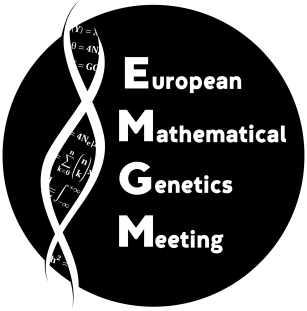
EMGM BREST 2025



EMGM Brest 2025 Book of Abtracts
53rd European Mathematical Genetics Meeting
Brest, France, April 8-9 2025

Edited by the Organisational Committee:

Anthony Herzig
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France
Gaëlle Marenne
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France
Aude Saint Pierre
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France
Ozvan Bocher
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France
Vincent Calvez
Laboratoire de Mathématiques de Bretagne Atlantique - LMBA (UMR 6205)
Equipe d'Analyse, Phénomènes Stochastiques et Applications
UBO, Brest, France
Hervé Perdry
Université Paris-Saclay
Centre de recherche en Epidémiologie et Santé des Populations, Villejuif, France

Abstracts were reviewed by the Scientific Committee:

Florian Privé
Aarhus University
Aarhus, Denmark
Myriam Brossard
Sinai Health System
Toronto, Canada
Simone Rubinacci
Institute for Molecular Medicine Finland (FIMM)
Helsinki, Finland
Anthony Herzig
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France

Gaelle Marenne
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France
Aude Saint Pierre
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France
Ozvan Bocher
Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies
Université Bretagne Occidentale, Brest, France
Hervé Perdry
Université Paris-Saclay
Centre de recherche en Epidémiologie et Santé des Populations, Villejuif, France

Invited Speakers:

Céline Bon
Maître de Conférences en génétique des populations, paléogénétique
UMR 7206, Paris, France
Presentation Title : Using Ancient DNA to Unveil the Genetic Diversity of Neolithic Populations in the Paris Basin: Interactions Between Farmers and Hunter-Gatherers
Beatriz C. D. Cuyabano
Statistical and computational methods to the research in quantitative genetics
INRAE, France
Presentation Title : The challenges of variance components estimation in the multi-omic era
Garrett Hellenthal
Genetics, Evolution & Environment
UCL, UK
Presentation Title : Leveraging haplotype sharing patterns to infer overlapping admixture events among populations

With special thanks to:

Marie-Aude Choque & René Vigouroux
Association Gaétan Saleun

Disclaimer: This volume includes the abstracts submitted by the authors, accepted following review by the Scientific Committee. Editorial changes were limited to formatting, stylistic harmonisation, standardisation of affiliations. The authors remain responsible for the content of their abstracts.

How to cite this volume:

Anthony Herzig, Gaëlle Marenne, Aude Saint Pierre, Ozvan Bocher, Vincent Calvez, and Hervé Perdry. *Book of Abstracts: 53rd European Mathematical Genetics Meeting.* Brest, France, 2026

Oral Presentations

Parent-of-Origin inference and its role in the genetic architecture of complex traits: evidence from ∼265,000 individuals

Hofmeister Robin[1,2], Cavinato Théo[1], Karimi Roya[3], Van Der Graaf Adriaan[1], Pajuste Fanny-Dhelia[2], Kronberg Jaanika[2], Taba Nele[2], Magi Reedik[2], Vaudel Marc[3], Rubinacci Simone[4], Johansson Stefan[3], Milani Lili[2], Delaneau Olivier[5], Kutalik Zoltan[1]

1 - University of Lausanne (Switzerland), 2 - University of Tartu (Estonia), 3 - University of Bergen (Norway), 4 - University of Helsinki (Finland), 5 - Regeneron Genetics Center (United States)

Parent-of-origin effects (POEs) occur when the impact of a genetic variant depends on its parental origin. Traditionally linked to genomic imprinting, these effects are believed to have evolved from parental conflict over resource allocation to offspring, which results in opposing parental genetic influences. Despite their potential importance, POEs remain heavily understudied in complex traits, largely due to the lack of parental genomes. Here, we present a multi-step approach to infer the parent-of-origin of alleles without parental genomes, leveraging inter-chromosomal phasing, mitochondrial and chromosome X data, and sibling-based crossover inference. Applied to the UK (discovery) and Estonian (replication) Biobank, we inferred the parent-of-origin for up to 221,062 individuals, representing the largest dataset of its kind. GWAS scans in the UK Biobank for more than 60 complex traits and over 2,400 protein levels contrasting maternal and paternal effects identified over 30 novel POEs and confirmed more than 50% of testable known associations. Notably, approximately half of our POEs exhibited a bi-polar pattern, where maternal and paternal alleles exert conflicting effects. These effects were particularly prevalent for traits related to growth (e.g., IGF-1, height, fat-free mass) and metabolism (e.g., type 2 diabetes, triglycerides, glucose). Replication in the Estonian Biobank and in 45,402 offspring from the Norwegian Mother, Father and Child Cohort Study validated over 75% of testable associations. Overall, our findings shed new light on the influence of POEs on diverse complex traits and align with the parental conflict hypothesis, providing compelling evidence for this understudied evolutionary phenomenon.

Detection of mosaic uniparental disomy from whole-exome and whole-genome sequencing data of single patient

Seeleuthner Yoann[1,2], Chaldebas Matthieu[1,2,3], Zhang Peng[1,2,3], Abel Laurent[1,2,3], Casanova Jean-Laurent[1,2,3,4,5], Bohlen Jonathan[6], Cobat Aurélie[1,2,3]

1 - HGID (France), 2 - Imagine Institute (France), 3 - St. Giles Laboratory of Human Genetics of Infectious Diseases (United States), 4 - Department of Pediatrics, Necker Hospital for Sick Children, Paris, France, EU (France), 5 - Howard Hughes Medical Institute [New York] (United States), 6 - Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität (Germany)

Uniparental isodisomy (UPD) are large genomic anomalies in which both copies of a chromosome or segment of a chromosome are inherited from a single parent. UPD can cause disease through parent of origin effect or homozygosity for a pathogenic variant underlying autosomal recessive diseases. UPD can be constitutional or somatic, affecting in the latter case only a fraction of the cells and resulting in somatic mosaicism. In the hematopoietic tissue, somatic mosaicism is frequent in the elderly and the prevalence increases with age. While constitutional UPD can be identified from whole-exome (WES) or whole-genome (WGS) sequencing data through the detection of runs of homozygosity (ROH), detection of mosaic UPD (mUPD) rely on SNP array. Here, we developed a two-states Hidden Markov Model (HMM) that uses the Minor-Reads Ratio (MRR) of variants at the heterozygous state from WES/WGS data of a single patient to identify chromosomal stretches with MRR values deviating from the 0.5 expected value. We performed large simulation studies to optimize the parameters of the HMM algorithm, and assessed its performance under various scenarios. Under H0, we analyzed 50 unrelated individuals with WES from our in-house database and found less than one event per individual, 90% of which were constitutional ROH. We further simulated telomeric mUPD with various MRR in those 50 individuals. For mUPD larger than 10Mb and simulated MRR from 0 to 0.30, the sensitivity of MOSUPD was higher than 95%. In conclusion, we developed an efficient and powerful tool to detect mUPD from single patient WES/WGS.

Detecting rare recessive variants involved in multifactorial diseases: validation and power of the Fantasio method

Foulon Sidonie[1,2], Truong Thérèse[1], Leutenegger Anne-Louise[2], Perdry Hervé[1]

1 - CESP Inserm U1018, Université Paris-Saclay, F-94807 Villejuif (France), 2 - Inserm Université Paris Cité, NeuroDiderot, Inserm U1141, Paris (France)

Genome-wide association studies (GWAS) aim to detect associations between genetic variants and multifactorial traits. They mainly use common variants and study them according to the additive genetic model. However, the genetic component of most multifactorial diseases is not yet fully elucidated. This could be partly due to the contribution of rare variants with recessive effects, which are difficult to identify in GWAS. We propose Fantasio, a method based on an excess of Homozygous-by-Descent (HBD) segments shared among cases compared to what is expected among controls. HBD segments, found in consanguineous individuals, are regions where rare recessive variants are more likely to be found. We present a simulation framework to assess the type I error and power of Fantasio, and the results. In these simulations, haplotypes from 1000 Genomes are shuffled to create new 'mosaic' haplotypes, allowing to control the consanguinity coefficient of simulated individuals. Some consanguineous cases are selected to carry rare recessive variants in a specific genomic region, while other cases and controls have varying degrees of consanguinity. The sample size, the percentage of cases linked to the rare recessive variants and the types of consanguinity are varied. Preliminary results show that the type I error is well controlled. For some genetic models (rare disease with 10% consanguineous cases), Fantasio achieves high power starting from a sample size of 250 cases. With these results, we are confident our method will be a good tool for studying rare recessive variants in more common multifactorial diseases, particularly with large sample sizes.

The effect of stratified type 2 diabetes genetic liability on non-cardiometabolic comorbidities

Arruda Ana Luiza[1,2,*], Bocher Ozvan[1,*], Taylor Henry[4,5,6,*], Cammann Davis[3,*], Yoshiji Satoshi[7,8], Yin Xianyong[9,10], Zhao Chi[11], Chen Jingchun[3], Wood Alexis[12], Suzuki Ken[13], Mercander Josep[7,14,15], Spracklen Cassandra[11], Meigs James[7,16,17], Vujkovic Marijana[18,19,20], Davey-Smith George[21], Rotter Jerome[22], Voight Benjamin[20,21,23,24], Morris Andrew[25], Zeggini Eleftheria[1,26]

1 - Institute of Translational Genomics, Helmholtz Munich, Neuherberg, 85764 (Germany), 2 - Technical University of Munich (TUM), School of Medicine and Health, Graduate School of Experimental Medicine, Munich, 81675 (Germany), 3 - Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, 4505 S. Maryland Parkway, Las Vegas, NV 89154 (USA), 4 - Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD (USA), 5 - British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge (UK), 6 - Heart and Lung Research Institute, University of Cambridge, Cambridge (UK), 7 - Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA (USA), 8 - Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto (Japan), 9 - Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing (China), 10 - Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI (USA), 11 - Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, MA (USA), 12 - USDA/ARS Children's Nutrition Center, Baylor College of Medicine, Houston, TX (USA), 13 - Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, University of Tokyo, Tokyo (Japan), 14 - Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA (USA), 15 - Harvard Medical School, Boston, MA (USA), 16 - Department of Medicine, Harvard Medical School, Boston, MA (USA), 17 - Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA (USA), 18 - Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA (USA), 19 - Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA (USA), 20 - Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA (USA), 21 - MRC Integrative Epidemiology Unit, University of Bristol, BS8 2BN Bristol (UK), 22 - Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA (USA), 23 - Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA (USA), 24 - Institute for Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA (USA), 25 - Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester (UK), 26 - TUM School of Medicine and Health, Technical University of Munich and Klinikum Rechts der Isar, Munich, 81675 (Germany)

Type 2 diabetes (T2D) is epidemiologically associated with a wide range of non-cardiometabolic comorbidities, yet their shared etiology remains underex-

plored. The T2D Global Genomics Initiative (T2DGGI) has defined eight non-overlapping mechanistic clusters of genetic risk variants from multi-ancestry T2D GWAS meta-analysis (2,535,601 individuals including 428,452 cases) that represent distinct pathways to disease (including beta-cell dysfunction and obesity). Using these T2D genetic clusters, we investigate putative causal links between cluster-stratified T2D liability and 21 non-cardiometabolic comorbidities through two-sample Mendelian randomization (MR) analyses. We find evidence of potential causal effects of T2D liability on 15 comorbidities, with most effects being driven by specific T2D genetic clusters. The obesity cluster is linked to 10 comorbidities and shows the strongest effects. For instance, we find evidence that the causal effect of T2D liability on cataracts (OR=1.06 [1.04, 1.08], P=5.82x10-9) is driven by the obesity cluster (OR=1.12 [1.07, 1.17], P=3.84x10-6). This effect persisted after adjusting for BMI in a multivariate MR analysis (OR=1.1 [1.06,1.16], P=6.48x10-5). We detect cluster-stratified T2D liability effects for osteoarthritis with divergent effect directions. We recapitulate the well-established link between T2D, osteoarthritis, and obesity by showing evidence of a causal effect of the obesity cluster on osteoarthritis (OR=1.23 [1.1.9, 1.27], P=1.03x10-37). Conversely, T2D liability linked to beta-cell dysfunction was found to be protective against osteoarthritis (OR=0.95 [0.93,0.98], P=1.7x10-4). Our findings provide insights into the biological mechanisms underpinning T2D liability and non-cardiometabolic comorbidities, offering a foundation for prevention strategies tailored to the genetic and multimorbidity profiles of T2D patients.

Using Genomic Structural Equation-Based Polygenic Scores to Improve Type II Diabetes Management

Koitmäe Merli[1,2], Fischer Krista[2,1], Läll Kristi[1], Möls Märt[2], Mägi Reedik[1]

1 - Institute of Genomics, University of Tartu (Estonia), 2 - Institute of Mathematics and Statistics, University of Tartu (Estonia)

Given the heterogeneous nature of type II diabetes (T2D), its pathophysiology, disease trajectory, and treatment responses vary substantially across individuals. Understanding the distinct biological mechanisms leading to a T2D diagnosis is therefore crucial for optimizing prevention and management strategies. To disentangle these complex pathways, we propose leveraging genomic structural equation modeling (genomic SEM) to identify latent genetic factors that drive distinct routes to disease onset. Genomic SEM integrates genome-wide association study summary statistics from multiple phenotypes to uncover shared and independent genetic architectures underlying T2D risk. By applying this approach, we aim to characterize biologically distinct subtypes of T2D and subsequently construct polygenic risk scores (PRS) for these latent factors. Using genomic SEM on T2D and its associated risk factors, we will delineate distinct genetic pathways contributing to disease development. We will then assess the predictive utility of these pathway-specific PRS in determining disease progression, treatment response, and comorbidity profiles, validating their efficacy in the Estonian Biobank dataset. Given that certain T2D subtypes may benefit more from early insulin intervention while others respond preferentially to lifestyle modifications, our findings could inform personalized strategies for diabetes prevention and management, ultimately advancing precision medicine in metabolic disorders. Our early findings reveal three distinct pathways shaping Type II diabetes: glucose homeostasis, insulin sensitivity and obesity related phenotypes. In summary, our study aims to uncover distinct genetic pathways in Type II diabetes, enabling better risk prediction and treatment stratification. These insights could pave the way for more effective and personalized interventions.

PRISM: a pleiotropy-driven framework to disentangle the effects of genetic variants in complex traits

Tournaire Martin[1], Nouira Asma[1], Favre-Moiron Mario[1], Rozenholc Yves[1], Verbanck Marie[1,2]

1 - UR7537-BioSTM, Biostatistique, Traitement et Modélisation des données biologiques (France), 2 - Inserm U900 (France)

Genome-wide association studies (GWAS) have uncovered countless genetic variants associated with complex human traits and diseases. Yet, association does not imply causality, and pinpointing which variants exert direct causal effects, and through which mechanisms, remains a formidable challenge. Although molecular markers such as eQTLs are often used to interpret GWAS findings, their limited overlap with disease-associated variants has recently sparked substantial concerns. In contrast, we present PRISM (Pleiotropic Relationships to Infer the SNP Model), a novel computational framework that leverages pleiotropy to disentangle direct variant–trait effects from pleiotropic effects in GWAS summary statistics. Drawing inspiration from Mendelian randomization, PRISM partitions significant associations into three categories: confounder-mediated, trait-mediated, and direct effects. By integrating results across a wide panel of traits, PRISM constructs robust and interpretable variant–trait pleiotropy networks. We first validated PRISM in a comprehensive GWAS simulation environment encompassing multiple complex scenarios. The method achieved high precision in distinguishing direct effects and accurately reconstructing variant networks. We then applied PRISM to 70 traits and diseases from the UK Biobank, revealing that direct effects represent less than 12% of significant variant-trait associations, yet are disproportionately enriched in heritability. Additionally, multiple lines of evidence suggest that PRISM variant networks reflect established biological pathways. By leveraging pleiotropy, PRISM offers a valuable approach for advancing the understanding of the genetic architecture underlying complex human traits and diseases.

Using multiomic integration to improve blood biomarkers of major depressive disorder: a case-control study

Mokhtari Amazigh, Ibrahim El Chérif, Gloaguen Arnaud, Barrot Claire-Cécile, Cohen David, Derouin Margot, Vachon Hortense, Charbonnier Guillaume, Loriod Béatrice, Decraene Charles, Yalcin Ipek, Marie-Claire Cynthia, Etain Bruno, Belzeaux Raoul[1], Lutz Pierre-Eric[2], Delahaye-Duriez Andrée[3]

1 - Institut de Neurosciences de la Timone (France), 2 - Institut des Neurosciences Cellulaires et Intégratives (France), 3 - NeuroDiderot U1141 (France)

Major depressive disorder (MDD) is a leading cause of disability, with a twofold increase in prevalence in women compared to men. Over the last few years, identifying molecular biomarkers of MDD has proven challenging, reflecting interactions among multiple environmental and genetic factors. Recently, epigenetic processes have been proposed as mediators of such interactions, with the potential for biomarker development. We characterised gene expression and two mechanisms of epigenomic regulation, DNA methylation (DNAm) and microRNAs (miRNAs), in blood samples from a cohort of individuals with MDD and healthy controls (n = 169). Case-control comparisons were conducted for each omic layer. We also defined gene coexpression networks, followed by step-by-step annotations across omic layers. Third, we implemented an advanced multiomic integration strategy, with covariate correction and feature selection embedded in a cross-validation procedure. Performance of MDD prediction was systematically compared across 6 methods for dimensionality reduction, and for every combination of 1, 2 or 3 types of molecular data. Feature stability was further assessed by bootstrapping. Results showed that molecular and coexpression changes associated with MDD were highly sex-specific and that the performance of MDD prediction was greater when the female and male cohorts were analysed separately, rather than combined. Importantly, they also demonstrated that performance progressively increased with the number of molecular datasets considered. Informational gain from multiomic integration had already been documented in other medical fields. Our results pave the way toward similar advances in molecular psychiatry, and have practical implications for developing clinically useful MDD biomarkers.

Deep Bayesian estimation of the intensity and timing of selection from a thousand ancient genomes of East Eurasians.

Laval Guillaume[1], Patin Etienne[1], Quintana-Murci Lluis[1]

1 - Human Evolutionary Genetics, CNRS UMR2000 (France)

Allele frequency trajectories have proven to be key for understanding the impact of natural selection. Here, we developed novel approximate Bayesian computation (ABC) and convolutional neural networks (CNNs) methods to estimate the intensity (s) and timing (T) of positive and negative selection, by directly using ancient and modern genotypes sampled over time. Through both computer simulations and analysis of empirical data, we showed that ABC and CNNs provide accurate, consistent predictions and can handle the scarcity of ancient samples for certain epochs. Our new predictions confirmed an increased frequency of recent selection in Europe, supporting a history of recent selection on host defenses against pathogens, consistent with previous work. We then applied our ABC and CNN algorithms, together with model-free methods based on the FST statistics, to a dataset comprising 1,176 ancient and 600 modern East Eurasians sampled over the past 10,000 years. We found a strong enrichment of coding mutations in selection signals (odds ratio for selection, OR=24, P<10-8), providing strong evidence for positive selection throughout the Holocene period in East Eurasia. We replicated the excess of recent selection found in Europe and identified both well-known and novel candidate genes for positive selection linked to metabolism, skin pigmentation and host defense against pathogens, including SLC44A5, ADH1B, ALDH2 and CYP2D6. By combining population genetics modeling, deep artificial neural networks, and extensive paleogenomic data, our study demonstrates the impact of natural selection on ancient and modern humans and shed light on the evolution of human diseases in East Eurasia.

Integrating Temporal and Spatial Dimensions in Genetic Selection Analysis: A Comprehensive Approach to Detecting Positive Selection in Human Genomes

Dina Christian[1], Bouvier Célia[2], Jolivet Zoé[2]

1 - Institut du thorax (IRS- Université de Nantes, 8 quai Moncousu, BP 7072, 44007 NANTES Cedex 1 France), 2 - Ecole Centrale Nantes (France)

Testing for genetic selection, particularly positive selection, involves identifying alleles that have increased in frequency due to evolutionary advantages. Modern and ancient DNA analyses provide powerful tools for this purpose. One approach reconstructs demographic histories using coalescent-based methods to understand population dynamics and then examines allele frequency trajectories over time. This method, exemplified by models like those developed by Li and Durbin (2011), helps identify alleles that have risen to prominence more rapidly than expected under neutral evolution, indicating positive selection. Another method, exemplified by Akbari et al., models allele frequency against time, similar to Genome-Wide Association Studies (GWAS), while accounting for genetic relatedness among individuals. This approach uses a genetic relationship matrix (GRM) to control for population structure and kinship, providing a more nuanced view of selection by distinguishing true selective sweeps from genetic drift or demographic effects. By integrating these methods, researchers can pinpoint genomic regions under selection, offering insights into human adaptation and evolutionary history. Building upon these foundations, we are extending the model to incorporate both temporal and spatial dimensions. This enhanced framework will allow us to track not only when selection events occurred but also where they took place geographically, providing a more comprehensive understanding of the selective pressures shaping human genomes.

Mutational signatures of deterministic and noise-induced evolutionary mechanisms

Insalata Ferdinando[1,2], Green Alistair[1,2], Lee Colman[1], Nastassia Bonetti[1,3], Fruet Cecilia[4], Jones Nick[1,5]

1 - Department of Mathematics, Imperial College London (United Kingdom), 2 - Department of Clinical Neurosciences, University of Cambridge (United Kingdom), 3 - School of Engineering, ENSTA Paris (France), 4 - Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (Switzerland), 5 - EPSRC Centre for Mathematics of Precision Healthcare (United Kingdom)

When two or more species compete, a typical problem in evolutionary biology is understanding what is the mechanism that could lead to one prevailing over the others, a necessary step in developing appropriate models. While it is intuitive to posit that the prevailing species has a higher overall growth rate, for example a higher replication rate, noise-induced selection mechanisms have attracted increasing attention in recent years. Models showing stochastic selection are often counterintuitive, as they can present identical growth rates for all species, but are widely relevant, given the intrinsically stochastic nature of biological systems. A crucial point is whether we can distinguish between different selection mechanisms with data that are available in a typical experiment. Here we develop a framework that compares the frequency distributions of randomly occurring neutral mutations in a spatially extended system where a species is expanding in a wave-like fashion. We find qualitative signatures of these frequency distributions that discriminate between explicit selection based on differences in growth rates and noise-induced selection, driven by differences in carrying capacity or in baseline turnover rates, but with identical growth rates. In addition, we find that standard statistical tests are able to detect these differences with sample sizes that are reasonable in an experimental setting. Our findings are applicable to current challenges in the field of evolutionary biology, as noise-induced selection has been repeatedly implied in debated phenomena, such as the spread of altruistic traits.

Characterizing the evolution and phenotypic impact of ampliconic Y chromosome regions

Lucotte Elise[1,2], Guðmundsdóttir Valdís Björt[3,4], M. Jensen Jacob[2], Skov Laurits[2], Coll Macià Moisès[2], Almstrup Kristian[5], H. Schierup Mikkel[2], Helgason Agnar[3,4], Stefansson Kari[3,6]

1 - ESE (France), 2 - Bioinformatics Research Centre, Aarhus University (Denmark), 3 - deCODE genetics [Reykjavik] (Iceland), 4 - Department of Anthropology, University of Iceland (Reykjavik) (Iceland), 5 - Department of Growth and Reproduction, Rigshospitalet, Copenhagen (Denmark), 6 - Faculty of Medicine, School of Health Sciences, University of Iceland (Iceland)

A major part of the human Y chromosome consists of palindromes with multiple copies of genes primarily expressed in testis, many of which have been claimed to affect male fertility. We examined copy number variation in these palindromes based on whole genome sequence data from 11,527 Icelandic men. Using a subset of 7947 men grouped into 1449 patrilineal genealogies, we infer 57 large scale de novo copy number mutations affecting palindrome 1. This corresponds to a mutation rate of 2.34x10-3 mutations per meiosis, which is 4.1 times larger than our phylogenetic estimate of the mutation rate (5.72x10-4), suggesting that de novo mutations on the Y are lost faster than expected under neutral evolution. Although simulations indicate a selection coefficient of 1.8% against non-reference copy number carriers, we do not observe differences in fertility among sequenced men associated with their copy number genotype, but we lack statistical power to detect differences resulting from weak negative selection. We also perform association testing of a diverse set of 341 traits to palindromic copy number without any significant associations. We conclude that large-scale palindrome copy number variation on the Y chromosome has little impact on human phenotype diversity.

Triangulating evidence to detect signatures of stabilizing selection acting on molecular traits in humans

Zanoaga Mihaela Diana[1], Kutalik Zoltan[1]

1 - University of Lausanne, Department of Computational Biology, Statistical Genetics Group (Switzerland)

Background: Evidence suggests that stabilizing selection shapes molecular trait evolution in primates, maintaining transcript and protein levels within optimal range. However, quantifying stabilizing selection remains challenging. Here, we propose a comprehensive approach that triangulates evidence from non-linear Mendelian Randomization (MR) and selection-aware GWAS models.
Data and Methods: We leveraged UK Biobank data from 337,386 unrelated white-British individuals, alongside association summary statistics for protein QTLs (3,000 proteins). First, we applied state-of-the-art methods, including GRM-MAF-LD and LDpred2, to infer stabilizing selection on protein levels. Second, we developed a robust non-linear Mendelian Randomization (MR) approach using linear and squared Polygenic Risk Scores (PRS) as instruments to investigate the presence of an inverted U-shaped causal relationship between molecular traits and fitness proxies (e.g., number of offspring), which would indicate stabilizing selection.
Results: Applying GRM-MAF-LD to 2,000 proteins, we identified significant selection estimates ($p \leq 0.01$), for 858 proteins, 94% of which were negative and therefore indicative of stabilizing selection. Non-linear MR identified 14 hits (at $p \leq 0.01$), with 71% being under stabilizing selection, out of these, 7 proteins were confirmed by the GRM-MAF-LD approach, and they were primarily linked to immune functions and lipid metabolism. PLA2G4A, a highly conserved gene across species and involved in inflammation signalling was identified to be under the strongest selection.
Conclusions: While selection-aware GWAS methods appear to be more powerful than MR, they provide orthogonal lines of evidence for stabilizing selection enabling robust prioritization of proteins most shaped by this evolutionary mechanism.

ChoruMM: a versatile multi-components mixed model for bacterial-GWAS

Frouin Arthur[1], Laporte Fabien[2,3], Hafner Lukas[4], Maury Mylène[4], Mccaw Zachary[5], Julienne Hanna[1], Henches Léo[1], Leclercq Alexandre[4,6], Chikhi Rayan[1], Lecuit Marc[4,6,7], Aschard Hugues[1,8]

1 - Department of Computational Biology, Institut Pasteur (France), 2 - Mer, Molécules, Santé, Université Catholique de l'Ouest - Angers (3 Place André Leroy, 49000 Angers France), 3 - Department of Computational Biology, Institut Pasteur (25-28 rue du Docteur Roux, 75015 Paris France), 4 - Institut Pasteur, Université Paris Cité, Inserm U1117, Biology of Infection Unit, Paris, France (France), 5 - Insitro [San Francisco] (United States), 6 - Biology of Infection Unit, National Reference Center and WHO Collaborating Center Listeria, Institut Pasteur, Inserm U1117 (France), 7 - Necker-Enfants Malades University Hospital, Department of Infectious Diseases and Tropical Medicine, Institut Imagine, AP-HP, Paris, France (France), 8 - Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA (United States)

Genome-wide Association Studies (GWAS) have been central in studying the genetics of complex human outcomes. In the last years there have been multiple efforts for implementing GWAS-like approaches to study pathogenic bacteria. Although a variety of methods have been proposed, it remains unclear how to appropriately model the complex population structure of bacterial cohorts. Here we examine the genetic structure underlying whole-genome sequencing data from 912 Listeria monocytogenes strains, and demonstrate that the standard human genetics model, commonly assumed by existing bacterial GWAS methods, is inadequate for studying such highly structured organisms. We leverage these results to develop ChoruMM, a robust and powerful multi-component linear mixed model, where components are inferred from a hierarchical clustering of the bacteria genetic relatedness matrix. We demonstrate through extensive simulations that our approach led to a diminution of false positive signals while maintaining a satisfying detection rate. Our ChoruMM package also includes post-processing and visualization tools that address the pervasive long-range correlation observed in bacterial genomes and allow for the assessment of type I error rate calibration. Transcript-level analyses of prfA, a establish Listeria virulence gene, demonstrated that ChoruMM effectively extracted relevant biological signals linked to Listeria virulence or the expression of its key virulence genes.

A robust liability-scale R-squared

Pedersen Emil[1], Steinbach Jette[1], Schork Andrew Joseph[2], Krebs Morten[2], Agerbo Esben[1], Privé Florian[1,3], Vilhjálmsson Bjarni[1,4]

1 - National Centre for Register-Based Research, Aarhus University, Aarhus (Denmark), 2 - Institute of Biological Psychiatry, Mental Health Services Sct. Hans, Roskilde (Denmark), 3 - Aarhus University (Denmark), 4 - Bioinformatics Research Centre, Aarhus University, Aarhus (Denmark)

Background: When predicting binary traits, the liability-scale R^2 is commonly used to assess the amount of variance explained by a given set of predictors. While the standard transformation from observed to liability-scale provided by Lee et al. (Gen Epi, 2012) accounts for case ascertainment, it relies on assumptions that may not be satisfied. Method: In this study, we compare the standard liability-scale transformation to an R^2 based on a weighted probit regression in both simulations and applications to the iPSYCH2015 case-cohort study. We consider generative models with combinations of continuous and discretely distributed predictors to represent common variables in genetics such as polygenic scores, family history measures, rare variants, and sex. Results: For a single predictor at a population prevalence of 1%, the common liability-scale R^2 transformation overestimates the true proportion of variance in liability explained by a family history indicator by at least 270%. The weighted probit regression R^2 shows much reduced bias (at most 12%). Using the Lee et al equations to estimate the total liability variance explained by multiple predictors can result in substantial bias (over- or underestimates by at least 25%) when non-normally distributed predictors are included, while the weighted probit R-squared is more robust (within 3%). Conclusion: Commonly applied transformation of the variance explained by genetic predictors to a liability-scale R-squared can be significantly biased by ascertainment in the sample and distributional properties of predictors. We recommend using a more robust probit regression when models contain predictors that are non-normal or skewed in distribution.

Search of the genetic predisposition for Hip Dysplasia, using the dog model through leveraging One Health and Open Science

Le Nézet Louis[1], Herzig Anthony[2], Machou Lauranne[1], Guyon Richard[1], Hedan Benoit[1], Derrien Thomas[1], Génin Emmanuelle[2,3], Quignon Pascale[1], FÉrec Claude[2,3], André Catherine[1]

1 - Institut de Génétique et Développement de Rennes (IGDR) UMR 6290, Rennes (France), 2 - Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest (France), 3 - CHU Brest, F-29200 Brest (France)

Hip dysplasia is a complex, polygenic disorder affecting both humans (Developmental Dysplasia of the Hip, DDH) and dogs (Canine Hip Dysplasia, CDH), leading to joint instability and osteoarthritis. Dogs constitute a unique spontaneous model for studying complex diseases due to their distinct genetic isolates across 400 breeds. Our study investigates CHD, which shares physiopathology with DDH, by identifying genetic loci and variants using Genome-Wide Association Studies (GWAS) in five predisposed breeds. We gathered clinical data from over 1000 dogs and sequenced 700 using cost-effective low-coverage (1X) whole-genome sequencing. These data were mapped on the CanFam-4 reference dog genome using nf-core/sarek and genotype imputation was performed using the nf-core/phaseimpute pipeline that we developed. Its modular design and nf-core integration ensure standardized, scalable, and transparent genomic analyses in diverse species. GWAS across and within breeds identified candidate loci near genes involved in skeletal development and joint formation, overlapping with previous canine and human studies, underscoring conserved genetic pathways influencing hip dysplasia. Ongoing analyses will further clarify genetic risk factors in dogs, with parallel investigations in human DDH cases. A key strength of our study is applying the same methodical approach to both human and canine sequencing data. We aim to compare canine GWAS data to human DDH and develop predictive genetic risk tests for dog breeding to reduce CHD prevalence. By enhancing our understanding of the genetic basis of CHD and providing open-source tools, our research contributes to the One Health initiative, bridging veterinary and human medicine to address shared genetic challenges.

Joint plasma pQTL analysis in the UK Biobank with efficient Bayesian inference

Li Yiran[1], Ruffieux Hélène[1], Whittaker John[1], Richardson Sylvia[1]

1 - MRC Biostatistics Unit, University of Cambridge (United Kingdom)

Plasma proteins play critical roles in biological processes and serve as biomarkers and drug targets. Genetic variants influencing plasma protein levels, termed protein quantitative trait loci (pQTLs), and especially pQTL hotspots (variants affecting multiple proteins), provide valuable insights into complex disease mechanisms and regulatory pathways. Despite its popularity, univariate screening for pQTLs has limited power when the number of tests is large and the associations are weak. To address this, we employ a scalable Bayesian joint-modelling method for pQTL hotspot discovery on the data of the UK Biobank Pharma Proteomics Project (Sun et al., 2023), which encompasses measurements of 2,923 blood plasma proteins in over 50,000 individuals. Our approach extends the hierarchical regression approach atlasQTL (Ruffieux et al., 2020) with a 'partial-update' variational inference algorithm that adaptively selects subsets of parameters to update in each iteration, enabling efficient analysis of Biobank-scale datasets. Simulation studies and pilot applications on chromosome 19 have demonstrated the algorithm's scalability and power. Our method promises to uncover a more comprehensive set of proteins regulated by hotspot loci, advancing the understanding of protein pathways and their roles in human health and disease. Sun, B. B., Chiou, J., Traylor, M., et al. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature, 622(329–338). https://doi.org/10.1038/s41586-023-06592-6. Ruffieux, H., Davison, A. C., Hager, J., Inshaw, J., Fairfax, B. P., Richardson, S., & Bottolo, L. (2020). A global-local approach for detecting hotspots in multiple-response regression. Annals of Applied Statistics, 14(2), 905–928. https://doi.org/10.1214/20-AOAS1332.

Summary Statistic-Derived Metabolite Ratio QTL Analysis Identifies 427 Otherwise Missed Loci and Reveals Catalyzing Enzymes

Rizi Sadegh[1], Kutalik Zoltan[2], Van Der Graaf Adriaan[2]

1 - University of Tehran (Iran), 2 - Université de Lausanne = University of Lausanne (Switzerland)

Background: Metabolic pathways are dysregulated in diseases like type 2 diabetes. Yet, quantitative trait locus (QTL) studies focus on one biomarker at a time. Metabolite ratios are established proxies for pathway activity, however, their use in QTL studies has been limited due to the quadratic computational burden of pairwise combinations.
Materials and Methods / Results: We developed a statistical framework to compute ratio QTL (rQTL) from existing single metabolite QTLs (mQTL) summary statistics. Validating on 299 classically derived ratios (from individual-level data), we find strong concordance (β: R2 = 0.87). After application to all (261,003) pairwise ratios of 723 metabolites we identified 17,454 rQTL with stronger significance than either constituent metabolite at Bonferroni significance ($P < 2 \times 10^{-13}$), of which 437 rQTLs were not study-wide significant ($P < 5 \times 10^{-11}$) in the original mQTL study. The closest genes to the novel rQTLs were enriched for lipid metabolism (Reactome: $P = 4 \times 10^{-14}$). Cis Mendelian randomization and colocalization analysis identifies 9 proteins significantly influencing 521 metabolite ratios (MR-link-2 $P < 1 \times 10^{-6}$), coloc $PP4 > 0.9$). All proteins were enzymes, including sulfotransferase 2A1 (SULT2A1) protein abundance as a causal regulator of the ratio between dehydroepiandrosterone sulfate and etiocholanolone glucuronide (novel rQTL $P = 2 \times 10^{-15}$): two testosterone metabolites. Conclusion: Our framework enables scalable and resource-efficient rQTL mapping from summary statistics, bypassing the need for individual-level data. We identify previously unknown metabolite pathway regulation, efficiently resolving rQTLs, otherwise understudied as phenotypes for QTL mapping.

Epigenetic signatures in developmental disorders: can signature patterns disclose enriched information on phenotypes?

Santini Amandine[1], May Angèle[2], Richard Anne-Claire[2], Nava Caroline[3], Cogné Benjamin[4], Thevenon Julien[5], Charenton Clément[6], Collaborators Anddi-Rares, Bonnet Céline[7], Depienne Christel[8], De Dieuleveult Maud[9], Nicolas Gaël[2], Charbonnier Camille[10]

1 - Univ Rouen Normandie, Normandie Univ, Inserm U1245, F-76000 Rouen, France (France), 2 - Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, Department of Genetics and Reference Center for Developmental Abnormalities, F-76000 Rouen, France (France), 3 - Sorbonne Université, Institut du Cerveau-Paris Brain Institute-ICM, Inserm, CNRS, Hôpital Pitié Salpêtrière, Assistance Publique- Hôpitaux de Paris (APHP), Département de Génétique, Paris, France (France), 4 - Nantes Université, CHU de Nantes, Service de Génétique médicale, CNRS, INSERM, l'institut du thorax, Nantes, France, laboratoire GCS SeqOIA, Paris, France (France), 5 - Service de Génétique, Génomique et Procréation, CHU Grenoble Alpes, Université Grenoble Alpes, INSERM U 1209, CNRS UMR 5309, Institut for Advanced Biosciences, Grenoble, France, GCS AURAGEN, Lyon, France (France), 6 - CNRS, Inserm, Université de Strasbourg, IGBMC UMR 7104- UMR-S 1258, Department of Integrated Structural Biology, IGBMC, Illkirch, France (France), 7 - Laboratoire de Génétique, CHRU de Nancy, INSERM NGERE U1256, Université de Lorraine, Vandœuvre-lès-Nancy, France (France), 8 - Institute of Human Genetics, University Hospital Essen, University Duisburg-Essen, Essen, Germany (Germany), 9 - INSERM U1163, Université de Paris, Imagine Institute, Paris, France (France), 10 - Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, department of Biostatistics and Reference Center for Developmental Abnormalities, F-76000 Rouen, France (France)

Due to heterogeneous and overlapping clinical signs, the diagnosis of neurodevelopmental disorders is complex. Many episignatures have been identified to help confirm diagnostic hypotheses or interpret candidate variants. However, few studies have investigated the fine-scale correlation between methylation and phenotype. We generated a database of 501 Illumina Epic array methylation profiles, including carriers of pathogenic variants in RNU4-2 (n=35), CHD3 (n=30), DNMT3A (n=36) and 86 age-matched normal controls. Following standard quality controls, normalization and differential analysis adjusting for confounding factors, we identified novel CHD3 and RNU4-2 episignatures. Methylation patterns for CHD3, RNU4-2 and DNMT3A (published signature) were compared to phenotype severity. CHD3 methylation patterns were compared to CHD7 (n=29) and CHD8 (n=32) published signatures. Beyond their ability to separate patients from controls, RNU4-2, CHD3 and DNMT3A signatures all showed heterogeneous methylation profiles among patients, driven by biological rather than technical variability. In particular, DNMT3A and RNU4-2 strengths of signature correlated with disease severity.

Joint analysis of CHD methylation profiles revealed two underlying components in the CHD7 signature. The most extreme component strongly mirrored the novel CHD3 signature as an echo to how CHD7 micro-phenotypes mirror CHD3 macro-phenotypes. These findings underline the potential of episignatures to enrich the understanding of phenotype heterogeneity within syndromes. Decrypting the fine-scale structure of signatures could help anticipate the severity of variants and improve patient medical care. To convert episignatures from binary biomarkers to enriched informative tools, larger sample sizes and the rigorous collection of phenotypes are required.

K-mer-based-genome-wide association studies of the gut microbiome

Malak Raphaël[1], Frouin Arthur[1,2], Henches Léo[3], Auvergne Antoine[1], Boetto Christophe[1], Chikhi Rayan[4], Aschard Hugues[1,5]

1 - Génétique Statistique - Statistical Genetics (France), 2 - Department of Computational Biology, Institut Pasteur (France), 3 - Génétique Statistique (France), 4 - Algorithme pour les Séquences Biologique (France), 5 - Harvard T.H. Chan School of Public Health (United States)

Genome-wide Association Studies (GWAS) have been central in studying the genetics of human phenotypes, and there is now growing interest in implementing GWAS-like approaches to assess the role of metagenome on human health. Previous works, focusing on GWAS of a single bacteria proposed as genetic variants k-mers, which are DNA-words of length k that can capture SNPs, insertions/deletions events, and presence/absence of genes. Here, we investigate the inference of a k-mer abundance matrix from gut microbiome metagenome shotgun sequencing, and the potential to conduct a taxonomy-free k-mer-based GWAS. We derived and quantify k-mers from the gut microbiome sequences of healthy participants in the Milieu Interieur cohort using kmtricks. We then reconstruct the gut taxonomic profiles of the cohort with Blast and compare it to the state-of-the-art taxonomic classifiers MetaPhlAn4 and Kraken2 to enhance the relevance of our approach. Finally, we replicated previous results by implementing GWAS. We build a 31-mer abundance matrix using microbiome data from N=938 individuals. After solving computational issues, about 97 million genetic variants remain. The taxonomic profile based on the k-mers using Blast shows a strong positive correlation with MetaPhlAn4's and Kraken2's. Furthermore, preliminary GWAS on Age, Sex and BMI replicate signals from species level association studies. K-mer abundance analysis tends to capture species abundance analysis, showing the suitability of our hypothesis. Our analyses show that informative k-mers can be derived from gut metagenome in large human cohorts, providing a mean toward microbiome GWAS without taxonomy reconstruction and more complex genetic variant construction (unitigs).

Integrating Multi-Omics and Machine Learning through Polygenic Risk Scores in Middle Eastern Populations for Cardiometabolic Traits

Shaar Abdullah[1], Mohamed Elshrif[1], Ullah Ehsan[1], Nemer Georges[2], Bashir Mohammed[3], Saad Mohamad[1]

1 - Qatar Computing Research Institute, Hamad Bin Khalifa University (Qatar), 2 - College of Health and Life Sciences, Hamad Bin Khalifa University (Qatar), 3 - Hamad Medical Corporation (Qatar)

Background: Cardiometabolic traits like Type 2 Diabetes (T2D) and lipids share complex interrelated pathophysiology. Polygenic risk scores (PRSs) have been developed mainly in European cohorts, often focusing on single traits and lacking multi-omics integration. Here, we evaluated the use of machine learning (ML) to integrate multi-omics data through PRS, and explored the importance of ensemble learning to develop multi-trait PRS in the Middle East.

Methods: Qatar Precision Health Institute dataset was used (13,994 whole genome sequence individuals, 2,906 with metabolomics and 2,845 with proteomics data). We studied 22 cardiometabolic traits (e.g., T2D, Hypertension, LDL-C, etc.). Local PRSs were developed, and 1,072 existing PRSs were evaluated. PRSs were combined linearly (ensemblePRS) or using ML models: support vector machines (SVM), random forests (RF), and xgboost.

Results: 948 of 1,072 PRS were associated with at least one trait ($P< 4.66x10-5$). For example, PGS000020 was the best for T2D (OR=1.75, CI=(1.54-1.99); AUC=0.60, CI=(0.58- 0.63); $P=5.14\times10-18$) and PGS003871 was the best for LDL-C (R2=0.12, CI=(0.10-0.14); $P=2.91\times10-137$). Local PRS slightly improved performance for 9 of 22 traits. EnsemblePRSs outperformed single PRSs (AUC=0.62 vs. 0.60 for T2D). Xgboost surpassed RF and SVM (AUC=0.64, 0.60, 0.61 for T2D, respectively). Multi-trait PRSs had an advan- tage over trait-specific PRSs, especially for lipids. Combining ensemblePRSs and xgboost predictions showed the best performance. Finally, adding omic scores showed marginal improvement for hypertension only.

Conclusion: Public PRSs transfer well to Middle Easterns for many cardiometabolic traits. Multi-trait and trait-specific PRSs combined with ML tools had the best performance

Characterisation of diverse global ancestries within participants of the UK Biobank

Pantring Fiona[1,2,3], Cavalleri Gianpiero L.[1,2,3], Gilbert Edmund[1,2]

1 - Royal College of Surgeons in Ireland (Ireland), 2 - The FutureNeuro Research Centre (Ireland), 3 - The SFI Centre for Research Training in Genomics Data Science (Ireland)

The UK Biobank (UKB) is a large dataset containing in-depth phenotype and genotype data of nearly 500,000 UK-based participants. Studies leveraging the UKB typically focus on a subset of participants with homogenous European ancestry according to self-identification and genotype-based principal component analysis. Here, we comprehensively characterise the remaining 78,573 UKB participants with diverse ancestries using population genetic approaches - identifying communities that reflect the population history of the UK. We developed a novel approach to characterise diverse ancestries in UKB by assigning individuals to primary continental-level ancestry clusters, and then fine-scale ancestry communities within those clusters. To determine continental-level ancestries, the machine learning algorithm XGBoost was trained using ADMIXTURE data from the 1000 Genomes, Human Genome Diversity and Simons Genome Diversity Projects and applied on ADMIXTURE data from the UKB to assign each individual to one of eight clusters. These continental clusters were further divided into fine-scale communities by applying Leiden community detection to a network of Identity-By-Descent sharing. We found that the UKB is a repository of diverse ancestries primarily of European-, African-, and South Asian-like descent. Within the eight continental ancestry clusters, over 250 fine-scale communities were detected. Whilst these ancestry communities capture worldwide diversity, they primarily reflect the demographic history of Great Britain and its Commonwealth in the 20th century. Therefore, these communities are not necessarily represented in other datasets and represent a critical resource for equitable research in Britain today as well as facilitating the detection of novel rare functional variation in otherwise understudied genetic communities.

Improved ancestry and admixture detection using principal component analysis of genetic data

Privé Florian[1]

1 - Aarhus University (Denmark)

The rapid expansion of genetic data from large-scale biobanks and genomic studies presents unprecedented opportunities to investigate the genetic basis of complex traits and diseases across diverse populations. While many national biobanks predominantly include individuals of similar genetic ancestry, they often also contain participants from diverse ancestries, enabling cross-population analyses. However, accurately identifying and characterizing genetic ancestry is challenging, especially in datasets where subtle population structure is obscured by the overrepresentation of one ancestry group. I present a novel method for ancestry and admixture detection that leverages principal component analysis (PCA) to enhance the separation of closely related ancestry groups. This approach clusters individuals into distinct ancestry groups while accommodating admixed individuals. To improve resolution, overrepresented groups can be subsampled, mitigating PCA distortion and allowing finer distinctions among ancestry groups. A subsequent application of this method within the refined PCA space enables further differentiation of closely related groups and uncovers detailed patterns of genetic structure. This method offers a robust framework for characterizing genetic diversity in biobanks and overcoming challenges posed by uneven ancestry representation. By enhancing the detection of subtle population structure, it advances genetic research and supports more equitable, precise analyses of genetic risk across ancestries. These insights are crucial for realizing the full potential of biobanks to deepen our understanding of the genetic underpinnings of human health and disease on a global scale.

Comparing the performance of clustering methods to understand fine-scale genetic structure using simulated data

Guivarch Mael[1], Herzig Anthony[1], Saint Pierre Aude[1], Génin Emmanuelle[1,2]

1 - Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest (France), 2 - CHU Brest, Brest, France (France)

With the exponential growth of genomic datasets across large geographic regions, understanding fine-scale population structure is essential for many applications in population genetics. Clustering algorithms are widely used to identify such structures by grouping individuals according to genetic similarities. However, choosing the most appropriate method can be particularly challenging. Additionally, determining the optimal number of clusters and resolving conflicting results across iterations adds another layer of complexity. We perform a comparative analysis of clustering methodologies, focusing on model-based algorithms specific to population genetics (FineSTRUCTURE), or contextless ones (Mclust). We also investigate non-probabilistic approaches using Leiden and kmeans algorithms. Our study highlights the impact of data preprocessing and cluster validation on resulting partitions. All clustering results were evaluated on spatially stratified and subsampled populations to approximate real-world conditions in genetics studies. Enhancing previous studies that focused on genetic structures at a large geographic scale (mostly between continents), we simulate local genetic data following spatial demographic scenarios. Specifically, we generate 27000 individuals across a 36-deme grid, allowing migration between adjacent demes. We fine-tune migration rates and coalescent times to reflect fine-scale genetic structures. Deme memberships provide a gold-standard partition used to evaluate clustering algorithm accuracy. In conclusion, our study provides guidance for interpreting fine-scale population structure and selecting suitable clustering algorithms. We highlight the performance of FineSTRUCTURE and Mclust, emphasizing the trade-off between computational time and accuracy. We confirm previous findings that haplotypic data outperform genotypes in clustering accuracy and underscore the importance of spatial subsampling in cluster detection.

Multi-ancestry fine-mapping accounting for ancestral and environmental heterogeneity improves resolution

Wang Siru[1], Ojewunmi Oyesola[2], Pirie Fraser[3], Motala Ayesha[3], Ramsay Michele[4], Morris Andrew[5], Fatumo Segun[2,6], Chikowore Tinashe[7,8,9], Asimit Jennifer[1]

1 - MRC Biostatistics Unit, Univerisity of Cambridge (United Kingdom), 2 - Precision Healthcare University Research Institute, Queen Mary University of London (United Kingdom), 3 - Department of Diabetes and Endocrinology, School of Clinical Medicine, University of KwaZulu-Natal (South Africa), 4 - Sydney Brenner Institute for Molecular Bioscience, Facultyof Health Sciences,University of the Witwatersrand (South Africa), 5 - Centre for Genetics and GenomicsVersus Arthritis, Centre for Musculoskeletal Research, The University of Manchester (United Kingdom), 6 - The African Computational Genomic (TACG) Research Group, MRC/UVRI and LSHTM, (Uganda), 7 - MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of the Witwatersrand (South Africa), 8 - Channing Division of Network Medicine, Brigham and Women's Hospital (United States), 9 - Harvard Medical School (United States)

Amongst diverse population groups, allelic heterogeneity is likely due to differences in genetic ancestry and environmental exposures. This heterogeneity impacts the power to detect genetic associations, and refinement of sets of potential causal variants underlying genetic associations, through statistical fine-mapping. Meta-regression of multi-ethnic genetic association (MR-MEGA) adjusts for and assesses heterogeneity due to genetic ancestry and the recently developed environment-adjusted MR-MEGA accounts for environmental exposures alongside genetic ancestry. In this work, we developed novel multi-ancestry fine-mapping methods, (env-)MR-MEGAfm, which allow for multiple causal variants in a genomic region. Employing a stepwise selection procedure, (env-)MR-MEGAfm integrates approximate conditional analyses with (env-)MR-MEGA to construct credible sets of potential causal variants. Both methods use genome-wide association study summary statistics and account for differing linkage disequilibrium (LD) from multiple cohorts, and env-MR-MEGAfm also accounts for summary-level environmental covariates. Additionally, (env-)MR-MEGAfm may be implemented with out-of-sample LD as a practical alternative when in-sample LD is unavailable. Through simulation studies, we showed that (env-)MR-MEGAfm had significant improvements in coverage, resolution and prioritization over the current multi-ancestry approaches. We also highlight env-MR-MEGAfm had improved SNP prioritisation over MR-MEGAfm in fine-mapping genetic associations with low-density lipoprotein cholesterol measured in twelve sex-stratified African cohorts comprising 19,000 individuals. In summary, (env-)MR-MEGAfm ac-

counts for cohort-level differences in genetic ancestry and environmental factors (for env-MR-MEGAfm) and allow for variants to be present in only a subset of cohorts. Finally, these methods only require summary-level data and allow for any number of cohorts, making them useful tools in consortia efforts.

Genetic study of multifactorial diseases: an abyssal drift

Clerget-Darpoux Françoise[1]

1 - INSERM (France)

The use of Polygenic Risk Scores (PRS) to predict the risk of a multifactorial disease was initiated by a 2007 founding paper. Since then, there has been a vertiginous explosion in the number of studies using these scores... and thereby accepting the hypotheses on which they are set: 1) for the disease under study, there is a Polygenic Additive Liability (PAL) explaining both its prevalence and familial recurrences, and 2) each genetic variant of this liability can be detected by GWAS between affected and unaffected individuals. Also based on the PAL model, heritability was brought back into focus with it. Very early on, in particular at the 2015 Brest EMGM, some challenged the validity of PRS and heritability estimates with decisive arguments. We have to admit that this has not slowed down their use. We think more important than ever to recall here the self-sufficient argument given by Falconer himself when he introduced the PAL model for non-monogenic diseases. The model cannot be applied when there are sub-groups of patients with different familial recurrences. This is the case for most – if not all – multifactorial diseases, which even often include monogenic subgroups. Despite differences between subgroups in individual or familial risk of being affected, the classification of individuals is carried out under the assumption of a uniform mode of transmission for all. Meaningless PRS are now provided as risk factors to clinicians, used by companies for embryo selection and justify this absurd proposal of "performing germline polygenic genome editing".

Posters

Exploration of non-coding and structural variations in early-onset Alzheimer disease patients: contribution of PacBio HiFi long-read sequencing

Abani Fatima-Zahra[1], Derambure Céline[1], Charbonnier Françoise[2], Vezain Myriam[1], Rousseau Stéphane[3], Schramm Catherine[1], Quenez Olivier[3], Nicolas Gaël[3]

1 - Univ Rouen Normandie, Normandie Univ, Inserm U1245, F-76000 Rouen, France. (France), 2 - Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, Department of Genetics, F-76000 Rouen, France. (France), 3 - Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, Department of Genetics and CNRMAJ, F-76000 Rouen, France. (France)

Early-Onset Alzheimer Disease (EOAD, onset $<$ 65 years), has a monogenic determin- ism in $<$ 15% of cases. For non-carriers of a pathogenic variant, this disease results from a combination of genetic and environmental factors. Among the genetic risk factors, genome wide association studies (GWAS) identified 80 loci associated with AD. Additionally, short- read exome and genome sequencing identified rare coding variants directly associated with a moderate to high risk of EOAD. Non-coding variants, repeats, and structural variations (SVs) remain under-explored due to detection challenges. We propose to explore these variations using PacBio's high-fidelity long-read DNA sequenc- ing, genome-wide, in an EOAD-cases/controls cohort with extreme phenotype sampling, to identify new genetic risk factors. We developed a bioinformatic analysis pipeline in order to identify all possible variations types (SVs, repeats, transposable elements). Our strategy focuses first on GWAS' loci to identify either the association-mediating signal or an independent signal at the same locus. Until now, 80 EOAD patients have been sequenced . We reached a mean coverage of 30x and a N50 of 30kb offering a high precision to identify variations. The results obtained in terms of numbers and types of variations are in line with the expectations of the literature. Comparison with previously sequenced exome data shows excellent concordance in coding regions. We will present the first results of rare variants in GWAS-defined loci. These preliminary results highlight the value of PacBio HiFi long-read sequencing in gener- ating sufficiently long reads to enable the detection of complex and non-coding variations, thereby enhancing our understanding of EOAD genetics.

Cross-methods GWAS summary statistics deconvolution

Aissa Sohane[1], Henches Léo[1], Julienne Hanna[1], Tern Courtney[2], Kalra Sean[2], Cho Michael[2,3,4], Aschard Hugues[1,5]

1 - Génétique Statistique - Statistical Genetics (France), 2 - Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital (United States), 3 - Harvard Medical School (United States), 4 - Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital (United States), 5 - Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health (United States)

Genome-wide association studies (GWASs) have identified thousands of genetic variants associated with quantitative traits and diseases, shedding light on pervasive pleiotropy across the genome. This finding has impelled the development of numerous methods for the deconvolution of complex pleiotropic associations and the inference of potential shared biological pathways across outcomes. Those methods vary broadly in their modelling, their implementation, their input data, and their scalability. As each of them rely on specific assumptions on the data and the multitrait genetic structure, no approach is expected to be universally better, and their potential concordance and discrepancies is undetermined. Here, we aim at comparing the performances of multiple approaches across a range of real data, and examine their joint informativeness for characterizing genetic structure underlying multiple traits. We considered an extensive class of methods, including i) descriptive approaches that examine genetic correlation at the genome-wide (e.g. LDSC), region-based (e.g. SUPERGNOVA) and single variants level ; ii) matrix factorization techniques that are typically applied to the complete genome-wide summary statistics (e.g. DEGAS, GLEANR, FactorGo) ; and iii) GWAS hits clustering (e.g. bNMF, MGMM, k-medoids). We applied all methods to multiple sets of outcomes pulled from a total of 100 GWAS summary statistics and covering molecular traits, biomarkers, and common diseases. We report differences and agreement between them across sets and examine solutions to merge their results into a single comprehensive framework.

Compression for Human Short-Read Sequence Data: An Empirical Comparison

Betschart Raphael O.[1,2], Sandberg Felicia[1], Blankenberg Stefan[1,3,4,5], Zoche Martin[6], Zeller Tanja[2,7], Ziegler Andreas[1,3,4,8]

1 - Cardio-CARE (Switzerland), 2 - Institute of Cardiogenetics, University of Lübeck (Germany), 3 - Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg (Germany), 4 - Centre for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg (Germany), 5 - German Center for Cardiovascular Research, Partner Site Hamburg/Kiel/Lübeck, Hamburg (Germany), 6 - Institute of Pathology and Molecular Pathology, University Hospital Zurich, Zurich (Switzerland), 7 - German Center for Cardiovascular Research, Partner Site Hamburg/Kiel/Lübeck, Lübeck (Germany), 8 - School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal (Pietermaritzburg South Africa)

Efficient data compression technologies are crucial to reduce the cost of long-term storage and file transfer in whole genome sequencing studies. This study compared the five compression tools DRAGEN ORA 4.3.4 (ORA), Genozip 15.0.62, repaq 0.3.0, Samtools 1.20, and SPRING 1.1.1 using genome-in-a-bottle samples that were sequenced 82 times on an Illumina NovaSeq 6000, with an average coverage of 35x. All tools provided lossless compression. ORA and Genozip achieved compression ratios of approximately 1:6 when compressing fastq.gz. repaq and SPRING had lower compression ratios of 1:2 and 1:4, respectively. repaq and SPRING took longer for both compression and decompression than ORA and Genozip. Genozip had an approximately 1:4 higher compression ratio for BAM files than SAMtools. However, the BAM compression of Samtools produces CRAM files, which are compatible with many software packages. ORA, repaq, and SPRING are limited to compressing fastq.gz files, while Genozip supports various file formats. Although Genozip requires an annual license, its source code is freely available, ensuring sustainability. In conclusion, short-read sequence data can be efficiently compressed. Commercial tools offer higher compression ratios than freely available software.

Deep Mendelian Randomization: explaining causality between traits at genome-wide scale

Favre-Moiron Mario[1], Verbanck Marie[1], Nouira Asma[1]

1 - Cancer et génome: Bioinformatique, biostatistiques et épidémiologie d'un système complexe (France)

Mendelian Randomization (MR) is a method that infers the causality between risk factors and diseases using genetic variants as instrumental variables. It has the potential to mimic drug target effects observed in clinical trials, paving the way for new therapeutic target discovery. However, MR faces biases such as pleiotropy, where a single variant influences multiple traits. To address these limitations, we propose an innovative approach that leverages artificial intelligence with the aim to 1) include a larger number of exposures and variants 2) incorporate a greater variability of omics data, and 3) integrate these data with a Double Machine Learning pipeline. We expect that this strategy will allow us to take advantage of the prediction capacities of ML algorithms to process the large amount of data in order to disentangle the pleiotropic effects of variants and therefore provide more accurate causal effect estimators. Our method is currently being tested in extensive simulation scenarios and will subsequently be applied to uncover intricate relationships between the immune system and cancer. The primary data in our pipeline are GWAS data, which are the basis of Mendelian Randomization analysis. A particular focus is placed on protein quantitative trait loci (pQTL) and expression quantitative trait loci (eQTL) data, given their significant potential for discovering therapeutic targets.

Body mass index, IGF1-related polymorphisms and risk of familial breast cancer in women with no BRCA1 or BRCA2 pathogenic variant

Fritsch Humblet Barbara[1,2,3,4,5], Jiao Yue[2,3,4,6], Eon-Marchais Séverine[2,3,4,6], Dondon Marie-Gabrielle[2,3,4,6], Le Gal Dorothée[2,3,4,6], Beauvallet Juana[2,3,4,6], Stoppa-Lyonnet Dominique[7,8,9], Andrieu Nadine[2,3,4,6], Lesueur Fabienne[2,3,4,6]

1 - Inserm, U1331 (France), 2 - Institut Curie (France), 3 - PSL-Research University (France), 4 - MinesParis-Tech (France), 5 - Université Paris-Saclay (France), 6 - Inserm, U1331 (France), 7 - Service de génétique, Institut Curie (France), 8 - Université Paris-Cité (France), 9 - Inserm, U830 (France)

Background. Body mass index (BMI) and some single-nucleotide polymorphisms (SNPs) associated with IGF1 metabolism are risk factors for breast cancer (BC) in the general population. No investigation has been performed so far in women presenting a familial predisposition to BC, except in women carrying a pathogenic variant in BRCA1 or BRCA2 (BRCA1/2). Therefore, we investigated the effect of BMI and of IGF1-related SNPs in high-risk women with no BRCA1/2 pathogenic variant. Methods. We conducted a case-control study in the French study GENESIS (1556 cases and 1546 controls). We assessed association between 639,470 SNPs associated with circulating IGF1 level or located in genes of KEGG pathways involving IGF1 and BC using logistic regression models. Results. A reduced risk of estrogen receptor (ER)-positive tumors was observed for overweight premenopausal women (OR=0.51, 95%CI:0.33-0.79). None of the SNPs was associated with BC except SNP rs117292219 located in STAT5A and associated with a reduced risk of ER-negative tumors (OR=0.41, 95%CI: 0.26-0.66). No interaction between BMI and any of the analyzed SNPs was observed. Conclusions. Our findings on BMI effect were consistent with that reported in the general population and in women carrying a BRCA1/2 pathogenic variant but we found few associations between IGF1-related SNPs and familial BC, even if variants at STAT5A locus warrant further investigation. Impact. This large case-control study does not support a major role of the genetic variability of IGF1 metabolism in familial BC risk in women with no BRCA1/2 pathogenic variant.

Fine-scale pharmacogenetic diversity in Europe : the example of France

Gros La Faige Marc[1], POPGEN Study Group[2], Génin Emmanuelle[1,3], Herzig Anthony[1]

1 - Inserm, Univ Brest, EFS, UMR 1078, GGB, F-29200 Brest (France), 2 - Inserm, Brest (France), 3 - CHU Brest, F-29200 Brest (France)

Pharmacogenetics is the study of genetic variants responsible for variable response to medication. These variants can explain alternate drug responses and understanding their effects thus represents a key public health issue. Previous studies have shown that some of these variants have frequencies that are stratified across human populations but little is known about their distribution at fine geographic scale within a country such as France. To study the diversity of pharmacogenes of interest in different regions of France, we used SNP-chip genotyping data on 9598 French individuals and associated spatial co-ordinates derived from the birthplaces of their ancestors collected as part of the POPGEN project. We derived different statistics commonly used in population genetics to identify pharmacogenetic variants with a heterogeneous frequency distribution and detected variant stratifications, such as gradients from north to south or east to west. We also found clusters of variants within specific sub-populations. We studied how these patterns could be explained by selective constraints by comparing their gene constraint metrics against those of other genes with similar sizes and we observed that certain pharmacogenes are significantly less constrained, which may explain their observed high levels of genotypes and phenotypes diversity. Overall, we identified some important pharmacogenes, like CYP2D6 or ABCG2, with fine-scale geographic specificities that have phenotype consequences for drug with prescribing recommendations. Exploring genetic diversity in pharmacogenes at finer geographic scales than previously done will improve our understanding of drug-gene interactions, while also informing potential benefits of personalized treatment based on pharmacogenetic variant data.

Causal relationships between gut microbiome and age-related traits

Grosso Federica[1], Zanetti Daniela[1], Sanna Serena[1]

1 - Istituto di Ricerca Genetica e Biomedica (Italy)

In the past 20 years, the involvement of gut microbiome in human health has received particular attention, but its contribution to age-related diseases remains unclear. We performed a comprehensive investigation of 4,033 potential causal relationships between 37 traits representing gut microbiome composition and function and 109 age-related phenotypes, through the causal inference method two-sample Mendelian randomization (MR). We used inverse variance weighted (IVW), as main method to assess significance and weighted median and MR-PRESSO methods as sensitivity analyses. For IVW we employed a false discovery rate (FDR) correction to control for multiple testing within each outcome. Pleiotropy (egger intercept), heterogeneity (Cochran's Q statistics), leave-one-out and reverse MR were considered as sensitivity analyses. Finally, we performed replication with independent datasets to ensure strength of results, along with a post-hoc power analysis. Five causal relationships remained significant after multiple testing correction and sensitivity analysis, specifically between two taxa of Coriobacteriales and the risk of developing age-related macular degeneration (pFDR=0.047), species Bifidobacterium adolescentis and levels of TNFSF12 protein in plasma (pFDR=0.0003), and the lactose-galactose degradation microbial I pathway and levels of IL-15Rα (pFDR=0.03) and TRAIL (pFDR=0.006) proteins in plasma. The causal relationship between the microbial pathway and TRAIL protein levels was further confirmed using independent data (p=0.01). These results support the role of gut microbiome in regulating the inflammatory circuit and the importance of rigorous methodologies and replication to establish causality in MR studies. However, future studies are needed to investigate the underlying biological mechanisms.

Towards new therapeutic strategies for protein x-deficient Triple-Negative Breast Cancers

Guichaoua Gwenn[1,2,3], Rodrigues-Ferreira Sylvie[4,5,6], Azencott Chloé-Agathe[1,2,3], Nahmias Clara[4,5], Stoven Veronique[1,2]

1 - Center for Computational Biology, Mines Paris, PSL Research University, 75006 Paris (France), 2 - Institut Curie, PSL Research University, 75428 Paris (France), 3 - INSERM U1331, 75005 Paris (France), 4 - Gustave Roussy Cancer Center, F-94800 Villejuif (France), 5 - INSERM U981, Université Paris-Saclay, F-94800 Villejuif (France), 6 - Inovarion, F-75005 Paris (France)

Triple-negative breast cancers (TNBCs) represent a clinically challenging subtype due to their aggressive nature and poor response to standard treatments. These tumours lack hormone receptors and HER2, limiting the effectiveness of targeted therapies. Among them, x-deficient TNBCs are associated with an even poorer prognosis and increased resistance to chemotherapy. Our goal is to identify novel therapeutic targets and predict active small molecules for these tumours.

To identify new therapeutic opportunities, we leveraged a combination of computational biology and experimental data. RNA-seq transcriptomic profiles from public cohorts of TNBC patients were integrated with in vitro data from engineered cell lines, allowing us to label samples according to their x status. Using differential expression and pathway enrichment analyses, we identified deregulated biological pathways and a set of transcription factors (TFs) differentially activated in x-deficient tumours. A key TF showed promise, although it was difficult to target directly.

To address this, we developed a chemogenomic pipeline combining large-scale data curation and machine learning. We introduced LCIdb, a large dataset of drug-target interactions covering a broader chemical and proteomic space than existing benchmarks. We then designed Komet, a scalable machine-learning algorithm optimized to predict drug-target interactions using LCIdb. Komet outperforms state-of-the-art methods in terms of speed and accuracy.

This combined approach, combining transcriptomic analysis and chemogenomics machine learning, helps develop personalized treatments for x-deficient TNBC.

GenEFCCSS: A resource for investigating genetic predispositions in in childhood cancers

Hamzaoui Ons[1,2,3], Bacq Delphine[4], Fresquet Marion[1,2,3], Zidane Monia[1,2,3], Hoarau Pauline[1,2], Deloger Marc[1], Boland-Augé Anne[4], Herzig Anthony[5], Haddy Nadia[1,2,3], Rubino Carole[1,2,3], Guerrini Lea[1,3], Dufour Christelle[1,3], Minard Véronique[1,3], Pacqument Hélène[6], Bourdeaut Franck[6], Winter Sarah[6], Adam-De-Beaumais Tiphaine[1], Lenez Laura[1], El-Fayech Chiraz[1], Blanché Hélène[7], Deleuze Jean-François[4], Génin Emmanuelle[5], Fresneau Brice[1,2,3], De Vathaire Florent[1,2,3]

1 - Gustave Roussy Institute, Villejuif (France), 2 - Inserm U1018, Villejuif (France), 3 - University Paris Saclay, Villejuif (France), 4 - CNRGH, Evry (France), 5 - Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest (France), 6 - Curie Institute, Paris (France), 7 - CEPH, Evry (France)

Introduction. Genetics is expected to play a significant role in the development of childhood cancer. This study examines germline variants associated with predisposition to primary and secondary neoplasms, as well as other related events, in children from the Extended FCCSS cohort (GenEFCCSS).
Material and Methods. GenEFCCSS cohort includes 8471 patients, of whom 2673 underwent whole-genome sequencing (WGS) using the NovaSeq X+ Illumina platform. Sequencing achieved an average depth of 30X. Baseline clinical characteristics were retrieved from hospital records. The sex ratio is approximately 1:1, with a median age at diagnosis of 6 years (IQR 2–12) and a median follow-up time of 28 years (IQR 19–36). The most common primary neoplasms include renal tumors (15%), neuroblastoma (12%), and Hodgkin's lymphoma (8%). Approximately 500 patients developed a second malignant neoplasm.
Raw FASTQ files received from CNRGH were preprocessed for quality filtering using Fastp. Subsequent analyses were performed using the nf-core/sarek pipeline for germline variant detection. Alignment was conducted with BWA-MEM2, and duplicate marking was performed with GATK MarkDuplicates. Variants were called using HaplotypeCaller, followed by joint caliing. ClinVar was used for annotation to identify pathogenic and likely pathogenic variants.
Results The distribution of pathogenic and likely pathogenic variants (exonic, exon-intron junctions, and deep intronic) in a curated list of 129 cancer predisposition genes, used in daily oncogenetics practice within the French Genomic Medicine Initiative, will be presented. A comparison of clinical characteristics between mutation carriers and non-carriers, particularly regarding the occurrence of second malignancies, will also be described.

Transitioning to DNAnexus

Henches Léo[1], Aschard Hugues[1,2], Benoit Gloria[1]

1 - Génétique Statistique (France), 2 - Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA (United States)

Recent years have witnessed significant adoption of cloud computing technologies, particularly among leading genetic research institutions, which are transitioning from traditional data sharing paradigms toward cloud-based architectures for data storage and computational processing. However, research laboratories face considerable challenges in adapting to this paradigm shift, as the additional layer of technical complexity impedes analysis workflows and necessitates careful resource management to maintain cost-effective operations in cloud environments. One strategy to solve this challenge is to focus on producing quality summary statistics in the cloud environment and download the results for local downstream analyses. As such, knowledge and experience with running Genome-Wide Association Study (GWAS) in the cloud is a valuable asset. In this work, we will compare multiple ways to run GWAS on DNAnexus, the new cloud infrastructure of the UK Biobank. Criteria such as computational cost, complexity of the tools and scalability will be measured. Additionally, we will discuss the drawbacks and advantages of the different datasets and data formats available on the platform.

Genome of Europe pilot studies: stepping stones towards a pan-European reference database

Herzig Anthony[1], Vicente Astrid[2,3], Martiniano Hugo[2,3], Rayner N. William[4], Genin Emmanuelle[1,5], Ray-Jones Helen[6], Van Rooij Jeroen[6], Uitterlinden André[6], Consortium The Genome Of Europe[7]

1 - Univ Brest, INSERM, EFS, UMR 1078, GGB, F-29200 Brest (France), 2 - Instituto Nacional de Saúde Doutor Ricardo Jorge, Av Padre Cruz, 649-016 Lisbon Portugal (Portugal), 3 - Biosystems and Integrative Sciences Institute (BioISI), Faculdade de Ciências da Universidade de Lisboa, Lisboa Portugal (Portugal), 4 - Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg (Germany), 5 - CHU Brest (France), 6 - Laboratory for Population Genomics, Erasmus MC, Rotterdam, The Netherlands (Netherlands), 7 - Genome of Europe Consortium, Digital Europe Program, EU (Netherlands)

Background: Within the 1 million genomes (1+MG) initiative, a digital Europe (DEP)-funded project called the Genome of Europe (GoE) will collect whole-genome sequencing data from >100,000 European citizens across 51 contributing partners from 29 countries. GoE aims to improve and disseminate understanding of the genetic dimension to public health in Europe, through interactions with the European Rare Disease Research Alliance (ERDERA), the European Genomic Data Infrastructure (GDI), and the upcoming Personalised Cancer Medicine (PCM) Joint Action. Methods: The GoE analyses will operate in a federated environment established by the GDI project, which raises technological and methodological challenges. Here we provide progress updates on the six use cases of GoE. Results: To inform aggregation of individual-level data, we have begun pilot studies to determine methodology for describing population structure in independent datasets, through techniques such as federated principal-component analyses. Functionalities for variant frequency look-up tools have been evaluated, including assessment of the potential impact of heterogeneity in sequencing technology and bioinformatic pipelines. We also assess possible approaches for federated imputation algorithms, in collaborations with the Helmholtz Imputation Server, as well as perspectives for generating polygenic (risk) score distributions, calibrated for ancestry. GoE will place particular attention on clinically-relevant variant frequencies in actionable-disease and pharmacogenetic genes. Structural variation, potentially of high impact, will also be investigated in GoE via long-read sequencing for >5,000 individuals. Conclusion: Progress has begun to leverage the prospective >100,000 whole genome-sequences of GoE for public health genomics; with an outlined roadmap for alignment with other key European initiatives.

Effect of study sample size and composition on supervised admixture models

Herzig Anthony[1], Guivarch Maël[1], Gros La Faige Marc[1], Marenne Gaelle[1], POPGEN Study Group[2], Genin Emmanuelle[1,3]

1 - Inserm, Univ Brest, EFS, UMR 1078, GGB, F-29200 Brest (France), 2 - Inserm, Brest (France), 3 - CHU Brest, F-29200 Brest (France)

Unsupervised admixture, a prevalent analysis in exploratory population genetics, is known to be highly sensitive to sample size and composition. The goal is to model a given sample as issuing from K hypothetical populations (with K to be chosen by the analyses), with each individual being assigned a set of K admixture components representing the contributions of the K hypothetical population to their genome. The distribution across the whole study of these components describes the broad population structure in the dataset. When genomes from well-established reference populations are in-hand, supervised admixture can be performed and this is argued to be a more robust approach that is more straightforward to interpret. In this circumstance, the composition of the reference datasets is known to greatly impact the final results. Here however, we examine the role of the composition and size of the study sample during supervised admixture by analysing 9,598 individuals from the general population in France either all at once, in groups, or one-by-one against reference populations from the 1000G and HGDP. We show that by changing the study sample composition, quite different results may be obtained, showing that internal population structure within the study sample play an important role, even when the admixture analysis is supervised.

Genetic correlations between asthma subtypes, neuropsychiatric disorders and lifestyle factors

Huang Haibo[1], Vernet Raphaël[1], Linhard Christophe[1], Suzuki Yuka[2], Julienne Hanna[2], Bouzigon Emmanuelle[1]

1 - Inserm, UMRS-1124, Université Paris Cité, group of Genomic Epidemiology of Multifactorial diSeases, Paris (France), 2 - Institut Pasteur, Université Paris Cité, Department of Computational Biology, Paris (France)

Asthma patients suffer more frequently than the general population from anxiety and depression. However, the link between asthma and neuropsychiatric disorders is poorly understood. It could potentially be mediated by shared lifestyle (e.g. smoking) or genetic factors. To clarify the interplay between neuropsychiatric disorders and asthma, we computed the global and local genetic correlations between 24 asthma, neuropsychiatric and lifestyle phenotypes by using LDSC (linkage disequilibrium score regression) and LAVA (local analysis of [co]variant association) on full GWAS summary statistics. Different asthma subtypes displayed varying degrees of global genetic correlation with neuropsychiatric disorders: adult-onset asthma showed significant and strong global genetic correlations (rho$>$0.2, FDR p-value$<$ 0.001) with major depressive disorder (MDD), anxiety, post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorder (ADHD), but childhood-onset asthma did not. We observed different patterns of association between traits when computing the local genetic correlations: childhood-onset asthma and adult-onset asthma have as many nominally significant local genetic correlations (p-value$<$ 0.05) with MDD, anxiety and PTSD. However, the percentage of positive local genetic correlations between childhood-onset asthma and neuropsychiatric traits is lower than for adult-onset asthma and neuropsychiatric traits, resulting in different global genetic correlations. To further understand their specific genetic links, we will 1) use multi-trait analysis to identify groups of variants with similar multi-trait genetic effects across asthma subtypes, neuropsy- chiatric disorders and lifestyle factors; 2) conduct a functional analysis to identify pathways and cell types tied to these groups of variants.

Multi-Trait GWAS across 52 Infectious Diseases: Uncovering the Role of Common Variants in Immune Response

Julienne Hanna[1], Kerner Gaspard, Ramos Jhonatan, Patarlageanu Andreea, Quintana-Murci Lluis, Patin Etienne

1 - Institut Pasteur, Université Paris Cité, Department of Computational Biology, F-75015 Paris, France (France)

Response and susceptibility to infectious diseases vary from person to person, owing to a combination of genetic and non-genetic factors. Recent Genome Wide Association Studies (GWAS) on COVID-19 susceptibility have highlighted the contribution of common genetic variants to this variability. However, GWAS data for other infectious diseases often consist of studies with limited sample size. In this context, multi-trait GWAS may be helpful to capture genetic evidence scattered across underpowered studies. Here, we leveraged the Joint Analysis of Summary Statistics (JASS) pipeline to analyze 91 GWAS summary statistics obtained from the GWAS catalog, and 23 and me. We i) performed a meta-analysis to regroup summary statistics by trait, resulting in 52 infectious traits (e.g. COVID-19, pneumonia, chronic sinus infection, ...), ii) conducted a multi-trait GWAS including all traits, and ii) detected 69 novel associations. The nearest genes of significant variants mapped to immune pathways such as: antigen processing and presentation (p-val=4.42e-17); positive regulation of T-cell activation (p-val=1.63e-14); and positive regulation of immune response (p-val=5.27e-13). This compendium of signals across infectious traits provides an opportunity to understand how genetic variants shape the immune response across a range of immune stimuli. By calculating the genome-wide genetic correlations across the 52 traits, we observed overall diffuse positive correlations (median correlation = 0.17, and 92% of positive correlations). To study pleiotropic effects across infectious diseases, we will search for patterns of multi-trait genetic effects using a range of dimensionality reduction techniques such as FactorGo, and Genetic Factor Analysis.

A simple demonstration of a privacy-preserving de-centralised genotype imputation workflow

Letaillandier Alban[1], Picard-Druet David[1], Ludwig Thomas E.[1,2], Marenne Gaëlle[1], Herzig Anthony F.[1]

1 - INSERM UMR1078 (France), 2 - CHU BREST (France)

Recently, a number of studies have looked at the problem of privacy and data-sharing restrictions in the context of missing genotype imputation servers. This relates to the most typical imputation pipelines which involve a whole-genome sequenced haplotype reference panel being compared to genotyped study individuals (who have missing data to be imputed). Hence, involving two datasets from separate sources coming together in one informatic environment, where relatively complicated statistical models are applied; specifically, hidden Markov modelling. We embarked on a thought experiment to provide a potential privacy-preserving approach involving federating the different internal tasks within the statistical methods used for imputation. This idea is relevant considering there is currently motivation for federated analyses platforms in Europe for making combined inference across multiple genomic data resources. This allows for very simple manipulations to protect sensitive individual level data, which enable imputation algorithms to complete on simple plain-text files. We provide here an illustration of how such a federated imputation server could be put in place, along with associated code, including a simple implementation of the Li-Stephens haplotype mosaic model to achieve the imputation of missing genotypes. We name our general framework ANONYMP for anonymised imputation. A demonstration of the concept is given involving simulated data generated with msprime. We show that dividing different parts of the required calculations for statistical imputation between several sites is a valuable new avenue in the field of privacy-preserving imputation server development.

Contribution of incorrect statistical methods to the excess of false-positive results in Mendelian randomization analyses

Mckeigue Paul[1], Old Tim[1], Iakovliev Andrii[1], Erabadda Buddhiprabha[1], Colhoun Helen[1], Spiliopoulou Athina[1]

1 - University of Edinburgh (United Kingdom)

Two-sample Mendelian randomization (2SMR) is a method for detecting causal effects of exposure on outcome by testing for association of genotype-outcome effects with genotype-exposure effects. Early enthusiasm for this approach has given way to disenchantment, expressed in recent commentaries that have noted a "deluge" of published studies that report support for causality. To investigate reasons for this apparent excess of positive results, we sampled 40 published 2SMR studies. Of the 30 studies that reported support for causality, 27 used the weighted median test, 15 used an outlier-removal procedure (MR-PRESSO), and 4 used a profile likelihood method (MR-RAPS). In simulations from a null model based on plausible assumptions about the distribution of pleiotropic effects, all three of these methods showed inflation of the variance of the test statistic and the Type 1 error rate. With the weighted median test, the Type 1 error rate was inflated more than 100-fold. A test based on marginalizing over the direct effects to compute the likelihood of the causal effect parameter had the lowest Type 1 error rate and the lowest Type 2 error rate, as we would expect. We conclude that the proliferation of 2SMR studies reporting evidence of causality is at least partly attributable to widespread use of incorrect statistical methods that are implemented in the MR-Base platform. Used with care, and with measures to control confounding of associations between instrument-exposure and instrument-outcome effects, 2SMR can support systematic causal inference.

Improved polygenic risk scores for out-population individuals.

Möls Märt[1]

1 - Tartu Ülikool = University of Tartu [Estonie] (Estonia)

Using polygenic risk scores (PRS) for people of different ancestry or mixed origin usually require ancestry-specific recalibration of risk based on multi-ancestry genome-wide association studies (GWAS). In scenarios where only GWAS data from a native population is available, estimating risk for individuals from different ancestries requires innovative solutions. For instance, while a GWAS might exist for an Estonian population, suitable data for Polynesian or Polynesian-Estonian mixed ancestry may not be available, complicating risk assessments for individuals in such groups (out-population individuals). We propose a novel method to assess the risk for out-population individuals using existing GWAS results and linkage disequilibrium (LD) structure for native population. GWAS estimates for SNP effects are biased due to omitted variables. For example, the other causal SNPs (sometimes correlated with the SNP of interest) are not included in the GWAS model. However, one can mostly cancel out these biases during PRS calculation. Using only one SNP per haploblock for PRS or correcting the estimated effects based on between SNP correlation (LD) leads to an analysis where the biases are cancelled out from the final PRS score. However, if an individual comes from a population with different LD structure, these methods can mishandle the bias and this may lead to inappropriate estimation of the total risk. However, a more careful treatment of GWAS bias can lead to improved PRS with less bias for out-population individuals. Some approaches to minimize the PRS bias for out-population individuals will be discussed and compared.

Predicting CART-T Cell Survival in Non-Hodgkin Leukemia: A Mathematical Approach.

Monsalve Gabriel[1], Mohammadnezhad Leila[2], Lavigne Aurore[1], Poulain Alexandre[1], Mitra Suman[2], Dabo-Niang Sophie[3]

1 - Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, F-59000 Lille (France), 2 - Inserm UMR1277, CNRS UMR9020 – CANTHER, Université de Lille, Lille University Hospital, Lille (France), 3 - Univ. Lille, CNRS, Inria, UMR 8524 – Laboratoire Paul Painlevé, F-59000 Lille (France)

In this poster, we present a preliminary guideline for a project that aims to establish a mathematical framework for the early prediction of the survival of CD19-specific CAR-T cells in immunotherapy. This framework will model the interactions between CD19-specific CAR-T cells and Non-Hodgkin leukemia cells.

Our methodology is based on the integration of spatial omics information with single-cell RNA sequencing (scRNA-seq) data. We will discuss two main challenges and strategies to enhance the effectiveness and precision of the cellular complexity interactions: interpretable factor identification and the integration of spatial and transcriptional data.

For the identification of biomarkers associated with survival outcomes, we use the SPECTRA algorithm due to its ability to predict generative clusters with interpretable factors. We identify two ways to improve computational costs: a Bayesian approach in factor identification and the optimization of the Expectation Maximization algorithm.

The CILCAD study: clinical and genetic characterization of a multi-generational p.Arg1231Cys mutation carrying pedigree from a Cilento founder population

Nutile Teresa[1], Ruggiero Daniela[1], Cennamo Pasqualina[1], Pizza Vincenzo[2], Lebenberg Jessica[3,4], Lambert Louis[4], Pluntz Matthieu[5], Cipriano Lorenzo[6], Di Pietro Andrea[7], Trojano Luigi[8], Tournier-Lasserve Elisabeth[5,9,10], Chabriat Hugues[3,4,9], Perdry Hervè[11], Leutenegger Anne-Louise[5], Ciullo Marina[1]

1 - Institute of Genetics and Biophysics A. Buzzati-Traverso, CNR, Naples (Italy), 2 - Dep. of Emergency and Time Dependent Networks, Neurology Unit, S.Luca Hospital, Vallo Della Lucania (Italy), 3 - Inserm Université Paris Cité, FHU Neuro-Vasc 2030 (France), 4 - Inserm Institut du Cerveau,Paris (France), 5 - Inserm Université Paris Cité, NeuroDiderot, Inserm U1141, Paris (France), 6 - Dep. of Molecular Medicine and Medical Biotechnology, University Federico II, Naples (Italy), 7 - Dep. of Neurology and Stroke Unit, AORN Sant'Anna e San Sebastiano, Caserta (Italy), 8 - Dep. of Psychology, University of Campania Luigi Vanvitelli, Caserta (Italy), 9 - APHP, Translational Neurovascular Centre and CERVCO, Hôpital Lariboisière, Paris (France), 10 - APHP, Service de génétique moléculaire Neurovasculaire, Hôpital Saint-Louis, Paris (France), 11 - Université Paris Saclay, Inserm CESP,Villejuif (France)

Background/Objectives: Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL, OMIM 125310) is an adult-onset, inherited small vessel disease caused by NOTCH3 mutations and characterized by a high phenotype variability. We have studied a large and multi-generational sample of individuals (connected in a 17-generation pedigree) from a founder population of South Italy carrying the widespread p.Arg1231Cys mutation to assess the phenotype variability in this sample and to identify additional genetic factors that could modulate CADASIL phenotype. Methods: Clinical data were compared between p.Arg1231Cys mutation carriers (CILCAD) and the NON-Carriers from the same population; also, MRI features in CILCAD were compared to those detected in UK-population samples. NOTCH3-region haplotypes were analyzed to define the minimum haplotype shared by the CILCAD and identify novel genetic contributions to the disease phenotype. Results: We found an increase in neurological manifestations in the CILCAD compared to NON-Carriers from the same population; however, our data suggest a milder disease phenotype associated with the p.Arg1231Cys mutation in the CILCAD. Moreover, we found that CILCAD subjects have lower total cholesterol and LDL levels than the rest of the population. Indeed, NOTCH3-region haplotype sharing analysis revealed

that the haplotype carrying the p.Arg1231Cys mutation in CILCAD also carries lowering cholesterol alleles. Conclusions: We hypothesize that additional genetic and/or environmental factors could contribute to a milder phenotype observed in CILCAD. We also believe that this sample could represent a useful resource for identifying novel genetic and environmental factors implicated in the disease, contributing to an improved understanding of CADASIL pathogenesis.

From DNA to GPA: a genetically informed study of the male disadvantage in schools

Ofstad Sverre[1], Demange Perline[1], Cheesman Rosa[1], Qin Qi[1], Torvik Fartein[1,2], Ystrøm Eivind[1]

1 - Department of Psychology, University of Oslo (Norway), 2 - Norwegian Institute of Public Health (Norway)

Boys are underperforming in schools compared to girls. Norway has one of the largest GPA gender gaps among OECD countries, with differences being most pronounced in middle school. In this article, we use a gene-environment interaction framework to investigate how schools might affect the association between genetic predispositions and GPA and whether this differs per gender. We calculated polygenic indexes for cognitive and non-cognitive skills for a sample of 32.000 middle school students in genotyped family trios from the Norwegian mother, father, and child study (MoBa). First, we leveraged the well-defined grouping of students into schools by examining gene-environment interactions through a random intercept and random slope model. Secondly, the random slopes were substituted by a wide range of population-representative school-level measures based on socio-economic conditions, demographic traits, and socio-behavioral characteristics. We find that the relationship between GPA and non-cognitive PGI is less pronounced in school environments that are associated with higher GPAs. The association of cognitive skills PGI with GPA show very little variance across school environments. Hence, school environments conducive to learning can compensate for genetic predispositions for non-cognitive skills. We also find that boys are considerably more influenced by school environments than girls. Although the gene-environment interaction terms are similar for both genders, the larger environmental influence for boys suggests that genetic variation in response to school environments is also larger for boys.

Decoding algorithms for Hidden Markov Models and detection of homozygosity by descent

Pluntz Matthieu[1,2], Foulon Sidonie[1,2], Perdry Hervé[2], Leutenegger Anne-Louise[1]

1 - Inserm Université Paris Cité, NeuroDiderot, Inserm U1141, Paris (France), 2 - CESP Inserm U1018, Université Paris-Saclay, F-94807 Villejuif, France (France)

Hidden Markov models (HMM) are widely used across biostatistics to model a sequence of observations informed by a sequence of unobserved dependent events, the hidden states. Inference of the unobserved events is known as decoding. The Viterbi algorithm is an efficient decoding algorithm which finds the most probable path of hidden states conditional on the observations. Another approach, posterior decoding, is the path of the most likely states at each step according to their marginal conditional probabilities, computed by the Forward-Backward algorithm. It has a different focus from the Viterbi decoding: while the state at each step is maximally likely, the transitions between states, and therefore the global path, might be unlikely or impossible. The posterior-Viterbi is a decoding which makes sure the decoded path has no impossible transition but otherwise coincides with the posterior decoding and may therefore produce unlikely transitions. We propose a broader intermediate family of decodings between the posterior and Viterbi decodings, which allow to choose a balance between focusing on local or global likelihood of the decoded path. In a simulation study, we evaluate those decodings to detect homozygous-by-descent (HBD) DNA segments in inbred individuals, where HBD status at a locus is unobserved and might randomly differ from HBD status at a neighbouring locus due to the recombination that occurred since the common ancestor. We compare the results of the decodings with the ones from the observational approach "Runs of homozygosity" (ROH).

Familial hypercholesterolaemia: prevalence and discrepancy between genotype and phenotype – results from the population-based Hamburg City Health Study

Riccio Cristian[1], Arnold Natalie[2,3,4], Koliopanos Georgios[1], Link Vivian[1], Guo Linlin[2,3,4], Betschart Raphael[1], Zeller Tanja[5], Blankenberg Stefan[1,2,3,4], Ziegler Andreas[1,2,3,6], Twerenbold Raphael[2,3,4]

1 - Cardio-CARE, Medicine Campus Davos, Davos, Switzerland (Switzerland), 2 - Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany (Germany), 3 - German Center for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Hamburg, Germany (Germany), 4 - Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany (Germany), 5 - Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany (Germany), 6 - School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa (South Africa)

Background: Familial hypercholesterolaemia (FH) is one of the most common monogenic diseases, affecting genes involved in low-density lipoprotein (LDL) metabolism. In Germany, there is limited population-based data on the prevalence of genetically confirmed FH (genFH) and genotype-phenotype associations. This study aimed to address this gap using whole-genome sequencing data from the Hamburg City Health Study. Methods: Pathogenic mutations in five FH-associated genes were analysed in 7,373 participants aged 45–74 years. Genotypes were compared to LDL cholesterol (LDL-C) levels, adjusted for lipid-lowering medication. Severe hypercholesterolaemia was defined as LDL-C $\geq$ 190 mg/dL. Results: Heterozygous FH was identified in 23 individuals (0.31%; 1 in 321), all due to LDLR mutations. Median treatment-adjusted LDL-C levels were higher in genFH individuals (191 mg/dL [149–210]) than in those without genFH (128 mg/dL [105–153], $p < 0.001$). Severe hypercholesterolaemia was observed in 6.5% of participants, of whom only 2.3% (n = 11) had genFH. LDL-C levels were below 130 mg/dL in 9.1% of genFH individuals, below 160 mg/dL in 31.8%, and only 50% had severe hypercholesterolaemia. Screening using LDL-C thresholds of $\geq$ 190 mg/dL, $\geq$ 160 mg/dL, or $\geq$ 130 mg/dL yielded sensitivities of 50.0%, 68.2%, and 90.9%, respectively, with 43, 98, and 175 individuals requiring screening to identify one genFH case. Conclusion: The prevalence of genFH in this cohort was 0.31%, aligning with global estimates. Significant discrepancies between genFH and LDL-C levels highlight the limitations of genetic FH screening.

Assessment of the functionality and usability of open source rare variant analysis pipelines

Riccio Cristian[1,2], Jansen Max[1,2], Sandberg Felicia[1,2], Koliopanos Georgios[1,2], Link Vivian[1,2], Ziegler Andreas[1,2,3,4,5]

1 - Cardio-CARE (Switzerland), 2 - Swiss Institute of Bioinformatics (Switzerland), 3 - Center for Population Health Innovation (POINT) (Germany), 4 - University Center of Cardiovascular Science & Department of Cardiology (Germany), 5 - School of Mathematics, Statistics, and Computer Science (South Africa)

Sequencing of increasingly larger cohorts has revealed many rare variants, presenting an opportunity to further unravel the genetic basis of complex traits. Compared with common variants, rare variants are more complex to analyze. Specialized computational tools for these analyses should be both flexible and user-friendly. However, an overview of the available rare variant analysis pipelines and their functionalities is currently lacking. Here, we provide a systematic review of the currently available rare variant analysis pipelines. We searched MEDLINE and Google Scholar until November 27, 2023, and included open source rare variant pipelines that accepted genotype data from cohort and case-control studies and group variants into testing units. Eligible pipelines were assessed based on functionality and usability criteria. We identified 17 rare variant pipelines that collectively support various trait types, association tests, testing units, and variant weighting schemes. Currently, no single pipeline can handle all data types in a scalable and flexible manner. We recommend different tools to meet diverse analysis needs. STAARpipeline is suitable for newcomers and common applications owing to its built-in definitions for the testing units. REGENIE is highly scalable, actively maintained, regularly updated, and well-documented. Ravages is suitable for analyzing multinomial variables, and OrdinalGWAS is tailored for analyzing ordinal variables. Opportunities remain for developing a user-friendly pipeline that provides high degrees of flexibility and scalability. Such a pipeline would enable researchers to exploit the potential of rare variant analyses to uncover the genetic basis of complex traits.

Fine-scale genetic structure of Armenians: an unsuspected diversity

Saint Pierre Aude[1], Tashjian Georges[2], Babikyan David[3,4], Hovhanessyan Kristine[5], Herzig Anthony[1], Guivarch Mael[1], Gazarian Aram[6], Maguesyan Pascal[2], Kévorkian Raymond H.[7], Isaian Anna[8], Bolland Anne[9], Deleuze Jean-François[9], Lathrop Mark[10], Sarkisian Tamara[3,4], Genin Emmanuelle[1], Hovnanian Alain[11,12,13,14]

1 - Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest (France), 2 - International Union of Land and Culture Organisations, Paris (France), 3 - Center of Medical Genetics and Primary Health Care, Yerevan (Armenia), 4 - Department of Medical Genetics, Yerevan State Medical University, Yerevan (Armenia), 5 - University of Liege, Department of Human Genetics, Liège (Belgium), 6 - International Union of Land and Culture Organisations, Paris (France), 7 - University Paris 8 Saint-Denis, Paris (France), 8 - Pediatrics Center of Excellence; Children's Medical Center, Department of Pathology; University of Medical Sciences, Tehran (Iran), 9 - Laboratory of functional Genomics, CNG, Atomic Energy and Alternative Energies Commission, Evry (France), 10 - Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montreal (Canada), 11 - INSERM U1163, Paris (France), 12 - Laboratory of Genetic Skin Diseases, Paris (France), 13 - University Paris Cité (France), 14 - Department of Genomic Medicine of Rare Diseases, Paris (France)

Introduction: Armenians have a strong and distinct ethnic and cultural identity that unites them as an ethno-national group. While the complex demographic history and varied geography are expected to have shaped the genetic make-up of the current Armenian individuals, they were notably under-represented in genomic research until recently. To investigate the fine-scale population genetic substructure of Armenians we used genotypic data (574,393 SNPs, Illumina) of 320 unrelated individuals, representative of the Armenian historical territories (Armenia, Anatolia, Cilicia, Karabagh, Iran) whose four grand-parents were born in the same region. Results: By combining our data with published Armenian samples, we provide evidence for a previously underappreciated fine-scale level of genetic structure within Armenians. The FineSTRUCTURE dendrogram shows that individuals from the Antioch region separate clearly from the other samples, followed by a second split formed by individuals from the mountainous Sasun region. All groups (k=6) diverged from one another within the last 150 generations consistent with the high genetic differentiation between clusters. Principal components analysis of Armenian individuals places them in a central position within a continuum joining Europe to Asia but highlights subtle genetic diversity within close geographic areas. Conclusions: Taken together, our results refine our current knowledge of genetic diversity and population structure of Armenians and contribute to our understanding of human history and health.

Developing a Polygenic Score (PGS) for Idiopathic Parkinson's Disease: Insights on Statistical Approaches for PGS Construction

Sendel Sebastian[1], Landoulsi Zied[2], Lohmann Katja[3], Laabs Björn-Hergen[4], König Inke R.[4], May Patrick[2], Klein Christine[3], Caliebe Amke[1]

1 - Institute of Medical Informatics and Statistics, Kiel University, University Medical Center Schleswig-Holstein, Campus Kiel, 24105 Kiel, Germany (Germany), 2 - Luxembourg Centre for Systems Biomedicine (University of Luxembourg, Esch-sur-Alzette, Luxembourg Luxembourg), 3 - Institute of Neurogenetics, University of Luebeck, University Medical Center Schleswig-Holstein, Campus Luebeck, 23538 Luebeck, Germany (Germany), 4 - Institute of Medical Biometry and Statistics, University of Luebeck, University Medical Center Schleswig-Holstein, Campus Luebeck, 23562 Luebeck, Germany (Germany)

Background: The etiology of Parkinson's disease (PD) remains poorly understood, with contributions from genetic and environmental factors. Although a polygenic score (PGS) for idiopathic PD has been developed using clumping and thresholding, newer methods like penalized regression and Bayesian inference might better capture SNP contributions. Aim: To perform methods comparison for PGS approaches and develop and validate an advanced PGS for idiopathic PD and assess its contribution to genetic PD manifestation. Methods: Using genotype data from 1,762 PD patients and 4,227 controls (European ancestry) from the ProtectMove cohort (www.protectmove.de), we compared five PGS tools based on three statistical approaches: clumping and thresholding (PRSice-2), penalized regression (lassosum2, LDAK) and Bayesian (LDpred2, PRS-CSx). The best-performing PGS was validated in two external datasets and applied to 771 PD-associated variant carriers (335 patients, 436 healthy individuals). Results: Bayesian PGS methods performed best, followed by penalized regression and clumping and thresholding. The best PGS (928,814 SNPs, LDpred2) achieved an area-under-curve (AUC) of 0.680 [0.665-0.695] (ProtectMove) and of 0.718 and 0.667 in the external datasets, outperforming existing scores. In genetic PD variant carriers, the PGS had an AUC of 0.639 for GBA1 and 0.594 for heterozygous PRKN variant carriers. Conclusion: Bayesian methods hold great promise for constructing PGS for complex diseases. Our PD-PGS demonstrates high performance and a potential for individual risk analysis in idiopathic PD. Further, it suggests that genetic risk factors for idiopathic PD also influence the manifestation of genetic PD forms.

HLA genotype combinations impact allele association with Multiple Sclerosis risk

Serova-Erard Anna[1,2], Faddeenkov Igor[2], Demuth Stanislas[2], Bourguiba-Hachemi Sonia[2], Vince Nicolas[2], Gourraud Pierre-Antoine[2,3], Cornélis François[1,2,4]

1 - Génétique – Oncogénétique Adulte – Prévention (GENOAP) (France), 2 - Team 5 : Neuroinflammation, mechanisms, therapeutic options (NEMO) (France), 3 - Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, INSERM CIC 1413 (France), 4 - UFR de Médecine et des professions paramédicales (France)

Associations between Multiple sclerosis (MS) and HLA involve 7 predisposing and 6 protective alleles (HLA-DRB1*03:01/*08:01/*13:03/*15:01, HLA-DQB1*03:02, HLA-DPB1*03:01, LTA-H51P and HLA-A*02:01, HLA-B*38:01/*44:02/* HLA-DQA1*01:01, HLA-DQB1*03:01, respectively). We investigated HLA--wide genotype combinations in MS associations. We analysed WTCCC HLA data for 11,376 MS-cases (2005 diagnosis criteria) and 18,872 controls. We performed principal component analysis for European ancestry selection. HLA alleles were imputed with HIBAG R package or inferred with proxy SNPs (rs2229092, rs9273912 and rs9277565) and recoded to consider only the 13 MS-associated alleles. Dataset was divided: 20% to search for MS-associated genotype combinations and 80% to test them for replication. Replicated combinations were assessed on the whole dataset. We retained 9,024 MS and 13,923 controls from European Ancestry. In the 20% sub-sample, we observed 41 nominally MS-associated genotype-combinations ($P < 0.05$) (out of 776). In the 80% sub-sample, 22 combinations were replicated (P corrected $< 0.05/41$). In the full dataset, they accounted for 23.61% of MS-cases with 14 predisposing combinations (OR 1.83-6.75) and 4.29% of MS-cases with 8 protective combinations (OR 0.30-0.57). Surprisingly, some predisposing combinations carried "protective" alleles and vice versa: HLA-allele MS-association depends on HLA-wide-genotypes. Strikingly, 4.29% of patients, diagnosed with MS prior to 2005, carried protective combinations: those combinations are to be investigated in patients whose former MS diagnosis would in 2025 be changed to Neuromyelitis optica spectrum disorder or Myelin oligodendrocyte glycoprotein antibody-associated disease. Those HLA-combinations could help clarifying disease heterogeneity and improve diagnosis.

Genetic regulation of protein expression in prediabetes and type 2 diabetes

Singh Archit[1], Ganslmeier Marlene[2], Bocher Ozvan[1,3], Tutino Mauro[1], Stefan Norbert[2,4,5], Fritsche Andreas[2,4,5], Jumpertz Von Schwartzenberg Reiner[2,4,5], Zeggini Eleftheria[1,6], Birkenfeld Andreas[2,4,5]

1 - Institute of Translational Genomics, Helmholtz Munich (Germany), 2 - Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Center Munich at the University of Tuebingen, Tuebingen (Germany), 3 - Univ Brest, INSERM, EFS, UMR 1078 GGB, F-29200 Brest (France), 4 - Department of Diabetology, Endocrinology, Nephrology, University of Tuebingen, Tuebingen (Germany), 5 - German Center for Diabetes Research (DZD E.V.), Neuherberg (Germany), 6 - TUM School of Medicine and Health, Technical University of Munich and Klinikum Rechts der Isar, Munich (Germany)

Prediabetes and type 2 diabetes (T2D) are characterized by insufficient insulin secretion and poor sensitivity, leading to complications and increased mortality. Genetic predisposition is known to play a key role in these metabolic disorders. Not all individuals with prediabetes develop T2D, highlighting the importance of understanding the molecular mechanisms involved in this progression. To investigate these differences, we utilized genomics and plasma proteomics data from 450 individuals enrolled in the prediabetes lifestyle intervention study. Through differential protein expression analysis, we identified 225 out of 2523 proteins to be differentially expressed in 318 individuals with prediabetes compared to 88 individuals with T2D. Using a mixture model framework, we identified 65 shared protein quantitative trait loci (pQTL) effects suggesting overlapping genetic regulation of proteins. Further, we identified four differential pQTLs i.e., genetic variants with significant effect in only one condition, near genes coding for proteins SELE, ACP5, SCLY, and GYS1, showing significant positive effects on protein levels in prediabetes but not in T2D. These proteins were also found to be differentially expressed in prediabetes compared to T2D. Notably, knockout studies of SCLY in mice link it to glucose and lipid homeostasis, implicating pathways associated with fatty liver disease and glucose intolerance. Additionally, we identified 20 shared pQTLs with opposite effects between prediabetes and T2D, underscoring differences in genetic regulation between the two states. In conclusion, our study provides new insights into the molecular mechanisms underlying progression from prediabetes to T2D and a better understanding of the role of their genetic regulation.

Statistical approach for quantifying and predicting the evolution of tumor heterogeneity in chronic lymphocytic leukemia (CLL)

Vidhi Vidhi[1], Champagnat Nicolas[2], Herbach Ulysse[1], Fritsch Coralie[1,2]

1 - Inria Nancy, Grand Est (France), 2 - Institut Élie Cartan de Lorraine (France)

Targeted therapies have significantly improved cancer treatment, yet their efficacy is limited by intra-tumor heterogeneity. In lymphomas and leukemias, clonal evolution is driven by VDJ recombination and somatic hypermutation, leading to subclonal diversity. High-throughput sequencing enables the reconstruction of this evolutionary history, providing insights into treatment resistance mechanisms. We present a probabilistic model for clonal reconstruction, utilizing the full VDJ profile to infer evolutionary trees. By leveraging VDJ data alone, we constructed more accurate and flexible trees that can adjust to new information, such as bulk temporal data. The next challenge is optimizing edge weights and incorporating unobserved clones in a robust manner. To address this, we developed a Variational Expectation-Maximization (VEM) algorithm to refine tree structure while ensuring computational efficiency. In future work, we aim to integrate variant call format (VCF) data, which contains mutation frequencies, into the model and adds critical information to the clonal evolution puzzle. This integration will enhance the accuracy and comprehensiveness of the tree reconstruction, marking a significant advancement over methods that only use one dataset. By providing a principled and computationally tractable approach, our work contributes to more accurate tumor modeling, with implications for understanding disease progression and treatment resistance.

Topological Representation in Polygenic Trait Variation of Chronic Diseases

Vomo-Donfack Kelly Larissa[1,2], Bousquet Guilhem[3,4], Falgarone Géraldine[3,5], Ginot Grégory[1], Morilla Ian[1,2]

1 - Laboratoire Analyse, Géométrie et Applications (LAGA) (France), 2 - University of Malaga, , Department of Genetics, MLiMO, 29010, Málaga (Spain), 3 - Equipe Synergie entre Cancer et Maladies inflammatoires Chroniques (France), 4 - Hôpital Avicenne, Service d'Oncologie Médicale, Bobigny (France), 5 - Unité de Médicine ambulatoire, Hôpital Avicenne, Bobigny (France)

Understanding the inheritance and expression of polygenic traits requires robust mathematical frameworks capable of preserving the intricate topological characteristics of genetic profiles. In this longitudinal study, we analyse genetic data from five members across different generations within eight families, focusing on the development and application of the Polygenic Topological Representation Theorem. This theorem enables a manifold-based characterisation of genetic profiles, capturing conserved polygenic variation within familial contexts. The current phase of the study centres on establishing a comprehensive topological framework, leveraging known polygenic markers such as MET, alongside other reference genes documented in the literature, to validate the method. These markers serve as anchors for identifying conserved variation clusters and ensuring the robustness of the representation. The polygenic traits analysed in this study are associated with cardiovascular, inflammatory diseases, and cancer, highlighting their relevance to chronic disease research. Future phases of the project will integrate optimal transport to transition between topological manifolds of family members, enabling the study of variation transmission pathways across generations. By anchoring this foundational work on empirical genetic data, our framework circumvents the need for synthetic augmentation and provides a strong basis for subsequent analyses. Our findings from this initial phase highlight the potential of topological representations to elucidate polygenic variation dynamics, offering a novel approach to mathematical genetics and advancing the understanding of complex trait inheritance.