FRANCE GENOMIQUE

Inserm
IT Génétique, Génomique, Bioinformatique

cnrs BIOLOGIE

International Symposium on Human Genomics

May 5-7, 2025

CICSU Auditorium, 4 place Jussieu, Paris, France

BOOK OF ABSTRACTS

ACTIVE MOTIF®
Enabling Epigenetics Research

SORBONNE UNIVERSITÉ

illumina®

10x GENOMICS®

Agilent
Trusted Answers

TaKaRa

SCALE biosciences

HAMILTON

Element Biosciences

Oxford Nanopore Technologies

PacBio

Table of contents

Instability of coding versus non-coding microsatellite sequences in mismatch repair deficient colon tumor cells: the fighting spirit

Alex Duval

Duval Alex (1)

1 - Inserm Team "Microsatellite Instability and Cancer", Sorbonne University, UMRS 938 – CRSA, Centre de Recherche Saint-Antoine), Hôpital Saint-Antoine, APHP, Paris, France (France)

Abstract

Microsatellite instability (MSI) due to mismatch repair deficiency (dMMR) is common in human cancer with 1 million new cases of MSI tumors per year worldwide. These cancers are known to be associated with common somatic frameshift and immunogenic mutations due to MSI affecting coding microsatellite-containing genes, while the noncoding pathophysiological impact of this genomic instability if any is yet poorly understood. In a recent study, we performed an analysis of both coding and noncoding MSI events at the different steps of MSI colorectal tumorigenesis by whole exome sequencing and investigated their consequences using RNA sequencing at the bulk-tumor and single-cell levels. At the DNA level, MSI was demonstrated to lead to hundreds of noncoding DNA mutations, notably at polypyrimidine U2AF RNA-binding sites endowed with cis-activity in splicing, very early prior to cell transformation in the dMMR colonic crypt and before the onset of coding mutations. At the RNA level, a consecutive exon skipping signature impairing colonic cell differentiation by notably affecting the expression of dozens of alternative exons encoding protein isoforms governing cell fate was observed in association in MSI cancer cells, while also targeting constitutive exons that make dMMR cells immunogenic in early stage before the onset of coding mutations. This signature is characterized by its similarity to the oncogenic U2AF1-S34F splicing mutation observed in several other non-MSI cancer. Through these and other observations, I will focus my talk on the role of coding and non-coding microsatellite instability in cancer, focusing on how their respective contributions can be synergistic or, on the contrary, antagonistic during tumor development, with large grey areas still remaining to be elucidated.

Keywords: Microsatellite instability, cancer

Genome-wide association study of survival in sepsis patients

Syphax Zeggane

Zeggane Syphax (1), Mslmane Alaa (1), Mambu Mambueni Hendrick (1), Annane Djillali (1), Fleuriet Jérôme (2), Heming Nicolas (1), Olaso Robert (3), Deleuze Jean-François (3), Boland Anne (3), Garchon Henri-Jean (1)

1 - UFR Sciences de la santé Simone Veil (UVSQ), INSERM U1173, laboratoire infection et inflammation chronique (2i) (France), 2 - RHU RECORDS (France), 3 - CEA, Centre National de Recherche en Génomique Humaine, Université Paris-Saclay (France)

Abstract

Sepsis is a life-threatening multifactorial disease resulting from an exaggerated inflammatory response to infection. It is manifested by organ dysfunction that can lead to death. Each year, 50 million people worldwide are affected, of whom 11 million are deceased. Of note, 42% of sepsis patients are children under the age of 5. Moreover, half of the survivors suffer from sequelae, mainly neurological and cognitive. In this context, the identification of genetic factors associated with sepsis outcomes is important to help identify high-risk patients who could benefit from personalized interventions. We conducted a genome-wide association study (GWAS) on whole genome sequencing data from 725 sepsis patients recruited in the University-Hospital Research Project (RHU RECORDS) and (IHU PROMETHEUS). Patients were drawn from three cohorts: APROCCHSS (n=254, double-blind randomized trial), RECORDS OBS (n=323, observational COVID-19 cohort), and TRIAL (n=148, ongoing adaptive double-blind randomized trial). Common single-nucleotide variants (SNVs, MAF > 1%) were analyzed following rigorous quality control procedures. To assess associations with 90-day survival, a Cox proportional hazards regression model adjusted for age, cohort, and the first two principal components was applied. Six variants reached genome-wide significance (5x10-8). Five of them were located in cis-regulatory modules (CRMs) and three were identified as expression quantitative trait loci (eQTLs) in GTEx database. Notably, the combined effect of two eQTL-CRM variants significantly improved survival probability modeling (p=7x10-16). Gene set enrichment analysis (GSEA), KEGG pathway, and REACTOME analyses revealed cardiac condition, hormonal secretion, neuronal development and morphogenesis, and calcium flux as the most recurrent terms associated with mortality in Sepsis patients. These findings highlight potentially key biological processes underlying sepsis progression and provide new insights into the genetic architecture of sepsis and potential targets for functional validation and precision medicine strategies. Nonetheless, they must be replicated in an independent cohort.

Keywords: Sepsis, GWAS, Survival analysis, eQTL, Genetics, Genomics

Unravelling the molecular mechanisms causal to type 2 diabetes across global populations and disease-relevant tissues

Ozvan Bocher

Bocher Ozvan (1, 2), Arruda Ana Luiza (2, 3), Yoshiji Satoshi (4, 5, 6), Zhao Chi (7), Su Chen-Yang (6), Yin Xianyong (8, 9), Camman Davis (10), Taylor Henry J. (11, 12), Chen Jingchun (10), Suzuki Ken (13), Mandla Ravi (4, 14), Huerta-Chagoya Alicia (4), Yang Ta-Yu (5), Matsuda Fumihiko (5), Mercader Josep (4, 14, 15), Flannick Jason (4, 16), Meigs James B. (4, 15, 17), Wood Alexis C. (18), Vujkovic Marijana (19, 20), Voight Benjamin (20), Spracklen Cassandra N. (7), Rotter Jerome I. (21), P. Morris Andrew (22), Zeggini Eleftheria (2, 23)

1 - Univ Brest, INSERM, EFS, UMR 1078 GGB, F-29200 Brest, France (France), 2 - Institute of Translational Genomics, Helmholtz Munich, Neuherberg (Germany), 3 - Technical University of Munich (TUM), School of Medicine and Health, Graduate School of Experimental Medicine, Munich (Germany), 4 - Broad Institute of MIT and Harvard (United States), 5 - Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto (Japan), 6 - McGill Genome Centre, McGill University, Montreal (Canada), 7 - Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst (United States), 8 - Nanjing Medical University (China), 9 - Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor (United States), 10 - Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas (United States), 11 - Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda (United States), 12 - University of Cambridge [UK] (United Kingdom), 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 (United States), 15 - Harvard Medical School, Boston (United States), 16 - Boston Children's Hospital, Boston (United States), 17 - Division of General Internal Medicine, Massachusetts General Hospital, Boston (United States), 18 - Baylor College of Medicine, Houston (United States), 19 - Corporal Michael J. Crescenz VA Medical Center, Philadelphia (United States), 20 - University of Pennsylvania Perelman School of Medicine, Philadelphia (United States), 21 - Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance (United States), 22 - Centre for Genetics and GenomicsVersus Arthritis, Centre for Musculoskeletal Research, The University of Manchester (United Kingdom), 23 - TUM school of medicine and health, Technical University Munich and Klinikum Rechts der Isar, Munich (Germany)

Abstract

Type 2 diabetes (T2D) is a prevalent disease that arises from complex molecular mechanisms, for which causality needs to be unravelled. Recent large-scale efforts by the T2D Global Genomics Initiative (T2DGGI) have generated novel insights into the genetic architecture of T2D. These findings can be subsequently used to pinpoint molecular mechanisms leading to T2D. In this work, we leverage the latest T2DGGI multi-ancestry genetic associations to identify causal molecular mechanisms in an ancestry- and tissue-aware manner. Using two-sample Mendelian randomization corroborated by colocalization evidence across four global ancestries, we analyse causal effects of 20,307 genetically-predicted gene expression and 1,630 protein levels on T2D risk using blood-derived cis-quantitative trait loci (QTL). We detect causal effects of 335 genes and 46 proteins on T2D risk, with 16.4% and 50% replication in independent cohorts, respectively. Increasing diversity enables identification of more causal molecular mechanisms due to the availability of additional instrumental variables. In our study, the causal effects of five and nine genes are detected only in the African and admixed American ancestry groups, respectively, and the causal effects of six proteins are detected only in the East-Asian ancestry group. Additionally, using gene expression cis-QTL derived from seven T2D-relevant tissues, we identify causal links between 676 genetically-predicted gene expression levels and T2D risk, including novel associations such as CPXM1, FAM20B or PTGES2. The identified causal effects are mostly shared across ancestries, but highly

heterogeneous across tissues. Our findings underscore the power of tissue-informed multi-omics causal inference analyses in uncovering molecular traits driving T2D.

Keywords: Type 2 diabetes, Mendelian randomization, QTL, Multi, ancestry

Identification of anti-TB therapy induced ADRs genetic markers using In-Silico approaches

Kamal Kishor

Kishor Kamal (1)

1 - IIPS (India)

Abstract

Introduction: Adverse drug reactions (ADRs) are associated with clinical morbidity and, in severe cases, even mortality. Globally billions of dollars are spent on managing these ADRs for common and uncommon diseases. Due to these reasons drug resistant strains have emerged and are now a serious challenge to TB eradication. To effectively deliver the available treatment regimen and ensure patient compliance it is important to manage ADRs more efficiently. Recent studies have demonstrated that drug outcomes are patient-specific and can, therefore be predicted. A few of these drugs, including a few administered for TB, have shown excellent correlation with response rates and development of ADRs. Method: ADRs selected based on frequency of occurrence (≥1%). Anti-TB drugs were reviewed to identify the candidate genes (DMETs, HLA). Genes analysed with different web tools and databases to extract their SNPs. MAF >0.01 shortlisted using NCBI Gene and dbSNP databases (built 141). SNPs which lay in a functional domain of the gene were prioritized using SNPinfo web server (www.snpinfo.niehs.nih.gov/). Additionally, same analysis was done for Indian population. Result: We identified 10 genes which maybe directly linked to ADRs to various anti-TB drugs, 4 of these have been documented earlier. Nearly 47 genes were identified for indirect association with ADRs by virtue of them being off-targets of the drugs. Lastly, 5 genes were reported for their allelic association with anti-TB DIH. To our knowledge, this is the first review reporting a list of possible genetic markers in context to TB ADRs to such a vast extent. Conclusions: New gens are identified that may be associated potentially with anti-TB drug ADRs. This would translate into not just patient welfare but would also save billions of dollars spent annually on managing drug induced ADRs.

Keywords: pharmacegemics, anti, TB Drug.

The Genome of Europe: Towards Implementing Genetic Information in Health Care and Prevention

André Uitterlinden

Uitterlinden André (1)

1 - Laboratory for Population Genomics, Erasmus MC, Rotterdam, The Netherlands (Netherlands)

Abstract

The 1 million genomes (1+MG) initiative is part of the Digital Europe Program (DEP) and was declared in 2018 by (now) 27 signatory EU countries aiming to make at least 1 million whole genome sequences (WGS) accessible for use in research, health care, and prevention. 1+MG Working Group 12 (WG12) named Genome of Europe (GoE), was started in 2019 with many country representatives to establish a European Reference Genome Database of >500k WGS (@30x coverage). From these discussions a proposal was formulated to sequence the first 100,000 genomes which was awarded for funding by DEP and started in October 2024. With a budget of 45 mio euro GoE has now >30 participating countries, 51 institutes and >200 scientists involved, and has defined a strategy to collect the first 100k genomes to be proportional and representative of the diverse European populations. The data collection will adhere to ELSI and ICT guidelines and be made accessible via the Genomics Data Infrastructure (GDI) project which has been previously funded as part of the DEP. Several GoE "use cases" were defined including variant look-ups, genetic diversity analyses, multi-ancestry imputation services, and recalibration of genetic risk profiles, and establish longer term (clinical) applications. GoE will stimulate European genomic research competitiveness, advancing personalized medicine and broader scientific and healthcare objectives and, apart from integrating with European initiatives in health care and prevention, also seek global collaboration with similar genome initiatives. In addition, I will discuss some (potential) applications of using genetic information in health care and prevention (such as mutation screening, PRS, pharmacogenetics (PGx), based on robust and cheap array genotyping technology.

Keywords: 1 million genomes initiative

Genomic insights into the evolutionary history and metabolic risk of Polynesians

Etienne Patin

Liu Dang (1), Rijo De Leon Gaston (1), Roux Maguelonne (1), Speidel Leo (2), Teiti Iotefa (3), Tessier Anita (3), Richard Vaea (3), Aubry Maite (3), Harmant Christine (1), Bisiaux Aurélie (1), Jaquaniello Anthony (1), Li Zhi (1), Endicott Phillip (4), Forster Annie (5), Hill Adrian V. S. (5), Ioannidis Alexander (6), Moreno-Estrada Andrés (7), Mentzer Alexander (5), Cao Lormeau Van Mai (3), Patin Etienne (1), Quintana-Murci Lluis (8)

1 - Human Evolutionary Genetics Unit, UMR 2000 (France), 2 - Genetics Institute, University College London (United Kingdom), 3 - Laboratory of Research on Infectious Vector-Borne Diseases, Institut Louis Malardé (French Polynesia), 4 - Evolutionary Biology Department, University of Tartu (Estonia), 5 - Centre for Human Genetics, University of Oxford (United Kingdom), 6 - Department of Biomolecular Engineering, University of California Santa Cruz (United States), 7 - National Laboratory of Genomics for Biodiversity, Advanced Genomics Unit, CINVESTAV (Mexico), 8 - Human Evolutionary Genetics Unit, UMR 2000 (France)

Abstract

Polynesians, a large ethno-linguistic group living in Oceania, exhibit some of the highest rates of metabolic disorders worldwide, attributed to natural selection favoring disease-causing alleles during periods of food scarcity. Here, we generated whole-genome sequences from 1,881 Western, Eastern and Outlier Polynesians to reconstruct their evolutionary history. We estimated that Polynesians descend from admixture between Austronesian- and Papuan-related populations ~2,200 years ago, followed by rapid diversification and strong bottlenecks. We found that Austronesian- and Papuan-related ancestry correlate with higher body mass index and darker skin pigmentation in Polynesians, respectively. We identified variants associated with several quantitative traits, including private HDL-associated variants near LDLR gene. Notably, risk variants for metabolic disorders showed no evidence of positive selection in Polynesians, suggesting that genetic drift has been the primary force shaping disease risk. These results highlight how reconstructing the genetic history of understudied populations can elucidate the evolutionary origins of human phenotypic variation.

Keywords: Metabolic disease, GWAS, genetic risk, understudied populations, Polynesia, evolutionary mismatch, polygenic selection

Exploring Rare Genetic Variants in French Centenarians: A Path to Understanding Longevity

Assia Benmehdia

Benmehdia Assia (1), Sahbatou Mourad (2), Sandron Florian (1), Bacq-Daian Delphine (1), Blanché Hélène (2), Le Floch Edith (1), How-Kit Alexandre (2), Zagury Jean-François (3), Deleuze Jean-François (1, 2, 4), Dandine-Roulland Claire (1)

1 - Centre National de Recherche en Génomique Humaine (CNRGH) (France), 2 - Centre d'Etude du Polymorphisme Humain (CEPH) (France), 3 - Laboratoire de Génomique, bio-informatique et chimie moléculaire (France), 4 - Centre de référence, d'innovation, d'expertise et de transfert (CREFIX) (France)

Abstract

Longevity is a complex phenotype shaped by interactions between genetic, environmental, and lifestyle factors. Within the AGENOMICS project, we specifically explore the genetic component, focusing on rare genetic variants with an allele frequency below 1% in the French population. Material and Methods: Whole-genome sequencing data were analyzed from over 1,000 French centenarians from the CEPH CHRONOS cohort. The samples consisted of peripheral blood mononuclear cells (PBMCs, 40%) and lymphoblastoid cell lines (LCLs, 60%), with 80% of participants being female. Variants were annotated and functionally predicted using the SnpEff tool and the Genome Aggregation Database (gnomAD). Results: 67 million genetic variants were identified, among which 2% are located in coding regions. Across all identified variants, 79% are single nucleotide polymorphisms (SNPs), and 21% are small insertions and deletions (INDELs). Notably, 22% of all detected variants were absent from the global gnomAD database, and nearly 60% of these novel variants had a frequency below 1%. Samples derived from LCLs exhibited a higher number of rare variants compared to those derived from PBMCs. Conclusion: The first results show that the genetic data of CHRONOS centenarians holds great potential for studying longevity, particularly rare genetic variants. We plan to further investigate the differences observed across age groups, sample collections, and sexes. To identify the unique genetic factors driving longevity in this cohort, we will compare our findings with data from the general French population. This will help distinguish variants specific to centenarians from those prevalent in the broader population. We thank AXA Mécénat Santé for their funding throughout all aspects of our study.

Keywords: Longevity, Rare variants, Centenarians

A critical comparison of clustering methods in structured populations under different spatial sampling schemes

Mael Guivarch

Guivarch Mael (1), Herzig Anthony (2), Saint Pierre Aude (1), Génin Emmanuelle (1, 3)

1 - Inserm, Univ Brest, EFS, UMR 1078, GGB, Brest, France (France), 2 - Inserm, Univ Brest, EFS, UMR 1078, GGB, Brest, France (France), 3 - CHRU Brest, Brest, France (France)

Abstract

With cross-sectional genomic studies now encompassing large geographic areas, fine-grained population structure analyses have become essential to correctly model such datasets. Identifying subpopulation using clustering algorithms is one way to detect such structure, which is particularly when allele frequencies in subgroups are to be estimated. However, selecting the optimal approach is challenging. Key considerations include choosing an appropriate algorithm based on data characteristics or the number of subpopulations to report while ensuring consistency in the clustering results. To address these challenges, we compare different clustering methodologies, including model-based approaches designed specifically for population genetics, such as FineSTRUCTURE, and more general methods like Mclust. Additionally, we assess non-probabilistic techniques such as Leiden and k-means, examining their effectiveness in clustering genetic data. Our study also explores how preprocessing steps and cluster validation strategies influence final results. To closely mimic real-world conditions in genetic studies, we add another layer of complexity by applying these methods to spatially stratified populations. We investigate in particular the impact of different study-sampling schemes as well as the impact of studying common genetic variations, haplotype-sharing, or rare variants afforded by whole-genome sequencing data. Unlike previous studies that focused on genetic structures at broad geographic scales (e.g., across continents), we simulate fine-scale genetic patterns under controlled spatial stepping-stone scenarios. We generate a dataset of 27,000 individuals across a 36-deme grid, allowing migration between adjacent demes. By fine-tuning migration rates and coalescent times, we approximate realistic continuous genetic landscapes. Our findings provide valuable guidance for interpreting fine-scale population structure and selecting suitable algorithms. We highlight the trade-offs between accuracy and computational time. We confirm that haplotypic data consistently improves clustering accuracy compared to genotypic data. Finally, we emphasize the crucial role of spatial subsampling in describing sub-populations, underscoring its importance in describing patterns of allele frequencies across stratified populations.

Keywords: clustering, spatial analysis, backward simulations

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

Florian Privé

Privé Florian (1)

1 - Aarhus University (Denmark)

Abstract

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.

Keywords: population structure, principal component analysis, biobank

Gene-environment interaction in human traits and diseases: a story of misconception

Hugues Aschard

Aschard Hugues (1, 2), Denis Linon (3), Mccaw Zachary (4)

1 - Institut Pasteur (France), 2 - Harvard School of Public Health (United States), 3 - Institut Pasteur (France), 4 - Insitro (United States)

Abstract

Gene-environment interaction typically refers to generative models where the effect of a genetic variation on an outcome depends on the value of a non-genetic risk factor. In a statistical model, an interaction is defined as the product of two variables (here G×E). By construction, this interaction term is correlated with the original variables, producing collinearity. For example, if E is a binary exposure with a frequency of 0.8 and G is a SNP with a minor allele frequency of 0.2, the expected correlation cor(G,G×E)=0.85. Such high collinearity can severely impact the interpretability of linear models. In the context of GxE, this has two consequences. First, estimates from interaction models are expected to have larger variance than estimates from marginal models, which explains the low statistical power of existing G×E interaction tests. Second, collinearity makes the assessment of each term's contribution to the outcome's variance challenging. The current standard, defined by mathematical convenience, consists in comparing the R-squared between the marginal (Y~G+E) and interaction models (Y~G+E+G×E), attributing most of the variance explained to the simpler model (marginal effects) and assigning what remains to more-complex terms (the interaction effect). We show that this approach systematically underestimates the contribution of GxE and poorly qualifies the expected change in genetic effect across the exposure spectrum. We propose an alternative metric, named the Pratt Index, that resolves this bias. When applied in real UK Biobank data, we measured variance explained by GxE up to five fold larger than suggested by the standard approach.

Keywords: gene, environment interaction, heritability, GWAS

The new biology revealed by single-molecule sequencing of the transcriptome

Ana Conesa

Conesa Ana (1)

1 - Institute for Integrative Systems Biology. Spanish National Research Council, Paterna, Spain (Spain)

Abstract

The advent of long-read sequencing technologies, such as PacBio and Oxford Nanopore, has revolutionized the ability to generate full-length transcript sequences. These advancements offer unprecedented insights into complex isoforms and transcript structures. As sequencing precision and depth improve, long-read methods are becoming indispensable for transcriptomics, enabling the identification of differential gene expression and isoform usage across conditions with robust replication. To accommodate the surge in long-read data, new algorithms for transcript reconstruction and quantification have emerged. However, the transition from short to long reads raises pressing questions about optimizing data preprocessing, experimental design, quantification, and normalization strategies. Key challenges include assessing the quality of transcript identification and quantification, building accurate long-read-based quantification tables, determining optimal replicate numbers and sequencing depth, addressing biases in transcript quantification, refining data analysis strategies tailored to long-read technologies and algorithms. In this presentation, I will share efforts from my lab to evaluate the quality and utilization of long-read transcriptomics data. Additionally, I will discuss the persistent challenges that must be addressed to fully transition from short reads to long reads in transcriptomics studies.

Keywords: long read RNAseq, analyses

DNA long-read sequencing, an interest for genetics predispositions to breast and ovarian cancer

Crystal Renaud

Renaud Crystal (1, 2), Chouteau Antoine (1), Aucouturier Camille (1, 2), Leman Raphaël (1, 2), Atkinson Alexandre (1), Lavole Thibaut (1), Sorreda-Ricou Agathe (1, 2), Boulouard Flavie (1, 2), Goardon Nicolas (1, 2), Krieger Sophie (1, 2), Castéra Laurent (1, 2)

1 - Laboratoire de Biologie et Génétique du Cancer [Centre François Baclesse] (France), 2 - Inserm U1245, Cancer and Brain Genomics (France)

Abstract

Current molecular diagnostic methods for Hereditary Breast and Ovarian Cancer (HBOC) are based on short-read sequencing (SRS) of exons and flanking regions. These approaches can fail to detect complex events. Emergence of third-generation sequencing (or long-read sequencing, LRS) enables identification of these events, no matter how complex. We propose here a complete LRS method: from DNA extraction to final bioinformatic analysis. We selected 200 families whose phenotype was in favour of HBOC syndrome and SRS results were negative. After extraction of high molecular weight DNA, sequencing was performed on the PromethION P2 solo using adaptive sampling technology on a panel of 161 genes linked to cancer. After demultiplexing and basecalling with Dorado, the pipeline performs alignment with minimap2, single nucleotide variants calling by Clair3 and annotation with SnpEff or VEP. Structural variants were called by Sniffles and annotated by AnnotSV and GreenVaran. Copy number variants were investigated using CNVkit. Two hundred patients were sequenced. The average coverage for our panel genes was 50 reads. Adaptive sampling gave an average coverage of 5 reads for the other regions, allowing for shallow sequencing. All positive controls were detected by the pipeline. Although the analyses are being processed, we already fully characterized 3 insertion of Alu elements and 2 tandem duplications. In conclusion, LRS makes allows the characterisation of complex events. Our approach facilitates the analysis of a large number of patients by LRS. The analysis of the 200 patients could explain part of the missing hereditary through the discovery of new variants.

Keywords: Long, read sequencing, analysis pipeline

wastewater-based epidemiology of human viruses by nanopore sequencing

Juejun Chen

Chen Juejun (1)

1 - Institut de pharmacologie moléculaire et cellulaire (France)

Abstract

Since COVID-19, wastewater-based epidemiology has been proven as a tool for tracking the evolution of SARS-COV-2 in the general population. Many studies have shown that viral levels in wastewater are linked to clinical patient data. The method is low-cost, efficient and anonymous, and it reflects well the spread in an area, including for asymptomatic patients. We optimized protocols to extract and sequence RNA from wastewater samples that were collected weekly in the Nice area between 2022 and 2024. Enrichment of sequences for 66 clinically-relevant viruses by hybrid-capture was performed with a panel of probes against these viruses. We developed an experimental and bioinformatics pipeline to automate the analysis process, in order to align and classify the sequences, then analyze the complexity of the viral metagenome. This approach revealed the circulation of more than 30 human viruses. A large fraction of the reads corresponded to intestinal viruses which were observed at a high and steady level throughout the year. Respiratory viruses, such as influenza viruses, were also detected, though at a lower level, but at periods that matched with the annual episodes of the disease. Variants of SARS-CoV-2 were also systematically identified. Sequence information provides a rationale to monitor the evolution of the group of viruses that are detectable in wastewater. This approach can be particularly useful to early detect the emergence of new variants of concern. An interest of the approach is also to document in one single sequencing experiment the possible evolutions of complex ecosystems that contain multiple viruses, and to document the diversity of viral communities. This study finally contributes to further establishing the interest of wastewater-based epidemiology for guidance of public health responses.

Keywords: Computational biology, astrovirus, influenza virus, wastewater, based epidemiology

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

Raphaël Malak

Malak Raphaël (1), Frouin Arthur (1, 2), Henches Léo (3), Auvergne Antoine (1), Boetto Christophe (1), Chikhi Rayan (2), Aschard Hugues (1, 4)

1 - Génétique Statistique - Statistical Genetics (France), 2 - Department of Computational Biology, Institut Pasteur (France), 3 - Department of Computational Biology, Institut Pasteur (France), 4 - Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115 (United States)

Abstract

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).

Keywords: GWAS, gut microbiome, metagenome studies, k, mers, biostatistics, bioinformatics, shotgun sequencing

Science in times of uncertainty: investigating and communicating on the origin of the COVID-19 pandemic

Florence Débarre

Débarre Florence (1)

1 - Institut d'écologie et des sciences de l'environnement de Paris (France)

Abstract

SARS-CoV-2, the virus causing COVID-19, was first detected in Wuhan, China in the end of December 2019. Five years later, its origin is still not known with certainty. While all the available data to date, and most published scientific articles, point to a natural origin linked to wildlife trade in a market, the idea of a "lab leak" is pervasive in the general population and in some media. In this talk, I will first give a brief overview of the available evidence, with an emphasis on genetic data. I will then discuss more generally the challenges of working on a topic that has become controversial, the differences between reasonable scientific doubt and rumors, between scientific inquiry and activism, and the responsibility of scientists in the public sphere.

Keywords: COVID, 19, debunking controversies

Genetics Of Deafness For Precision Medicine

Christine Petit

Sophie Boucher[1], Salim Aiche[2], Samia Abdi[3,4], Malak Salame[5], Mirna Mustapha[6], Ahmed Houmeida[5], Abdelhamid Barakat[7], Cherine Charfeddine [8], Rahma Mkouar[8], Fatima Ammar Khodja[9], Mohamed Makrelouf[3], Akila Zenati[3], Paul Avan[2], Sonia Abdelhak[8], Crystel Bonnet[2] & Christine Petit[2,10]

[1]Service d'ORL et chirurgie cervico-faciale, CHU d'Angers, Equipe Mitolab, Institut Mitovasc, CNRS UMR6015, UMR Inserm 1083, Université d'Angers, Angers, France.
[2]Université Paris Cité, Institut Pasteur, AP-HP, INSERM, CNRS, Fondation Pour l'Audition,
Institut de l'Audition, IHU reConnect, F-75012 Paris, France
[3]Laboratoire de Biochimie Génétique, Service de Biologie - CHU de Bab El Oued, Université d'Alger 1, Algiers, Algeria
[4]Laboratoire Central de Biologie, CHU Frantz Fanon, Faculté de Médecine, Université Saad Dahleb, Blida, Algeria
[5]Unité de Recherche sur les Biomarqueurs dans la Population Mauritanienne, Université des Sciences, de Technologie, et de Médecine (USTM), Nouakchott, Mauritania
[6]Department of Biomedical Science, University of Sheffield, UK
[7]Laboratoire de Génétique Moléculaire Humaine, Département de Recherche Scientifique, Institut Pasteur du Maroc, Casablanca, Morocco
[8]Institut Pasteur de Tunis, LR16IPT05, Biomedical Genomics and Oncogenetics Laboratory, Tunis, Tunisia
[9]Equipe de Génétique, Laboratoire de Biologie Cellulaire et Moléculaire, Faculté des Sciences Biologiques, Université des Sciences et de la Technologie Houari Boumédiène (USTHB), El Alia, Bab-Ezzouar, Algiers, Algeria
[10]Collège de France, Paris, France

Abstract

Today, genetic data dominate precision medicine for auditory sensory disorders. More generally, our knowledge of the molecular physiology and pathophysiology of the auditory system is based almost entirely on genetic information and, as a corollary, current therapeutic research focuses principally on gene therapy. With the aim of improving the molecular diagnosis of deafness and developing gene therapy approaches, we addressed the genetic architecture of adult-onset sensorineural deafness (presbycusis) and Usher syndrome type 1 (USH1), the most severe inherited multisensory disorder (profound deafness, balance defects and retinitis pigmentosa).

Presbycusis. Presbycusis, or age-related hearing loss (ARHL), is a major public health issue. About half the phenotypic variance has been attributed to genetic factors. We assessed the contribution of monogenic forms by considering ultrarare variants predicted to be pathogenic as probably causing Mendelian forms. We focused on severe presbycusis without environmental or comorbidity risk factors. We performed whole-exome sequencing on multiplex family cases of age-related hearing loss (106 independent mARHL cases) and simplex/sporadic cases of age-related hearing loss (177 sARHL cases), together with controls with normal hearing (127 controls). In a first cohort and a replication cohort for which studies are ongoing, ultrarare variants (allele frequency [AF] < 0.0001) of 42 genes responsible for autosomal dominant early-onset forms of deafness, predicted to be pathogenic, were detected in 27% of mARHL and 28% of sARHL cases vs. only 7.5% of controls (P = 0.001); half these variants were previously unknown (AF < 0.000002). *TECTA*, *MYO7A* and *PTPRQ* mutations were present in 7.2% of ARHL cases (accounting for 26% of the solved ARHL cases) but less than 1% of controls. Evidence for a causal role of variants in presbycusis was provided by pathogenicity prediction programs, documented haploinsufficiency, three-dimensional structure/function analyses, cell biology experiments, and reported early effects. These results demonstrate that the genetics of presbycusis is shaped not only by polygenic risk factors

of small effect size revealed by common variants, but also by ultrarare variants probably resulting in monogenic forms that could potentially be treated by emerging inner ear gene therapies.

Usher1 syndrome/DFNB. We set up a cohort of 450 unrelated deaf individuals from families with severe-to-profound bilateral prelingual HI (hearing impairment) in Tunisia, Jordan, Algeria, Morocco, and Mauritania (TJAMM cohort). Consanguinity rates are high in all these countries. Genome analysis by panel-based DNA sequencing (HearPanel-IdA-1) fully resolved 90% of cases of autosomal recessive forms of isolated deafness (DFNB forms). Biallelic mutations in *USH/DFNB* genes underlying Usher syndrome or DFNB forms were detected in 13% of the patients, raising the crucial question of how to distinguish mutations causing USH from those causing DFNB. Through an initial study focusing on the *USH1/DFNB* mutations of the cohort subsequently extended to hundreds of cases, characteristic features of mutations underlying USH1, the most severe form of Usher syndrome (deafness associated with vestibular dysfunction and retinitis pigmentosa), and DFNB were identified. These results should significantly improve the clinical interpretation of genotypes. Furthermore, a comparative analysis of the two types of mutations provided clues to the retinal pathophysiology of USH1.

Keywords:

Graph neural networks reveal digenic disease candidates through biological network analysis

Romain Nicolle

Nicolle Romain (1, 2), Malan Valérie (2, 3), Rausell Antonio (1, 2)

1 - Laboratoire de Bioinformatique Clinique (France), 2 - Service de Médecine Génomique des Maladies Rares (France), 3 - Laboratoire de Génétique des Troubles du Neurodéveloppement (France)

Abstract

More than 4,500 genes are associated with monogenic diseases, including ~400 haplosensitive genes. Yet, ~50% of patients with developmental disorders lack a molecular diagnosis. Yeast studies and recent human research suggest digenism may explain unresolved cases. Gene pairs with loss-of-function (LoF) variants could lead to disease by disrupting compensatory pathways. To uncover such scenarios, we hypothesized that pairs of genes sensitive to complete or heterozygous loss-of-function variants would present, respectively, similar characteristics to individual essential and haplosensitive genes in the context of biological networks. To test this, we built species-specific knowledge-based networks for yeasts and humans, integrating protein interactions, gene expression, signaling pathways, and gene features like sequence conservation. We trained Graph Neural Networks (GNNs) on these networks to predict essential and haplosensitive genes and evaluated their ability to identify pathogenic gene-pair inactivations. We first validated the approach in yeasts, where most double-gene mutants have been studied. A GNN trained to classify n=1311 essential from n=5268 non-essential genes accurately distinguished 72884 synthetic lethal from 72884 non-lethal gene pairs (AUROC = 0.70). To do so, gene pairs were represented as new nodes in the network, inheriting connections from both genes, and predictions relied solely on relational and neighboring gene features. In humans, a GNN trained to classify n=10654 known digenic disease-causing pairs from n=11000 random pairs achieved AUROC = 0.80. Additionally, a GNN trained to discriminate n=2907 haplosensitive from n=16482 non-haplosensitive genes reached a cross-validation AUROC = 0.90. Finally, when training a GNN model for the identification of human cell-line synthetic lethal pairs, analogous to the one in yeasts, a non-random AUROC > 0.6 was achieved. Further validation is underway using sequencing data from an intellectual disability cohort. Our results demonstrate that biological network properties extracted from GNNs can predict inactivation intolerance, enabling digenic disease hypothesis generation for unresolved rare disorders.

Keywords: Bioinformatics, Deep Learning, Graph Neural Network, Digenism

VIOLA: Variant PrIOritization using Latent spAce to improve mitochondrial diseases diagnosis

Justine Labory

Labory Justine (1, 2), Boulaimen Youssef, Singh Jasmine, Paquis-Flucklinger Véronique, Ait-El-Mkadem Saadi Samira, Bannwarth Sylvie, Bottini Silvia (3)

1 - Université Côte d'Azur (France), 2 - Institut Sophia Agrobiotech (France), 3 - INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France (France)

Abstract

Mitochondrial diseases (MDs) are rare and highly heterogeneous disorders caused by variants in nuclear or mitochondrial genes involved in mitochondrial functions. Next-Generation Sequencing technologies enable the rapid identification of all individual variants, yet the challenge remains to identify the disease-causing variant among thousands. Variant prioritization, which ranks variants based on their relevance to a phenotype, is a powerful approach for this task. However, most existing methods rely on large cohorts and predefined rules, which are often unsuitable for rare diseases where responsible variants exhibit unique characteristics. We present VIOLA (Variant prIOritization using Latent spAce), a novel approach based on the hypothesis that disease-causing variants are rare and have unique properties. VIOLA assumes that these variants are outliers in an individual's variant distribution. It first collects 32 scores spanning from conservation to splicing to describe variants properties. Then uses a variational autoencoder to extract latent features from these properties. At the level of the latent space, a density-based clustering technique is used to retrieve the outliers' variants. A series of stringent filters on variant quality, coverage and type are applied. Finally, VIOLA provides two scores: the VIOLA score (VS) which combines the Mahalanobis distance metric, transcriptomic and phenotypic features from HPO terms and MD-related properties, and the combined VIOLA score, which integrates the VS and the Exomiser pathogenicity score. We also added a third rank to prioritize variants compatible with autosomal recessive mode of inheritance. We tested VIOLA on 20 Whole Exome Sequencing datasets, including four diagnosed and 16 undiagnosed patients. VIOLA selects only 1% of input variants, considerably reducing the list of candidates to be inspected. It correctly identified and ranked the known causal variant in the top 20 for all diagnosed patients. VIOLA is a patient-specific tool integrating genomics, phenotypics, and transcriptomics data to improve MDs diagnosis.

Keywords: Mitochondrial disease, Variant prioritization, Machine learning, Omics

Metanalysis of germline whole exome sequencing in 1,435 cases of testicular germ cell tumour to evaluate disruptive mutations under dominant, recessive and X-linked inheritance models

Zeid Kuzbari

Kuzbari Zeid (1), Rowlands Charlie F. (1), Wade Isaac (2), Garrett Alice (1, 3), Loveday Chey, Choi Subin (1), Torr Beth (1), Litchfield Kevin (4), Reid Alison (1), Huddart Robert (1), Broderick Peter (1), Houlston Richard S. (1), Turnbull Clare (1)

1 - The institute of cancer research [London] (United Kingdom), 2 - University of Oxford (United Kingdom), 3 - St George's University Hospitals (United Kingdom), 4 - University College, London (United Kingdom)

Abstract

Background and objective: Testicular germ cell tumour (TGCT) is the most common cancer in young men and over half of its high estimated heritability is unexplained. Our objective: to identify rare pathogenic germline variation driving TGCT susceptibility. Methods: This study is a case-control meta-analysis of whole-exome sequencing data from three datasets (ICR, TCGA and UK Biobank). Following quality control, we retained unrelated male individuals of European ancestry comprising 1,435 TGCT cases and 18,284 cancer-free controls. We performed gene-level association testing of protein-truncating variants and nonsynonymous disruptive variants across genes associated with cancer susceptibility, ciliary function, sex development, sperm dysfunction, DNA repair and in linkage disequilibrium with GWAS candidate loci. We then widened the scope to analyse 19,355 genes exome-wide under dominant and recessive models, including X-linked genes. Results and limitations: No individual gene-disease association was identified following multiple testing correction. However, functional gene-set analyses identified an excess of associations with genes involved in microtubular/ciliary pathways (p=1.69x10-8). Our study was well powered to detect genes bearing rare variants of moderate to high effect size (OR≥5), but power diminished rapidly for more modest effect sizes (OR

Keywords: Cancer susceptibility genes, Gene association testing, Germ cell tumour, Germline mutations, Meta analysis, Testicular cancer, Whole exome sequencing

Approaches to prioritize non-coding disease risk variants

Steven Gazal

Gazal Steven (1)

1 - University of Southern California (United States)

Abstract

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with human diseases, primarily in non-coding regions. However, identifying the underlying causal variants remains challenging, limiting the translation of GWAS findings. This presentation will introduce novel approaches for prioritizing these non-coding variants by accurately predicting their deleterious and functional effects. Firstly, I will introduce a new machine-learning framework aiming at leveraging GWAS polygenic signal and a hundred of functional annotations to optimally predict disease effect sizes for all common and low-frequency variants. Second, I will describe how to optimally leverage two of the most exciting recent resources to predict the deleterious and functional effect of non-coding variants: evolutionary constraint and population-scale single-cell datasets.

Keywords: GWAS, machine learning

Rare variant aggregate association analysis using imputed data is a powerful approach

Suzanne M. Leal

Leal Suzanne M. (1), Liu Tianyi (1), Auer Paul L (2), Wang Gao (1), Naderi Elnaz (1), Dewan Andrew T. (3)

1 - Columbia University (United States), 2 - Medical College of Wisconsin (United States), 3 - Yale University (United States)

Abstract

Imputation can cost-effectively generate genotypes for millions of variants. The power of performing rare variant aggregate association tests using imputed genotypes was evaluated. White Europeans from the UK Biobank with exome sequence and genotype array data were analyzed. Using the genotype type array data from the UK Biobank, imputation was performed using the HRC r1.1 (N=64,976 Haplotypes) and TOPMed r3 (N=267,194 haplotypes) reference panels. Simulations were used to compare the power of performing rare variant aggregate association analysis using sequence and imputed data. The number of genes with >2 rare variants (missense, nonsense, splice site) was approximately the same for exome sequence and TOPMed imputed data, but HRC imputed data had ~10% fewer genes. For the imputed data, using a less stringent R2 threshold (i.e. >0.3 vs. >0.8) led to greater power to detect aggregate associations due to additional rare variants included in the test. Exome sequence data provided the highest power for rare variant aggregate association testing, with TOPMed imputed variants usually having less than a 20% reduction in power. HRC imputed variants provided substantially less power. We also performed rare variant aggregate association analyses using UK Biobank phenotype, exome sequence data, and imputed variants for PCSK9 and low-density lipoprotein (N=159,904 study subjects) and APOC3 and triglyceride levels (N=160,036 study subjects). For these analyses even when ultra-rare variants (minor allele frequency

Keywords: Imputation, Complex traits, Rare variant aggregate association tests, Sequence data

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

Sidonie Foulon

Foulon Sidonie (1, 2), Truong Thérèse (3), Leutenegger Anne-Louise (2), Perdry Hervé (1)

1 - CESP Inserm U1018, Université Paris-Saclay, F-94807 Villejuif, France (France), 2 - Inserm Université Paris Cité, NeuroDiderot, Paris (France), 3 - CESP Inserm U1018, Université Paris-Saclay, F-94807 Villejuif, France (France)

Abstract

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. In 2013, Génin et al. proposed the HBD-GWAS method, which relies on Homozygous-by-Descent (HBD) segments. HBD segments, found in consanguineous individuals, are regions where rare recessive variants are more likely to be found. HBD-GWAS performs a linkage analysis based on excess HBD segments to highlight genomic regions linked to the disease. Here, 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. 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.

Keywords: GWAS, Rare recessive variants, Consanguinity, Multifactorial diseases, Statistical power

LDAK-PBAT: A Novel Pathway-Based Analysis Tool for Decoding the Genetics of Complex Diseases

Takiy Berrandou

Berrandou Takiy (1), Speed Doug (2)

1 - Université Paris Cité, Paris Cardiovascular Research Center, Inserm, Paris, France (France), 2 - Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark. (Denmark)

Abstract

LDAK-PBAT is a novel, computationally efficient pathway-based analysis tool that quantifies overall heritability enrichment by aggregating variant contributions within biological pathways using only GWAS summary statistics and a reference panel. Designed to advance personalized medicine, LDAK-PBAT rapidly evaluates thousands of pathways within minutes. In simulation studies, LDAK-PBAT demonstrated robust Type I error control: using 100 null phenotypes across four p-value thresholds, false positives were well below expectations, with zero significant pathways at $p = 10^{-5}$ and the Bonferroni level. Further simulations with over 1,125 phenotypes across 20 pathways confirmed high accuracy, achieving RMSE values as low as 0.00093 and correlation coefficients up to 0.98 in estimating pathway heritability. Additionally, LDAK-PBAT achieved an F1 score of 0.734—significantly outperforming MAGMA (0.636) and hypergeometric testing (0.570)—across genetic architectures with heritability levels of 0.1, 0.3, and 0.5 and causal SNP counts from 2,000 to 10,000. In large-scale applications, LDAK-PBAT analyzed 6,000 pathways across 37 traits from the UK Biobank, MVP, and PGC, identifying 4,861 significant pathways at a Bonferroni-adjusted threshold versus 97 detected by MAGMA. For example, it identified 283 significant pathways for height, 403 for Alzheimer's disease, and 857 for LDL cholesterol, while alternative methods produced markedly lower counts—demonstrating consistent power across diverse conditions. Moreover, pathway significance was validated across increasing UK Biobank sample sizes, with confirmed pathways rising from 348 at 50k individuals to 108 at 200k, and up to 101 confirmed by other methods. Sensitivity analyses confirmed that enrichment estimates remain highly concordant despite variations in SNP density and reference panels. By providing precise, scalable, and robust pathway heritability estimation, LDAK-PBAT offers unprecedented insights into the genetic architecture of complex diseases and stands as an indispensable tool for large-scale genetic research and personalized medicine. These comprehensive evaluations underscore its broad applicability and potential to transform genetic research.

Keywords: complex traits, GWAS, gene, set/pathway, based association test, summary statistics

rcRS algorithm: Incorporating complex genetic model into risk estimation

Fabien Laporte

Laporte Fabien (1), Perdry Hervé (2), Herzig Anthony (3), Dina Christian (1), Génin Emmanuelle (4), Redon Richard (1)

1 - Nantes Université, CNRS, Inserm, l'institut du thorax, F-44000 Nantes, France (France), 2 - CESP Inserm U1018, Université Paris-Saclay, F-94807 Villejuif, France (France), 3 - Inserm, Univ Brest, EFS, CHU Brest, UMR 1078, GGB, F-29200 Brest, France (France), 4 - Inserm, Univ Brest, EFS, CHU Brest, UMR 1078, GGB, F-29200 Brest, France (France)

Abstract

Since the 2000's and the appearance of genome-wide association studies, researchers studied the effect of genetic onto diseases at a fine level through association, mainly for common variants. Observing large number of associations with human diseases later allowed development of polygenic risk scores, which have multiple clinical applications as risk stratification for disease prevention. Most PRS algorithms are based on a weighted sum of the effect of risk variants. Nevertheless, genetic architecture of diseases could be more complex and rare variant mutations have to be accounted for risk stratification. Thanks to the development of machine learning algorithms, a few PRS algorithms try to catch some epistasis effects. Here we present the rcRS (Rare and Common Risk Score) algorithm, a PRS algorithm which combines non-linear effects across the genome for both common and rare variant mutations. Firstly, we validate the proposed non-linear PRS algorithm on a simulated population with rare and common variant mutations which impact disease probability. Genotypes are simulated using mosaic of haplotypes through the packages R Mozza, sampling from 1000 genomes haplotypes. Then phenotypes are simulated using the HAPGEN2 and the deepRVAT simulation algorithms, separately for common and rare variants respectively. Secondly, we apply rcRS to Intracranial Aneurysm (IA) dataset composed of 300 IA carriers and 300 controls. Results are promising regarding the simulated datasets. rcRS perform better than classical PRS algorithms regarding non-linear genetic models with an increase of the Area Under the Curve (AUC) (0.78 and 0.69 respectively). Regarding the real dataset, rcRS performs equally as classical PRS algorithms when weigths are freely estimated in the analysed dataset. The small numbers of individuals used in this case does not allow to catch correctly the genetic effect. With a higher number of individuals, we expect an increase of the AUC.

Keywords: Genetics, Genetic Risk Score, NonLinear model, Simulation, Intracranial Aneurysm

IMGT® Population Analysis of the Human IGH Locus: Unveiling Novel Polymorphisms and Copy Number Variations Across Diverse Genome assemblies

Ariadni Papadaki, Maria Georga

Papadaki Ariadni (1), Georga Maria (1), Jabado-Michaloud Joumana (1), Folch Géraldine (1), Giudicelli Veronique (1), Duroux Patrice (1), Kossida Sofia (1)

1 - IMGT (France)

Abstract

Unraveling the genetic complexity of the human immunoglobulin heavy chain (IGH) locus provides valuable insights into the mechanisms that contribute to the efficacy and remarkable specificity of the adaptive immune response. Despite its crucial role, IGH locus remains insufficiently characterized, with its allelic diversity and polymorphisms particularly across diverse populations inadequately investigated. In this study, we present a holistic population-level analysis of the human IGH locus, expanding upon existing references by analyzing 16 human genome assemblies from varied ancestries, including African, European, Asian, Saudi, and Mixed (African, European, Native American) backgrounds. Through our examination of both maternal and paternal haplotypes, we uncover novel IGH alleles, copy number variations (CNVs), and polymorphisms, notably within the variable (IGHV) and constant (IGHC) gene regions. Our findings reveal extensive previously unexplored genetic variability in the constant region and distinct CNV forms across individuals. Notably, this research contributes to an enriched and updated IMGT® reference IGH dataset and lays the groundwork for a comprehensive IMGT® haplotype database that can support future studies in population-specific immune profiles and autoimmune susceptibility. Such a resource promises to propel personalized immunogenomics forward, with exciting applications in cancer immunotherapy, COVID-19, and other immune-related diseases. By illuminating the evolutionary forces that sculpt immune diversity across human populations, it opens new avenues for understanding and tailoring immune responses.

Keywords: IMGT, immunogenetics, immunoinformatics, immunoglobulin (IG), antibody, germline, IGH repertoire, copy number variation (CNV), haplotype, population patterns, allotype

Combinatorial DNA-Pools targeted-sequencing as a robust cost-effective method to detect rare variants: analysis strategy and application to dilated cardiomyopathy genetic diagnosis.

Claire Perret

Perret Claire (1), Proust Carole (2), Ader Flavie (1, 3), Has Jan (4), Pruny Jean-François (5), Isnard Richard (1, 6), Richard Pascale (1, 3), Trégouët David-Alexandre (2), Charron Philippe (1, 6), Villard Eric (1)

1 - INSERM UMRS1166- ICAN (France), 2 - INSERM UMRS1219 (France), 3 - UF cardiogénétique et myogénétique moléculaire et cellulaire, Hôpital Pitié Salpêtrière (France), 4 - Department of Internal Medecine III- University of Heidelberg- Germany (Germany), 5 - Département de génétique, centre de référence maladies cardiaques héréditaires - Hôpital Pitié-Salpêtrière (France), 6 - Département de cardiologie - Hôpital Pitié Salpêtrière (France)

Abstract

Background: We developed a cost-effective NGS method based on combinatorial DNA pooling to efficiently detect rare variants. The originality of our approach lies in the development of a specific data analysis method that accurately discriminates true rare variants from false positives within pools. The experimental workflow and data analysis were validated by comparison with standard simplex sequencing technology. This method was applied to the genetic diagnosis of dilated cardiomyopathy (DCM). Methods: 96 non-indexed 8-DNA pools were constituted from 384 DNAs, each being included in a unique pair of concordant pools (sequenced twice though). Exonic regions of a cardiomyopathy gene panel were captured and sequenced by NGS at 500X per pool. Variant calling was performed using Freebayes algorithm, and the resulting QUAL value was compared between variants detected in concordant pools (true variants) and non-concordant pools (false positives) to accurately identify true variants. To benchmark performance, 50 DNA samples were sequenced using standard simplex NGS. Results: Freebayes parameters QUAL=12 was found to be discriminative between false-positives and true variants. Using this threshold, our pool-sequencing method successfully detected 96% of rare variants identified in DNA-simplex. Applied to 384 DCM patients, this approach identified 100 pathogenic ACMG class 4 and 5 variants (26%). Conclusion: We report an innovative, cost-efficient (4-fold cost reduction) NGS pooling method with a dedicated variant analysis strategy that enables accurate detection of rare variants. This approach might be applied more broadly for cost-effective genetic diagnosis in various rare diseases.

Keywords: pool_DNA, NGS, cardiomyopathie

Long-Read RNA sequencing in cardiomyopathies: a new approach for genetic diagnostic with strong potential ?

Laëtitia Rialland

Rialland Laëtitia (1, 2), Legrand De Milleville Elisabeth (1), Madry Hélène (3), Mohand Oumoussa Badreddine (3), Pruny Jean-François (4), Charron Philippe (1, 4), Richard Pascale (1, 2), Villard Eric (1)

1 - INSERM UMRS1166, Faculté de Médecine (France), 2 - Centre de génétique et de cytogénétique, Unité fonctionnelle de cardiogénétique et myogénétique moléculaire et cellulaire, Hôpital Pitié Salpetrière (France), 3 - UMS PASS P3S (France), 4 - Département de génétique, centre de référence maladies cardiaques héréditaires (France)

Abstract

Background: Dilated cardiomyopathies (DCM) are inherited diseases, for which genetic diagnostic conditions patient follow-up and family care (major gene TTN). However, current DNA Short Read (SR-DNA-Seq) based sequencing diagnostic yield is only 15-25% suggesting methodological limitations. This lack might be mitigated by RNA Long-Read sequencing (LR-RNA-Seq) which detects elusive DNA rearrangements and cryptic aberrant splicing causative mutations. We aim to evaluate LR-RNA-Seq for detection of Single Nucleotide Variants (SNVs) in coding exons and short/complex DNA Structural Variants (SV). Material and methods: Cardiac RNA from 26 DCM cases were sequenced on a PromethION24 (ONT). SNVs and SV were called using a LR dedicated bioinformatic tool (Clair3). True variants calling thresholds were inferred from replication using DNA Sanger. Results: Samples were covered with 11.6M reads (mean size:1094pb). Among 33 LR-RNA-seq variants (Clair3 Quality Score :1- 20), 11 were confirmed in DNA, allowing to define key parameters witnessing for high prior probability variants. All 3 causal SNV already diagnosed were detected. We detected 2 new causative truncating variants in the TTN gene in 2 other patients. Finally, we identified rare monoallelic variants overlapping with TTN locus, suggesting hemizygosity confirmed by Genomic DNA qPCR. Overall, the diagnostic yield of LR-RNA seq was 23% (6/26), at the upper limit of the reported yields for DCM. Conclusion: LR-RNA-seq appears to be an efficient method to detect coding variants in expressed genes, as well as SV. Ongoing analysis of splicing and NMD-related variants in our LR-RNA-seq data might further increase diagnostic yield, potentially representing a key strategy to complement SR-DNA-Seq in genetic diagnosis.

Keywords: Long Read, RNA, seq, Genetic diagnostic, Cardiomyopathies

Innovative insights on the genetic architecture of the human plasma proteome through meta-analysis of English and Italian protein Quantitative Traits Loci studies

Solène Cadiou

Cadiou Solène (1), Mapelli Alessia (1), Pontali Giulia (1), König Eva (2), Filosi Michele (2), Ghasemi-Semeskandeh Dariush (1), Ferolito Brian R. (3), Massi Michela (1), Cuccuru Gianmauro (1), Jiang Xiyun (4), Rainer Johannes (2), Pramstaller Peter (2), Pattaro Cristian (2), Pereira Alexandre (3), Di Angelantonio Emanuele (1, 4), Giambartolomei Claudia (1, 3), Fuchsberger Christian (2), Butterworth Adam S. (4)

1 - Human Technopole (Italy), 2 - Institute for Biomedicine, Eurac Research (Italy), 3 - Veterans Affairs Healthcare System (United States), 4 - University of Cambridge (United Kingdom)

Abstract

Background: Circulating plasma proteins are critical disease markers and drug targets, yet the genetic factors influencing their inter-individual variation are not fully understood. Here we provide an updated characterization of the genetic architecture of the human plasma proteome, leveraging the 7k aptamer-based protein platform at scale for the first time. Methods: Plasma protein levels from 7143 proteins were measured using the SomaLogic 7k assay in two European cohorts: the UK-based INTERVAL study (9251 samples) and the Italian CHRIS cohort (Cooperative Health Research in South Tyrol, 4194 samples). Genome-wide protein quantitative trait locus (pQTL) analyses were conducted independently in each study, followed by meta-analysis. We identified regional associations, performed conditional analysis to select independent signals, and tested them for colocalization across proteins. To understand trans mechanisms shared across proteins, we spatially characterize signal densities across chromosomes to identify hotspots and represented within-hotspots colocalization results with protein networks. Community detection algorithms disaggregate the hotspots in different protein modules potentially linked to common regulation mechanisms. Mendelian Randomization (MR) and colocalization with 2011 traits were additionally conducted using an external dataset comprising meta-analyses of European ancestry samples from the Million Veteran Program, UK Biobank, and FinnGen biobanks. We tested our cis-pQTLs for colocalization with expressionQTL in various tissues. Results: We identified 7870 significant pQTL (1784 cis and 6086 trans), including 3171 novel associations. Newly assessed proteins were less detectable and less likely to show cis associations compared to proteins assessed previously. 7994 significant MR associations were found, with 729 possible new opportunities of drug retargeting. 5472 cis-pQTLs showed at least one significant colocalization with cis-eQTLs. Conclusion: Our study, the first of its scale across European populations, provides new actionable insights in the genetic architecture of the human proteome and gives perspective on the benefits of expanding proteomic assays for genomic research.

Keywords: meta, analysis, pQTL, proteome, Mendelian Randomization, colocalization

Lifting the veil on Challenging Medically Relevant Genes

Victor Grentzinger

Grentzinger Victor (1, 2), Palmeira Leonor (2), Durkin Keith (1, 2), Artesi Maria (1, 2), Charloteaux Benoit (2), Dideberg Vinciane (2), Bours Vincent (1, 2)

1 - GIGA - Human Genetics (Belgium), 2 - Centre Hospitalier Universitaire de Liège (Belgium)

Abstract

While the cost of DNA sequencing has never been cheaper, a number of genetic diseases remain difficult to diagnose. Nearly 400 medically relevant genes are still challenging to characterize due to the complex nature of their sequence. This complexity can arise from a variety of factors, such as the existence of at least one pseudogene, a large Short Tandem Repeat region or a Variable Number Tandem Repeat region. These genes are classified in the so-called NGS and/or Sanger dead zone. As such, the access to reliable and cost-effective genetic tests is limited. To resolve this issue, we decided to focus on improving the characterization of the following genes by using long-read sequencing: PKD1/PKD2, responsible for Autosomal Dominant Polycystic Kidney Disease, and FLG, involved in Atopic Dermatitis. For PKD1/PKD2 genes, we amplified their sequence by long-range PCR before sequencing the products by Oxford Nanopore Sequencing. We were able to retrieve all variants previously confirmed by Sanger sequencing on 34 samples with ADPKD. For FLG, while investigating the 23 publicly available PacBio HiFi data of the 1000 Genome project, we identified new undescribed alleles in African samples. To determine if these variations are population specific, we analyzed 1111 additional public samples with long-read data. We discovered more than 15 novel alleles mostly from Sub-Saharan populations. With the knowledge of their existence, we hope to better characterize the FLG gene, especially for African and African-descent populations. We also investigated, in our cohort of public data, the MUC1 and SMN1/SMN2 genes, responsible respectively for Autosomal Dominant Tubulointerstitial Kidney Disease and Spinal Muscular Atrophy. Our next goal is to design cost efficient techniques to improve the sequencing of these challenging medically relevant genes in a clinical setting.

Keywords: Challenging Medically Relevant Genes, genetic diagnosis, long, read sequencing, kidney diseases, eczema, spinal muscular atrophy, bioinformatics

Searching for biologically consequential and inconsequential miRNA/target interactions using the evolutionary history of vertebrate miRNA genes

Hervé Seitz

Seitz Hervé (1)

1 - Institut de génétique humaine (France)

Abstract

MicroRNAs ("miRNAs") repress target mRNAs in a sequence-specific manner: targets tend to be recognized when they exhibit a perfect match to the miRNA's "seed" (nt 2–7 of the miRNA), especially if the seed match is located in the 3´ UTR. Because the seed is so short, seed matches are very frequent in the transcriptome. In order to predict true miRNA targets, current methods select seed matches which have been conserved in evolution. But even phylogenetically conserved seed matches are very common, and each mammalian miRNA is predicted to target hundreds of distinct genes. The current consensus therefore states that miRNAs control biological phenotypes through the coordinated regulation of large gene networks. Yet several observations suggest that this view is inaccurate. In vivo genetics rather show that, when experimentally assessed, just a very limited number of targets (1 or 2 targets) appear to control most of the phenotype. Other targets are really bound and repressed by the miRNA at the microscopic scale, but their repression appears inconsequential at the macroscopic scale. In order to solve that paradox, we interrogated the causes of the phylogenetic conservation of miRNA seed matches in vertebrate 3´ UTRs. When a miRNA family disappears in a given clade, its 3´ UTR seed matches tend to still be selectively conserved relatively to the rest of the UTR. These data indicate that most conserved miRNA seed matches are actually conserved for miRNA-independent reasons. This explanation reconciles molecular biology, comparative genomics and in vivo genetics. It is also likely to be generalizable to other gene regulators besides miRNAs (e.g., transcription factors, RNA-binding proteins), whose molecular specificity and whose repressive effect are comparable. These notions therefore question the definition of "gene regulation", which is commonly described at the molecular level, but could be more aptly described at the macroscopic level.

Keywords: miRNA

Impaired RNA Polymerase II Elongation Reveals Novel Molecular Mechanisms in Multiple Sclerosis

Christian Muchardt

Muchardt Christian (1)

1 - Institut de Biologie Paris-Seine (IBPS), CNRS UMR 8256, Biological Adaptation and Ageing, Sorbonne Université, 75005, Paris, France. (France)

Abstract

Multiple sclerosis (MS) is an autoimmune and inflammatory disease with a largely unknown etiology. Notably, several single nucleotide polymorphisms (SNPs) associated with increased MS risk affect genes encoding regulators of RNA Polymerase II (RNAPII) pause-release and elongation, including subunits of the integrator complex, NELF, and DSIF. However, the functional connection between RNAPII elongation dynamics and MS pathology remains unexplored. To address this, we performed high-depth RNA-sequencing on monocytes from 19 MS patients and 7 symptomatic control donors. Our analysis revealed that a significant subset of MS patients exhibited transcriptional signatures indicative of reduced integrator complex activity. Specifically, we observed increased enhancer RNA (eRNA) length and accumulation, potentially explaining the elevated levels of human endogenous retrovirus (HERV) transcripts frequently reported in MS patients. Additionally, integrator dysfunction correlated with increased RNAPII initiation but impaired elongation, leading to the preferential expression of short genes and downregulation of longer genes. This transcriptional imbalance aligns with the elevated expression of short pro-inflammatory genes and the disrupted expression of long genes, such as those encoding solute carrier (SLC) proteins involved in transmembrane transport. Our findings thus reveal a unifying mechanism allowing to explain multiple transcriptional manifestations of MS.

Keywords: Multiple sclerosis, RNA, seq

Identifying causal cell types for human diseases and risk variants from candidate regulatory elements

Artem Kim

Kim Artem (1), Gazal Steven (1)

1 - University of Southern California (United States)

Abstract

The SNP-heritability of human diseases is extremely enriched in candidate regulatory elements (cREs) from disease-relevant cell types. A critical next step is to infer if these enrichments are driven by multiple causal cell types, and to understand if individual variants impact disease risk through a single or multiple of these. Here, we propose CT-FM and CT-FM-SNP, two methods accounting for cREs sharing across cell types to identify independent sets of causal cell types for a trait and for its candidate causal variants, respectively. We applied CT-FM to 63 GWAS summary statistics (average N = 417K) using nearly one thousand cRE annotations, primarily coming from ENCODE4. CT-FM inferred 79 sets of causal cell types with corresponding SNP-annotations explaining a high fraction of trait SNP-heritability (~2/3 of the SNP-heritability explained by the largest cRE ressources), and identified 14 traits with multiple independent sets of causal cell types uncovering previously unexplored cellular mechanisms in schizophrenia, auto-immune diseases, height and BMI. We applied CT-FM-SNP to 39 UK Biobank traits, and predicted high confidence causal cell types for 2,798 candidate causal non-coding SNPs. Our results suggest that most SNPs impact a phenotype through a single set of cell types, while pleiotropic SNPs might target different cell types depending on the phenotype context. Altogether, CT-FM and CT-FM-SNP shed light on how genetic variants act collectively and individually at the cellular level to impact disease risk.

Keywords: GWAS, fine, mapping, heritability, complex traits

Multi-modal learning methods for single-cell data integration

Laura Cantini

Cantini Laura (1)

1 - Paris Artificial Intelligence Research Institute (France)

Abstract

Single-cell RNA sequencing (scRNAseq) is revolutionizing biology and medicine. The possibility to assess cellular heterogeneity at a previously inaccessible resolution, has profoundly impacted our understanding of development, of the immune system functioning and of many diseases. While scRNAseq is now mature, the single-cell technological development has shifted to other large-scale quantitative measurements, a.k.a. 'omics', and even spatial positioning. In addition, combined omics measurements profiled from the same single cell are becoming available. Each single-cell omics presents intrinsic limitations and provides a different and complementary information on the same cell. The current main challenge in computational biology is to design appropriate methods to integrate this wealth of information and translate it into actionable biological knowledge.In this talk, I will discuss two main computational directions for multi-omics integration, currently explored in my team: (i) joint dimensionality reduction to study cellular heterogeneity simultaneously from multiple omics and (ii) multilayer networks to integrate a large range of interactions between the features of various omics and isolate the regulators underlying cellular heterogeneity.

Keywords: scRNA

pyROMA, a python software for representation and quantification of module activity from single cell and bulk transcriptomic data.

Altynbek Zhubanchaliyev

Zhubanchaliyev Altynbek (1, 2), Bonnet Eric (3), Najm Matthieu (4), Martignetti Loredana (5)

1 - Computational Systems Biology group (France), 2 - FIRE PhD Doctoral School (France), 3 - Université Paris-Saclay, CEA, CNRGH (France), 4 - Computational Systems Biomedicine Lab (France), 5 - Computational Systems Biology group (France)

Abstract

High-dimensional biological data, such as single-cell transcriptomic profiles, offer unprecedented insights into tissue biology and disease mechanisms through their high resolution and precision. However, such data also poses significant analytical challenges due to noise and high dimensions. A crucial issue is to translate these complex data into actionable biological insights, for instance by quantifying the activity of a set of genes related to a biological pathway. Furthermore, the quantification needs to be assessed with a robust statistical significance to be useful. While recent advances in neural networks have demonstrated impressive power for various single-cell tasks, linear methods continue to serve as essential workhorses due to their reliability and straightforward interpretability (1). In this work, we introduce pyROMA, a single-cell adaptation of the original ROMA (Representation and Quantification of Module Activity) framework (2). Initially developed for bulk transcriptomics (3), ROMA quantifies pathway activity (e.g., Reactome pathways) and compares that activity against a null distribution of random gene sets of the same size, providing accurate estimates of statistical significance. We have reimplemented this approach in Python, optimizing and adapting it for single-cell datasets while ensuring seamless integration with widely used analysis pipelines such as scanpy (4). To validate our method, we applied pyROMA to different datasets, including datasets related to the study of cystic fibrosis at single-cell resolution (5). Our preliminary results reveal significant dysregulation in both inflammatory pathways and those activated due to chronic inflammation in cystic fibrosis samples, offering novel insights into the disease's molecular mechanisms. We show that pyROMA produces results that are consistent with previous implementations and identifies key disease-relevant pathways, while efficiently processing large-scale single-cell data. Collectively, these findings position pyROMA as a powerful, interpretable, and reliable computational tool for analyzing single-cell transcriptomic data. *References are in the supplementary file.

Keywords: Single, cell transcriptomics, Human genomics, Module activity quantification, Python software, Bulk transcriptomics, Pathway analysis, Computational biology, Cystic fibrosis, Gene set

Imagine the Medicine of the Future Now

Mickaël Ménager

Ménager Mickaël (1)

1 - Imagine - Institut des maladies génétiques (IHU) (France)

Abstract

In this talk, we will discuss the current state of the art regarding the use of single-cell multi-OMICs data, in research projects links to pathologies and diseases, with a specific focus on two rare genetic diseases, both characterized by an uncontrolled and excessive production of type I IFN: Aicardi Goutières Syndrome (AGS) and Sting Associated Vasculopathy with onset in Infancy (SAVI). We will go through the positive and yet to be improved aspects of single-cell data both at the wet (bench) and dry (computational analyses). We will also discuss on more direct application of single-cell multi-OMICs data towards diagnostic, prognostic and therapeutic: What can already be done, what is missing to start using single-cell data with direct feedback to clinicians and patients and how to integrate molecular data with already existing methods of diagnostic. There will be a special focus on the needs of spatial multi-OMICs data and on how to implement them into personalized medicine.

Keywords: single cell, multiomics

Early COPD single-cell and spatial transcriptomics

Morgane Fierville

Fierville Morgane (1, 2)

1 - Institut de pharmacologie moléculaire et cellulaire (France), 2 - Signal, Images et Systèmes (France)

Abstract

Chronic obstructive pulmonary disease (COPD) is one of the main causes of death in world (WHO), either alone or through an increased disposition to lung cancer. We built COPD dataset based on human biopsies collected in patients at early stages of the disease. Single cell and spatial transcriptomics approaches were used in order to better delineate the molecular drivers of the bifurcation toward the disease. Spatial transcriptomics (ST) analyzed large tissue sections (up to 3cm2), representing a total of about one million cells on which the expression of 5000 genes was assessed at a single-cellular resolution. By measuring cell-cell interactions and subcellular RNA expression in each cell type, we were able to define several hallmarks of the disease, such as the imbalance of secretory versus multiciliated cells in the airways. Individual cells were annotated using scMusketeers, a modular deep learning model that was keen at the identifying rare cell types and reducing batch effect, based on a reference annotation from the Human Lung Cell Atlas (HLCA). We analyze more particularly airway zones, after developing a method to unfold curved local structures. We derived from this approach a quantification of the spatial distribution of the cells and transcripts along or perpendicular to the two principal axes of the bronchi.

Keywords: spatial transcriptomics, human airways, copd, imaging, bioinformatics

Single-nucleus transcriptomic analysis of ageing in the mouse lemur prefrontal cortex

Clémence Su

Su Clémence (1), Dupuis Léo (2), Derbois Céline (1), Petit Fanny (2), Garcia Lolie (3), Deleuze Jean-François (1), Hirbec Hélène (4), Dhenain Marc (3), Bonnet Eric (1)

1 - Centre National de Recherche en Génomique Humaine (France), 2 - Laboratoire des Maladies Neurodégénératives - UMR 9199 (France), 3 - Laboratoire des Maladies Neurodégénératives - UMR 9199 (France), 4 - Institut de Génomique Fonctionnelle - Montpellier GenomiX (France)

Abstract

Age-related diseases will be a major challenge of the 21st century. Transcriptomic analyses are key to identify novel mechanisms leading to pathological aging. However, studies in primates, which serve as better translational models for human pathologies, remain sparse. The gray mouse lemur (Microcebus murinus) is a small lemur primate with a maximum life span of 12 years. It represents a unique model for studies related to aging. Here, we performed single-nucleus RNA sequencing (snRNA-seq) of prefrontal cortex samples from middle-aged (4 year-old, n=2 males) and aged (10 year-old, n=5 males, 2 females) animals, using the 10X Genomics platform. After quality control, we obtain a total of ~85,000 nuclei, with a range of ~6,500 to ~12,500 nuclei per sample and a median number of genes detected per nuclei of ~1900. Different algorithms tested for data integration provided similar and consistent results. Annotation was performed using a combination of marker genes, previous Microcebus murinus annotations and prediction algorithms. Neuronal types were dominant, but glial cells (microglia, astrocytes and oligodendrocytes) were also well represented, with an approximate neuron-to-glia ratio of 70/30. Differential gene expression analyses comparing young and aged animals identified dozens or hundreds differentially expressed genes (DGEs) depending on the cellular type. Typical age-related pathways such as oxidative stress and inflammatory responses were enriched in DGEs.

Keywords: Mouse lemur, Prefrontal cortex, Aging, single nucleus RNA, seq, Bioinformatics.

Session: Genetics and Pathologies

Poster #01: A Narrative Review on BRCA Gene Mutations in the Bangladeshi Breast Cancer Patients

Mahin Hasan

Hasan Mahin (1, 2), Tabassum Nuzhat (2), Maruf Abdullah Al (2)

1 - Department of Genetic Engineering and Biotechnology, University of Dhaka (Bangladesh), 2 - College of Pharmacy, University of Manitoba (Canada)

Abstract

Background: In Bangladesh, breast cancer is the second most frequent type of cancer and is growing alarmingly among women. About 5–10% of breast cancer are inherited and linked to BRCA1 and BRCA2 gene mutations. A mutational profile is essential for personalized treatment. Despite their utility, BRCA genetic tests are rarely done in Bangladesh, a least-developed country, prompting this review to summarize the BRCA gene mutations in this population. Methodology: We searched the literature for the spectrum of BRCA1 and BRCA2 gene mutations in the Bangladeshi population with breast cancer. We narratively summarized published research on the topic. Results: Six articles were included for data extraction. Methods used to detect mutations included polymerase chain reaction-restriction fragment length polymorphism, targeted sequencing using an Illumina Miniseq sequencer, and validation through Sanger sequencing. BRCA1 (rs80357713, rs80357906) and BRCA2 (rs11571653) polymorphisms were found to be significantly associated with breast cancer in the Bangladeshi population in a case-control study that included 310 breast cancer patients and 250 healthy controls. Another study that analyzed BRCA1 exon2 had observed a wild-type sequence and concluded that BRCA1 185delAG mutation may not have a strong recurrent effect on breast cancer. Insertions, deletions, and single-nucleotide substitution were observed on BRCA1 exon2 in another study on 50 breast cancer patients. Three nonsynonymous and two synonymous mutations on BRCA1 exon 2 (n = 50) and three novel mutations on exon 11 were observed (n = 65) in this population. Finally, two frameshift deletions in BRCA2 were observed in 43 breast cancer patients. Conclusion: With drug authorities allowing medications designed to target particular genetic abnormalities, personalized medicine is becoming more important in cancer treatment. However, low-resource nations like Bangladesh struggle because they lack the infrastructure, investment, and skilled workers needed to do genetic testing, which compromises the vital care that cancer patients need.

Keywords: Breast Cancer, BRCA, Bangladeshi Population

Poster #02: A Needle in a Haystack: Improving Genetic Analysis of Challenging Medically Relevant MUC1 Gene

Victor Grentzinger

Grentzinger Victor (1, 2), Palmeira Leonor (2), Durkin Keith (1, 2), Artesi Maria (1, 2), Charloteaux Benoit (2), Dideberg Vinciane (2), Bours Vincent (1, 2)

1 - GIGA - Human Genetics (Belgium), 2 - Centre Hospitalier Universitaire de Liège (Belgium)

Abstract

Autosomal Dominant Tubulointerstitial Kidney Disease (ADTKD) is a progressive tubulointerstitial fibrosis and tubular atrophy leading to End-Stage Kidney Disease. The disease has a prevalence of 0.7 to 4 per million in the United States and Ireland. Multiple mutations on various genes can induce this disease, with specific symptoms and development at various stages of life. One of these genes, MUC1, is known for its frameshift variant in exon 2 leading to an early stop codon, inside a Variable Number Tandem Repeat (VNTR) region of a 60-mer repeat. This translates into a truncated, misfolded Mucin-1 protein whose accumulation leads to ADTKD. Due to the complex VNTR region, the usual method for ADTKD-MUC1 genetic diagnosis is to screen exclusively for this frameshift mutation. However, this method does not characterize the VNTR region, and other potential pathogenic variants may be overlooked. To ensure that we will not miss any of them, we want to improve the characterization of MUC1. To that end, we analyzed publicly available long-read data from the 1000 Genome project, to determine if long-read can overcome the limitation of short-read. To do so, we generated multiple consensus sequences of various sizes of the MUC1-VNTR, and aligned our data against them, to determine if we could have a correct characterization of the number of repetition for each sample. The next step is to develop a method to sequence the DNA of ADTKD patients with long-read platforms, and accurately characterize their MUC1 sequence to identify pathogenic variations in a clinical setting.

Keywords: MUC1, Challenging Medically Relevant Genes, genetic diagnosis, long, read sequencing, kidney disease, bioinformatics

Poster #03: Cellular functional tests of ARX variants provide further insights into a better understanding of genotype-phenotype correlations in male and female patients

Rasha Faraj

Faraj Rasha (1, 2), Farrugi Audrey (3, 4), Dubos Aline (4), Schalk Audrey (5), Voisset Cecile (1, 2), Friocourt Gaëlle (1, 2, 6)

1 - Inserm UMR1078, Faculté de Médecine et des Sciences de la Santé, Université de Bretagne Occidentale, Brest, France (France), 2 - Inserm UMR1101, Faculté de Médecine et des Sciences de la Santé, Université de Bretagne Occidentale, Brest, France (France), 3 - Institut de Médecine Légale de Strasbourg, Fédération de Médecine Translationnelle de Strasbourg (FMTS), Université de Strasbourg, 11 Rue Humann, Strasbourg, France (France), 4 - Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CNRS UMR7104, INSERM U1258, Université de Strasbourg, Illkirch, France (France), 5 - Laboratoire de Diagnostic Génétique, Nouvel Hôpital Civil, Hôpitaux Universitaires de Strasbourg, Strasbourg, France (France), 6 - Centre de Référence des Déficiences Intellectuelles et polyhandicaps de causes rares. CHRU de Brest - Bâtiment 5 - 2 Avenue Foch, 29609 Brest Cedex, France. (France)

Abstract

Intellectual disability (ID) is a major neurodevelopmental challenge, with mutations in the X-linked ARX gene being a frequent cause. ARX, a master regulator of brain development, encodes a transcription factor orchestrating the delicate symphony of key target genes' expression essential for proper GABAergic neuron development and function. Variants in ARX cause a broad clinical spectrum, from asymptomatic female cases to severe neurodevelopmental disorders (NDDs) in male and female patients. This study aims to investigate the functional consequences of 15 ARX variants, including likely pathogenic and some novel de novo variants with atypical clinical presentations, such as cases of sudden infant death. Missense and frameshift variants were transiently expressed in N2a cells. Functional assays, including Luciferase reporter, Western blot, Immunofluorescence, and RT-qPCR, assessed the level of expression of the produced proteins, their localization, their interaction with known corepressors TLE1 and CtBP1, and their transcriptional activity on a few ARX-known targets. Results obtained showed that all variants tested exhibited disrupted transcriptional regulation, often linked to altered expression or localization. Interestingly, our findings reveal that some variants exert dominant negative effects, potentially offering a novel additional explanation for the skewed X-inactivation and the unexpectedly severe phenotypes observed in female patients. Altogether, these findings shed valuable light on the molecular mechanisms driving ARX-related NDDs and their remarkable phenotypic diversity. Through a comprehensive review of published ARX variants, we aim to establish clear genotype-phenotype correlations based on the nature of the variant (frameshift or missense), its localization, as well as the sex of the patient.

Keywords: Interneuronopathies, Epileptic Encephalopathy, non mediated mRNA decay

Poster #04: PFMG2025 - Integrating genomic medicine into the national healthcare system in France

Frédérique Nowak

Nowak Frédérique (1), Contributors Pfmg2025 (1)

1 - Institut Thématique Technologies pour la Santé, Inserm (France)

Abstract

Integrating genomic medicine into healthcare systems is a health policy challenge that requires continuously transferring scientific advances into clinics and ensuring equal access for patients. France was one of the first countries to integrate genome sequencing (GS) into clinical practice at a nationwide level, with the ambition to provide more accurate diagnostics and personalized treatments. In 2016, the French government launched the 2025 French Genomic Medicine Initiative (PFMG2025) which has so far focused on patients with rare diseases (RD), cancer genetic predisposition (CGP) and cancers. In accordance with the French Health Technology Assessment Agency, 77 pre-indications (68 for RD/CGP and 9 for cancers) have been selected. National guidelines were drawn to ensure optimal prescriptions and standardize medical practices. For each pre-indication, a flowchart was designed to define the eligibility criteria for GS. As of December the 31st 2024, the GS clinical laboratories (FMGlabs) received 29,779 prescriptions for RD/CGP patients with 21,700 results returned to prescribers (72.8 %), and 6,235 prescriptions for cancers patients with 5,833 results returned to prescribers (93.6 %). For RD/CGP, the diagnostic yield was 30.6%. PFMG2025 was also designed to provide a continuum between research and care, with the objective of sharing genomic data both at a national and international level. Secondary use of PFMG2025 data for research is one of the key objectives of the national facility for secure data storage and intensive calculation (Collecteur Analyseur de Données – CAD). Furthermore, France is part of the European '1+ Million Genomes' initiative which aims at sharing genomic data on a European scale. We outline the proactive implementation of GS in clinical practice highlighting the key elements of feasibility and accessibility of such a national initiative for French citizens and we present the main deliverables and the expected upcoming challenges, both for healthcare and research.

Keywords: Genomic medicine, PFMG2025, French genomic medicine initiative, Rare diseases, Cancer genetic predisposition, Cancers, Genome sequencing

Poster #05: Reclassifying NOBOX variants in Primary Ovarian Insufficiency cases with a corrected gene model and a quantitative framework

Sandrine Caburet

Caburet Sandrine (1, 2)

1 - Institut Jacques Monod (Université Paris Diderot, Bât. Buffon, 15 rue Hélène Brion, 75205 Paris cédex 13 France), 2 - UFR Sciences du Vivant [Sciences] - Université Paris Cité (France)

Abstract

The NOBOX gene, encoding a gonad-specific transcription factor with a crucial role in early folliculogenesis, is one of the major genes implicated in Primary Ovarian Insufficiency (POI). We refined and corrected the human NOBOX gene model using recent genomic and RNA-seq data from human fetal and adult gonads and 16 other mammalian species. The corrected NOBOX gene model led to the invalidation of two transcripts, including the one currently annotated as canonical. The two correct isoforms were expressed in fetal ovaries, and only the longest was present in adult testes. Rejecting the transcript considered until now as canonical implies that the variants reported in POI patients were incorrectly evaluated for pathogenicity. To reclassify all NOBOX variants, we set up a comprehensive and quantitative framework with updated variant frequencies from GnomAD4 and POI-adjusted parameters following improved ACMG/AMP guidelines. The reclassification of the 44 NOBOX variants reported in POI cases showed that only 14 variants are potentially causative for POI. This implies that the POI cases considered as solved because of wrongly-classified variants should be reanalyzed. Furthermore, functional assays performed using the proteins derived from the obsolete transcripts can logically be deemed as irrelevant. We also classified the 117 NOBOX variants reported in ClinVar and the 2613 ones present in GnomAD4. Contrary to the current idea, our results indicate that NOBOX should be considered as a recessive gene for POI. Our quantitative classification framework addresses the need for disease-specific and quantitative application of ACMG/AMP guidelines, and can be used to efficiently assess variants in other POI genes. In general terms, combining available and recent genomic and transcriptomic data from several species with a quantitative scoring framework provides an improved and efficient approach for validating gene models and accurately classifying gene variants, enabling a better molecular diagnosis of Mendelian disorders.

Keywords: NOBOX, Primary Ovarian Insufficiency, Female infertility, Ovary, Variant classification, Variant pathogenicity

Poster #06: Strategies for identifying causal mosaic mutations in rare diseases

Jules Lepont-Richez

Lepont-Richez Jules (1), Letexier Mélanie (1, 2), Gras Margaux (1), Aho Glele Ahouefa Printil (1), Viari Alain (1, 3), Turon Violette (1, 2), Eychenne Thomas (1), Deleuze Jean-François (1, 2)

1 - Centre de référence, d'innovation, d'expertise et de transfert (France), 2 - Centre National de Recherche en Génomique Humaine (France), 3 - Synergie Lyon Cancer-Platform of Bioinformatics-Gilles Thomas (France)

Abstract

Background/Objectives: Mosaic mutations in rare diseases are characterized by highly variable and potentially very low allele frequencies and heterogeneous tissue distribution. This often preclude new mutations detection and thus, the vast majority of hospital protocols aim at targeting known mutations with panel sequencing or digital PCR. Consequently, the France Genomic Medicine Plan 2025 (FMG2025) has highlighted the need for a standardized protocol for the discovery of new mosaic mutations in non-cancer diseases. Following this plan's request, The Reference Center for Technology, Innovation and Transfer (CRefIX) investigated the possibility of identifying low-frequency mutations using high depth whole-exome sequencing (WES). Materials & Methods: We designed artificial mosaicism samples using the HAP-1 cell line by diluting two CRISPR knockout cell lines for two known genes with the wild type isogenic cell line. We thus obtained samples of these two known mutations at frequencies of 10%, 5%, 1%, 0.5%, and non-diluted standards for each mutation and the wild type. All samples were used to perform WES and were sequenced on NovaSeq 6000, to target 600X coverage. Variant calling was performed by Mutect2 (GATK v4.5.0.0) and position screening by samtools mpileup (samtools v1.18). Results & Conclusions: The results showed the presence of both mutations down to frequencies of 1% and 0.5% for one of them when using mpileup on the precise positions. However, Mutect2 is only able to identify them at frequencies of 5% and 10%, whether in tumor-only mode or using the wild-type sample as a comparison. We have therefore demonstrated the technical feasibility of detecting mutations down to allelic frequencies of 1% with WES average 300X effective coverage. However, bioinformatics tools adapted to mosaic mutations are needed to identify them in real-life practice. We therefore plan to use this standardized data to benchmark available bioinformatics tools. Funding: ANR-18-INBS-0001 (French National Research Agency)

Keywords: mosaicism, low frequency mutations, high depth whole exome sequencing, PFMG2025

Poster #07: Targeted mRNA sequencing helps to classify variants affecting splicing in Hypertrophic Cardiomyopathies

Laëtitia Rialland

Rialland Laëtitia (1, 2), Blin Emilie (2), Perret Claire (1), Ader Flavie (2), De La Grange Pierre (3), Charron Philippe (4), Villard Eric (1), Richard Pascale (1, 2)

1 - INSERM UMRS1166, Faculté de Médecine (France), 2 - Centre de génétique et de cytogénétique, Unité fonctionnelle de cardiogénétique et myogénétique moléculaire et cellulaire, Hôpital Pitié Salpetrière (France), 3 - GenoSplice [Paris] (France), 4 - Département de génétique, centre de référence maladies cardiaques héréditaires (France)

Abstract

Background: Hypertrophic cardiomyopathies (HCM) are inherited cardiac diseases with an autosomal dominant transmission, among which MYBPC3 is the main causal gene responsible for haplo-insufficiency. RNA splicing variants in MYBPC3 appears to be a prevalent cause of HCM, however RNA analysis is not developed for these diseases because of cardiac tissue unavailability. Thus, these variants are often classified as Variant of Unknown Significance (VUS) and can't be use for clinical purposes. We propose an enrichment method to detect splicing aberrations in MYBPC3 cDNA causing cardiomyopathies from blood cells mRNA. Material and method: We selected 26 variants (16 intronics and 10 exonics) detected on DNA potentially affecting splicing. polyA+ RNA purified from venous blood cells was retro-transcribed, captured with optimized design and sequenced on NextSeq550. A specific bio-informatic pipeline was developed to automatically detect splicing events. Results: The gene MYBPC3 was very well covered and interpretable (RPM~6809, Mean depth~2947X). We detect a splice aberration for 19/26 (73%) of cases, consistent with their respective predictive score, among which 6 (32%) creates a novel junction; 8 (42%) modifies the proportional usage of annotated junctions and 5 (26%) leads to the retention of the entire intron. Then, we perform bio-informatic screening that was able to detect all pathogenic events, including intron retention even when no abnormal junction is associated. Conclusion: Targeted mRNA sequencing from blood cells helps to classify splice affecting variants, thus improving the yield of molecular diagnostic. This method, associated with a specific bio-informatic pipeline can be used as a screening approach and expanded to other cardiomyopathy genes.

Keywords: mRNA, seq, cardiomyopathies, targeted, blood

Poster #08: Title: GenEFCCSS: A resource for investigating genetic predispositions in in childhood cancers

Brice Fresneau, Florent De Vathaire

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 (2), Haddy Nadia (1, 2, 3), Rubino Carole (1, 2, 3), Guerrini Léa (1, 3), Dufour Christelle (1, 3), Minard Veronique (1, 3), Pacqument Hélène (5), Bourdeaut Franck (5), Winter Sarah (5), Adam-De-Beaumais Tiphaine (1), Lenez Laura (1), El-Fayech Chiraz (1), Blanché Hélène (6), Deleuze Jean-François (4), Génin Emmanuelle (2), Fresneau Brice (1, 2, 3), De Vathaire Florent (1, 2, 3)

1 - Institut Gustave Roussy (France), 2 - INSERM (France), 3 - University paris saclay (France), 4 - Centre National de Recherche en Génomique Humaine (France), 5 - Centre de recherche de lÍnstitut Curie [Paris] (France), 6 - CEPH (France)

Abstract

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 with available blood or saliva samples 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. Among those sequenced, 1518 received radiotherapy, and 2130 underwent chemotherapy for childhood cancer. 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 calling to generate a final VCF file for all samples. Variant annotation was performed using the ClinVar database 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.

Keywords: Germline Variants Gene Cancer Childhood

Poster #09: Understanding the link between autism and preterm births

Selin Korkmaz

Korkmaz Selin (1, 2, 3, 4), Cliquet Freddy (1), Leblond Claire (1, 2), Bourgeron Thomas (1, 2)

1 - Génétique Humaine et Fonctions Cognitives (France), 2 - Université Paris-Cité (France), 3 - R2D2 (France), 4 - ED3C (France)

Abstract

Autism is a neurodevelopmental disorder (NDD) characterized by repetitive patterns of behavior and interests and difficulties in social communication. Its prevalence approaches 2% of the population, with a substantial genetic contribution, accounting for over 87% heritability. We know that there are over 200 genes associated to autism and over 1,800 genes involved in neurodevelopmental disorders that may be implicated in autism. These genes are involved in various pathways related to synaptic function and chromatin remodeling. Preterm birth, defined as delivery before 37 weeks of gestation, affects approximately 10% of the population. Interrupted brain maturation from preterm birth increases the consequences of altered functional development. A meta-analysis revealed that preterm infants have a 30% higher risk of autism compared to those born full-term. Moreover, the prevalence of autism in preterm infants increases with the severity of prematurity. Preterm birth has been linked with increased likelihood of autism, but the causality remains unclear. Prematurity could be one independent adversity factor that will act in addition to genetic factors associated with autism. Another hypothesis is that the context of prematurity will reveal new susceptibility factors not yet detected in individuals born at full-term. Preterm individuals were excluded from the initial cohorts recruited to identify genes for autism. Thus, there are no study comparing the profiles of autistic individuals with and without prematurity. Considering the increased likelihood of autism in preterm birth, it is essential to study autism in regard to prematurity. The aim of this study will be to highlight the clinical and genetic differences between individuals with autism only, individuals with prematurity only, and individuals with both prematurity and autism. In this study, we are using the large-scale SPARK cohort, which comprises 351,401 individuals. Additionally, deep phenotyped cohorts such as LEEP and InovAND will be used, particularly to investigate impacts through imaging.

Keywords: autism/premature births/human genetics/bioinformatics

Poster #10: Identification and characterization of novel non coding transcripts in sepsis patients

Alaa Mslmane

Mslmane Alaa (1), Mambu Mambueni Hendrick (1), Fleuriet Jérôme (2), Heming Nicholas (3), Annane Djillali (4), Garchon Henri Jean (5), Records Rhu

1 - Inserm, UMR 1173, infection et inflammation, université Paris-Saclay, UVSQ, 78180 Montigny-le-Bretonneux, France (France), 2 - Department of Intensive Care, AP-HP University Versailles Saint Quentin-University Paris Saclay, Garches, France (France), 3 - General Intensive Care Unit, Hopital Raymond-Poincare, Garches, France (France), 4 - General Intensive Care Unit, Hopital Raymond-Poincare, Garches, France (France), 5 - Inserm, UMR 1173, infection et inflammation, université Paris-Saclay, UVSQ, 78180 Montigny-le-Bretonneux, France (France)

Abstract

Long non-coding RNAs (lncRNAs), a growing class of RNA molecules of length > 200 nucleotides, are involved in various biological processes in health and disease, including in regulation of gene expression, chromatin remodeling and protein relocation. The field of lncRNAs is constantly evolving, with new RNAs being discovered notably in stress conditions such as cancer and myocardial infraction. In this context, we explored a potential role of lncRNAs in sepsis. Sepsis refers to a dysregulated immune response triggered by host infection, characterized with life-threatening organ dysfunctions. Despite being a major cause of global mortality and morbidity, specific diagnostic and therapeutic tools remain challenging. The main aim of our study was to identify and characterize novel lncRNAs in sepsis patients to ultimately discover useful biomarkers, and to contribute to the development of new therapies and improve patient outcomes. We performed bulk RNA sequencing on whole blood samples from hospitalized sepsis patient infected with SARS-COV-2. Samples (n=159) were collected at days 0, 7, and 14 post infection. Novel lncRNAs, or SEpsis-Associated Transcripts (SEATs), were identified, characterized (transcript length, number of exons) and classified based on genomic location, relative to protein-coding genes, cis-regulatory elements (CRM) and repetitive sequences. Finally, we conducted differential gene expression analysis to identify specific transcriptomic signatures between deceased patients and those surviving at day 90 following admission in ICU. We identified 32 SEATs corresponding to 13 genes, either mono- or multi-exonic. Expression of two of them was significantly upregulated in deceased patients. Two other SEATs were convergent with protein-coding genes. Moreover, we revealed an overlap between the newly identified SEATs and CRMs known to be involved in steroid hormone signaling. Finally, we identified 50 significant differentially expressed genes, among which 3 SEATs were found downregulated in surviving patients.

Keywords: Sepsis, Long non coding RNA, Transcriptomic

Poster #11: Deep Mendelian Randomization: explaining causality between traits at genome-wide scale

Mario Favre-Moiron

Favre-Moiron Mario (1), Nouira Asma (1), Verbanck Marie (1)

1 - U1331 - Oncologie computationnelle , Epidémiologie génétique des cancers (France)

Abstract

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.

Keywords: Mendelian Randomization, Pleiotropy, Cancer, Epidemiology, Causal Inference

Poster #12: Assessing the phenotypic variability of CADASIL cerebral angiopathy due to NOTCH3 p.R1231C mutation by comparing data from UK Biobank, an isolated population, and a hospital cohort

Matthieu Pluntz

Pluntz Matthieu (1, 2), Lambert Louis (3), Lebenberg Jessica (3, 4), Nutile Teresa (5), Ruggiero Daniela (5), Hervé Dominique (1, 4), Perdry Hervé (2), Tournier-Lasserve Elisabeth (1, 4, 6), Chabriat Hugues (3, 4, 7), Ciullo Marina (5), Leutenegger Anne-Louise (1)

1 - Inserm Université Paris Cité, NeuroDiderot, Paris (France), 2 - CESP Inserm U1018, Université Paris-Saclay, F-94807 Villejuif, France (France), 3 - Inserm UMRS1127 Paris Brain Institute (ICM), Paris (France), 4 - APHP, Translational Neurovascular Centre and CERVCO, Hôpital Lariboisière, Paris (France), 5 - Institute of Genetics and Biophysics A. Buzzati-Traverso, CNR, Naples (Italy), 6 - APHP, Service de génétique moléculaire Neurovasculaire, Hôpital Saint-Louis, Paris (France), 7 - Université Paris Cité, FHU Neuro-Vasc 2030, Paris (France)

Abstract

CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy) is an adult-onset hereditary small vessel disease caused by mutations in the NOTCH3 gene. Studies suggest that the mutations located on EGFr domains 1-6 are more severe than those on EGFr domains 7-34. The mutation p.R1231C, located on EGFr domain 31, is particularly prevalent in populations of European or Central South/West Asian descent. CADASIL's phenotypic expression varies both across and among populations. This study aims to compare clinical and imaging characteristics of p.R1231C mutation carriers identified in the UK Biobank with those from an isolated population in Italy and a hospital-based cohort in France. By integrating genomic, clinical, and neuroimaging data, we assess differences in disease characteristics across these groups as well as between carriers of p.R1231C and other NOTCH3 mutations. Clinical, imaging and genomic data are collected from three sources: UK Biobank, a long term prospective study of 500,000 individuals across the United Kingdom including N=255 p.R1231C carriers (among 973 NOTCH3 mutation carriers); a cohort of patients followed at CERVCO rare disease referral center in Lariboisière Hospital including N=17 p.R1231C carriers (among 357 index cases); and an isolated founder population in Cilento, southern Italy, including N=35 NOTCH3 mutation carriers, all carrying p.R1231C. This study leverages the large-scale UK Biobank dataset to contextualize findings from the more selected cohorts. Adjusting for vascular risk factors, we compare stroke occurrence, cognitive function, and imaging markers such as white matter hyperintensities, lacunar infarcts, and brain atrophy across populations. This will help understand the variability of CADASIL phenotypes. Grants: AARG-NTF-20-683992, ANR-20-CE17-0031. UKBiobank application #92392.

Keywords: UK Biobank, CADASIL, Phenotypic variability, Brain imaging, Stroke

Poster #13: Building Regionally Anchored French Population Genomic Panels for Better Insights into the Genetic Architecture of Diseases

Anthony Herzig

Herzig Anthony (1), Le Folgoc Gaëlle (1), Blanché Hélène (2), Marenne Gaëlle (1), Saint Pierre Aude (1), Zins Marie (3), Nowak Frédérique (4), Dina Christian (5), Redon Richard (6), Deleuze Jean-François (7, 8, 9), Study Group Popgen, Consortium Francegenref, Genin Emmanuelle (1)

1 - GGB (France), 2 - CEPH (France), 3 - Inserm-Paris Saclay University, University of Paris, Villejuif, France (France), 4 - Institut Thématique Technologies pour la Santé, Inserm (France), 5 - Institut du Thorax (France), 6 - Institut du Thorax (France), 7 - Centre National de Recherche en Génomique Humaine (CNRGH, France), 8 - Centre d'Etude du Polymorphisme Humain (CEPH, France), 9 - Centre de référence, d'innovation, d'expertise et de transfert (CREFIX, France)

Abstract

With the decreasing cost of whole-genome sequencing (WGS), several countries have begun developing national reference panels reflecting the genetic diversity of their local populations. In Europe, such efforts align with the Genome of Europe (GoE) project, which aims to provide access to >100,000 WGS representative of populations living in Europe. Prior to the GoE launch, two pilot projects were conducted in France to explore fine-scale genetic diversity within and between regions. Both projects prioritized individuals with all four grandparents born within a geographic area of less than 100 km. The first pilot, FranceGenRef, involved WGS of 856 individuals sampled from existing biobanks, focusing on limited French regions. To achieve broader geographic coverage, the second pilot, POPGEN, leveraged the Constances cohort. A total of 15,000 volunteers received salivary DNA extraction kits, with 10,250 successfully included and 9,772 genotyped using Illumina GSA. Based on genotyping data and ascendants' birthplaces, 4,000 individuals were selected for WGS. These pilot studies, along with new samples including minority populations via the France Genomic Medicine Initiative, will constitute the 17,000 WGS contribution to GoE. Here, we evaluate the effectiveness of this sampling design in capturing local genetic variation compared to publicly available databases. Furthermore, with support of exploratory analyses of haplotypes sharing, we discuss how such sampling approaches could aid in distinguishing local neutral variants from pathogenic variants, improving the diagnosis of rare diseases. These findings could provide valuable insights for the GoE project, refining sampling strategies and serving as a model for implementing similar pilots across Europe.

Keywords: genetic diversity, haplotype sharing, local ancestry, genetic architecture

Poster #14: ChoruMM: a versatile multi-components mixed model for bacterial-GWAS

Arthur Frouin

Frouin Arthur (1), Laporte Fabien (2), Hafner Lukas (3), Maury Mylène (3), Mccaw Zachary (4), Julienne Hanna (1), Henches Léo (5), Leclercq Alexandre (3, 6), Chikhi Rayan (1), Lecuit Marc (3, 6, 7), Aschard Hugues (8, 9)

1 - Department of Computational Biology, Institut Pasteur (France), 2 - Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France (France), 3 - Institut Pasteur, Université Paris Cité, Inserm U1117, Biology of Infection Unit, Paris, France (France), 4 - Insitro (United States), 5 - Department of Computational Biology, Institut Pasteur (France), 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 - Génétique Statistique (France), 9 - Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA (United States)

Abstract

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.

Keywords: GWAS, Linear Mixed Model, Bacteria, Non Human Genetic

Poster #15: Cross-methods GWAS summary statistics deconvolution

Sohane Aissa

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)

Abstract

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 decon- volution of complex pleiotropic associations and the inference of potential shared biological pathways across outcomes. Those methods vary broadly in their modelling, their implemen- tation, 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 molecu- lar 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.

Keywords: GWAS, Multi trait Studies, Pleiotropy, Genetic Correlation, Deconvolution, Matrix Factorization

Poster #16: Diversity of pharmacogenes in the different French regions

Marc Gros La Faige

Gros La Faige Marc (1), Study Group The Popgen (2), Génin Emmanuelle (2, 3), Herzig Anthony (2)

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

Abstract

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.

Keywords: Pharmacogenetics

Poster #17: GOLDogs: Association and impact of genomic point mutations and structural variations on canine longevity

Dimple Adiwal

Adiwal Dimple , Labadie Karine (1), Mottier Stéphanie (2), Cruaud Corinne , Guyon Richard (3), Houel Armel (4), Le Nézet Louis (2), Cadieu Édouard (4), Hoffmann Nicolaï (2), André Catherine (5), Aury Jean-Marc , Hedan Benoit (5), Derrien Thomas

1 - Genoscope - Centre national de séquençage [Evry] (France), 2 - Institut de Génétique et Développement de Rennes (IGDR) UMR 6290 (35000 Rennes France), 3 - Institut de Génétique et Développement de Rennes (UMR 6290, France), 4 - Institut de Génétique et Développement de Rennes (IGDR) UMR 6290 (35000 Rennes France), 5 - Institut de Génétique et Développement de Rennes (IGDR) UMR 6290, Rennes, France (France)

Abstract

The domestic dog (Canis lupus familiaris) represents an exceptional model for studying genotype-phenotype relationships due to its unique evolutionary history and breed diversity. While breed average lifespan is known to inversely correlate with body weight, the genetic basis of longevity variation remains poorly understood. The GOLDogs project aims to create an extensive comprehensive catalog of genetic variations across multiple dog breeds harbouring different longevity expectencies by combining innovative sequencing approaches and analyzing their impact on longevity. This large-scale genomic study employs a two-step strategy: First, low-pass (1X) whole genome sequencing of 540 aged dogs (20 dogs from each of 27 breeds) will enable genome-wide association studies (GWAS) to identify genomic regions linked to longevity, controlling for factors such as body weight, sex, and inbreeding. Second, long-read Oxford Nanopore/PacBio sequencing of 100 dogs (4 per breed) from 25 breeds will catalog structural variants, particularly focusing on transposable elements that may influence aging. Finally, the project will generate 25 breed-specific reference genomes using a combination of high depth long-read sequencing providing high resolution of breed-specific genomic architecture. The study leverages the unique Cani-DNA Biological Resource Center, created and managed by the team, containing over 32,000 samples from more than 300 breeds, with 3,530 samples having documented dates of death. This exceptional resource enables the selection of dogs with extreme and median longevity within breeds, allowing for robust statistical analyses. The project will create a valuable genomic resource for the scientific community while specifically investigating the genetic basis of aging in dogs. Understanding the genetic factors influencing canine longevity may provide insights into aging mechanisms conserved across mammals or species-specific, potentially benefiting both veterinary and human medicine and fundamental research.

Keywords: Longevity, Whole genome sequencing, Dog, model, Long, read sequencing, Structural variants, Aging, GWAS

Poster #18: Psoriasis: A Case Study on Using Biological Networks for Gene Discovery

Giann Karlo Aguirre Samboní

Aguirre Samboní Giann Karlo (1), Azencott Chloé-Agathe (1), Massip Florian (1), Lemoine Gwenaëlle (1), Molineros Julio (2)

1 - Mines Paris - PSL (École nationale supérieure des mines de Paris, France), 2 - Johnson and Johnson Innovative Medicine (United States)

Abstract

Biological networks are essential for illustrating interactions among nodes, such as genes and proteins. They enhance post-Genome-Wide Association Study (GWAS) analyses by facilitating the discovery of susceptibility loci with functional relationships. This study investigates the applicability of various network methods (including Hierarchical HotNet, SigMod, Heinz, and dmGWAS) on a biobank scale, utilizing Whole Exome Sequencing and SNP microarray data to improve gene identification in post-GWAS analyses. Each method combines p-values of individual variants into gene-level p-values using MAGMA, which aggregates variant-level association statistics. After obtaining gene-level p-values, each method analyzes a pre-existing biological network, here a protein-protein interaction network from BioGRID, to detect sub-networks significantly associated with psoriasis. While all methods aim to identify such sub-networks, they differ in their mathematical frameworks and approaches, including clustering, network propagation, tolerance inclusion, topology, and random walks. Our goal was to study the ability of these methods to discover genes associated with psoriasis, a chronic condition characterized by excessive keratinocyte proliferation and immune cell infiltration. We used genotype data from 500,000 patients in the United Kingdom Biobank (UKB), along with clinical information. To address the heterogeneity of the solutions from the network-based methods, we examined their overlap and calculated a stable consensus, termed the "Stable Consensus Network (SCN)," through a subsampling procedure. The resulting SCN allowed us to identify common susceptibility loci and pathways across different methods, providing a more comprehensive understanding of the genetic basis of psoriasis.

Keywords: GWAS, Biological Networks, Stable Consensus Network, Psoriasis, Whole Exome Sequencing

Poster #19: Scaling up variant prioritisation in the dark genome to improve rare disease molecular diagnosis

Gaëlle Marenne

Ogloblinsky Marie-Sophie C (1), Bocher Ozvan (1), Gros La Faige Marc (1), Génin Emmanuelle (1, 2), Marenne Gaëlle (1)

1 - Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest (France), 2 - Centre Hospitalier Régional Universitaire de Brest, F-29200 Brest (France)

Abstract

Large genetic heterogeneity of rare diseases and limited knowledge of the non-coding genome are major challenges for molecular diagnosis. The PSAP-genomic-region method was proposed to improve prioritisation of causal variants across the whole genome [1], leveraging pathogenicity scores and observed frequencies in the general population. The method was evaluated on different types of variants and its performance evaluated globally for coding and non-coding variants. In this work, we focused on the non-coding variants to explore perspectives and further improve their prioritisation. A total of 271 non-coding ClinVar variants (176 with dominant inheritance, 96 with recessive inheritance) were inserted in the whole-genome data of 533 non-Finnish-European from the 1000 Genomes Project phase 4. PSAP null distribution were derived on genomic regions defined all over the genome [2] based on observed variant frequencies in different populations from gnomAD-v3. Two different pathogenicity scores were compared: CADD score v1.6 and an adjusted version of the CADD score (ACS) that considers genomic function. We found that using the ACS highly improved the prioritisation of non-coding variants, especially the dominant ones: only 25% of the dominant non-coding ClinVar were prioritised in the top10 using CADD scores while 57% reached top10 using ACS. The performance at top10 for recessive non-coding ClinVar improved also but less: from 79% to 83%. This improvement was emphasised for splicing variants. Partitioning the genome based on function thus represents a promising whole-genome strategy. By providing a much-needed method to assess the likelihood that a non-coding variant identified in a patient's whole-genome data is more deleterious than variants observed in the general population, PSAP genomic-region addresses an important gap in rare disease research. This approach could enable the integration of the dark genome into molecular diagnosis, helping to uncover the role of previously overlooked variants. References: [1] 10.1002/gepi.22593; [2] 10.1371/journal.pgen.1009923

Keywords: rare disease, molecular diagnosis, variant prioritisation, non coding genome

Poster #20: Transitioning to DNAnexus

Gloria Benoit

Benoit Gloria (1), Henches Léo (1), Aschard Hugues (1, 2)

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

Abstract

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. In this work, we will compare multiple ways to run a classic Genome-Wide Association Study (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.

Keywords: GWAS, Cloud computing, Whole genome sequencing

Poster #21: Bioinformatic tools for the analysis of antibody repertoires

Gael Millot

Millot Gael (1, 2), Chappert Pascal (3), Bruhns Pierre (4)

1 - Bioinformatics and Biostatistics Hub (France), 2 - Unit of Antibodies in Therapy and Pathology (France), 3 - INSERM U1151 (France), 4 - Unité des Anticorps en Thérapie et en Pathologie, Institut Pasteur, UMR1222 Inserm (France)

Abstract

The immune system generates a vast repertoire of antibodies in response to infection or immunization. Exploration of such repertoire is made possible by sorting antigen-specific memory B cells by FACS, or, as recently published by our group, by sorting antigen-specific antibody-secreting plasmablasts and plasma cells by droplet-based microfluidic sorting (drop-seq). Such technologies allow to pair antigen specificity with antibody gene sequences (VH-JH and VL-JL) but imply to tackle ever growing datasets for the bioinformatic analysis of antibody repertoires. In agreement with the Adaptive Immune Receptor Repertoire (AIRR) standards and recommendations, we have developed several bioinformatic tools for analyzing the output of such repertoires, freely accessible to the scientific community and upgraded to integrate the complexity of VH-VL pairing in the analyses, tree generation and antibody maturation analyses. The one currently in development of the version 2 is Repertoire Profiler (https://gitlab.pasteur.fr/gmillot/repertoire_profiler). It benefits from the R immcantation solution (https://immcantation.readthedocs.io/en/stable/index.html) and performs: (1) annotation of mRNA sequencing of the immunoglobulin heavy (VH) or light variable (VL) region, (2) clustering of the annotated sequences into clonal groups, (3) visualization of the clonal groups in trees integrating potential metadata, like affinities for the antigen, (4) statistical description of the repertoire. Output files include the matrix count of the V and J allele usage (repertoire), that can be analyzed with our Comat tool (https://gitlab.pasteur.fr/gmillot/comat). Comat performs the 2 by 2 statistical comparison of repertoires of same dimension in batch and regroup the results of the batch into a multidimensional scale analysis. Sci Transl Med. 2024 doi:10.1126/scitranslmed.ado4463.

Keywords: Lymphocytes, Single Cell mRNA sequencing, Immunoglobulin Variable Region, repertoire

Poster #22: Identification of genetic susceptibility to develop invasive pneumococcal disease in children by whole-exome sequencing.

Morgane Gélin

Gélin Morgane (1, 2), Leman Claire (1, 3), Morin Martin (4), Durand Axelle (1), Rousseau Olivia (1), Gras-Leguen Christèle (2, 5, 6, 7), Lorton Fleur (2, 5, 6, 7), Toubiana Julie (8, 9), Thomas Caroline (10, 11), Martin Jérôme (1, 10, 12), Limou Sophie (4), Vince Nicolas (1), Launay Elise (2, 5, 6, 7)

1 - Centre de Recherche en Transplantation et Immunologie - Center for Research in Transplantation and Translational Immunology (France), 2 - Department of Pediatrics and Pediatric Emergency, Hôpital Femme Enfant Adolescent (France), 3 - Department of Nephrology and dialysis (France), 4 - Centre de Recherche en Transplantation et Immunologie - Center for Research in Transplantation and Translational Immunology (France), 5 - Center of Clinical Research Femme Enfant Adolescent (France), 6 - Obstetrical, Perinatal, and Pediatric Epidemiology Research Team (Epopé), Center of Research in Epidemiology and Statistics (France), 7 - Nantes Université (France), 8 - Department of General Pediatrics and Pediatric Infectious Diseases (France), 9 - Pasteur Institute, Biodiversity and Epidemiology of Bacterial Pathogens (France), 10 - Centre De Référence constitutif Des Déficits Immunitaires Primitifs, French national registry of patients with primary immunodeficiencies (France), 11 - Department of hematology and pediatric immunology, Hôpital Femme Enfant Adolescent (France), 12 - Laboratoire d'immunologie, CIMNA (France)

Abstract

Invasive pneumococcal diseases (IPDs) remain severe infections in the pediatric population. A deeper understanding of the molecular factors involved in the development of IPDs could lead to improved treatment and prevention, leading to better outcome. Our study aimed to identify genetic factors implicated in the development of IPDs through a rarely used large-scale genetic strategy by performing whole-exome sequencing (WES) on children admitted to the pediatric intensive care unit with IPD, but without any known innate error of immunity. We included 36 patients and 70 controls. We identified the NUPR2 gene as significantly more frequently mutated in cases than in controls (p=1.7e-06) using a burden test method including both rare and common variants (n=1,779,391). Interestingly, this gene is implicated in the NFkappaB pathway which regulates inflammatory processes. We then restricted our analysis only to very rare and possibly disease-causing variants (n=2,641). We identified the NUPR2, ZNF697, and FZD2 genes as more frequently mutated in cases than in controls (p=6.5e-14, p=2.8e-07, and p=1.2e-05, respectively). FZD2 is a component of the Wnt/b-catenin pathway, which also interacts with the NFkappaB pathway. However, ZNF697 is not described in the literature as having immune functions. Overall, we used an untargeted WES approach on a pediatric population with extreme phenotypes of pneumococcal infection and identified 3 potential novel genes associated with IPDs, including 2 with known involvement in inflammatory and immune processes. To our knowledge this is the first study using WES on patients with such a severe form of pneumococcal disease. These unprecedented findings require further validation through replication studies, sequencing of parents, and functional analyses.

Keywords: Whole, exome sequencing, invasive pneumococcal diseases, pediatrics, innate errors of immunity, rare variants, extreme phenotypes

Poster #23: Identification of genetic variants of interest in individuals with severe neurological or hematological Events Supposedly Attributable to Vaccination or Immunization (ESAVI) following administration of the ChAdOx1 nCoV-19 vaccine from Brazil

Roberta Soares Faccion

Soares Faccion Roberta (1), Gomes Lopes Milena Regina (1), Drumond Piazi Marcelle (1), Azamor Da Costa Barros Tamiris (2), Prado Cunha Daniela (1), Nunes Da Silva Agonigi Bruna (1), Ribeiro Fernandes Alexandre (3), Regla Vargas Fernando (4), Palheiro Mendes De Almeida Daniela (2), Vasconcelos Zilton (1, 5)

1 - Instituto Nacional de Saúde da Mulher, da Criança e do Adolescente Fernandes Figueira [Rio de Janeiro] (Brazil), 2 - Instituto de Tecnologia em Imunobiológicos (Bio-Manguinhos) [Rio de Janeiro] (Brazil), 3 - Universidade Federal Fluminense [Rio de Janeiro] (Brazil), 4 - Instituto Oswaldo Cruz = Oswaldo Cruz Institute [Rio de Janeiro] (Brazil), 5 - Fundação Oswaldo Cruz (Fiocruz, Brazil)

Abstract

Introduction: The COVID-19 pandemic boosted the production and commercialization of vaccines in record time. Thus, pharmacovigilance activities have been intensified to detect rare events supposedly attributable to vaccination or immunization and ensure their efficiency and safety. Severe adverse events associated with COVID-19 vaccines from different platforms have been identified. Rare neurological and hematological Severe Adverse Events (SAE) cases of Guillain-Barré syndrome (SGB), acute disseminated encephalomyelitis (ADEM), transverse myelitis (MT) and Vaccine-induced Immune Thrombotic Thrombocytopenia (VITT) have been detected after the use of the ChAdOx1 nCoV-19 vaccine in Brazil and worldwide. The rarity of these immune-mediated phenomena highlights the importance of understanding the genetic factors underlying their occurrence. Objective: To identify possible genetic variants with biological relevance to the neurological adverse events associated with ChAdOx1 nCoV-19 vaccine. Methods: This is a case series study with an associative objective. Whole blood samples from 32 cases of the SAEs above-mentioned – gathered throught the notifications to the pharmacovigilance of Bio-manguinhos/Fiocruz and from the database of the Brazilian National Immunization Program – and 39 family members were collected for DNA extraction and whole genome sequencing (WGS). The DNA libraries were sequenced on the NovaSeq 6000 platform (Illumina) and the data from the WGS was analyzed on the Franklin platform. Results: A total of 40 genes with variants of interest (VI) were found. These genes were previously associated with the above-mentioned SAEs, and/or Live attenuated vaccines SAEs, and/or with Inborn Erros of Immunity. The next steps of the study are to confirm the impact on the VI through functional assays, transcriptomics, and proteomics. Conclusions: SAEs associated with vaccines are extremely rare and might be associated with equally rare genetic variants. Vaccine pharmacovigilance studies are key to elucidate SAEs causes, identify individuals in potential risk, and reassure the general public on the safety of approved vaccines.

Keywords: ChAdOx1 nCoV, 19 vaccine, Events Supposedly Attributable to Vaccination or Immunization (ESAVI), Genetic susceptibility, Autoimmunity, Inborn errors of immunity.

Poster #24: ANNEXA: A comprehensive pipeline for extending genome annotations using long-read transcriptome sequencing

Nicolaï Hoffmann

Hoffmann Nicolaï (1), Cadieu Edouard (1), Besson Aurore (1), Lorthiois Matthias (1), Le Bars Victor (1), Houel Armel (1), Hitte Christophe (1), André Catherine (1), Hedan Benoit (1), Derrien Thomas (1)

1 - Institut de Génétique et Développement de Rennes (IGDR) UMR 6290 (35000 Rennes France)

Abstract

In recent years, the advent of long-read transcriptome sequencing (LR-RNAseq) has enabled the sequencing of full-length transcripts, thus enhancing genome annotation by providing an unfragmented view of the transcriptome and a better definition of repeated regions. However, leveraging these LR sequencing technologies necessitates the development of specialized bioinformatics tools. To this end, we have developed ANNEXA, an all-in-one reproducible pipeline written in the Nextflow workflow management language that extends user-provided reference annotations from long-read sequencing data. ANNEXA works by using only three input files: a reference genome, a reference annotation, and a file listing the LR-RNASeq data. It reconstructs transcripts and quantifies their abundance, using two possible transcript reconstruction tools (Bambu or StringTie) before predicting whether novel identified transcripts are coding (mRNAs) or non-coding RNAs (lncRNAs). Novel transcripts are further filtered based on their likelihood to correspond to full-length transcripts using deep learning models, trained to validate novel transcription start sites. A final quality control process produces a graphical report, allowing users to quickly obtain key metrics, such as the number of novel genes, isoforms and exons for both lncRNAs and mRNAs. To demonstrate the robustness of ANNEXA, we have tested its ability to extend reference annotations (Ensembl/Gencode) using Oxford Nanopore Technologies (ONT) sequencing data. In a dog/human comparative oncology study, we have applied ANNEXA on ONT LR-RNASeq from 2 human and 4 dog mucosal melanoma cell lines. In the 2 human cell lines, ANNEXA identified 66 novel genes and 427 novel transcripts, absent from the Gencode reference annotation. In the 4 canine cell lines, ANNEXA identifies 75 novel genes and 237 novel transcripts, including 4 lncRNA genes exclusively expressed in mucosal melanoma lines, which may serve as potential biomarkers. Overall, our work presents a new bioinformatic pipeline to automatically reconstruct and characterize mRNAs and lncRNAs from long-read transcriptome data.

Keywords: Genome annotation, Long read sequencing, lncRNA, Nextflow

Poster #25: CITE-seq Workflow for Multimodal Single-Cell Analysis

Jovana Brocic

Brocic Jovana (1), Guégan Justine (1), Bielle Franck (2), Coutelier Marie (1)

1 - Data Analyses Core (France), 2 - BRIGHT Hétérogénéité, immunité et thérapie des tumeurs cérébrales (France)

Abstract

Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) enables simultaneous profiling of RNA and surface protein expression, offering unparalleled resolution of cellular heterogeneity. We describe an optimized workflow that combines state-of-the-art computational tools to maximize data quality and downstream interpretability in CITE-seq experiments. The preprocessing begins with Cell Ranger [1], a robust pipeline for transforming raw sequencing data into gene-barcode count matrices, both for RNA data and antibody-derived tag (ADT) data. These can be processed separately with dimensionality reduction techniques, such as UMAP, and clustering analyses, to reveal distinct cellular subsets. In the Seurat [2] workflow, ADT data is normalized using the centered log-ratio transformation (CLR) method. Alternatively, the denoised and scaled by background (DSB) method [3] normalization enhances protein signal accuracy by adjusting for ambient RNA and technical noise, offering improved reliability in protein expression profiling. Clusters obtained after this normalization however show little overlap with RNA-based clusters. For a comprehensive multimodal analysis, we thus employ Weighted Nearest Neighbor (WNN) integration [2] to combine RNA and ADT modalities. By assigning modality-specific weights to each cell, WNN integration enables precise identification of cell types and states, leveraging the complementary strengths of transcriptomic and proteomic data. Our workflow confirms the added value of combining both assays by using these data processing approaches, from normalization to integration. It demonstrates the reliability and reproducibility of these methods in multimodal genomic studies.

Keywords: CITE, seq, DSB normalization, Seurat, Weighted Nearest Neighbor, multimodal analysis

Poster #26: Single Cell DNA methylomes from multiple tissues demonstrates tissue heterogeneity and target enrichment as a driver of read utility

Samuel Laborde

Laborde Samuel (1), Holmes Austin (1), Jansen Camden (1), Publication Eric (1), Khare Sanika (1), Skinner Dominic (1), Alves Bryce (1)

1 - Scale Biosciences (United States)

Abstract

Comparative analysis of single-cell gene expression technologies aids in selecting the right platform for experiments. However, emerging single-cell (sc) technologies like scDNA methylation are not yet widely available for such comparisons. Instead of evaluating one tissue across platforms, we present a single technology analyzed across multiple tissues and sequencing depths using target enrichment. By integrating bulk-DNA methylation with single-cell data, we transform methylation metrics to a gene level, providing guidance on maximizing data per cell without excessive sequencing. Single-cell DNA methylation enhances current genetic and epigenetic analyses by offering an orthogonal dataset that further characterizes functional genomics. Scale Bio provides a kitted platform for generating and analyzing thousands of single-cell methylomes. While bulk DNA methylation is a well-established gene regulatory mechanism, large-scale single-cell datasets remain limited. Early-stage single-cell technologies often require excessive sequencing and cell numbers for basic analyses like clustering, annotation, and differential methylation. At these high depths, biologically relevant sample multiplexing becomes challenging. We analyzed six single-cell methylomes (solid tumor, blood tumor, young PBMCs, old PBMCs, human brain, and mouse brain) using the Scale Bio scMet kit. Tumor samples underwent copy number variation analysis alongside down-sampled read depth and cell numbers. Across all samples, we evaluated clustering and annotation at varying sequencing depths, finding that brain tissue tolerates lower read depths. Libraries were enriched using the Twist Human Methylome panel, demonstrating that regulatory region enrichment enables lower read depths while maintaining significant DMR detection. Our findings highlight that tissue heterogeneity and enrichment of known DMRs are critical when determining sequencing depth and cell numbers. Single-cell methylomes continue refining known regulatory elements while uncovering de novo DNA methylation events at the single-cell level.

Keywords: single cell, methylomes, DNA methylation, target enrichment

Poster #27: Impact of joint Dimension Reduction methods for survival prediction - extension of a multi-omics benchmark study

Vincent Le Goff

Le Goff Vincent (1), Guillemot Vincent (2), Philippe Cathy (3), Mendes Gwendoline (1), Deleuze Jean-François (4), Le Floch Edith (5), Gloaguen Arnaud (5)

1 - Centre National de Recherche en Génomique Humaine (France), 2 - Hub of Bioinformatics and Biostatistics (France), 3 - Building large instruments for neuroimaging: from population imaging to ultra-high magnetic fields (France), 4 - CEA, Centre National de Recherche en Génomique Humaine, Université Paris-Saclay (France), 5 - Centre National de Recherche en Génomique Humaine (France)

Abstract

With multi-omics studies came the hope that looking jointly at several molecular layers would unravel the underlying dysregulations. However, analyzing multi-omics datasets poses challenges due to their high dimensionality and heterogeneity. In a benchmark of survival methods on cancer datasets, (Herrmann & al., 2021) demonstrated that models accounting for the inherent group structure in these multi-omic datasets can enhance their prediction performances, yet it remains unclear whether integrating molecular and clinical data improves predictive power over clinical data alone. Extending this benchmark, we aim to explore methods that extract links between omics data blocks by employing 4 joint Dimension Reduction (jDR) techniques: Regularized Generalized Canonical Correlation Analysis (RGCCA), Joint and Individual Variations Explained, integrative Non-Negative Matrix Factorization and Multi-Omics Factor Analysis. Our approach uses a sequential modeling strategy, initially estimating a reduced space from molecular data alone via unsupervised or supervised techniques, followed by survival prediction using a Cox model trained on this joint reduced space, with and without clinical data. Our results indicate that jDR methods combined with a Cox model outperform traditional techniques like penalized regression, boosting, or random forests. When analyzing clinical and omics data jointly, jDR methods perform significantly better than the reference, which is not the case for non jDR ones. However, when predicting solely from omics data, most methods significantly underperform compared to the clinical-only model, with some exceptions that outperform the reference, notably supervised versions of RGCCA. Future research will extend this benchmark to incorporate additional jDR methods compared in (Cantini & al., 2021), enabling us to identify a robust base model for subsequent methodological advancements that integrate biological knowledge. Herrmann & al. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief Bioinform 2021 Cantini & al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun 2021

Keywords: multi, omics, benchmark, joint dimension reduction

Poster #28: Improved bioinformatics analysis of second and third generation sequencing approaches for accurate length determination of short tandem repeats and homopolymers

Yimin Shen

Shen Yimin (1), I. Jeanjean Sophie (2), M. Hardy Lise (2), Daunay Antoine (2), Delépine Marc (3), Gerber Zuzana (3), Alberdi Antonio (4), Tubacher Emmanuel (1), Deleuze Jean-François (1, 2, 3), How-Kit Alexandre (2)

1 - Laboratory for Bioinformatics, Foundation Jean Dausset – CEPH, Paris, France (France), 2 - Laboratory for Genomics, Foundation Jean Dausset – CEPH, Paris, France (France), 3 - Centre National de Recherche en Génomique Humaine (CNRGH), CEA, Institut François Jacob, Evry, France (France), 4 - Technological Platform of Saint-Louis Research Institute (IRSL), Saint-Louis Hospital, University of Paris, Paris, France (France)

Abstract

Microsatellites are short tandem repeats (STRs) of a motif of 1 to 6 nucleotides that are ubiquitous in almost all genomes and widely used in many biomedical applications. However, despite the development of next-generation sequencing (NGS) over the past two decades with new technologies coming to the market, accurately sequencing and genotyping STRs, particularly homopolymers, are still very challenging today due to several technical limitations. This leads in many cases to erroneous allele calls and difficulty in correctly identifying the genuine allele distribution in a sample. In the present study, we assessed several second and third NGS approaches in their capability to correctly determine the length of microsatellites using plasmids containing A/T homopolymers, AC/TG or AT/TA dinucleotide STRs of variable length. Standard PCR-free and PCR-containing, single Unique Molecular Index (UMI) and dual UMI 'duplex sequencing' protocols were evaluated using Illumina short-read sequencing, and two PCR-free protocols using PacBio and ONT long-read sequencing. Several improved bioinformatics algorithms were developed to correctly identify microsatellite alleles from sequencing data, including four and two modes for generating standard and combined consensus alleles, respectively. We provided a detailed analysis and comparison of these approaches and made several recommendations for the accurate determination of microsatellite allele length.

Keywords: Microsatellite, short tandem repeat STR, homopolymer, next, generation sequencing NGS, long, read sequencing, short, read sequencing, bioinformatics analysis

Poster #29: Integrative multiparametric analysis of circulating cell-free nucleic acids of plasma during aging

Nicolas Tessier

Tessier Nicolas (1), M. Hardy Lise (1), Mauger Florence (2), Daunay Antoine (1), Daviaud Christian (2), Horgues Caroline (2), Sahbatou Mourad (1), Le Buanec Hélène (3), Deleuze Jean-François (1, 2), How-Kit Alexandre (1)

1 - Laboratory for Genomics, Foundation Jean Dausset – CEPH (France), 2 - CEA, Centre National de Recherche en Génomique Humaine, Université Paris-Saclay (France), 3 - Saint-Louis Research Institute, INSERM U976 - HIPI Unit, University of Paris (France)

Abstract

Plasma circulating cell-free nucleic acids (ccfNAs) provide an exceptional source of information about an individual's health, yet their biology in healthy individuals during aging remains poorly understood. Here we present the first integrative multiparametric analysis most types of plasma ccfNAs in 139 healthy donors aged between 19 and 66 years, focusing on quantity, integrity and DNA methylation. We showed a highly significant increase in ccfnDNA levels during aging ($p < 0.001$), associated with a decrease in its integrity ($p < 0.05$), while no significant change was detected in mtDNA levels and ccfDNA methylation. Moreover, a significant increase in ccfmRNA, ccfrRNA ($p < 0.05$) and miR-483-5p ($p < 0.001$) levels was detected during aging, but without any changes in ccfRNA integrity. ccfDNA and ccfRNA levels were correlated ($p < 0.001$) and a similar pattern was observed between ccfmtDNA and ccfRNA levels, suggesting a possible common release, maintenance and/or clearance mechanism

Keywords: Plasma, circulating cell, free nucleic acids, circulating cell, free DNA, circulating cell, free RNA, aging, DNA methylation, miRNA

Poster #30: Missense Variant Mapping onto Reference Proteome Structures

Lucas Chataigner

Chataigner Lucas (1), Aschard Hugues (1), Julienne Hanna (1)

1 - Department of Computational Biology, Institut Pasteur, Université Paris Cité, 25–28 Rue du Dr Roux, 75015 Paris, France (France)

Abstract

A significant portion of phenotypic variation remains unexplained by common variants, driving interest in the contribution of rare variants to missing heritability. Traditional association studies often lack the power to detect rare variant effects, leading to the use of burden tests that aggregate variation within a gene, sacrificing granularity for statistical strength. To regain granularity and characterize individual rare variant effects, there is growing interest in combining genomic data with structural biology to understand how coding variants, particularly missense variants, affect protein function, stability, and pathogenicity. Missense variants, the second most prevalent form of genomic variation after synonymous substitutions, directly impact gene function through single amino acid changes. Two main approaches are typically leveraged to visualize missense variant impact on gene function. The first maps features onto canonical structures, exemplified by tools like AlphaMissense, which uses deep learning to predict pathogenicity and generate position-based marginal values. The second approach modifies protein structures to evaluate substitutions, as seen in tools like the PSnpBind database, which examines variants' effects on protein–ligand interactions. Minor allele frequency—a fundamental metric for understanding genetic variation—has been overlooked in current structural mapping. To address this, we developed tools to map frequency data from 807,162 individuals in the gnomAD v4 database onto protein structure predictions for 20,000 UniProt canonical isoforms. These tools enable visually evocative and detailed subgroup analyses (such as ancestry-specific investigations). This work offers opportunities to enhance our understanding of genotype-phenotype relationships, with applications in personalized medicine.

Keywords: Missense, Structural Biology, Rare Variant Effects, Multi Ancestry