Towards personalized medicine in allergy: integrating multi-omics data with advanced machine and deep learning techniques

Type of Fellowship: Short Term Research Fellowship
Home Institution: San Pablo-CEU University (Madrid, Spain)
Host Supervisor: Dr. Rocio Rebollido-Rios
Host Institution: University Hospital of Cologne, Department I of Internal Medicine
University of Cologne (Germany)
Duration: 3 months (September 15th – December 15th, 2024)

1 Background

Severe allergic inflammation is increasingly recognized as a significant clinical challenge due to its association with profound mucosal epithelial remodeling and heightened inflammatory responses. These responses are characterized by increased activity of type 2 T-helper (Th2) cells and extensive cellular infiltration in affected tissues [1–5]. The rising prevalence of severe allergies, often resistant to conventional therapies such as corticosteroids, immunotherapy and biologics, underscores the need for deeper insights into the mechanisms underlying these diseases and the development of innovative treatment strategies. Patients with severe allergic conditions experience a reduced quality of life, marked by frequent exacerbations, hospitalizations and comorbidities such as allergic rhinitis, atopic dermatitis and allergic asthma [6–9]. These challenges not only impact patient well-being but also place significant strain on healthcare systems, highlighting the urgency of advancing our understanding of these conditions.

Recent advances in omic technologies have transformed allergy research by offering a holistic perspective on the molecular landscape of allergic diseases. Multi-omics approaches combine data from genomics, transcriptomics, proteomics and metabolomics, allowing researchers to explore intricate networks and molecular pathways involved in severe allergic phenotypes [10, 11]. This integrative approach is essential for uncovering the complex mechanisms driving allergic inflammation and identifying novel biomarkers for patient stratification and therapeutic targeting. Consequently, there is a growing emphasis on personalized medicine, where interventions are tailored based on individual molecular profiles to improve patient outcomes and optimize resources.

Integrating omics data is crucial not only for understanding the molecular underpinnings of severe allergies but also for developing advanced treatment approaches. By identifying key molecular players and uncovering interactions among immune system components, researchers aim to create innovative therapeutic strategies that extend beyond traditional treatments. This focus on personalized, data-driven medicine holds significant potential to transform allergy management by offering more precise and effective interventions tailored to individual patient needs. The application of multi-omic technologies in allergy research thus represents a promising frontier for developing novel approaches to treat and manage severe allergic diseases, ultimately improving the quality of life for affected individuals.

In addition to these technological advances, systematically extracting data from scientific studies and leveraging insights from the literature remain essential in allergy research. Manual curation of information, such as genes, proteins and sample sources, enables detailed annotation of key biological entities and relationships, forming a robust foundation for hypothesis generation and contextualizing novel findings. This approach facilitates the identification of patterns, trends and gaps in the literature, providing valuable insights into the molecular networks driving allergic diseases. Furthermore, such data extraction strategies can bridge distinct fields, such as allergy and oncology, by uncovering shared pathways and mechanisms. The emerging field of AllergoOncology highlights this interdisciplinary synergy, investigating the intersection of allergic responses and cancer. By integrating knowledge across domains, researchers can uncover novel connections that deepen understanding of allergy and other immune-related conditions.

Ultimately, combining omic data with systematic literature analyses offers a powerful strategy to advance allergy research. This comprehensive approach enhances understanding of the molecular networks underlying severe allergic diseases and uncovers opportunities to establish connections between allergy and other immune-related conditions, fostering the development of innovative therapies and improving patient care.

2 Objectives

The ultimate goal of this project is to bridge the gap between molecular findings and clinical features, leading to advancements in personalized medicine for allergy patients.

To achieve this, the following specific aims were proposed:

1. Integration of multi-omics and clinical data from large-scale studies.
2. Elucidation of common pathways and mechanisms in allergy.
3. Development of a machine/deep learning classifier for patient stratification.

3 Methods

3.1 Data source

This study was conducted as part of the BIOGRIAL project (PI19/00044). A total of 1,392 patients were recruited from nineteen hospitals across Spain. Comprehensive data, including clinical variables and omics datasets (metabolomics and proteomics), were collected and organized in a REDCap database, ensuring data integrity and accessibility for subsequent analyses.

3.2 Data preprocessing

Targeted proteomics was performed on serum samples using OLINK Target 96 Inflammation Panel, which quantifies 92 proteins involved in the inflammatory and immune response processes. These experiments were conducted at the Institute of Applied Molecular Medicine (IMMA) at San Pablo-CEU University (Madrid, Spain).

The samples were processed in three experimental batches. These batches encompassed 574, 796 and 82 samples, respectively. To address batch-specific variability, bridge samples were selected from batch 1 and included in batches 2 and 3. These bridge samples, critical for normalizing data across experiments conducted at different time points, were chosen based on criteria such as high detectability, successful quality control and comprehensive representation of the dataset.

Data preprocessing involved the following steps (Figure 1):

1. Identification and removal of duplicated samples.
2. Exclusion of protein NPX (Normalized Protein eXpression; log2 transformed) values failing quality control.
3. Substitution of NPX values below the limit of detection with NA.
4. Visualization of data using the `ComplexHeatmap` R package [12, 13].

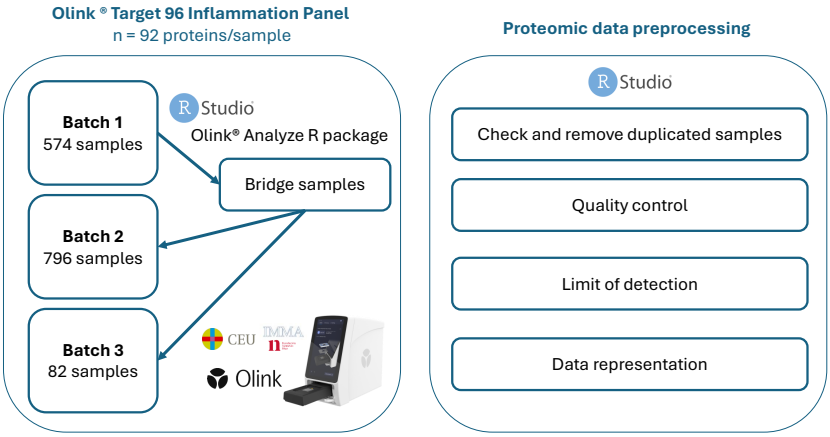

Figure 1: Pipeline for preprocessing Olink proteomic data.

In addition to these steps, the 92 proteins included in the Olink panel were annotated using the UniProt database via the `UniProt.ws` R package [14]. These annotations, encompassing protein families and other relevant features, were curated and tailored to enhance data representation and interpretation.

4 Results

The initial weeks of this research stay were dedicated to gaining proficiency in the analysis of Olink proteomic data. Communication with the Olink support team was essential to address and resolve specific issues related to data processing.

Within each batch, no duplicate samples were detected. However, inter-batch duplicates (bridge samples) were included to facilitate normalization across batches. During quality control, a total of 12384 proteins from various samples were excluded due to failing quality thresholds. Additionally, 42151 proteins were identified as being below the limit of detection and were subsequently labeled as missing values. Protein annotation revealed that 76 out of the 92 proteins in the Olink panel had associated protein family annotations in the UniProt database.

Normalized Protein eXpression (NPX) values were visualized in a heatmap (Figure 2), where rows (proteins) and columns (samples) were hierarchically clustered to identify patterns and relationships within the dataset. The heatmap was enriched with additional annotations, including UniProt protein family classifications and patient allergen sensitization profiles, providing a multidimensional view of the data. This visualization not only facilitated an intuitive understanding of the distribution of NPX values but also highlighted potential connections between protein families and clinical profiles.

Preliminary analysis revealed three distinct clusters of proteins based on NPX value distributions: proteins with low NPX values or no detection, proteins with moderate NPX values and proteins with high NPX values. These clusters may represent biologically meaningful groupings, such as proteins with varying levels of expression or relevance in the context of inflammatory or immune

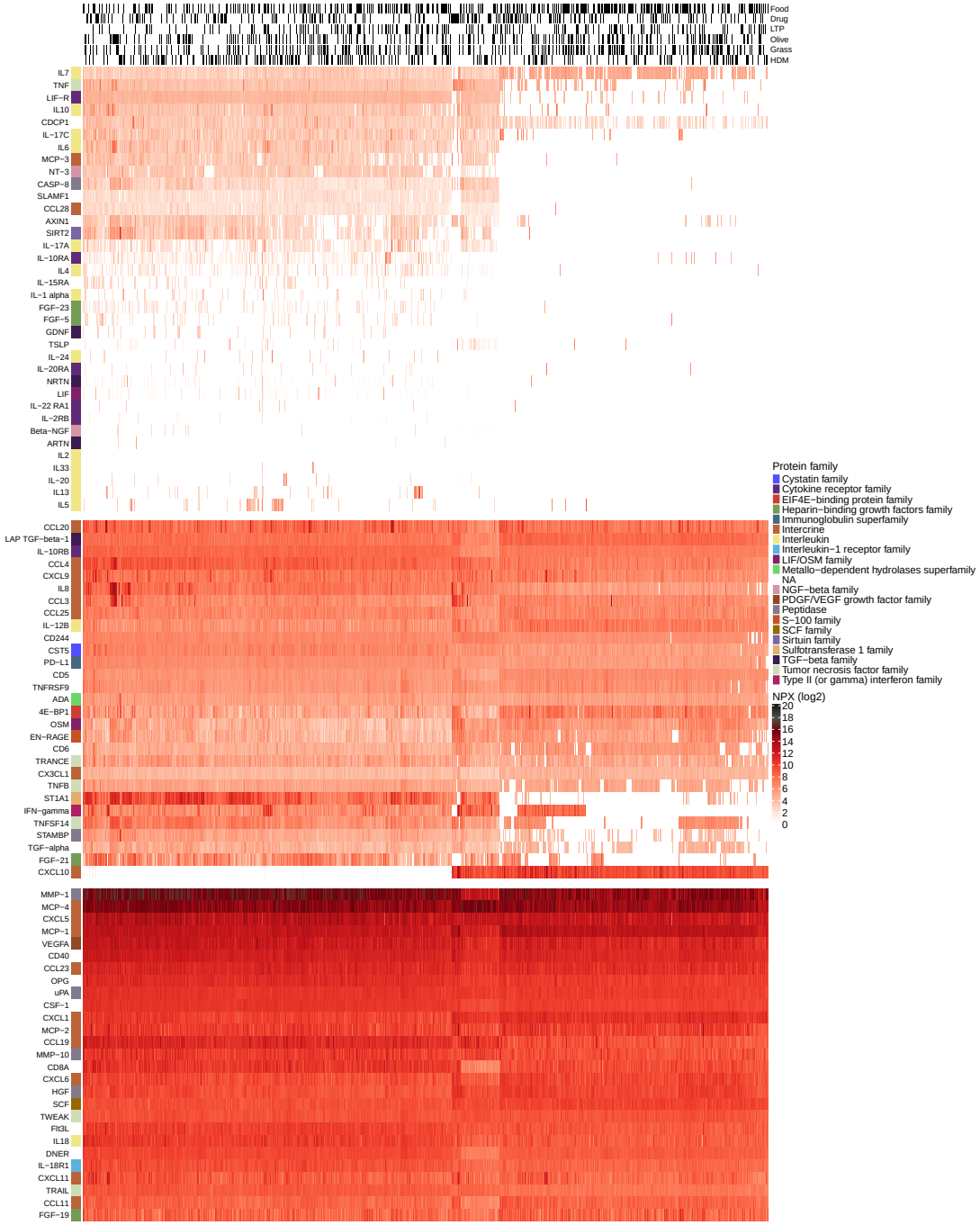

Figure 2: Heatmap of log2-transformed protein expression levels (NPX values) for the 92 analyzed proteins. Top annotations display the allergen sensitization profiles of patients, while side annotations indicate protein families categorized according to the UniProt database.

responses. Proteins with low or undetectable NPX values could indicate limited expression or absence in the analyzed samples, potentially reflecting proteins with minor roles in the studied conditions. Conversely, proteins with high NPX values might signify key players in the underlying pathological mechanisms or robust biomarkers for specific allergic or inflammatory processes.

Although these findings are preliminary, they already suggest intriguing patterns that require further exploration. The identification of clusters offers valuable insights into protein behavior across patient samples and lays the foundation for further investigation into the molecular mechanisms that drive allergic responses. These initial results emphasize the potential of integrating proteomic data with clinical annotations to uncover meaningful biological insights and guide the development of targeted therapeutic strategies.

5 Adaptations from the research plan

Despite thorough preparation, the project encountered unexpected delays due to challenges in obtaining the required clinical and metabolomic data. These data were essential for integrating clinical and omic datasets to correlate molecular profiles with clinical phenotypes. However, the absence of clinical data significantly hindered the planned analyses and computational modeling, limiting progress to the preprocessing of proteomic data.

In response to these setbacks, the project scope was adapted to focus on alternative avenues of research. Specifically, efforts were redirected toward contributing to the interdisciplinary field of AllergoOncology, which investigates the intersection of allergic responses and cancer. This work, conducted as part of the European Academy of Allergy and Clinical Immunology (EAACI) AllergoOncology Working Group, aims to explore the molecular roles of immune cells, particularly monocytes and macrophages, in allergic diseases and cancer. Preliminary findings highlight the dual roles of these cells in promoting inflammation in allergy whereas contributing to immune evasion in cancer, with implications for both disease progression and therapeutic development.

The adapted project focuses on the identification of altered molecular profiles in monocytes and macrophages from patients with allergic diseases using scientific literature. These molecular profiles were analyzed to reveal shared pathways and mechanisms that link allergy and cancer. Preliminary findings have already provided promising insights, underscoring the potential to uncover novel biomarkers and therapeutic targets with relevance to both conditions.

5.1 Methods

The article selection and analysis process involved systematic screening, detailed curation and manual annotation. This approach facilitated the extraction of key data, including diseases, sample sources, genes, proteins and metabolites. Raw text from annotations was preprocessed using a standardized text preprocessing pipeline, which included data cleaning and formatting steps such as removing punctuation, special characters, extra whitespaces and irrelevant information, as well as correcting spelling errors. Annotated molecules, including proteins and genes, were standardized to gene symbols and subsequently mapped to their corresponding ENTREZ IDs using the `limma` [15] and `AnnotationDbi` [16] R packages, respectively. This procedure identified a total of 451 unique molecules, comprising both coding and non-coding genes. Disorders were categorized using their Medical Subject Headings (MeSH) unique IDs to ensure standardized classification.

To uncover functional and disease-related relationships, semantic similarity analyses were conducted. Genes were first grouped by their associated diseases and a semantic similarity analysis of Disease Ontology (DO) gene clusters was performed using the `DOSE` R package [17]. This analysis applied the Wang measure and the Best Match Average (BMA) methods, as implemented in the `clusterProfiler` R package [18]. Genes were then categorized by cell types, distinguishing

between monocytes and macrophages based on the focus of the articles. For each cell type, a semantic similarity analysis was conducted on gene clusters annotated with Gene Ontology (GO) Molecular Function (MF) terms using the `GOSemSim` R package [19, 20], using the same Wang measure and BMA methods. These analyses provided valuable insights into functional similarities within and across cell type-specific gene clusters.

Complementing the semantic similarity analyses, a Reactome pathway enrichment analysis was performed using the `ReactomePA` [21] R package. This analysis encompassed all 451 genes, applying a stringent false discovery rate (FDR) threshold of 0.001, calculated using the Benjamini-Hochberg (BH) method. This rigorous approach ensured robust and reliable results, highlighting significantly enriched pathways and offering key insights into the molecular mechanisms underlying the studied conditions.

5.2 Results

A total of 451 molecules were identified and annotated across all allergic disorders. To assess their relevance, the frequency of occurrence of each molecule (i.e., the number of articles studying each molecule) was analyzed across the collection of articles. Among the 20 most frequently studied molecules, surface protein markers such as CD14, CD68, CD206 (MRC1) and HLA-DRA stood out. Pro-inflammatory cytokines and chemokines closely linked to monocyte and macrophage activity—IL12A/B, TNF, IL1A/B, CCL17, CXCL8, IL6 and IL18—were prominent, alongside the anti-inflammatory cytokine IL10. Molecules specifically associated with allergic conditions, such as the IL4 receptor (IL4R) and the low-affinity immunoglobulin gamma Fc region receptor III-A (FCGR3A), were also highlighted.

Pathway enrichment analysis revealed significant involvement of immune system-related pathways, underscoring the critical role of cytokine signaling in allergic inflammation. Key pathways identified included "Signaling by Interleukin-10," "Neutrophil degranulation," "Interleukin-4 and Interleukin-13 signaling," and "Interferon-gamma signaling." In addition to these immune pathways, signal transduction processes such as "Chemokine receptors binding chemokines," "Class A/1 (Rhodopsin-like receptors)," and "GPCR ligand binding" were enriched, highlighting their importance in cellular communication and responses to environmental stimuli. Furthermore, pathways related to cellular responses to external stimuli and programmed cell death were enriched, suggesting that regulation of cell survival and apoptosis plays a key role in shaping the allergic monocytic profile.

Semantic similarity analyses provided additional insights into the molecular relationships between allergic diseases. For monocytes, high functional similarities (greater than 0.75) were observed between allergic dermatitis, allergic asthma and allergic rhinitis based on Gene Ontology (GO) Molecular Function terms. A similar pattern was observed for macrophages, further underlining the shared molecular characteristics of these cell types in allergic conditions. Disease Ontology-based analyses revealed even higher similarities (above 0.95) among allergic asthma, allergic rhinitis and atopic dermatitis, reinforcing the interconnectedness of these disorders at a molecular level.

Despite these advances, critical gaps and biases remain in the current literature, particularly regarding the presence of comprehensive studies using human samples that include healthy controls. The application of advanced omics technologies holds promise for providing a more holistic understanding of the molecular mechanisms underlying allergic responses.

The findings of this project represent a promising but preliminary step in exploring the molecular landscape of allergic diseases. The molecules and pathways identified are likely only a fraction of the broader molecular networks involved, emphasizing the need for further research to uncover novel therapeutic targets. The outcome of this work will be published this year in a high-impact

journal and is integrated into a custom Shiny app designed for interactive exploration of the study results, showcasing the potential for significant contributions to the field of allergy research.

6 Acknowledgments and personal reflections

I would like to begin by expressing my heartfelt gratitude to Dr. Rocio Rebollido-Rios for supporting my application and hosting me over the past three months in her Computational Biomedicine and Bioinformatics research group at the University Hospital of Cologne. I am deeply thankful for everything I have learned from her, for the time she dedicated to mentoring me and for her invaluable guidance both within and beyond this project. I also want to extend my gratitude to Mina, whose support has been truly remarkable. Rocio and Mina, thank you both so much. I truly felt at home during my stay.

During this research stay, I had the opportunity to gain experience in omics data processing, data extraction and data visualization methods. I also attended weekly group seminars and result discussion meetings, which provided with learning experience, not only in Bioinformatics but also in many broader scientific areas. These experiences have been instrumental in my growth as a scientist and have allowed me to build meaningful relationships with my host institution and its members.

Lastly, I would like to express my sincere thanks to EAACI for awarding me this Short-Term Research Fellowship and for making this enriching experience possible. This fellowship enabled me to collaborate with exceptional scientists in the field of Bioinformatics.

References

[1] Saglani et al. 2015, The European respiratory journal, 46, 1796.

[2] Fahy, J. V. 2015, Nature reviews. Immunology, 15, 57.

[3] Divekar et al. 2015, Current opinion in allergy and clinical immunology, 15, 98.

[4] Benamar et al. 2023, Seminars in Immunology, 70, 101847.

[5] Bellinghausen et al. 2022, Frontiers in Immunology, 13.

[6] Eder et al. 2006, The New England journal of medicine, 355, 2226.

[7] Holgate, S. T. 1999, Nature, 402, B2.

[8] Delgado-Dolset et al. 2022, Allergy, 77, 1772.

[9] for Asthma, G. I. 2024, Global strategy for asthma management and prevention, accessed on 2024-11-01.

[10] Devonshire et al. 2023, World Allergy Organization Journal, 16, 100777.

[11] Vetrano et al. 2022, Frontiers in Immunology, 13.

[12] Gu et al. 2016, Bioinformatics.

[13] Gu, Z. 2022, iMeta.

[14] Carlson et al. 2024, UniProt.ws: R Interface to UniProt Web Services, r package version 2.44.0.

[15] Ritchie et al. 2015, Nucleic Acids Research, 43, e47.

[16] Pagès et al. 2024, AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor, r package version 1.66.0.

[17] Yu et al. 2015, Bioinformatics, 31, 608.

[18] Xu et al. 2024, Nature Protocols, 19, 3292.

[19] Yu, G. 2020, Methods in Molecular Biology, 2117, 207.

[20] Yu et al. 2010, Bioinformatics, 26, 976.

[21] Yu et al. 2016, Molecular BioSystems, 12, 477.