic biases and functional inference methods. Shallow shotgun sequencing is a cost-effective and taxonomically high-resolution alternative to 16S-based methods, but functional inference results are limited by low sequencing depth. In well-studied microbiomes, mapping reads to biome-specific genome databases results in very few unassigned reads.

Methods: We devised a functional inference method where reads are mapped against biome-specific databases to predict functional profiles. We used three datasets from red junglefowl-, mouse-, and human-gut, containing matched deep-shotgun metagenomic and 16S rRNA metabarcoding data for taxonomic and functional benchmarking. Additionally, human-gut deep sequencing data was subsampled to 1 M reads to replicate previously obtained results with BioSIFTR.

Main results: The taxonomic and functional profiles produced by BioSIFTR with sequencing depth of >0.5M reads closely agreed with the results of deep sequencing data (inference using all biomes). The profiles produced with metabarcoding sequencing were highly dissimilar to both the BioSIFTR and the deep-sequencing results. Furthermore, BioSIFTR was able to replicate the results of a previous study on functional differences in the fecal microbiome between high and low trimethylamine N-oxide producing participants, using only <2% of the original sequencing data.

Limitations and wider implications: Shallow-shotgun sequencing combined with BioSIFTR (biome-specific catalogue), can be a viable alternative to 16S rRNA metabarcoding-based analyses, providing an increased taxonomic and functional resolution, and lower bias. However, shallow-sequenced data may also suffer the same inherent bias associated with deep-sequenced data, namely that of poor representation of least abundant taxa.

27

LEVERAGING GENOME-WIDE ASSOCIATION STUDY SUMMARY STATISTICS TO IDENTIFY POSSIBLE DRUG TARGETS FOR UTERINE FIBROIDS

Lisette Haug[1], Natàlia Pujol Gualdo[2], Jelisaveta Džigurski[2], Valentina Rukins[2], Estonian Biobank Research Team[2], Reedik Mägi[2], Triin Laisk[2]

[1]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia
[2]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

E-mail address of presenting author: lisette.haug@ut.ee

Background: Uterine fibroids are benign tumors affecting many women of reproductive age, causing various symptoms that reduce the quality of life. Existing treatments such as hormone therapy and surgery often have side effects, variable success, and do not reliably prevent fibroid recurrence.

Study question: Can genome-wide association study (GWAS) data be used to identify potential drug targets for uterine fibroids?

Methods: This study analysed summary statistics from a GWAS of data from the Estonian Biobank and FinnGen R7 cohorts, including 51,378 cases (ICD-10 code D25) and 243,913 controls. Using FUMA, significant SNPs were annotated and gene prioritization was performed through three approaches: coding variant mapping, Open Targets Genetics V2G scoring, and Mouse Genome Informatics phenotype mapping. Candidate gene sets of 333, 44, and 57 genes were assessed with the GREP tool to identify potential drug targets and repositioning candidates.

Main results: We identified 194 lead SNPs significantly associated with uterine fibroids ($p<5\times10^{-8}$), including 29 novel risk loci. GREP analysis revealed 13 potential drug targets, with ESR2 and KDR being most promising. Several associated drugs such as raloxifene and estradiol are already used, while others, including antineoplastic agents like ramucirumab, present novel but potentially risky avenues.

Limitations: Using different approaches for gene prioritization results in different gene sets with unique biological implications and can therefore lead to varied outcomes in follow-up analyses, including drug target identification.

Wider implications: This study offers new leads for fibroid treatment through drug repositioning and ultimately supports the utility of GWAS-based gene prioritization for drug target discovery in gynecological diseases.

Maternal Mood Disorders Are Associated With Increased Proportion of Placental Immune Cell Types Driving Gene Expression Differences

Soile Hytti[1], Taru Tukiainen[1,2], Darina Czamara[3], Marius Lahti-Pulkkinen[1,4], Polina Girchenko[1,5], Jari Lahti[1], Cristiana Cruceanu[3,6], Linda Dieckmann[3], Elisabeth B Binder[3], Katri Räikkönen[1,7]

[1]Department of Psychology, University of Helsinki, Helsinki, Finland
[2]Institute of Clinical Medicine, School of Medicine, University of Eastern Finland, Kuopio, Finland
[3]Department Genes and Environment, Max Planck Institute of Psychiatry, Munich, Germany
[4]Centre for Cardiovascular Science, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, United Kingdom
[5]Clinical Medicine Research Unit, MRC Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland
[6]Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden
[7]Department of Obstetrics and Gynecology, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

E-mail address of presenting author: soile.hytti@helsinki.fi

Background: Maternal mood disorders increase the risk of neurodevelopmental adversities in children. Altered placental function may play a role but remains largely unexplored.

Study question: How are maternal mood disorders reflected in placental molecular signatures through gestation?

Methods: In the Finnish InTraUterine Sampling in Early Pregnancy study, we investigated differential gene expression in first-trimester chorionic villi (n=267) and birth placentas (n=493) in relation to maternal history of mood disorder diagnoses, antidepressant and anxiolytic medication purchases, and depressive and anxiety symptoms during pregnancy. We also examined differences in inferred abundancies of placental cell types and whether they explained potential gene expression differences.

Main results: Maternal mood disorders and depressive and anxiety symptoms above clinical cutoffs, but not medication purchases, were together associated (pFDR<0.1) with 285 differentially expressed genes (DEGs) in placenta. These gene expression changes correlated moderately with those in chorionic villi, where no DEGs were detected. The DEGs upregulated in mothers with mood disorders and symptoms were enriched in inflammation pathways and expression in placental immune cells. Moreover, placentas from mothers with mood disorders and symptoms showed higher relative abundances of

immune cell types, which largely explained the observed DEGs. Further, expression of DEGs correlated positively with inflammatory marker GlycA levels in both maternal and cord blood plasma.

Limitations: While our sample size is larger than in the previous studies, study population comprises women from high-resource Nordic settings, which limits global generalization.

Wider implications: We identified a robust placental molecular signature implicating inflammation and immune response, offering insights into mechanisms linking maternal mood disorders with neurodevelopmental adversities in children.

MULTI-ANCESTRY, TRANS-GENERATIONAL GWAS META-ANALYSIS OF GESTATIONAL DIABETES AND GLYCAEMIC TRAITS DURING PREGNANCY REVEALS LIMITED EVIDENCE OF PREGNANCY-SPECIFIC GENETIC EFFECTS

Valentina Rukins[1,*], Caroline Brito Nunes[2,*], Aminata H. Cisse[3,*], Frédérique White[4,*], Nancy McBride[5,6], Gunn-Helen Moen[2,7,8], and Reedik Mägi[1] on behalf of the GENetics of Diabetes In Pregnancy (GenDiP) Consortium

[*]Joint first authors
[1]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[2]Institute for Molecular Bioscience, The University of Queensland, St Lucia, Queensland, Australia
[3]Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, UK
[4]Département de Biologie, Faculté des Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada
[5]MRC Integrative Epidemiology Unit at the University of Bristol, UK
[6]Population Health Science, Bristol Medical School, University of Bristol, UK
[7]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[8]University of Queensland Frazer Institute, University of Queensland, Woolloongabba, Australia

E-mail address of presenting author: valentina.rukins@ut.ee

Background: Gestational diabetes mellitus (GDM) affects ~14% of pregnancies worldwide, causing adverse maternal and fetal outcomes and increasing maternal type 2 diabetes (T2DM) risk. Despite being a significant concern for public health, GDM remains understudied.

Study question: The GenDiP Consortium sought to investigate the impact of maternal and fetal genetic risk factors on GDM and pregnancy glycemic traits by conducting the largest multi-ancestry genome-wide association study (GWAS) meta-analyses of these traits to date.

Methods: We leveraged data from 30 cohorts across African, East Asian, South Asian, European, and Hispanic populations. Our GWAS meta-analyses included: i) GDM diagnosis ($N_{maternal}$=38,375 cases, 776,075 controls; N_{fetal}=3,126 cases, 90,877 controls), ii) Fasting glucose ($N_{maternal}$=55,371; N_{fetal}=15,855), iii) 1-hour glucose post- oral glucose tolerance test (OGTT) ($N_{maternal}$=38,439; N_{fetal}=9,365), iv) 2-hour glucose post-OGTT ($N_{maternal}$=46,401; N_{fetal}=15,046), and v) HbA1c ($N_{maternal}$=9,724; N_{fetal}=7,035).

Main results: We identified 37 GDM-associated loci (19 novel) and five novel loci for glycemic traits, all mediated through the maternal genome. Most GDM loci overlapped

with T2DM and non-pregnant glycemic traits, with limited evidence for pregnancy- specific effects. Notably, we observed ancestry-specific effects at the fasting glucose locus *ABCB11*, with opposite directions in European and East Asian populations.

Limitations: This study has several limitations that should be considered, including sample size imbalance across ancestry groups, with European and East Asian populations being predominant; reduced power for HbA1c and fetal analyses due to small sample sizes; and varying GDM diagnostic criteria across cohorts.

Wider implications: Our findings provide insights into the genetic background of GDM and highlight the importance of large, ancestrally diverse studies in enhancing genomic research.

GENETIC RISK FACTORS OF VASOMOTOR SYMPTOMS

Mia Golob, Jelisaveta Džigurski, Triin Laisk

University of Tartu, Tartu, Estonia

E-mail address of presenting author: miagolob19@gmail.com

Background: Menopausal vasomotor symptoms are one of the most common and bothersome symptoms that women go through during menopausal transition. While hormone replacement therapy is effective in managing these symptoms, not all women can or want to use it due to health concerns. More research is needed to find alternative treatment options and understand the underlying biological mechanisms. The primary purpose of this study is to shed light on the underlying biology of menopausal vasomotor symptoms by: a) conducting a genome-wide association study (GWAS) meta-analysis, combining data from the Estonian Biobank with publicly available datasets, and b) performing post-GWAS analyses.

Study question: What genetic risk factors for menopausal vasomotor symptoms can be identified through the largest-to-date GWAS meta-analysis, combining summary statistics from the Estonian Biobank, FinnGen, and the UK Biobank in approximately 200,000 women of European ancestry?

Methods: GWAS offers a hypothesis-free approach to identify genetic variants associated with traits or conditions. We conducted GWAS meta-analysis by combining summary statistics from three large biobanks: the Estonian Biobank, the FinnGen, and the UK Biobank, with a total sample size of 291,738 women (cases: 62,139, controls: 229,599). The cases were defined using ICD (International Classification of Diseases) N95 four-character subcategory codes (N95.1, N95.3, N95.8 and N95.9), and the controls were women without those codes. The final meta-analysis results were restricted to variants present in all three cohorts and with a minor allele frequency (MAF)>1%. Post-GWAS analyses, such as annotation and gene prioritisation, were carried out using the Functional Mapping and Annotation (FUMA) web-based tool.

Main results: Our meta-analysis revealed one significant locus on chromosome 4 (rs74827081, p-value=3.09×10^{-34}) in the intronic region of the *TACR3* gene, which codes for the neurokinin 3 receptor. Functional annotation and gene prioritisation further supported *TACR3* as a strong candidate gene. Tissue expression analysis mapped to the brain and pituitary gland tissue, although these results were not statistically significant after correction for multiple testing.

Wider implications: The findings from our study provide strong evidence for an association between *TACR3* and vasomotor symptoms, validating previous results and advancing our understanding of their genetic background. This may support the development of personalised, non-hormonal treatments. However, as most participants were of European ancestry, generalisability is limited, and future research should include diverse ancestry groups to ensure broader applicability.

31

MULTI-OMIC RISK PREDICTION SCORES FOR RHEUMATOID ARTHRITIS IN THE ESTONIAN BIOBANK

Galadriel Velázquez, Kristi Läll, Triin Laisk, Elin Org, Reedik Mägi, Ene Reimann

Institute of Genomics, University of Tartu

E-mail address of presenting author: galadriel.lucia.velazquez.silva@ut.ee

Background: Rheumatoid arthritis (RA) is a chronic autoimmune disease of unknown aetiology, for which early diagnosis and treatment are essential to improve long-term outcomes. As RA is highly heritable, genetic research plays a key role in understanding its pathogenesis and detecting novel genetic predictors. Additionally, integrating genetic and additional omic layers may improve early detection and risk stratification over traditional models based on conventional risk factors (CRFs).

Study question and methods: This study aimed to evaluate whether multi-omic data integration can improve RA risk prediction in the Estonian Biobank (EstBB) beyond traditional models. Therefore, a genome-wide association study (GWAS) meta-analysis combining data from the Estonian Biobank (EstBB) and 27 additional European cohorts and a total of ~1 million participants was performed. Potential new genomic risk loci were defined, and genes within these regions were prioritised to identify new potential genetic predictors. In parallel, Polygenic risk scores (PRSs) were constructed and compared with existing scores. To improve prediction further, RA-associated metabolites are currently being identified for integration into a multi-omic model.

Main results: We identified seven potential novel RA causal genes. The best-performing PRS was enhanced by CRFs such as sex, age, smoking and BMI, leading to improved predictive performance. Integration of the metabolomic data layer is ongoing and expected to enhance model performance.

Limitations: However, our study has some limitations. Identified genetic associations could be influenced by population-specific effects, as all data were of European ancestry.

Wider implications: In conclusion, this multi-omic framework may support early RA prediction and will be implemented in the EstBB Participant Portal to enhance preventive strategies.

No evidence for a causal effect of thyroid function and disease on risk for attention-deficit/hyperactivity disorder: A two-sample Mendelian randomization study

Triinu Peters[1,2,3], Lars Dinkelbach[3,4], Luisa Rajcsanyi[1,2,3], Anke Hinney[1,2,3], Raphael Hirtz[5]

[1]Section of Molecular Genetics in Mental Disorders, University Hospital Essen, Essen, Germany
[2]Center for Translational Neuro- and Behavioral Sciences, University Hospital Essen, University of Duisburg-Essen, Essen, Germany
[3]Institute of Sex and Gender-Sensitive Medicine, University Hospital Essen, Germany
[4]Department of Pediatrics III, University Hospital Essen, University of Duisburg-Essen, Essen, Germany
[5]Helios University Medical Centre Wuppertal – Children's Hospital, Witten/Herdecke University, Wuppertal, Germany

E-mail address of presenting author: triinu.peters@uni-due.de

Background: Attention-deficit/hyperactivity disorder (ADHD) is the most prevalent neurodevelopmental disorder in children, characterized by inattention, hyperactivity, and impulsivity. These symptoms overlap with those of thyroid dysfunction, especially hyperthyroidism. Previous epidemiological and clinical studies suggest associations between thyroid status and ADHD risk but may be affected by confounding and reverse causation.

Study question: Is there a causal effect of thyroid function or thyroid diseases on ADHD risk?

Methods: We performed two-sample Mendelian randomization (MR) using genetic instruments from large GWAS on thyroid-stimulating hormone (TSH), free T3 (FT3), free T4 (FT4) (Sterenborg et al., 2024), hypothyroidism, hyperthyroidism, Hashimoto's thyroiditis, and Graves' disease (Sakaue et al., 2021). The ADHD GWAS (Demontis et al., 2023) included 38,691 cases and 186,843 controls. Eight MR methods were computed to assess possible violations of assumptions.

Main results: No causal effect of any thyroid trait or thyroid disease on ADHD risk was found.

Limitations: The analysis was sufficiently powered for TSH and FT4. Thyroid GWAS were conducted in adults, while ADHD typically begins in childhood. Although ADHD's genetic architecture appears consistent across sex, age, and symptom domains, data on thyroid genetics across the lifespan are lacking. Thus, the validity of adult-based instrumental

variables for children cannot be confirmed. MR assumes linearity, potentially missing non-linear effects. The mainly European ancestry of samples limits generalizability.

Wider implications: Despite overlapping symptoms between ADHD and thyroid dysfunction, no causal relationship was found. Limited power for some thyroid diseases precludes definitive conclusions. Overall, these findings do not support routine thyroid function testing in ADHD patients without clinical signs of thyroid dysfunction.

PREV-GEN – PREVENTATIVE OUTREACH WITH GENETIC TESTING

Kalle Pärn[1,2], Minja Pehrsson[1], Olli Mikael Carpen[1,2,3]

[1]Helsinki Biobank
[2]University of Helsinki, Faculty of Medicine
[3]Helsingin ja Uudenmaan sairaanhoitopiiri (HUS)

E-mail address of presenting author: ext-kalle.parn@hus.fi

Smarter implementation of genomic information in prevention of common diseases, such as cardiovascular diseases and cancers, is one of the unutilized possibilities in healthcare in Finland and other countries. In a recent nationwide pilot project, we successfully applied genomic information originating from biobank research projects (FinnGen) in screening of individuals with high genetic cancer risk.

Building on the pilot project results, our new PREV-GEN initiative aims to build and validate a scalable model, enabling healthcare providers and associated parties to offer personalized disease risk screening in a cost-effective way, shifting healthcare from reactive to preventive.

The initiative takes a hybrid approach, with one part focusing on analysis of additional disease-predisposing variants, especially those leading to cancers (Lynch syndrome) and cardiometabolic diseases (familial hypercholesterolemia) in the existing biobank data, and the other on the expansion of the sample number in a prospective setting to prove the feasibility and cost-effectiveness of the screening approach for cancer-risk variants from the pilot (BRCA,PALB2) and for additional high-risk genes. In a parallel approach, we will utilize comprehensive HUS hospital electronic registries to identify individuals, whose clinical phenotypes suggest a genetic background, and select their biobank specimens for genomic analysis.

A chip-based approach will be used in initial screening, followed by NGS-based clinical validation. The findings will be returned to donors and healthcare providers, for implementation as part of routine healthcare. Cost-benefit analysis will be carried out to define the economic value of this initiative and a participant survey on general attitudes will be conducted.

Association between meat intake and subclinical atherosclerosis in the population-based Swedish CardioPulmonary BioImage Study

Getachew Arage[1], Luka Marko Rašo[1], Ulf Hammar[1], Koen F. Dekkers[1], Ulrika Ericson[2], Susanna C Larsson[3,4], Karl Michaëlsson[3], J Gustav Smith[5,6,7], Gunnar Engström[2], Johan Ärnlöv[8,9,10], Marju Orho-Melander[2], Lars Lind[11], Tove Fall[1], Shafqat Ahmad[1,12,13]

[1]Molecular Epidemiology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden
[2]Department of Clinical Sciences in Malmö, Lund University, Malmö, Sweden
[3]Medical Epidemiology, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden
[4]Unit of Cardiovascular and Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[5]The Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University and the Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden
[6]Department of Cardiology, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden
[7]Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Lund University, Lund, Sweden
[8]Division of Family Medicine and Primary Care, Department of Neurobiology, Care Science and Society, Karolinska Institutet, Huddinge, Sweden
[9]School of Health and Social Studies, Dalarna University, Falun, Sweden
[10]Center for Clinical Research Dalarna, Falun, Uppsala University, Sweden
[11]Clinical Epidemiology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden
[12]Preventive Medicine Division, Harvard Medical School, Brigham and Women's Hospital, Boston, United States
[13]School of Natural Sciences, Technology and Environmental Studies, Södertörn University, Sweden

E-mail address of presenting author: getachew.debebe@uu.se

Background: The relationship between meat consumption and atherosclerotic cardiovascular disease remains unclear, and underlying mechanisms are not fully understood.

Study question: Are different types of meat intake associated with subclinical atherosclerosis, and are these associations mediated by cardiovascular (CVD) risk factors, plasma metabolites, or gut microbial diversity?

Methods: This cross-sectional study included 8,943 participants aged 50–64 years from the Swedish CArdioPulmonary BioImage Study, free from coronary heart disease at baseline. Coronary atherosclerosis was assessed using computed tomography angiography (plaque presence and segment involvement score), and carotid atherosclerosis by ultrasound. Meat intake (unprocessed red, processed red, and white meat) was measured via a web-based food frequency questionnaire. Logistic and ordinal regression estimated associations per 20 g/day increase in meat intake. Mediation analyses quantified the contribution of CVD risk factors, plasma metabolites, and gut microbial diversity.

Main results: Intake of unprocessed and processed red meat was associated with 4% (OR=1.04, 95% CI: 1.01–1.07) and 5% (OR=1.05, 95% CI: 1.00–1.10) higher odds of coronary atherosclerosis, respectively. Processed red meat was also associated with carotid atherosclerosis (OR=1.04, 95% CI: 1.00–1.09). BMI mediated 38% of the unprocessed red meat–coronary atherosclerosis association, followed by lipids (20.6%), glycaemic measures (18.9%), and mannose metabolite (8.75%).

Limitations: The cross-sectional design limits causal inference, and residual confounding cannot be excluded.

Wider implications: Our findings highlight the role of meat intake in cardiovascular risk and suggest that targeting cardiometabolic pathways, including BMI and lipid metabolism, could help mitigate diet-related atherosclerosis.

Integrating Polygenic Risk Scores into OMOP Common Data Model using the Estonian Biobank: A Preliminary Study of AI-Driven Cardiovascular Disease Prediction in Atherosclerosis

Djeane Debora Onthoni[1], Marek Oja[1], Sirli Tamm[1], Ami Sild[1], Kerli Mooses[1], Reedik Mägi[2], Kristi Läll[2], Estonian Biobank research team[2], Sulev Reisberg[1,3], Raivo Kolde[1], Jaak Vilo[1,3]

[1]Institute of Computer Science, University of Tartu, Tartu, Estonia
[2]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[3]STACC OÜ, Tartu, Estonia

E-mail address of presenting author: djeane.onthoni@ut.ee

Background and study question: The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) lacks native support for genetic data such as Polygenic Risk Score (PRS), crucial for precision medicine. This study asks whether integrating PRS into the OMOP CDM using custom concept IDs can enhance data completeness and enable its utility through AI-driven cardiovascular disease (CVD) risk modeling.

Methods: Using the Estonian Biobank, we calculated 72 CAD PRSs from PRS Catalog's open database[1] and integrated them into the OMOP CDM. We trained a Lasso Logistic Regression model to predict CVD (610 cases) within an atherosclerosis population (10,445 individuals), using conditions, measurements, and PRSs data, while adjusting for age and sex. Survival analysis was also performed to examine associations between the selected PRSs and time-to-event CVD events.

Main results: The PRSs were successfully integrated into the OMOP CDM alongside other measurements. Our model combining PRSs (72 features), conditions (38 features), and measurements (28 features) achieved the highest AUC of 0.69 (95% CI: 63.83–73.55), significantly outperforming model using only conditions and measurements (DeLong's test p=0.03; 95% CI: 0.002–0.063). Model interpretation of the selected 19 features identified 12 conditions, 1 measurement, and 6 PRSs. Only one PRS showed a significant association with high CVD risk: the top 25% PRS quartile had a 1.33-fold higher hazard (95% CI: 1.11–1.60) compared to the average risk group (25th-75th percentile).

Limitations and wider implications: This work shows the feasibility of integrating PRS into the OMOP CDM, but standardized concept ID for each PRS is needed. Addressing this would support tools and enable data sharing across institutions.

References

1. Lambert SA, Wingfield B, Gibson JT, Gil L, Ramachandran S, Yvon F, Saverimuttu S, Tinsley E, Lewis E, Ritchie SC, Wu J. Enhancing the Polygenic Score Catalog with tools for score calculation and ancestry normalization. Nature Genetics. 2024 Oct;56(10):1989-94

IDENTIFYING PREDICTORS OF WEIGHT LOSS FROM A DIVERSE SET OF BIOLOGICAL, BEHAVIOURAL, AND PSYCHOLOGICAL FACTORS

Birgit Malken[1], Kadri Arumäe[2], Ivan Kuznetsov[1], Priit Palta[1], Uku Vainik[1,2]

[1]Institute of Genomics, University of Tartu
[2]Institute of Psychology, University of Tartu

E-mail address of presenting author: birgit.malken@ut.ee

Background: The number of individuals who are overweight or obese has doubled since 1990 and shows no signs of decreasing or stabilization; the prevalence of severe obesity (6%) is even predicted to increase by 130% over the upcoming 20 years (WHO, 2024; Lv et al, 2017).

Study question: We aim to investigate which variables from a diverse set of factors predict weight loss.

Methods: We predicted the percentage of weight lost, calculated as the person's minimum weight following their highest recorded weight, in an Estonian Biobank subsample (N=59,482) of non-weight-stable people with BMI≥25 kg/m^2 at baseline (i.e., at their maximum weight), with at least one follow-up weight measurement available between six months and four years after baseline. We used 202 polygenic scores (PGS) calculted by using SbayesR based on summary statistics from Genome-wide association studies in the UK Biobank cohort, as well as 249 metabolites from Nightingale and 198 personality nuances from a 100NP personality questionnaire.

Main results: Preliminary results show that a higher weight loss percentage is predicted by BMI, psychiatric traits, inflammatory biomarkers and dietary choices.

Limitations: Limitations of our study include non-standardized weight measurement protocols and the inclusion of only individuals of European ancestry.

Wider implications: Weight change as a phenotype is currently still very much understudied. Our research contributes to the understanding of the genetic mechanisms of why people react differently to treatments of obesity. We will use this as an insight for developing a weight loss questionnaire and work on studying genetic causality, which will serve as a roadmap for designing behavioral interventions.

References

1. World Health Organization. (2024, November 23). Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
2. Lv, N., Azar, K. M. J., Rosas, L. G., Wulfovich, S., Xiao, L., & Ma, J. (2017). Behavioral lifestyle

interventions for moderate and severe obesity: A systematic review. Preventive medicine, 100, 180–193. https://doi.org/10.1016/j.ypmed.2017.04.022

PREDICTING BLOOD METABOLITE PROFILES USING MULTI-TARGET NEURAL NETWORKS

Merve Nur Güler[1], Luca Pagani[1,2], Lili Milani[1], Burak Yelmen[1]

[1]Institute of Genomics, University of Tartu, Estonia
[2]Department of Biology, University of Padova, Italy

E-mail address of presenting author: guler@ut.ee

Background: Understanding the genetic basis of complex traits and obtaining predictive genetic risk scores are essential objectives in genomics. Non-linear relationships, interactions, and high-dimensional patterns in genetic data can be difficult to model by common linear methods. Such limitations hinder accurate prediction of complex biological phenotypes, including metabolite profiles.

Study question: This study investigates whether genetic information and deep learning (DL) can be leveraged to predict individual blood metabolite profiles, quantitative biomarkers of metabolic health.

Methods: We designed multi-target DL models for the prediction of 249 metabolites measured by NMR spectroscopy, using 11,935 GWAS-prioritised SNPs along with covariates (age, sex, BMI, and top 10 genetic PCs). The models were trained on 70,708 and validated on 8,838 individuals from the Estonian Biobank.

Main results: We compared traditional linear regression DL approaches for predicting metabolites. A single-target linear model using all covariates and SNPs achieved strong performance (r=0.58 for HDL_size), which was closely matched by our multi-target DL model (r=0.60) trained to predict all metabolites simultaneously. This shows that predicting many traits together can be done without losing accuracy. Our ongoing research focuses on improving DL architectures to enhance performance and disentangle covariate and SNP effects.

Limitations: Prediction accuracy is ultimately limited by the heritability of each metabolite, meaning there is a biological ceiling no model can surpass.

Wider implications: DL offers an efficient way to predict many traits at once, making it a valuable tool for large-scale genomic studies. Beyond metabolites, our approach will be applied to traits like craniofacial measurements and extended to other biobanks.

CVDLINK: Federated AI for Advancing Predictive Models of Cardiovascular Disease

Urmo Võsa[1], Oliver Aasmets[1], CVDLINK Consortium

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

E-mail address of presenting author: urmo.vosa@ut.ee

Background: Biobanks containing genetic, omics, and electronic health records data are valuable resources for improving predictive models for cardiovascular diseases (CVD). However, collaborative machine learning using these data is hindered by technical, legal, and data protection challenges. Federated learning addresses some of these issues by enabling multi-center analyses without sharing sensitive personal information.

Study question: The EU-funded CVDLINK project aims to improve the predictive models for CVD by federated learning on multi-layered data from multiple biobanks.

Methods: A collaborative international effort between clinicians, machine learning specialists, genomics researchers, and biobanks results in a federated AI platform and analysis pipelines that will be used on data from 7 biobanks to train models predicting several CVDs. Models and developed tools are validated in prospective validation studies in multiple countries.

Main results: The AI platform requirements, common data model, and quality control plan have been established. Pilot analyses in the Estonian Biobank demonstrated that baseline health variables, BMI, smoking status, and education predicted sudden cardiac death, suggesting their potential utility in predictive models.

Limitations: The features showing associations should be further cross-replicated and incorporated into multi-center machine learning analyses. Because the annual incidence rate of some target cardiovascular diseases is modest, the performance of some models can only be evaluated at a later stage of the project.

Wider implications: The groundwork laid by federated machine learning projects such as CVDLINK has great potential to improve disease risk prediction and benefit clinicians in planning preventive interventions.

EASIGEN-DS: DESIGNING A DISTRIBUTED RESEARCH INFRASTRUCTURE ON ADVANCED GENOMICS TECHNOLOGIES

Mireia Vaca-Dempere[1], Mònica Bayés[1], Tuuli Reisberg[2], Kadri Raav[2], **Kristiina Tambets**[2] & EASIGEN Consortium

[1]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain
[2]Institute of Genomics, University of Tartu, Tartu, Estonia

E-mail address of presenting author: kristiina.tambets@ut.ee

Innovation in the genomics field is highly dynamic and demands substantial up-front investments. The toolbox of sequencing platforms—each with unique features requiring specialised training and extensive cross-functional expertise—is rapidly expanding.

Here, we propose a design study to establish a new European Research Infrastructure for Advanced Genomics Technologies: **EASIGEN**. The consortium unites a network of eleven genomics facilities, each with distinct application specialities and a combined sequencing capacity of over 450,000 human genomes per year. It also partners with experts in key areas such as data stewardship, ethical, legal and social implications, socio-economic impact, and innovation.

To develop an excellent scientific, technological, and operational design, we will conduct landscape studies, stakeholder consultations, and community surveys. In parallel, we will seek the support and commitment of national and EU stakeholders, as well as user communities, through a multifaceted dissemination plan.

We envision that EASIGEN will encompass:

1. **Technology Platform** – offering state-of-the-art genomic analysis methods, including single-cell genomics, spatial genomics, long-read sequencing, and population-scale genomics capacity.

2. **Support Platform** – providing standardisation tools, training, co-innovation with industry, and consulting services for smaller facilities, including those in widening countries.

Both platforms will serve a broad community of users across the health, biomedical, and life sciences sectors.

This new infrastructure will reinforce the European Research Area, foster interconnections within the ESFRI landscape, and catalyse innovation in the biotechnology-based industry.

THE GENOME OF EUROPE: A REFERENCE DATASET OF GENOMES

Jeroen van Rooij[1], Helen Ray-Jones[1], Erwin Bovenkamp[1], the Genome of Europe consortium[2], André Uitterlinden[1]

Presented on behalf of Genome of Europe consortium by **Andres Metspalu** and **Merit Kreitsberg**

[1]Department of Internal Medicine, Erasmus MC, Rotterdam, the Netherlands
[2]Genome of Europe Consortium, Digital Europe Program, EU

E-mail address of presenting author: merit.kreitsberg@ut.ee

The 1 million genomes (1+MG) initiative aims to make at least 1 million whole genome sequences (WGS) accessible for use in research, health care, and prevention. Working Group 12, named Genome of Europe (GoE), aims to establish a European Reference of >500k WGS (@30x coverage) to better capture European genetic diversity. A digital Europe program (DEP)-funded initiative (also called GoE) will collect 100k samples in line with the 1+MG framework, and via the Genomics Data Infrastructure (GDI). In 2024, we introduced GoE at ESHG.

We here outline the initiated GoE activities involving 51 institutes from 29 European countries. We have generated a sampling scheme for the 40 most common ancestral populations across the participating countries. We have further mapped the sequencing technologies used for this data collection and identify eligible legacy data. Three work packages establish prerequisites for GoE data collection: WP2 (Data quality & standards), WP3 (Ethical & legal) and WP4 (Data Infrastructure). WP5 will perform the data collection. WP6 will demonstrate scientific value through six use cases. WP7 handles communication, and WP1 coordinates.

The GoE project has been well received, and is looking to expand, both by enabling additional countries to join, and by collaboration with other genome datasets. GoE will contribute to setting up national genome programs across Europe and assist in the implementation of genetic information in health care and prevention in a broad sense.

41

GENETIC ORIGINS OF THE KIRITIMATI POPULATION FROM CENTRAL-EASTERN MICRONESIA

Maximilian Larena[1], Afifa Enam Chowdhury[1], Ma. Junaliah Tuazon Kels[1], **Kai Tätte**[2], Mait Metspalu[2], Carina M. Schlebusch[1,3,4], Ralph Garcia-Bertrand[5], and Rene J. Herrera[5]

[1]Human Evolution, Department of Organismal Biology, Uppsala University, SE-752 36 Uppsala, Sweden
[2]Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, 51010, Estonia
[3]Palaeo-Research Institute, University of Johannesburg, Johannesburg, South Africa
[4]SciLife Lab, Uppsala, Sweden
[5]Department of Molecular Biology, Colorado College, Colorado Springs, CO 80903, USA

E-mail address of presenting author: kai.tatte@ut.ee

Background: The migration of Austronesian-speaking populations throughout Oceania has long intrigued researchers. The Kiribati islands, situated along the boundaries of Micronesia and Polynesia, provide a crucial link in this migration.

Methods and study question: We analysed the genome-wide SNP data of the Kiritimati population of Kiribati to uncover their genetic origins and connections with other Oceanian groups.

Main results: Our study reveals that the Kiritimati population primarily exhibits Remote Oceanian-related ancestry associated with ancient Lapita and present-day Polynesian populations. In addition, they harbour about 28% of Papuan-related ancestry. The admixture between East Asian (Austronesian) and Papuan-related component is estimated to have happened around 1830 years before present. Surprisingly, despite having gone through a colonial era, European genetic input is minimal or non-existent. The genetic links between Kiritimati, ancient Lapita, and modern Polynesians underscore the shared ancestry and continuous gene flow across these regions. This genetic continuity and ongoing links are supported by linguistic and cultural evidence, illustrating a complex history of migration and admixture in Oceania.

Limitations: The analyses depend on available reference groups, which have some sampling gaps. Furthermore, ancient samples are especially sparse in this area.

Wider implications: The study clarifies Oceania's migration routes and relationships between populations. It also contributes to identity and heritage awareness among Kiribati people. This is particularly important today, as the rising sea levels threaten the homelands of Pacific island populations and communities begin to relocate across the globe.

“Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.”