To the EAACI Vice President Science- the EAACI Headquarters,

Project Title
Data analysis to unravel the biological mechanisms underpinning acute FPIES for improved diagnostics

Name, Country
Julián Montoto-Louzao, Spain

Type, Duration and Location of the Fellowship
EAACI Research Fellowship
3 months (February 16th – May 17th, 2025)
Imperial College London, United Kingdom

Host Institution and Supervisor Name
Host Institution: Imperial College London
Host Supervisor: Dr. Marta Vázquez-Ortiz
Collaborating Researchers at Host Institution:
Dr. Myrsini Kaforou, Dr. Sara Fontanella, Dr. Jethro Herberg

Home Institution and Supervisors
GENVIP Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago de Compostela, Spain
Dr. Antonio Salas, Dr. Federico Martinón-Torres, Dr. Alberto Gómez-Carballa

1. Informative Description of the Activity During the Fellowship

During my three-month EAACI Research Fellowship at Imperial College London, I joined the bioinformatics group led by Dr. Myrsini Kaforou, based at the Department of Infectious Diseases, South Kensington Campus. My role focused on transcriptomic data analysis within the multi-center BIO-FPIES project, aimed at identifying diagnostic signatures for Food Protein-Induced Enterocolitis Syndrome (FPIES). I also collaborated with the Complex Data Analysis (CDA) group led by Dr. Sara Fontanella and Prof. Adnan Custovic, based at the Hammersmith Campus. My work was supervised by Marta Vázquez-Ortiz, co-IP of the BIO-FPIES project.

February – March 2025: Data Quality Control and Normalization

In the early weeks of the fellowship, I focused on the preprocessing of batch 2 RNA-seq data, generated at GENVIP, with the goal of aligning it to the analytical standards of batch 1, which had been processed at Imperial. I reproduced Imperial's pipeline, based on STAR (v2.7.10a), Ensembl v109 annotations, and featureCounts, with only minor adaptations, ensuring compatibility while maintaining reproducibility. This

phase was carried out in close collaboration with the bioinformatics team of the Department of Infectious Disease (DOID), who originally developed the batch 1 pipeline.

Despite general consistency, quality control revealed differences that required deeper investigation and joint discussion. Batch 2 showed higher proportions of ribosomal and intronic reads, likely due to slightly less efficient ribosomal depletion and increased sequencing depth, which, while not compromising overall data quality, introduced shifts in exon to intron ratios and gene-level distributions. In addition, batch 2 had a modest increase in intergenic reads, potentially reflecting either minimal genomic DNA contamination or expression from unannotated transcripts such as lncRNAs or miRNAs. These findings, brought to light through collaborative QC interpretation with DOID's team, led to constructive exchanges around how to best summarize counts, interpret annotations, and plan integration.

We ultimately concluded that differences in read length, sequencing depth, and RNA quality could explain the observed discrepancies. As a result, we made final adjustments to the preprocessing pipeline and reprocessed the batch 2 samples to reduce those differences to a minimum. These discussions also helped define the strategy for downstream normalization, confirming that combining both batches would require filtering for genes expressed across datasets and reinforcing the importance of harmonizing preprocessing choices prior to statistical modeling.

The next phase focused on the deep integration of RNA-seq data from two batches generated at different sites: batch 1 sequenced at Imperial, and batch 2 sequenced in Spain (GENVIP). Ensuring consistency and eliminating technical biases between them was crucial. I conducted extensive QC and filtering to harmonize the datasets and participated in key meetings to align approaches between both institutions.

Challenges included:

- Mismatches between sample identifiers in count matrices and metadata
- Missing phenotypic annotations in batch 2
- Inconsistencies across recruiting centers
- Classification discrepancies in the severity of FPIES reactions

These issues required on-site discussions with clinicians and data scientists to refine patient inclusion and classification logic.

I worked closely with Dr. Myrsini Kaforou's bioinformatics team at South Kensington and with the CDA group at Hammersmith. We evaluated several batch correction strategies through joint analyses and regular discussions. After testing ComBat-seq

and other approaches, we selected RUVSeq (k=2) based on PCA inspection, which showed no residual batch effect and highlighted meaningful biological variation.

I also contributed to filtering out transcripts that could confound analyses, such as ribosomal and globin genes, and reviewed QC metrics across batches to confirm the success of the preprocessing pipeline.

April 2025: Differential Expression and Integration Planning

In April, we initiated the differential expression analysis (DEA) using DESeq2, exploring comparisons between FPIES reactors and non-reactors. Although initial plans included more complex comparisons (e.g., baseline vs reaction), we strategically decided to limit the first analysis to leverage on-site collaboration. These extended analyses will be completed upon my return to GENVIP.

Preliminary results revealed a set of differentially expressed genes (DEGs) that are coherent with:

- Early proteomic findings from the CDA group
- Transcriptomic patterns identified in earlier phases of the BIO-FPIES project

In parallel, I participated in planning discussions for the integration of transcriptomic and proteomic data, including the design of upcoming validation experiments using Nanostring technology. These meetings also addressed the strategy for harmonizing external infection data with the FPIES dataset to improve classification models.

May 2025: External Cohort Integration and Signature Drafting

In May, my focus shifted toward the integration of external datasets from the EUCLIDS-PERFORM-DIAMONDS consortium to distinguish FPIES from clinical mimics such as sepsis or viral gastroenteritis. This required collaboration with the bioinformatics preprocessing team at Imperial to align pipelines, sample annotations, and QC thresholds across cohorts.

I also began drafting diagnostic models using FS-PLS and LASSO, in collaboration with specialists from Dr. Kaforou's team with experience in high-dimensional biomarker selection. These models will be further developed and validated after the fellowship.

Throughout the entire fellowship, I actively participated in:

- Weekly meetings with Dr. Kaforou's bioinformatics team
- Department of Infectious Diseases seminar series
- CDA group's statistical methodology seminars

- CDA journal clubs

These interactions greatly enriched my understanding of multi-omic analysis in translational research and supported the methodological choices applied to the FPIES dataset.

2. What Questions Were Addressed and Why?

The core scientific question addressed during this fellowship as part of the BIO-FPIES project was:
Can we define transcriptomic signatures that distinguish acute FPIES reactions from non-reactions and from other acute illnesses such as sepsis or gastroenteritis, and thereby contribute to the development of less invasive and more rapid diagnostic tools?

This question is of major clinical relevance, as acute FPIES currently lacks a diagnostic test. Diagnosis relies on oral food challenges, which are invasive, resource-intensive, and carry inherent risks. Developing a molecular signature would represent a breakthrough in clinical allergy and pediatric emergency care.

To answer this, we pursued the following specific research questions for this fellowship:

1. How can multi-batch RNA-seq data from different sequencing centers be harmonized for robust analysis of FPIES gene expression profiles?
   This question addressed the technical challenge of integrating transcriptomic data generated across institutions (Imperial College London and GENVIP), ensuring analytical consistency and interpretability.
2. Which genes are differentially expressed between FPIES reactors and non-reactors during oral food challenges?
   This analysis aimed to identify molecular markers that characterize the acute reaction phase in FPIES.
3. How can gene expression data from FPIES patients be distinguished from data from children with acute infectious diseases?
   By integrating transcriptomic profiles from the EUCLIDS-PERFORM-DIAMONDS consortium, we aimed to explore whether diagnostic classifiers could differentiate FPIES from clinically similar conditions like bacterial sepsis or viral gastroenteritis.
4. What are the most appropriate statistical and machine learning methods (e.g., RUVSeq, DESeq2, FS-PLS, LASSO) for this high-dimensional dataset?

This question reflects the analytical aspect of the project — identifying the best tools for batch correction, differential expression, and biomarker discovery in complex, multi-omic datasets.

5. What design elements are needed for a robust validation strategy using Nanostring technology in a real-world diagnostic context?
   By contributing to the early planning of the Nanostring validation phase, the fellowship also addressed practical aspects of translating omics data into clinically applicable tests.

3. What Was the Nature of the Research?

This was a computational and translational research project within the BIO-FPIES study, focused on identifying transcriptomic biomarkers for acute FPIES. The work involved high-dimensional RNA-seq data analysis from pediatric blood samples collected during oral food challenges.

Key tasks included:

- Preprocessing and harmonization of sequencing data from two centers (Imperial and GENVIP)
- Normalization and batch correction using RUVSeq amongst other methods
- Differential expression analysis with DESeq2 to compare reactors vs non-reactors
- Initial development of predictive models (FS-PLS, LASSO)
- Integration planning with external infection datasets
- Collaboration in designing validation experiments using Nanostring technology

The project combined advanced bioinformatics techniques with close interdisciplinary teamwork, bridging data science, infectious diseases, and clinical allergy.

4. What Was the Result?

The fellowship led to several key outcomes:

- A clean, normalized RNA-seq dataset integrating two sequencing batches, ready for downstream analyses.

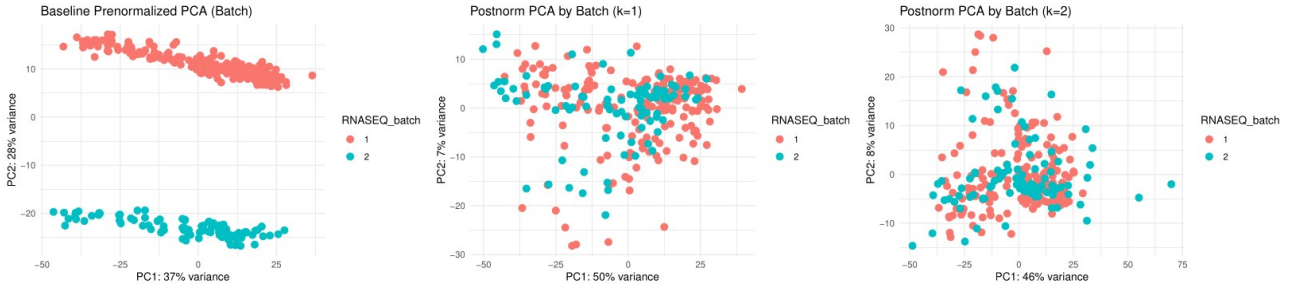

Figure 1: Principant Component Analysis plot of batch 1 and batch 2 samples, pre-normalized (left), normalized k=1 (middle) and normalized k=2 (right)

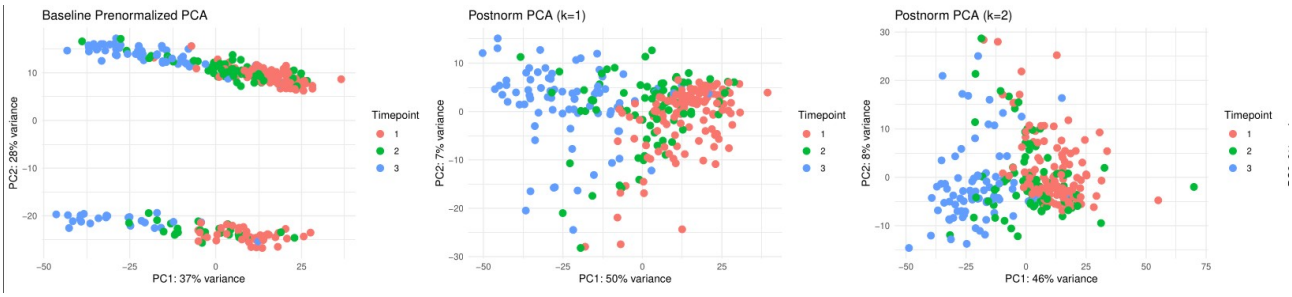

Figure 2: Principant Component Analysis plot of timepoint 1, 2 and 3 samples, pre-normalized (left), normalized k=1 (middle) and normalized k=2 (right)

- Preliminary differential expression results identifying genes distinguishing FPIES reactors from non-reactors, aligning with proteomic trends and previous transcriptomic findings.
- Harmonization strategy defined for integrating external infectious disease cohorts, enabling future comparisons with sepsis and gastroenteritis.
- Initial diagnostic models using FS-PLS and LASSO were drafted, with further refinement planned post-fellowship.
- Contributed to the design of the Nanostring validation phase, including gene panel discussions and sample prioritization.

These results lay the foundation for the next stage of the BIO-FPIES project: multi-omic integration and validation of diagnostic signatures.

While no publications were submitted directly during the fellowship, the transcriptomic work built upon previous project stages currently under peer review.

Also, I have been invited to speak at PAAM 2025, where future insights and outputs from this project will be presented.

5. How Will the Findings Impact Future Research?

The outcomes of this fellowship provide a solid launchpad for the validation and refinement of transcriptomic diagnostic signatures for acute FPIES. The cleaned dataset, analytical pipeline, and draft models will directly support:

- The Nanostring validation phase, which I will lead at GENVIP
- Multi-omic integration with proteomic data to enhance diagnostic accuracy
- Development of clinically applicable tools to differentiate FPIES from infections in emergency settings

Additionally, the collaborative workflows established between Imperial and GENVIP will streamline future research in allergy and immune-mediated disorders, with shared expertise in high-throughput analysis and translational design.

6. Personal Reflection

This fellowship has been one of the most enriching experiences of my research career so far. Being at Imperial, surrounded by experts in bioinformatics, biostatistics, infectious diseases, immunology and allergy, allowed me to grow both technically and personally. I felt part of a truly collaborative and interdisciplinary environment, where discussions flowed naturally and every idea was met with interest and constructive feedback.

I had the chance to improve my skills in areas like RNA-seq normalization, multi-batch integration and diagnostic model development, while also learning how to work closely with clinical teams to make sure our analyses could truly benefit from a focused perspective on patient care. Joining regular team meetings, seminars and journal clubs helped me see how research questions are tackled from multiple angles, and gave me a broader view of how data science can contribute to translational medicine.

What I value the most is the sense of continuity that this experience created. The connection between GENVIP and Imperial has grown stronger, and I return to Spain with clearer goals, practical knowledge, and the confidence to lead the next steps of the BIO-FPIES project. I’m really grateful to everyone who welcomed me into their teams and shared their expertise with generosity and patience.

7. Acknowledgements

I would like to warmly thank Dr Myrsini Kaforou, Dr Sara Fontanella, Dr Mike Levin, Dr Jethro Herberg, Dr Adnan Custovic, Dr Antonio Salas, Dr Federico Martinón-Torres, and Dr Alberto Gómez-Carballa for their guidance throughout this project. I also want to thank Dr Darije Custovic, Dr Anhar Ullah, Dr. Sabrina Kapur and the rest of the CDA team, as well as Dr Amedine Duret, Dr Claire Dunican, Dr Rebecca Worsley, Dr Lara Oberski, Dr Dominic Coote, Dr Aubrey Cunnington and everyone in the Department of Infectious Diseases team, for making me feel part of the group from day one.

Special thanks to Dr Marta Vázquez-Ortiz for her kind guidance and her invaluable help throughout every step of the fellowship.