nature communications

Article

https://doi.org/10.1038/s41467-024-54603-5

Adaptation in human immune cells residing in tissues at the frontline of infections

Received: 28 September 2023

Accepted: 14 November 2024

Published online: 28 November 2024

Check for updates

Irepan Salvador-Martínez[1,5], Jesus Murga-Moreno[2,5], Juan C. Nieto[1], Clara Alsinet[1], David Enard[2] ✉ & Holger Heyn[1,3,4] ✉

Human immune cells are under constant evolutionary pressure, primarily through their role as first line of defence against pathogens. Most studies on immune adaptation are, however, based on protein-coding genes without considering their cellular context. Here, using data from the Human Cell Atlas, we infer the gene adaptation rate of the human immune landscape at cellular resolution. We find abundant cell types, like progenitor cells during development and adult cells in barrier tissues, to harbour significantly increased adaptation rates. We confirm the adaptation of tissue-resident T and NK cells in the adult lung located in compartments directly facing external challenges, such as respiratory pathogens. Analysing human iPSC-derived macrophages responding to various challenges, we find adaptation in early immune responses. Together, our study suggests host benefits to control pathogen spread at early stages of infection, providing a retrospect of forces that shaped the complexity, architecture, and function of the human body.

The need to respond to ever-changing challenges posed by pathogens makes host-pathogen interactions a clear example of the evolutionary arms race. In the last couple of decades, many population genetics and molecular evolution studies, which study the change in the genetic composition of populations and species, have shown that the immune system is under constant evolutionary pressure and provided multiple examples of genes evolving under increased rates of adaptation. The evolution of human immune genes has even been associated with specific diseases and pathogens, such as malaria[1,2] and the Black Death pandemic that killed 30–50% of the Afro-Eurasian population in the XIV century[3]. At a deeper evolutionary time scale, studies on primates and mammals showed that a significant amount of adaptation is due to the arms race between viruses and their hosts[4]. Such a competition not only plays a selective pressure leading to the rapid fixation of advantageous amino acid changes on multiple antiviral protein-coding genes but also in non-antiviral protein-coding genes. As an example, members of the APOBEC family, PKR and TRIM5, antiviral molecules that directly attack viral molecules, as well as TFRC and ANPEP, non-antiviral proteins involved in iron transportation and viral receptors respectively, are well-known examples with unusual rates of adaptation in several organisms including humans[4–9]. However, previous knowledge of specialised immune genes is required to make the connection between the host's positive selection process and pathogens, and such well-understood examples represent only a small fraction of all adaptation events in immune genes[4].

Technically, previous studies on immune gene evolution have focused on measuring ω (the ratio of non-synonymous, d_N, to synonymous, d_S, substitution per site)[10–12], to determine pervasive or episodic adaptive evolution of the immune system[4–7,9]. Nonetheless, non-synonymous changes may alter the function and protein-coding structure, so they tend to be deleterious and removed from the population by the action of purifying selection, which restricts the ability of ω to identify positive selection and tends to mask the impact of adaptive substitutions. To circumvent these limitations, the McDonald Kreitman test (MK-test[13]) combines both divergence and polymorphic data to summarise the rate of adaptation with the quantity α (the proportion of adaptive non-synonymous substitution)[14]. Because adaptive mutations are supposed to contribute almost exclusively to divergence and not polymorphism, polymorphic data is used as a background to calibrate the expected

[1]CNAG, Centro Nacional de Análisis Genómico, Barcelona, Spain. [2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [3]Universitat de Barcelona (UB), Barcelona, Spain. [4]ICREA, Barcelona, Spain. [5]These authors contributed equally: Irepan Salvador-Martínez, Jesus Murga-Moreno. ✉e-mail: denard@arizona.edu; holger.heyn@cnag.eu

divergence rate under the neutral model in the presence of purifying selection[15,16], increasing the power of detecting positive selection. However, during the last decades, it has been repeatedly shown that weakly selected polymorphism can attain low and high frequencies[17–20], depending on the underlying Distribution of Fitness Effects (DFE)[21], biasing MK-test and α inferences. To account for these biases, we used ABC-MK, an MK-test extension proposed by Uricchio et al.[22] and recently extended by Murga-Moreno et al.[23]. This approach is particularly useful when the proportion of weakly advantageous alleles segregating in the Site Frequency Spectrum is relatively high and in the presence of the evolutionary process known as background selection (BGS), present also in the human genome[22]. BGS decreases neutral genetic diversity when neutral variants are removed together with the deleterious variants they are linked to and can also impact positive selection signatures, especially at high frequencies[22]. ABC-MK is an Approximate Bayesian Computation (ABC) which takes advantage of $\alpha_{(x)}$ summary statistics proposed by Messer and Petrov [24]. By modelling $\alpha_{(x)}$ with multiple DFE random parameter combinations along with BGS, ABC-MK exploits $\alpha_{(x)}$ patterns to infer α while providing flexibility to analyse heterogeneous datasets from the human genome, where BGS strength is variable and weakly adaptive variants are an essential component of adaptation[22,23]. In addition to quantifying adaptation with α, we also estimate ω_a, the rate of adaptive non-synonymous substitution relative to the mutation rate, and we explore the relationship between adaptation and evolutionary gene age in the human immune system by a gene phylostratigraphy approach. This method involved ranking genes into different Phylostrata based on their evolutionary emergence time[25].

Other important challenges, when studying adaptation in the immune system, are the spatio-temporal dynamics and cell-type specificity of immune gene expression that have not been taken into account in previous studies, mainly due to a lack of comprehensive transcriptomic reference maps of the immune system at the cell-type level. In recent years, the Human Cell Atlas (HCA) has created comprehensive transcriptome reference maps of healthy human organs and tissues at different developmental and adult stages, with the ultimate objective of characterising every cell type in the human body[26]. As part of this effort, the Human Developmental Cell Atlas bionetwork generated a reference map of the developing immune system[27] and a complementary study provided an Adult Multi-tissue Immune Cell Atlas[28]. Together, both atlases represent the most comprehensive transcriptomic characterisation of the human immune system and provide the basis for a better understanding of immune protein adaptation at cell-type resolution.

In here, we used ABC-MK along with the Developmental and Adult Immune cell atlases, to quantify adaptation at the cell-type level since humans split from chimpanzees. This helped us to understand immune protein adaptation within human immune cells and to identify immunological challenges driving immune adaptation. By leveraging spatio-temporal single-cell resolved reference maps of >1.2 million cells, we pinpoint immune cell types and states under high adaptation rates. Observing a high adaptation in tissue-resident-memory T cells, we expanded our analysis integrating single-cell and spatial transcriptomics data from the Lung Cell Atlas[29] to provide a detailed map of adaptation of immune cells in barrier tissues. Finally, to better understand the selective pressures, immune response timing and the role of natural selection on human immune cells, we analysed the innate immune response to different stimuli that mimic various inflammatory and pathogen exposures, using activated induced pluripotent stem cell (iPSC)-derived macrophages, a successful pathogen infection model. Altogether, our analyses of adaptive evolution charted the landscape of cell types contributing to the arms race between the human host and its pathogens and helped to clarify the origin of abundant host immune protein adaptation, allowing us to hypothesise immune functions and pathogen classes that drove immune adaptation during human evolution.

Results

Building an Immune Atlas of Adaptation

The Human Developmental Cell Atlas bionetwork generated a reference map of the developing immune system with ~900,000 cells from 25 donors and more than 100 annotated cell states. The Developmental Atlas includes data from haematopoietic organs (yolk sac, liver, bone marrow), lymphoid organs (thymus, spleen and lymph node), non-lymphoid tissues (skin, kidney and gut) and spans 4–17 weeks after conception[27]. The Adult Multi-tissue Immune Cell Atlas comprised ~300,000 cells from 21 donors and 16 tissues. Selected tissues included primary (bone marrow) and secondary (spleen and lung-draining and mesenteric lymph nodes) lymphoid organs, mucosal tissues (gut and lung), as well as blood and liver[28]. In the Developmental Immune Cell Atlas, the immune compartment represented hematopoietic stem cell (HSC) and hematopoietic progenitors, erythro-megakaryocytes as well as the major myeloid and lymphoid cell lineage populations. The spatio-temporally resolved reference captured the emergence of immune lineages from hematopoietic progenitors, characterising immune cell development before tissue seeding to the developing lymphoid organs, such as the thymus, spleen, lymph nodes, and peripheral non-lymphoid organs. On the other side, the Adult Immune Tissue Atlas provided the characterisation of mature immune cells in circulation and specialised to tissue microenvironments, such as tissue-resident macrophages and memory T cells. Together, both datasets provide an annotated resource that represents a complete diversity map of human immune cells in a steady state from the embryo to the adult.

In order to identify human immune cell types with signatures of selection in development and adulthood and to generate a comprehensive Immune Atlas of Adaptation (Fig. 1), we retrieved the annotated single-cell datasets from both resources and de novo computed cell-type-specific differentially expressed genes (DEG). Specifically, after removing cell clusters reported as cycling (no cell-type specificity), cell doublets or low-quality cells, we performed a differential gene expression test for each cell population compared to all remaining cells in the same compartment (Wilcoxon test, FDR $< 1 \times 10^{-4}$, logFC > 1). For each cell population, we used the top 500 DEG sorted by logFC value (or all DEG for populations with <500 genes) to normalise differences in the number of differentially expressed genes between cell populations, while maximising the number of genes required to perform accurate MK-test estimations in human lineage following Murga-Moreno et al.[30] and Murga-Moreno et al.[23] results (see "Methods" section for further information). In the Developmental Atlas, some cell types from the progenitors compartment are also present in an additional compartment. For example, Pre-ProB cells and Promonocytes are also part of the Lymphoid and Myeloid compartments, respectively. For these cell types, as the DEG were obtained independently for each compartment, we report the ABC-MK-test results obtained for each compartment.

To determine the proportion of adaptive non-synonymous substitution (α) driven by weakly and strongly beneficial alleles ($\alpha = \alpha_S + \alpha_W$), we ran ABC-MK independently for each cell population[22,23]. Each estimation used random parameter combinations to estimate 10^5 different $\alpha_{(x)}$ combinations accounting for multiple Distributions of Fitness Effects (DFE) and background selection while sampling empirical data and accounting for partial recombination (see "Methods" section, Supplementary Data 1). We used `ABCreg`[31] to perform ABC comparing ABC-MK sampled $\alpha_{(x)}$ and empirical datasets. Because α prior values range from 0 to 1 to measure the proportion of adaptive substitutions, we used the mode of the posterior distribution

https://doi.org/10.1038/s41467-024-54603-5

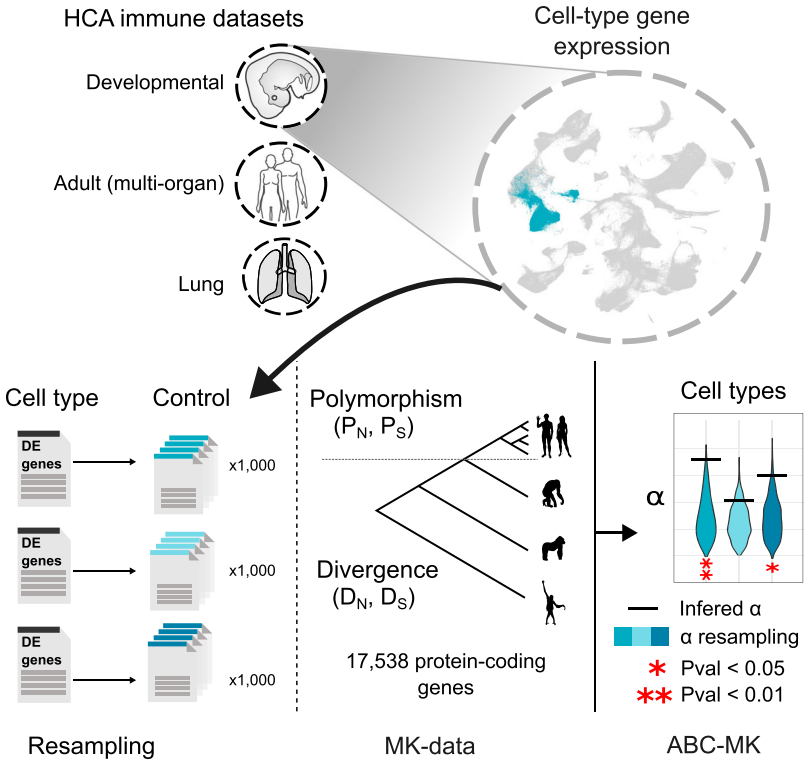

Fig. 1 | Methodological abstract. Single-cell transcriptomic datasets from the Human Developmental Immune Cell Atlas[27], the Adult Multi-tissue Immune Cell Atlas[28] and the Lung Cell Atlas[29] were retrieved and the top 500 differential expressed genes (DEG) for each immune cell type were computed de novo. For significance tests, we compare each dataset with 1000 control datasets having similar average values of multiple confounding factors (see "Methods" section). We ran ABC-MK to measure the adaptation rate using data from the Thousand Genomes Project[83] as polymorphic data and human reciprocal orthologs on chimpanzee, gorilla and orangutan genome assemblies as divergent data. Dataset icons were modified from the Reactome icon library[100] licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Species icons are available from and.PhyloPic (http://www.phylopic.org).

to quantify α, α_S and α_W, better accounting for cases where adaptation is low or absent (posteriors distributions displaced to 0 values).

To determine a significant increase in adaptation, we build a null distribution accounting for 1000 control datasets matching the size of the analysed immune genes subset for each cell population, following a resampling approach[32,33]. The procedure allowed us to match, resample and compare genes with similar overall average values for many confounding factors that could affect the adaptation rate and mislead the comparison (see "Methods" section for further information)[32,34]. We evaluated each cell population estimation of its corresponding null empirical distribution and estimated the associated *p*-values ("Methods" section). This process ensures statistically significant outcomes surpassing the empirical distribution primarily from positive selection related to the cell populations' immune function, rather than other confounding factors (see "Methods" section for a detailed list of possible confounders). However, this process also relies on a limited supply of control genes, which can bias the variance of the null controls ("Methods" section). To address this limitation, we also provide an unbiased measure of the type I error ("Methods" section), which allowed us to estimate the number of cell types below an error threshold by chance (Supplementary Data 10). Throughout the manuscript, we report the raw resampling *p*-values but also provide a more stringent analysis based on the type 1 error estimation and respective FDR values from multiple testing corrections (Benjamini–Hochberg correction[35]) (Supplementary Data 10).

To determine whether significant α values directly rely on the action of positive selection and are not due to differences in detection power, we further quantified the relationship between the number of DEG and the number of cells to the α values, defining empirical *p*-values through linear regression analysis. Here, we did not detect an effect of the number of between the DEG or cells on α or the empirical *p*-values (Supplementary Fig. 1). In addition, we tested cell counts impact on the α estimation by progressively downsampling cell numbers counts (100%, to 50%, 20%, 10%, 5%, and 1%), without detecting systematic biases (Supplementary Fig. 2). In addition, to evaluate the possibility that the significant adaptation signal is shared between cell types, we checked the total number of shared genes. Here, we identified primarily pairs of lineage-related cell types and, hence, do not expect any shared genes to drive adaptation globally, making the test analysis statistically independent (see Supplementary Fig. 3).

Finally, we replicated the analyses by estimating ω_a, the rate of adaptive non-synonymous substitution relative to the mutation rate, to provide further evidence of protein adaptation (see Supplementary Figs. 4–7 and Supplementary Data 2–5). Our results showed a picture of highly elevated adaptation, similarly on α and ω_a estimations, in diverse immune cell populations in both the developing embryo and in adult human tissues.

Most foetal progenitor cell types have a high adaptation rate

In the Developmental Immune Cell Atlas, we found eight hematopoietic stem and progenitor states, nine lymphoid, three myeloid and one Megakaryocyte population with significantly elevated adaptation rates (Fig. 2, Supplementary Fig. 4, Supplementary Data 2, 6). Intriguingly, we found Hematopoietic Stem Cells (HSCs) and seven Hematopoietic Progenitor Cells (HPCs) with a high adaptation rate in the progenitor compartment. Specifically, five out of seven myeloid progenitors (MEMP, CMP, GMP, Promonocytes and early Megakaryocytes) and two out of four lymphoid progenitors (LMPP and PreProB) showed a high adaptation rate. HSCs give rise to all cell lineages of the immune system and are the only hematopoietic cells with self-renewal and full

https://doi.org/10.1038/s41467-024-54603-5

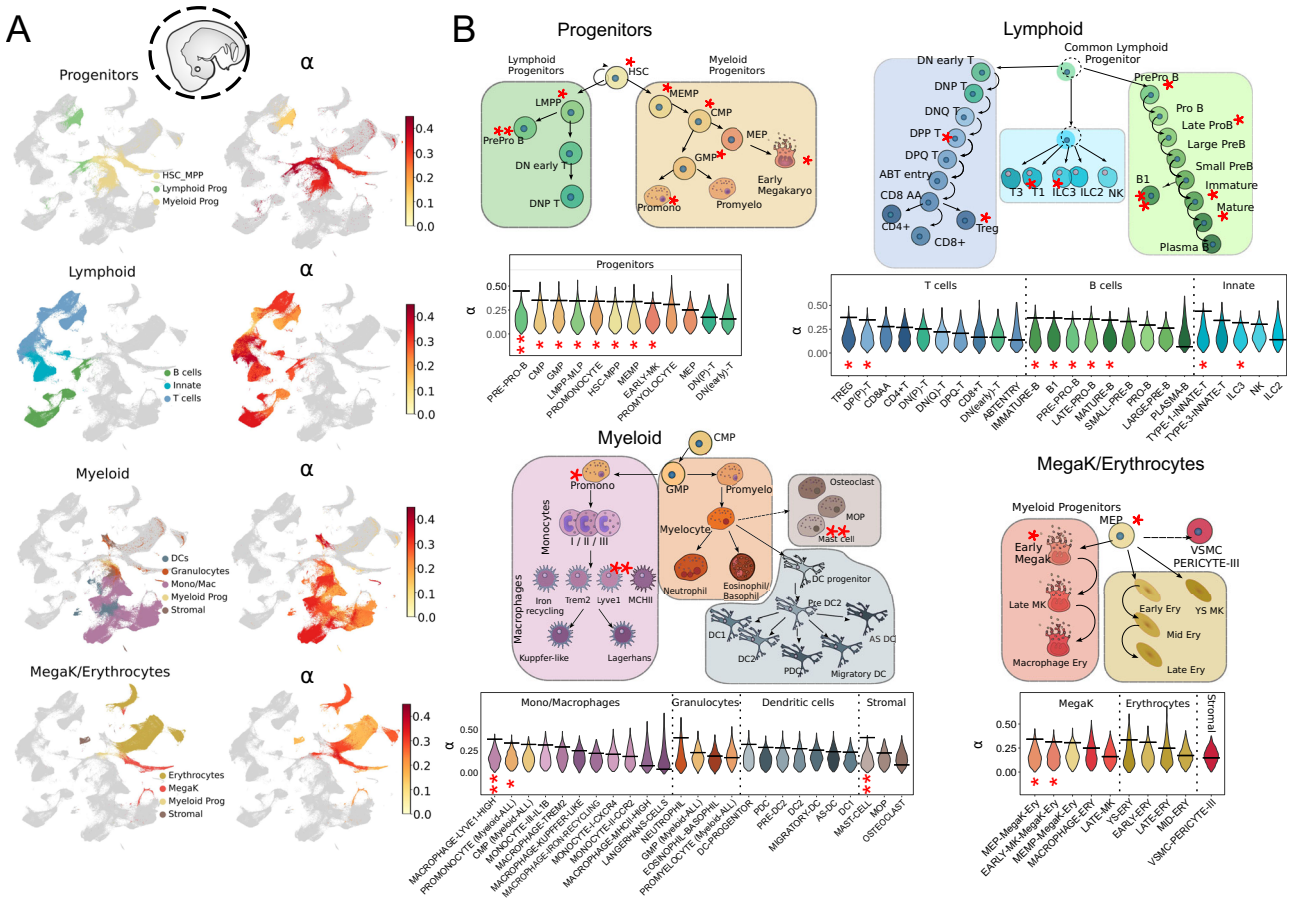

Fig. 2 | Protein adaptation in the Developmental Atlas. A On the left, each compartment is highlighted in the global UMAP (from ref. 27) with a colour code that corresponds to specific lineages. On the right, the inferred α value for each compartment is projected on the same UMAP. Note that the α value is inferred at the cell-type level. **B** A scheme with the putative lineage relationships between the cell types within each compartment is shown. Lymphoid cells with a dashed outline are not part of the analysed dataset and are only depicted to clarify lineage relationships. Under each scheme we show the inferred α value of each cell type as a horizontal bar and the α value distribution of its control resampling dataset ($n = 1000$) as a violin plot. The asterisks highlight the cell types with a significantly high α (p-values) were estimated through one-sided α empirical distribution. *p-value < 0.05, **p-value < 0.01 and all significant cell lines account for a type 1 error < 0.05. Exact p-values are provided in Supplementary Data. Embryo and cell-type icons were modified from the Reactome icon library[100] licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

differentiation capacity through lineage progenitor stages (e.g., Lympho-myeloid primed progenitor and Megakaryocyte-erythroid progenitor, LMPP and MEMP). HSCs are, however, also a sensory hub of the immune system, sensing and responding to molecular cues and physical interactions with their niche's complex multicellular network[36]. For example, depletion of certain blood cell populations activates HSCs and subsequently HPCs to undergo differentiation[37]. The adequate readout of signals from the hematopoietic stem and progenitor niche is therefore crucial to ensure haematopoietic homoeostasis in development and throughout life[36]. Additionally, HSCs respond to infections and inflammatory signals through different mechanisms, including sensing pathogen-derived products and pro-inflammatory cytokines produced after pathogen infection[37]. Considering the above, we hypothesise that the high adaptation rate inferred in foetal HSCs and HPCs is related to their key position as a guardian of the hierarchy controlling homoeostasis and immunomodulation in the developing embryo.

Tregs, ILC3, B1 and B cells under maturation among foetal lymphoid cells with a high adaptation rate

The lymphoid compartment is subdivided into three lineages: T cells, B cells and Innate lymphoid cells. T cells and B cells are the major components of the adaptive immune system in humans. Noteworthy, adaptive immunity is a shared feature of all vertebrates with jaws (gnathostomes) and is thought to have evolved more than 400 million years ago in the placoderms, the first jawed vertebrates[38]. T and B cells use cell surface receptors to recognise antigens of pathogens that trigger cytotoxic activation and the production of antibodies[39]. Two additional lymphoid cell types are devoid of antigen receptors: Natural Killer (NK) cells and Innate lymphoid cells (ILCs). Functionally, NK cells complement cytotoxic T cells in killing infected cells, whereas ILCs share many functions of T cells, such as sensing stress signals, microbial compounds and the tissue cytokine milieu[40].

Consistent with the finding in the haematopoietic stem cell and progenitor compartment, significant signals of adaptation were found in lymphoid progenitor states. Specifically, we found five out of nine lymphoid cell subtypes in the B lineage (Pre ProB, Late ProB, B1, Immature and Mature B cells), two out of ten in the T cell lineage (DPP and Treg) and two out of five in the Innate lymphoid cell subtypes (ILC3 and Type 1 Innate T) with a high adaptation rate. From the spatio-temporal perspective, the high adaptation of Immature and Mature B cells coincides with the migration of Immature B cells in the foetal bone marrow or liver to the spleen, where immature cells transition to mature B cells. This transition is evident in the observed organ distribution of annotated cell types in the Developmental Atlas, as the proportion of Immature and Mature B cells in the liver decreases from

36% to 6%, whereas it increases in the spleen from 43% to 81% (Supplementary Fig. 8). During this maturation process, a checkpoint is reached where potentially autoreactive B cells are eliminated[41]. Failure to eliminate autoreactive cells can lead to the development of autoimmune diseases later in life, such as Systemic Lupus Erythematosus and Rheumatoid Arthritis[41]. Similarly, regulatory T cells (Treg) control the immune response and help prevent autoimmune diseases[42]. In contrast, the function of B1 cells and innate lymphoid cells is to provide protection from external challenges. B1 cells are a subclass of B cells that are involved in the humoral immune response and have been exclusively described during embryo development[43,44]. Although B1 cells are not part of the adaptive immune system (lack memory response), they perform similar roles as other B cells, such as the production of antibodies against antigens and acting as antigen-presenting cells. On the other hand, ILC3 participates in innate mechanisms on mucous membranes, contributing to host-commensal mutualism and pathogen clearance. They are also involved in the activation of NK cells, including cell lysis and secretion of pro-inflammatory cytokines, and protect the intestinal mucosa from infections of various pathogens[40].

Lyve1+ macrophages and Mast cells among foetal myeloid cells with a high adaptation rate

Innate immunity, as displayed by myeloid cells, is a more ancient form of host defence against infection that involves pattern recognition systems (e.g., RLRs and TLRs), whose molecular components are shared by all multicellular animals[45]. The myeloid compartment is subdivided into four main lineages: Monocytes/Macrophages, Granulocytes, Dendritic cells and Stromal cells (Fig. 2). Myeloid cells are highly plastic, i.e., changing cellular states depending on external stimuli, and are recruited to tissues after pathogen invasion via chemokine receptors. Their role, however, is not limited to immune responses; myeloid cells are known to play an important role in supporting tissue homoeostasis and development[46]. In the myeloid compartment, we found two out of eleven cell states in the Monocyte/Macrophage lineage (Macrophage Lyve1 High and MEMP) and one out of three cell types in the stromal compartment (Mast cell) with significantly elevated adaptation rate (Fig. 2B).

Macrophages with high Lyve1 expression have been found to localise in a specific niche surrounding blood vessels[47] and to have an important role in maintaining homoeostatic vascular tone by regulating collagen production[48]. These tissue-support functions are also reflected in the GO terms enriched in their DEG, like “endothelial cell migration" and “mesenchyme development" (Supplementary Fig. 9 and Supplementary Data 11). Suo et al.[27] showed that the Lyve1- subpopulation of macrophages in the yolk sac exhibits an increased self-renewal potential across various organs. Mast cells, with a long history predating the development of adaptive immunity, are primarily renowned as significant sources of mediators responsible for acute allergic reactions and IgE-dependent immediate hypersensitivity reactions[49]. Furthermore, both early macrophages and Mast cells were noted for their roles in processes such as angiogenesis, tissue morphogenesis, and the maintenance of tissue homoeostasis before adopting their more traditional immunological functions. We, therefore, hypothesise that the functions under adaptation in Mast cells and Lyve1 macrophages are related to their role in tissue support and homoeostasis.

Finally, in the Megakaryocyte and Erythroid compartment, we only found two out of ten cell types with high adaptation rates: early megakaryocytes and MEP cells, progenitor cells of more differentiated Megakaryocytes and Erythrocytes.

Long-lived resident cells show a high adaptation rate in adults

In contrast with the Developmental Immune Cell Atlas, where we found mostly progenitor populations with elevated adaptation rates, the picture was very different in the Adult Tissue Immune Atlas. Here, we found only four cell types with a high adaptation rate, belonging to the most differentiated and long-lived immune populations (Fig. 3, Supplementary Fig. 5 and Supplementary Data 3, 7). In the T cell lineage (Fig. 3B), the cell types with the highest adaptation were tissue-resident-memory effector memory CD8+ T cells (Trm_em_CD8) and resident-memory Th1 and Th17 (Trm_Th1_Th17). Trm T cells are long-lived cells that are resident at sites of prior antigenic exposure, i.e., epithelial barrier tissues (lung, skin, gut), to support tissue surveillance and as protective immunity to reinfection[50,51]. For example, influenza-specific lung Trm T cells localise in clusters close to lung epithelial cells, responding rapidly to reinfection through the direct release of cytotoxic mediators, cytokines and chemokines[51]. In humans, higher frequencies of CD8+ Trm cells in the airway have been associated with improved viral control and reduced symptom severity. Trm T cells express different marker genes with potential differences in selection during host-pathogen evolutionary pressure[51]. We, therefore, reasoned that Trm T cell populations with high protein adaptation could point to selective processes specific to the barrier tissue and pathogen entry site. Intriguingly, both Trm cells with a high protein adaptation were enriched in lymphoid tissue and also in non-lymphoid tissue including liver and lungs, in contrast to Trm cells without signal for adaptation (Trm_Tgd and Trm_gut_CD8) that mostly localised in the gut (Supplementary Fig. 10). Noteworthy, the presence of the Trm T cell populations with a high protein adaptation in the liver and lungs had been validated in situ through targeted spatial profiling[28]. These results suggest a high adaptation rate related to a tissue-specific cellular defence against respiratory pathogens, to provide a rapid response against airborne viruses and bacteria, thereby limiting their growth and spread through the organism.

Consistently, the other cell types showing a high adaptation rate in adults are also found in barrier tissues. CD56-bright CD16-negative Natural Killer cells (NK_CD56bright_CD16-; NK CD16neg) also show tissue-resident profiles and are specialised in cytokine production, in contrast to NK CD16pos cells with cytotoxic activity mostly found in peripheral blood[52]. Their role as immune response mediators is also evident by the GO term enrichment in their DEG, such as “Response to the molecule of bacterial origin" (GO:0002237; Supplementary Fig. 9 and Supplementary Data 12). The role of NK CD16neg cells has been described as regulatory, influencing both innate immunity through cytokine production, but also contributing to the local adaptive immune response[52]. Dendritic cells (DCs) are also found in barrier tissue, such as the lung mucosa, where they sample antigens for their presentation to adaptive immune cells in tissue-draining lymph nodes. Together, we found a high adaptation rate in adult immune cells mainly in cell types localised to barrier tissue as a first line of defence against pathogens, leading us to speculate that the main selective pressure in adult humans comes from external challenges, potentially respiratory pathogens entering the lung epithelium, and that host adaptation is most efficient when it affects early stages of infection before pathogens spread broadly.

Adaptation of immune resident cells in lung barrier tissues

In order to investigate the adaptation of lung-resident immune cells further, we expanded our analysis to the Human Lung Cell Atlas[29]. In contrast to the Adult Immune Cell Atlas, where cells derive from different organs/tissues, the Lung Cell Atlas provides in-depth transcriptomic profiles of immune cells in healthy human lungs (82, 180 immune cells), thus, allowing a systematic view of the adaptation of lung-resident cells. Further, the Lung Cell Atlas includes five proximal-to-distal locations, profiled with single-cell and spatial transcriptomics, allowing to test for variances in the distribution of tissue-resident cell types (Fig. 4B). Strikingly, the results from the ABC-MK-test confirmed the high adaptation rate of CD8 TRM and NK CD56-bright cells (named “Trm em CD8" and “NK CD56-bright CD16-" in the Multi-tissue Adult

https://doi.org/10.1038/s41467-024-54603-5

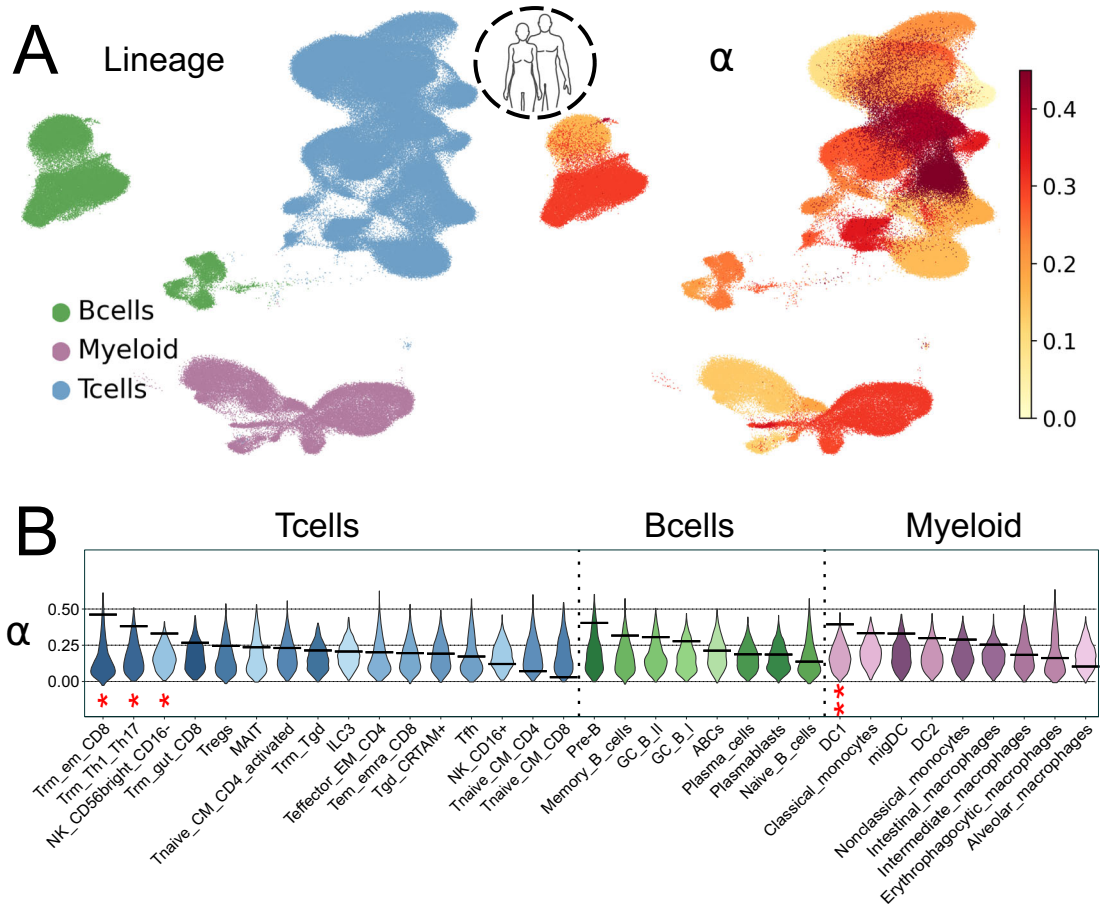

Fig. 3 | Protein adaptation in the Adult Atlas. A On the left, each compartment is highlighted in the global UMAP (from ref. 28). On the right, the inferred α value is projected on the same UMAP. Note that the α value is inferred at the cell-type level. Human icons were modified from the Reactome icon library[100] licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). **B** For each cell type the inferred α value is shown as a horizontal bar and the α value distribution of its control resampling dataset ($n = 1000$) is shown as a violin plot. The asterisks highlight the cell types with a significantly high α (p-values were estimated through one-sided α empirical distribution. *p-value < 0.05, **p-value < 0.01 and all significant cell lines account for a type 1 error <0.05. Exact p-values are provided in Supplementary Data).

Adult Atlas), providing independent validation of the positive selection of these tissue-resident immune cells in the lung tissue barrier against inhaled pathogens (Fig. 4C). By leveraging the spatial transcriptomics information of the lung dataset[29], we were further able to interrogate the tissue location of CD8 TRM and NK CD56-bright cells. Specifically, we used the cell-type loadings after data deconvolution[29,53] and normalised the mean loadings per cell type across all samples ($N = 11$) to get a weighted value, representing the relative abundance at distinct locations in the lung. Clustering weighted values grouped cell types by their lung tissue distribution, revealing CD8 TRM and NK CD56-bright cells and other cell types with high adaptation to be preferentially localised in barrier layers of the lung ("Small airways" and "Multilayer epithelium"; Fig. 4B). Leveraging the resolution of the Lung Cell Atlas, we also identified additional lung-resident immune cell types in the T cell (Tregs and NK CD16 high), myeloid (Macrophages CX3CR1+, DC activated, and DC2) and B cell compartment (B naive and B plasma IgG; Fig. 4B). Altogether, our results in the lung support our hypothesis of high adaptation localised in barrier tissues.

Pathogen-associated early response in macrophages shows a high adaptation rate

To further delineate the type of selective pressure and the immune response timing leading to adaptation in human immune cells, we charted the dynamics of different stimuli that mimic different pathogen exposures. To this end, we retrieved data from the Human Induced Pluripotent Stem Cells Initiative (HipSci), a project generating human induced pluripotent stem cells (iPSC) from hundreds of individuals, as a resource for directed cell-type differentiation and targeted perturbation to chart disease mechanisms[54]. More specifically, we used data from a study involving >200 iPSC lines differentiated into macrophages and exposed to 10 stimuli that emulate responses to pro/anti-inflammatory cytokines and bacterial or viral infections at two different time points (Fig. 5A, B). This data allowed us to test the protein adaptation rate of different macrophages activation states, one of the most plastic immune cell types[55]. To identify processes and timing under adaptation, we evaluated the DEG from macrophages stimulated for 6 h and 24 h vs control unstimulated samples. For each stimulus, we categorised responses as "early" (exclusively after 6 h), "late" (exclusively after 24 h) or "sustained" (found at both time points)[56,57]. This classification simulates the sequential release of inflammatory mediators by macrophages in response to pathogen products, which constitutes a crucial part of the innate response against invading pathogens. It is worth noting that while an initial inflammatory response is vital, restoring macrophages to their resting state is equally important, as prolonged macrophage activation leads to tissue damage[58]. Our results (Fig. 5C, Supplementary Fig. 7, Supplementary Data 5, 9) show high protein adaptation mainly in the early responses, specifically to stimulation with IFNG, R848 and CIL. IFNG is known to induce a pro-inflammatory response in Macrophages associated with intracellular pathogens, including bacterial, protozoal or viral

https://doi.org/10.1038/s41467-024-54603-5

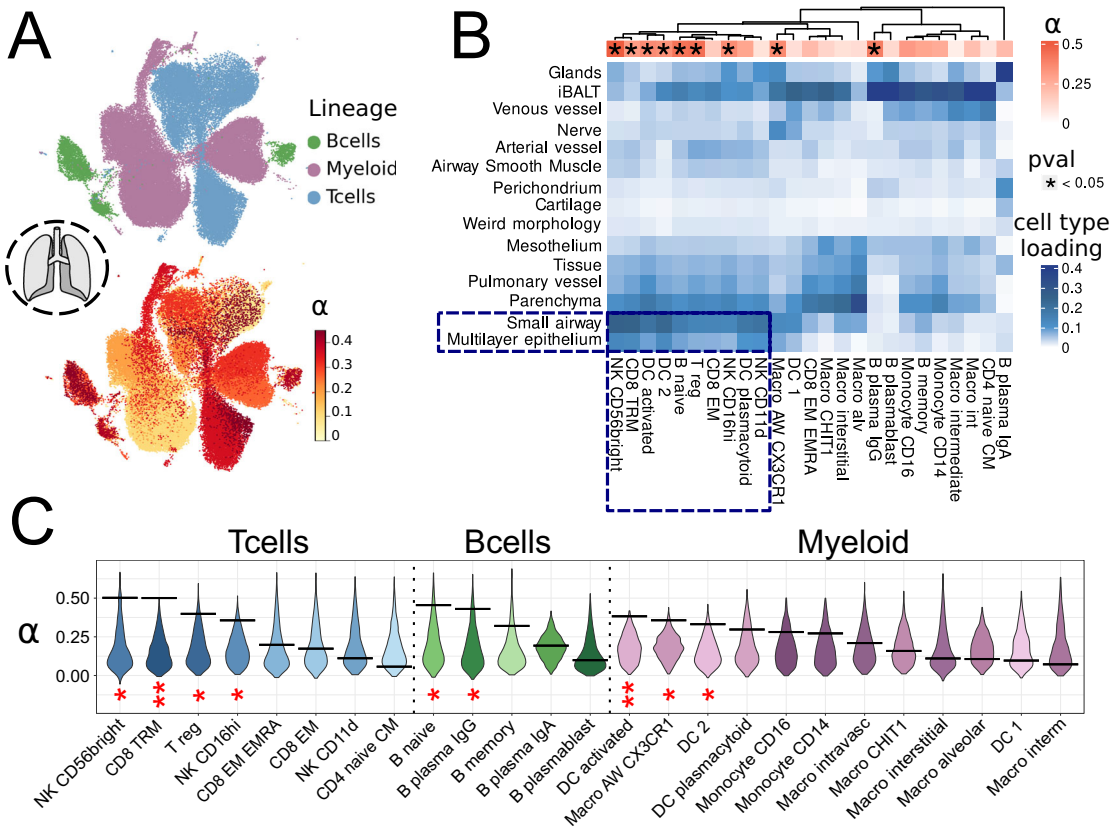

Fig. 4 | Protein adaptation in the Lung Atlas. A Lung immune cells UMAP showing the different compartments (top) and the inferred α value (bottom). Note that the α value is inferred at the cell-type level. Lung icon was modified from the Reactome icon library[100] licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). **B** Cell-type loadings heatmap. Normalised cell-type loading values (in blue) represent the relative abundance of a given cell type in the different tissues. At the top, a dendrogram shows the cell types (columns) clustering, to show cell types with a similar tissue distribution. The alpha value for each cell type is shown (in red) with asterisks highlighting cell types with a significantly high α (*p-value < 0.05). Blue dotted areas highlight that the majority of cells with significant alpha localise preferentially to "Small airway" and "Multilayer epithelium". **C** For each cell type the inferred α value is shown as a horizontal bar and the α value distribution of its control resampling dataset ($n = 1000$) is shown as a violin plot. The asterisks highlight the cell types with a significantly high α (p-values were estimated through one-sided α empirical distribution. *p-value < 0.05, **p-value < 0.01 and type 1 error < 0.05. Exact p-values are provided in Supplementary Data).

infections. R848 is an agonist of Toll-like receptor (TLR) 7 and 8 that mimics a viral infection, whereas CIL represents a combination of stimuli (CD40 ligand + IFNG + sLPS) to mimic response to bacterial infection. In addition to adaptation signals in the early responses, we found the sustained response to Interleukin-4 (IL4) to have a high adaptation rate. IL4 is generally viewed as a T helper 2 cytokine capable of polarising macrophages into an anti-inflammatory phenotype[59].

Immune cells are enriched with Eutelostome genes

Having charted the immune types with significant adaptation rates, we then focused on the evolutionary history of genes expressed in the human immune cells. Previous studies[60,61] have shown that evolutionarily younger genes have a higher adaptation rate than older genes. Thus, we reasoned that cell types with high adaptation to be similarly enriched and expressing evolutionary younger genes. To test this hypothesis, we used a phylostratigraphy approach[25], where genes are ranked in different categories or phylostrata (PS) based on their evolutionary emergence time. For humans, nineteen phylostrata have been defined: the first phylostratum (ps1) contains the oldest genes shared by all living organisms, while the last phylostratum (ps19) harbours genes that appeared in the primate lineage and are shared by humans, chimpanzees and marmosets, among others. Using PS value estimates for all human genes[62], we tested for PS enrichment in DEGs of each immune cell type using a hypergeometric test. Our results showed that most immune cell types, in both Developmental and Adult atlases, are enriched with genes of the Euteleostome phylostrata (PS 12, Fig. 6). Euteleostomes (sometimes referred as "bony vertebrates") is a clade that emerged >400 Mya, which includes jawed fish and their descendants: bony fish, amphibians, reptilians, birds and mammals. Interestingly, it is believed that the adaptive immune system originated in the first jawed fish (the placoderm) as jawed fish descendants have T cell receptors (TCR), B cell receptors (BCR) as well as Major histocompatibility complex (MHC) genes as a consequence of major macroevolutionary events: two whole genome duplications and the invasion of the RAG transposon[63]. Only a few innate immune cell types showed enrichment of younger PS. In the Developmental Atlas, Macrophages MHCII high, Promonocytes (progenitors), GMP (progenitors), Type 1 innate T cells, and NK cells were enriched with Amniote genes (PS 14) and NK cells were the only cell type enriched with PS 17 genes (subgroup of placental vertebrates that include hoofed and pawed animals). In the Lung Atlas, NK CD16high, NK CD56-bright and gamma-delta T cells (gdT) were the only cell types showing enrichment of genes that evolved after the Eutelostomes (PS ≥ 15).

Discussion

We combined the spatio-temporal transcriptomic dataset of the human immune system and state-of-the-art methods to infer adaptation at the protein sequence level. Looking at the developing and mature immune system through the lens of adaptation provided intriguing insights into the selective pressures that shaped our

https://doi.org/10.1038/s41467-024-54603-5

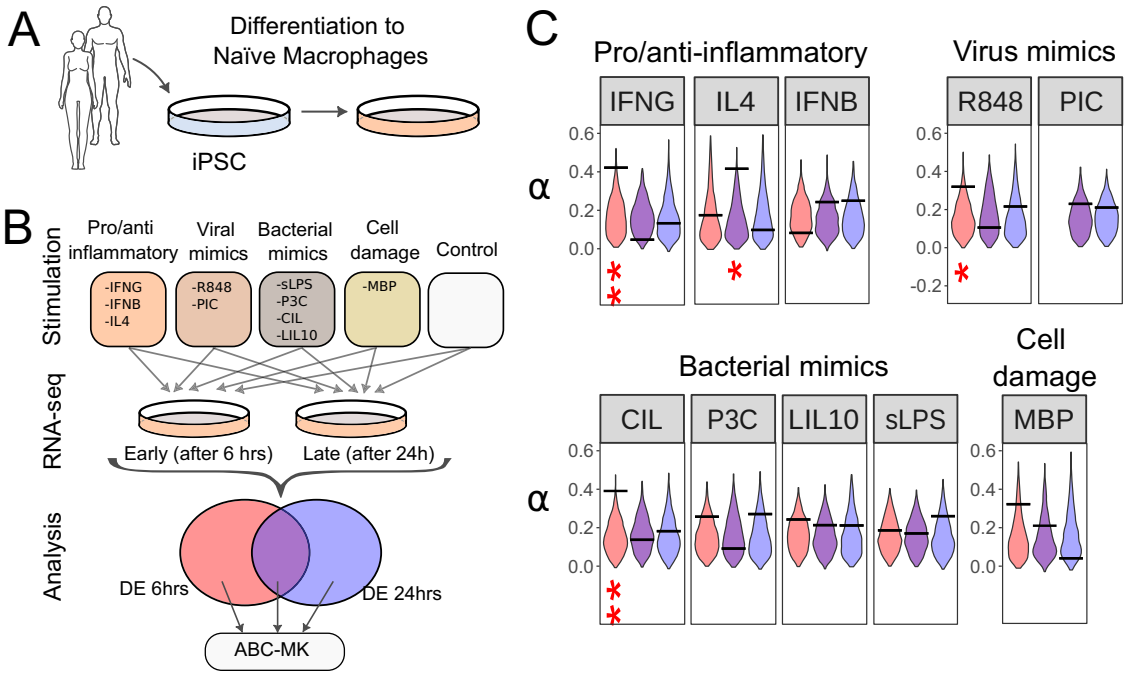

Fig. 5 | Protein adaptation in activated macrophages. A Overview of the experimental workflow followed by Panousis et al.[55]. More than 200 iPSC lines were derived from unrelated healthy donors and differentiated into naive macrophages. Human icons were modified from the Reactome icon library[100] licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). **B** Naive macrophages were stimulated and their transcriptomes were sequenced by RNA-seq at two post-stimulation time points (6 h and 24 h). A differential expression analysis was performed between each timepoint/stimuli and a control. For each stimulus, we further categorised responses as "early" (exclusively after 6 h), "late" (exclusively after 24 h) or "sustained" (found at both time points). **C** For each stimulus and time-specific response we show the inferred α value (horizontal bar) and the distribution of the control resampling dataset ($n = 1000$) as a violin plot. The asterisks highlight the responses with a significantly high α (p-values were estimated through one-sided α empirical distribution. *p-value < 0.05, **p-value < 0.01 and type I error < 0.05. Exact p-values are provided in Supplementary Data).

immune system, including the evolutionary arms race between the human host and its pathogens. Diverging from prior research, we discerned not genes, but cells exhibiting an adaptive signature. Generating an Atlas of Human Adaptation at the cell level, enabled to pinpoint cell types and states under selective pressure and forces that shaped the human innate and adaptive defence since humans split from chimpanzees.

Although the main function of the immune system is defence against pathogens (e.g., viral and bacterial), immune cells also play a crucial role in other processes, such as homoeostasis and development, potential additional targets of selection. Such functions are orchestrated by highly specialised states at different time points and locations during human development and throughout our lifetime. Given the wide phenotypic and functional diversity of human immune cells with their unique importance towards forming a universal defence mechanism, one might have expected evolutionary pressure to be ubiquitous. However, we found the highest adaptation signals at the very extreme ends of lineage commitment: in early foetal progenitors and in adult tissue-resident cells that form immune memory and represent the first line of defence against repeated infections. These results suggest that selective pressure is strongest at the foundation and executive roles, i.e., cells ensuring immune cell renewal and homoeostasis, and at the battlefront (barrier tissues), where rapid responses against pathogens are needed.

During development, the foetus is inside an immune-privilege environment, protected by the placenta that serves as an immunological barrier between the foetus and the mother. The immune system is still relatively immature in newborns and matures throughout childhood until reaching a mature state in adults. The foetal environment is not sterile however and foetal immune cells are exposed to a diverse range of immune-stimulatory molecules, from semi-allogeneic antigens from maternal cells to bacteria. Intriguingly, although stem and progenitor cells are present throughout our lifetime, we only detected high adaptation rates in the foetal population, further underlining the value of the spatio-temporal design and the respective efforts to generate atlases of the developing and adult human body. It further highlights the fact that challenges and tasks of immune cells change continuously during our lives to tackle intrinsic and external challenges[64]. In line with this, the response to stimuli in the foetus can be different from the adult. For example, both adult and foetal DCs respond to toll-like receptor stimulation, but in a different way; in contrast to adult DCs which induce effector T cells, foetal DCs induce Treg cells in a proposed mechanism of tolerance and immune suppression[65].

We detected genes expressed in foetal HSCs and most HPCs to have a high protein adaptation rate, a surprising result considering their major role in fueling lymphoid and myeloid lineages. HSCs, however, also function as an integrating sensory hub of the immune system, responding to environmental cues within their niche, which is highly distinct in foetal and adult stages[66]. Foetal HSCs arise in different organs at different developmental stages, including the yolk sac, placenta and foetal liver, before colonising the bone marrow in postnatal stages[67]. It has been suggested that these dramatic changes in local environments during the foetal period influence HSC activity[67]. Thus, the requirement of the HSCs and lineage progenitors to respond to a changing environment and challenges in a strictly controlled manner is a plausible candidate as a trait under selection. Additionally, early developmental events can determine later effector responses, as shown by a study in mice, where foetal immune cells responded to prenatal inflammation by activating a transient lymphoid-biased HSC commitment that reshaped not only the foetal but also postnatal immune output[68]. This opens the possibility that early differentiation

https://doi.org/10.1038/s41467-024-54603-5

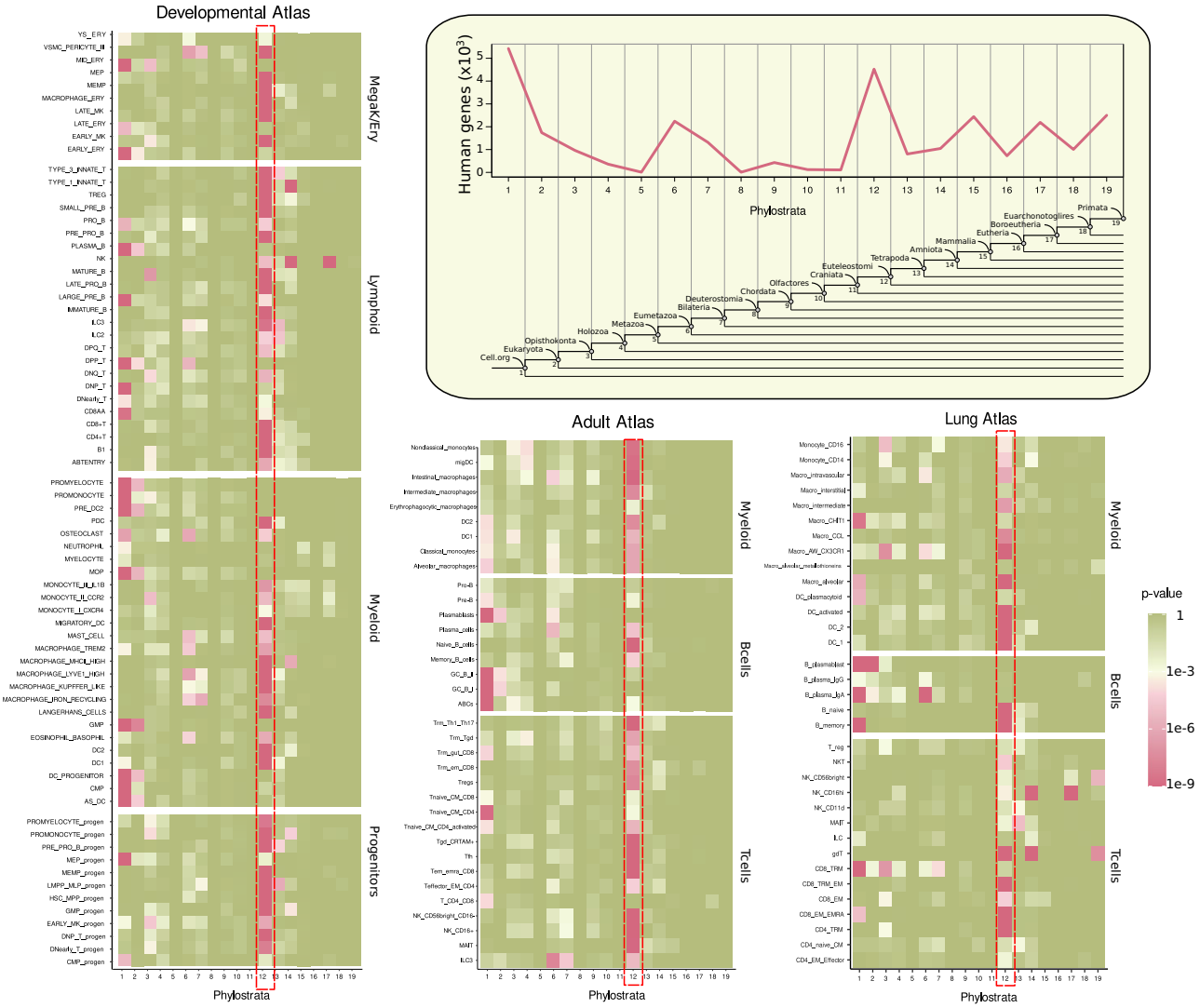

Fig. 6 | Phylostrata enrichment in immune cell types. Heatmaps showing PS enrichment in the top differentially expressed genes of each immune cell type (hypergeometric test, p-value < 0.001) in the Developmental (left), Adult (centre) and Lung (right) Cell atlases. The red dotted rectangles highlight PS 12 (Eutelostomes), enriched in most immune cell types. In the inset box, a line plot shows the number of human genes ranked by their phylostratum (evolutionary time of origin); PS values from ref. 62. At the bottom, a phylogenetic tree is shown with the clades that correspond to each phylostratum (PS1-PS19).

events in progenitor cells might be under selection if these regulate the frequency of immune cell types later in life. In line, the relative proportions of immune cell types found in homoeostasis in adults significantly differ in humans compared to non-human primates[69].

At the other end of immune cell differentiation, we found adult long-lived tissue-resident-memory (Trm) T and CD16neg NK cells with a high rate of protein adaptation in the Multi-organ Adult Cell Atlas. This finding was corroborated by focusing on the lung as an organ with a particular function as a barrier against inhaled pathogens. Moreover, we found these Trm and NK populations preferentially localised in the barrier compartments of the lung, specifically the multilayer epithelium and the small airways. It is indeed known that Trm T cells localise in barrier tissues (sites of recurrent infections) and leverage their immune memory from initial encounters with pathogen-derived antigens to provide rapid response to future reinfections. Functionally, the role of Trm T cells is tightly linked to IFNG signalling. In mice, IFNG produced by Trm cells has been shown to enhance leucocyte recruitment to the infection site and tissue-wide antiviral responses like the upregulation of type I IFN signalling pathway factors[51]. Airway CD8 Trm cells produce IFNG faster than systemic effector CD8 T cells and IFNG-deficient CD8 Trm cells were less effective at controlling pathogen load after viral infection[70]. Similarly, natural Killer CD16neg cells (another lymphoid resident cell type with a high adaptation rate) also produce a high amount of IFNG when exposed to bacterial infections[52]. IFNG produced by T and NK cells in turn activates Macrophages, which secrete a high level of inflammatory cytokines and further promote an inflammatory innate response. IFNG activated Macrophages, also known as “classically activated Macrophages", have a well-documented role in host defence to intracellular pathogens including bacterial, protozoal or viral infections (reviewed in ref. 71).

The high adaptation of resident-memory IFNG-producing cells, together with our observation that the highest adaptation in activated Macrophages was related to the early response to IFNG, suggests that tissue-resident cell types responding to pathogens via IFNG signalling have been under strong adaptive evolution in the human lineage in the millions of years since humans split from chimpanzees. Based on these results, we hypothesise that host immune adaptation is most efficient in cells residing in recurrent infection sites, to facilitate a prompt immune response before invading pathogens spread, mutate and counter-adapt.

A caveat of our analysis is that, although we found several cell types with a high adaptation value (i.e., with raw *p*-values and type I error lower than 0.05), many of these showed an FDR > 10% (mean FDR = 13.2%; range 0.1%–35.2%) after correcting for multiple testing. Increased FDR values in cell types with elevated rates of adaptation could be a consequence of low statistical power as well as the use of controls that match multiple covariates. We strongly believe, however, our results to be meaningful as they are supported by biological processes of adaptation to pathogen exposure and were validated through in vitro experiments and in independent datasets (i.e., Lung Cell Atlas). It is worth noting that developmental MACROPHAGE_LYVE1_HIGH, TYPE_1_INNATE_T, PRE_PRO_B and MAST_CELL, macrophages CIL stimuli at an early stage and lung DC_activated and CD8 TRM cell types accounted for type I error values lower or equal to 0.005 and FDR < 10%, being the lowest among the analysed dataset and indicating that the increase in adaptation rate compared to the base null distribution is a consequence of the action of positive selection.

Finally, our phylostratigraphy results suggest that the considerable protein adaptation found in various human immune cell types not to be a consequence of the recruitment of new genes (which tend to evolve and adapt faster), but rather due to adaptive mutations in immune genes that have been present in our ancestors since the emergence of the Immune Adaptive System, more than 400 million years ago.

Methods

Gene expression data

scRNAseq and snRNAseq annotated data was downloaded in the AnnData format from https://developmental.cellatlas.io/fetal-immune for the Developmental Atlas, from https://www.tissueimmunecellatlas.org/ for the Adult Atlas and from https://5locationslung.cellgeni.sanger.ac.uk/ for the Lung Atlas. Specifically, for the Developmental Atlas, available H5AD files for the four main immune compartments (HSC/progenitor cells, Megakaryocyte/Erythroid cells, Myeloid cells and Lymphoid cells) were downloaded and used for downstream analysis. For the adult Atlas, available H5AD files for the three main immune compartments (Myeloid, T and Innate lymphoid cells and B cells) were downloaded for downstream differential expression (DE) analyses. For the Lung Atlas, the global dataset was downloaded as an H5AD file, from where the different immune compartments (T cells, B cells and Myeloid) were subset and used for downstream DE analysis. The global datasets for both Developmental and Adult atlases were also downloaded as H5AD files only to produce the UMAP plots in Fig. 2A and Fig. 3A.

scRNAseq/snRNAseq pre-processing and DE analysis

For each compartment analysed, we discarded cells originally annotated as proliferating or "cycling" (as a cell in a proliferating state expresses a battery of genes related to the cell division process that are not cell-type specific), doublets or low quality. We performed normalisation and log-transformation of the raw data using the `scanpy.pp.normalize_per_cell()` and `scanpy.pp.log1p()` functions of the scanpy package (version 1.9.1;[72]). For differential expression of the different cell types within a compartment, we ran a DE test using the `scanpy.tl.rank_genes_groups()` function of the `scanpy` package. More specifically, we considered genes upregulated in one cell type versus the rest of the cells in the same compartment using the Wilcoxon test and selected the genes with a Benjamini–Hochberg FDR corrected *p*-value < 1×10^{-4} and a minimum log2fc value of 1. Finally, we sorted the DEG by their log2fold change values. These thresholds were chosen in an effort to ensure the selection of clearly differentially expressed genes in each cell type. Although there is no consensus on what threshold should be used for DE analysis in scRNAseq, widely used packages like Seurat[73] report DEG using pvals_adj < 5×10^{-2} and logFC > 0.1.

We exported the top 500 DE genes of each cell type (or all genes in case there were fewer than 500) to be analysed with the ABC-MK along with the resampling software. For the Developmental Atlas, we converted the gene ID to ENSEMBL IDs using the `ensembldb` Bioconductor package (version 2.18.1[74]). This was not necessary for the Adult Atlas, as the genes were already annotated using ENSEMBL IDs. We also provide the inferred α at the compartment level for the three different cell atlases (see Supplementary Fig. 11), inferred using the same strategy described before for the cell types.

Spatial mapping of lung immune cells

In order to interrogate the spatial distribution of immune cells we downloaded the Lung Cell Atlas Cell2location output provided as a H5AD file at https://5locationslung.cellgeni.sanger.ac.uk/. More specifically, we used the non-negative matrix factorisation (NMF) factor loadings computed by Cell2location. NMF factor loadings (referred to as "cell-type loadings") can be interpreted as the proportion of cells of each cell-type present in each tissue. As in ref. 29, we normalised the mean cell-type loadings per cell type across all the ST samples that passed QC ($N = 11$) to get a weighted cell-type loading representing the relative abundance of a given cell type in the different tissues (e.g. parenchyma, epithelia) across the ST slides. Then, we produced a heatmap using the ComplexHeatmap R package (version 2.10.0[75]) clustering the cell-type weighted values in order to group cell types with a similar tissue distribution.

Activated macrophages data

Macrophage activation differentially expressed genes were obtained from the study by Panousis et al.[55], available as Supplementary material. As described in the study, iPSCs were differentiated to macrophages[76] and stimulated for 6 h and 24 h with 10 ng/mL recombinant human interleukin-10 (Peprotech, 200-10-2), 10 ng/mL recombinant human interferon-b (Peprotech, 300-02BC-5), 20 ng/mL recombinant human interleukin-4 (Peprotech, 200-04-5), 50 ng/mL P3C (Pam3CSK4) (Tocris, 4633/1), 20 ng/mL recombinant human interferon-γ (Peprotech, 300-02-20), 10 ng/mL lipopolysaccharides from Escherichia coli O127:B8 (Sigma Aldrich, L3129), 40 ng/mL human recombinant tumour necrosis factor alpha (Peprotech, 300-01A-10), 100 ng/mL R848 (Resiquimod) (Invivogen, tlrl-r848), or 5 ng/mL recombinant human sCD40 Ligand (Peprotech, 310-02-10), 27.5 μg/ml poly I:C (Invivogen, tlrl-pic, diluted in Lipofectamine -Thermo Fisher, L3000001- and Opti-MEM media -Thermo Fisher, 31985062-) as described in ref. 55. RNA libraries were obtained using the low-input bulk RNA-seq preparation for control and stimulated macrophage samples[55]. Differentially expressed genes for each condition were obtained using the Wald test from DESeq2[77]. Benjamini–Hochberg procedure was used for multiple testing correction (FDR).

From these DE gene lists, we only considered genes with a log2fc > 1 and with a *p*-value < 0.05. For each stimulus, we further categorised the genes as "early" (found exclusively in the DE gene list after 6 h), "late" (found exclusively in the DE gene list after 24 h) or "sustained" (found at both time points).

GO term analysis

We performed GO term analyses using the Biological Process (BP) ontology for each set of DE genes used for the ABC-MK-test, using the R packages clusterProfiler (version 4.2.2), enrichplot (version 1.14.2[78]), GO.db and org.Hs.eg.db (BioConductor version 3.14). The results, separately for each cell atlas, are provided in Supplementary Data 11–13.

Phylostratigraphy enrichment analysis

We obtained the Phylostrata (PS) values for human genes from the study of Litman and Stein[62], where the authors determine consensus

ages for human protein-coding and noncoding genes, using publicly available databases. The PS data was downloaded from the supplementary material provided with their original paper.

To test for PS enrichment, we used the hypergeometric test considering all the genes with PS values as the population ($N = 27{,}974$) and the DE genes of each cell type (obtained as described above) as the sample ($n \leq 500$). For each PS (1–19) this test uses the hypergeometric distribution to determine the statistical significance of having drawn a sample of k genes of a given PS (out of n genes of the sample), from a gene population of size N containing K number of PS genes. We considered significant p-values $\leq 1 \times 10^{-3}$. The test was performed using the `phyper()` function of the `R stats` package (R version 4.1.0).

Divergence and polymorphic data

Classical and derived MK-test approaches combine both divergence (D) and polymorphism (P) data comparing alleles that are likely to have fitness effects (putatively selected; N) to those less likely to be under selection (putatively neutral, S). To perform MK-test estimations on human immune cell lines, we obtained associated information on polymorphism and divergence data treating synonymous and non-synonymous mutations as putatively neutral and selected respectively in the protein-coding sequences of the human hg38 assembly and Ensembl release 109 annotations[79].

Divergence counts were obtained by estimating fixed substitution in the human branch based on human, chimpanzee, gorilla, and orangutan alignments. Transcript sequences were obtained by aligning human CDS transcripts from Ensembl v109 on panTro6, gorGor6, and ponAbe3 assemblies from UCSC using `pblat`[80]. Best-reciprocal orthologs were re-aligned using `MACSE v2`[81], an explicit multiple sequence alignment software accounting for the underlying codon structure, frameshift and stop codons. The number of non-synonymous and synonymous fixations on the human branch was estimated using `Hyphy v2.5`[82].

1000GP phase 3 high-coverage phased data[83] across seven African ancestry populations were used to retrieve polymorphic sites and derived allele frequencies. We annotated non-synonymous and synonymous polymorphism using VEP[79] and estimated ancestral and derived allele frequencies using Ensembl v109 human ancestral allele information from EPO multi-alignments. In total, 17,538 orthologues were included in the analysis, counting sites in multiple isoforms only once.

Because we are interested in the long-term adaptation of the human immune system, we filtered the differentially expressed genes using mammalian orthologs. Gene age can significantly impact the rate of adaptation in populations since populations further away from their optimal conditions take larger evolutionary steps, and those closer to their optimum conditions take smaller steps[61]. This effect is known as the adaptive walk model of gene evolution, exemplified by Moutinho et al.[61] using Drosophila melanogaster and Arabidopsis thaliana polymorphic and divergence empirical data. We filter differentially expressed genes with the ortholog gene list described in ref. 34 to test for long-term adaptation while avoiding confounding signals due to gene age. This list contains Ensembl v99 human genes with orthologs in a minimum of 251 out of 261 mammalian genome assemblies. As exposed in ref. 34, the assemblies were extracted from NCBI Genome and had a minimum N50 contig size of 30 kb to reduce the number of possible truncated genes. The orthologs for mammals were defined based on the best-reciprocal hits using `Blat`[84] alignments, resulting in 13,495 orthologs.

Quantifying adaptation using ABC-MK

We measured the long-term rate of protein adaptation using ABC-MK, an Approximate Bayesian Computation extension of the MK-test[22,23,34]. Such a rate is usually estimated as the proportion of adaptive non-synonymous substitution and can be derived from classical MK-test by the quantity α[14]

$$\alpha = 1 - \left(\frac{D_S}{D_N}\frac{P_N}{P_S}\right) \quad (1)$$

where D_S and D_N are synonymous and non-synonymous fixed differences, and P_S and P_N are synonymous and non-synonymous polymorphic sites. When α is close to 1, positive selection is the predominant determinant of molecular divergence. If α is close to 0, drift dominates sequence divergence. Nonetheless, α approximation from the MK-test has multiple drawbacks that could bias the estimation. During the last decade, empirical and theoretical studies have mainly argued that the presence of selected alleles segregating can attain at both low and high frequencies leading to an underestimation of α[22,85–89].

ABC-MK use a generic Approximate Bayesian Computation (ABC) to infer the rate of adaptation: first, samples the parameter values from prior distributions; second, simulates a model using these random parameter values; and third, calculates informative summary statistics and compares them to observed empirical data. ABC-MK procedure computes the analytical expectation of α by frequency categories (x) on the Site Frequency Spectrum (similar to ref. 24), quickly estimating $\alpha_{(x)}$ for several combinations of prior distribution values (see Supplementary Data 1 to check prior values). Parameter combinations allow flexibility on the underlying DFE accounting for different amounts of deleterious, weakly and strongly adaptive mutations and background selection (BGS) to finally match the observed $\alpha_{(x)}$ with the best-fitting ones among the analytically estimated $\alpha_{(x)}$. It is worth noting that weakly adaptive mutations do not reach fixation quickly. Therefore, the existence of non-synonymous polymorphism in the site frequency spectrum (SFS) cannot be ignored, as it can be in the case of strongly adaptive mutations[22]. Moreover, as shown in ref. 22, the strength of BGS can also remove weakly adaptive alleles, so linked selection can heavily affect the $\alpha_{(x)}$ and the expected amount of fixations. Both disturb the shape of the $\alpha_{(x)}$, which translates into a downward trend of $\alpha_{(x)}$ at higher frequencies. Accounting independently for weak and strong adaptation along with BGS, ABC-MK can estimate the proportion of adaptation driven by weak and strong adaptation in any virtual DFE scenario, depending on prior values. Note that the mean strength and shape of the gamma deleterious DFE were set from uniform distributed values while accounting for the inferred values in ref. 90, creating random deleterious DFE shapes and an amount of weakly deleterious polymorphism. Here we use these functionalities as previously shown in ref. 32 and ref. 22, to quantify α as the contribution of weak and strong adaptive mutation in the presence of BGS: $\alpha = \alpha_W + \alpha_S$. In addition, we also estimated ω_a, the rate of adaptive non-synonymous substitution relative to the mutation rate, following Galtier [88] to support evidence of protein adaptation.

For each analysis, we subset 10^5 values from prior distributions to simulate and estimate $\alpha_{(x)}$ summary statistics, which allows flexibility for the underlying DFE and BGS strength. Posterior distributions were estimated using ABCreg[31] comparing ABC-MK summary statistics and the empirical datasets. We set tolerance in ABCreg to record 2500 acceptances from the rejection method and post-adjusted parameters using local linear regression[31,91]. Because ABC-MK is not intended to perform MK-test analyses at the gene level or in small pools of genes, we followed Murga-Moreno et al.[23] recommendations to obtain robust and accurate α estimation when analysing DEG sets. Consequently, we excluded the datasets accounting for less than 500 polymorphic sites and $\alpha_{(x)}$ summary statistics from the analytical estimations and empirical data were compared using a subset of used $\alpha_{(x)}$ values at Derived Allele Counts (DAC) $x = 2, 4, 5, 10, 20, 50, 200, 661$, avoiding singletons since they are particularly sensitive to sequencing errors and distortions due to demographic processes. In addition, note that we exclude any polymorphic site above frequency $x > 0.7$ (DAC > 925)

from empirical and analytical estimations. Based on the findings of ref. 92 about mispolarization regarding the human genome, it can lead to a significant decline in the $\alpha_{(x)}$ at higher frequencies, particularly when considering positive selection. Nonetheless, Supplementary Data 14 provide the DEG cases as well as the polymorphic, divergent and confounding factors data in support of future analyses at the gene level (see ref. 30 and ref. 23 for further discussion).

Resampling and false discovery rate

To highlight evolutionary patterns associated with an increased adaptation rate on differentially expressed genes on immune human cell lines, we compare each dataset with 1000 control datasets while matching the total number of analysed genes for each line. Control datasets represent null distribution, accounting for the same average values of multiple confounding factors that can also determine the rate of adaptive evolution compared to the rest of the genome. The resampling test uses the same algorithm described by Di et al.[32], Enard and Petrov [33] to build the cases from the top 500 DEG and control sets. Note that the pipeline excludes human leucocyte antigen (HLA) and histone genes from the 500 DEG since extreme nucleotide polymorphism level can lead to artefact biassed MK-results as well as to avoid spurious signals due to balancing selection[93]. The procedure involves adding control genes by pairs to the control set as it grows, ensuring that it has the same range of confounding factor values as the gene set being tested. This approach avoids matching set genes individually with control genes, which would limit the potential gene pool. Instead, we progressively match the gene set with pair controls and set a 5% matching limit over the gene set average confounding values. Additionally, we restrict each control gene from appearing more than three times in each control set as we increase the control pool. We matched a total of 14 confounding factors between set and control genes:

- Ensembl v99 longest CDS length for the gene (between isoforms).
- Average GC content for each coding sequence.
- Average GC1 content at the first codon nucleotide position for each coding sequence.
- Average GC2 content at the second codon nucleotide position for each coding sequence.
- Average GC3 content at the third codon nucleotide position for each coding sequence.
- Average GTEx v8 TPM (Transcripts Per Million) mRNA expression across 53 tissues (in log base 2)[94].
- Average expression (in log base 2) in testis from GTEx v8 tissues[94].
- Number of protein-protein interactions (PPIs) in the human protein interaction network[95]. We use the log (base 2) of the number of PPIs.
- Proportion of GERP conserved segments in a 50 kb window centred on the Ensembl gene (halfway between gene start and gene end)[96].
- Proportion of GERP conserved segments in a 500 kb window centred on the Ensembl gene (halfway between gene start and gene end)[96].
- Average GERP conservation score in a 50 kb window centred on the Ensembl gene[96].
- Average GERP conservation score in a 500 kb window centred on the Ensembl gene[96].
- McVicker's B estimate of background selection in a 50 kb window centred on the Ensembl gene[97].
- Average recombination rate in a 500 kb window centred on the Ensembl gene, from the deCode 2019 recombination map[98].

Although such a resampling approach allows us to search for a significant increase in adaptation in each cell line, the number of confounding factors restricts the number of genes that can be used as controls, which could ultimately lead to an increase in the false positive rate. Note that a smaller set of controls means that the same control genes are probably resampled multiple times, which inevitably causes a higher variance of the control null distribution and a higher variance of the overall average of the null distribution, compared to the ideal theoretical case where there would an infinite number of control genes to sample from. Although we restricted the control genes from appearing more than three times in each control set, it still could inflate the rate of false positive tests if the average of the small control sets is far from what the average would be for control sets built from an infinite number of control genes. Hence, our resampling approach, even controlling for multiple genomic features that can highly correlate with higher adaptation signals, cannot be directly used to obtain unbiased type I errors. To solve this problem, we followed Enard and Petrov[33] and Di et al.[32] to measure type I errors in each cell type by repeatedly performing ABC-MK analyses in size-matched Differentially Expressed Genes (DEG) sets but creating 1000 datasets by shuffling DEG and control genes. Such random sets represent a null distribution of estimates of adaptation without the impact of the differentially expressed genes. Since the sample size and the genomic features do not differ for the null distribution of randomised genes because of the resampling approach, re-running the entire analysis reproduces the same sampling size biases for the randomised genes as the actual tested cell line. Finally, by measuring the proportion of results with α greater than the analysed case, we effectively estimate the type I error, which then allows us to assess the expected number of cell types below a given type I error threshold by chance, over all cell lines. It is important to note that the current type I error measure differs from ref. 33 and ref. 32 (where we called the estimations of type I errors estimations of False Positive Rates). Unlike ref. 33 and ref. 32, the signatures of evolution exploited by our test of selection are independent from one gene to another. Therefore, type I error measures do not require an elaborate approach to take the clustering of genes in selective sweeps into account (as was the case in those previous publications), but only require simple permutations. In addition, we apply multiple testing corrections and adequately estimate the FDR by applying Benjamini–Hochberg correction[35] from the type I error measures for the 162 analysed cell types.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The scRNAseq and snRNAseq data used in this study are available at https://developmental.cellatlas.io/fetal-immune, https://www.tissueimmunecellatlas.org/ and https://5locationslung.cellgeni.sanger.ac.uk/ for the Developmental Atlas, the Adult Atlas and Lung Atlas databases respectively.

Code availability

We provide the scripts necessary to replicate our analysis in Zenodo repository[99] as well as the following GitHub repository: https://github.com/irepansalvador/Immune_Adaptation_Atlas_2023.

References

1. Tishkoff, S. A. et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).
2. Ko, W. Y. et al. Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *Am. J. Hum. Genet.* **88**, 741–754 (2011).
3. Klunk, J. et al. Evolution of immune genes is associated with the Black Death. *Nature* **611**, 312–319 (2022).

4. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
5. Sawyer, S. L., Emerman, M. & Malik, H. S. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol.* **2**, e275 (2004).
6. Sawyer, S. L., Wu, L. I., Emerman, M. & Malik, H. S. Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. *Proc. Natl Acad. Sci. USA* **102**, 2832–2837 (2005).
7. Elde, N. C., Child, S. J., Geballe, A. P. & Malik, H. S. Protein kinase R reveals an evolutionary model for defeating viral mimicry. *Nature* **457**, 485–489 (2009).
8. Demogines, A., Abraham, J., Choe, H., Farzan, M. & Sawyer, S. L. Dual host-virus arms races shape an essential housekeeping protein. *PLoS Biol.* **11**, e1001571 (2013).
9. Jacquet, S. et al. Adaptive duplication and genetic diversification of protein kinase R contribute to the specificity of bat-virus interactions. *Sci. Adv.* **8**, eadd7540 (2022).
10. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).
11. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
12. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. K. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
13. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652–654 (1991).
14. Smith, N. G. C. & Eyre-Walker, A. Adaptive protein evolution in Drosophila. *Nature* **415**, 1022–1024 (2002).
15. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
16. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
17. Balloux, F. & Lehmann, L. Substitution rates at neutral genes depend on population size under fluctuating demography and overlapping generations. *Evolution* **66**, 605–611 (2012).
18. Lanfear, R., Kokko, H. & Eyre-Walker, A. Population size and the rate of evolution. *Trends Ecol. Evol.* **29**, 33–41 (2014).
19. Galtier, N. & Rousselle, M. How much does ne vary among species? *Genetics* **216**, 559–572 (2020).
20. Rousselle, M. et al. Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. *PLoS Genet.* **16**, e1008668 (2020).
21. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
22. Uricchio, L. H., Petrov, D. A. & Enard, D. Exploiting selection at linked sites to infer the rate and strength of adaptation. *Nat. Ecol. Evol.* **3**, 977–984 (2019).
23. Murga-Moreno, J., Casillas, S., Barbadilla, A., Uricchio, L. & Enard, D. An efficient and robust ABC approach to infer the rate and strength of adaptation. *G3* https://doi.org/10.1093/g3journal/jkae031 (2024).
24. Messer, P. W. & Petrov, D. A. Frequent adaptation and the McDonald-Kreitman test. *Proc. Natl Acad. Sci. USA* **110**, 8615–8620 (2013).
25. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).
26. Lindeboom, R. G. H., Regev, A. & Teichmann, S. A. Towards a human cell atlas: taking notes from the past. *Trends Genet.* **37**, 625–630 (2021).
27. Suo, C. et al. Mapping the developing human immune system across organs. *Science* **376**, eabo0510 (2022).
28. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
29. Madissoon, E. et al. A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nat. Genet.* **55**, 66–77 (2023).
30. Murga-Moreno, J., Coronado-Zamora, M., Casillas, S. & Barbadilla, A. impMKT: the imputed McDonald and Kreitman test, a straightforward correction that significantly increases the evidence of positive selection of the McDonald and Kreitman test at the gene level. *G3 Genes Genomes Genet.* **12**, jkac206 (2022).
31. Thornton, K. R. Automating approximate Bayesian computation by local linear regression. *BMC Genet.* **10**, 35 (2009).
32. Di, C., Murga Moreno, J., Salazar-Tortosa, D. F., Lauterbur, M. E. & Enard, D. Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants. *eLife* **10**, e69026 (2021).
33. Enard, D. & Petrov, D. A. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philos. Trans. R. Soc. B: Biol. Sci.* **375**, 20190575 (2020).
34. Di, C., Murga-Moreno, J. & Enard, D. Stability evolution as a major mechanism of human protein adaptation in response to viruses. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.01.518739 (2022).
35. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
36. Pinho, S. & Frenette, P. S. Haematopoietic stem cell activity and interactions with the niche. *Nat. Rev. Mol. Cell Biol.* **20**, 303–320 (2019).
37. King, K. Y. & Goodell, M. A. Inflammatory modulation of HSCs: viewing the HSC as a foundation for the immune response. *Nat. Rev. Immunol.* **11**, 685–692 (2011).
38. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11**, 47–59 (2010).
39. Sonnenberg, G. F. & Hepworth, M. R. Functional interactions between innate lymphoid cells and adaptive immunity. *Nat. Rev. Immunol.* **19**, 599–613 (2019).
40. Eberl, G., Colonna, M., Di Santo, J. P. & McKenzie, A. N. J. Innate lymphoid cells: a new paradigm in immunology. *Science* **348**, aaa6566 (2015).
41. Melchers, F. Checkpoints that control B cell development. *J. Clin. Investig.* **125**, 2203–2210 (2015).
42. Tripathi, S., Tsang, J. S. & Park, K. Systems immunology of regulatory T cells: can one circuit explain it all? *Trends Immunol.* https://doi.org/10.1016/j.it.2023.08.007 (2023).
43. Tangye, S. G. To B1 or not to B1: that really is still the question! *Blood* **121**, 5109–5110 (2013).
44. Wong, J. B. et al. B-1a cells acquire their unique characteristics by bypassing the pre-BCR selection stage. *Nat. Commun.* **10**, 4768 (2019).
45. Iwama, R. E. & Moran, Y. Origins and diversification of animal innate immune responses against viral infections. *Nat. Ecol. Evol.* **7**, 182–193 (2023).
46. Graham, D. B. & Xavier, R. J. Conditioning of the immune system by the microbiome. *Trends Immunol.* **44**, 499–511 (2023).
47. Wang, Z. et al. An immune cell atlas reveals the dynamics of human macrophage specification during prenatal development. *Cell* https://doi.org/10.1016/j.cell.2023.08.019 (2023).
48. Lim, H. Y. et al. Hyaluronan receptor LYVE-1-expressing macrophages maintain arterial tone through hyaluronan-mediated

regulation of smooth muscle cell collagen. *Immunity* **49**, 326–341.e7 (2018).
49. Galli, S. J., Gaudenzio, N. & Tsai, M. Mast cells in inflammation and disease: recent progress and ongoing concerns. *Annu. Rev. Immunol.* **38**, 49–77 (2020).
50. Szabo, P. A., Miron, M. & Farber, D. L. Location, location, location: tissue resident memory T cells in mice and humans. *Sci. Immunol.* **4**, eaas9673 (2019).
51. Paik, D. H. & Farber, D. L. Anti-viral protective capacity of tissue resident memory T cells. *Curr. Opin. Virol.* **46**, 20–26 (2021).
52. Michel, T. et al. Human CD56bright NK cells: an update. *J. Immunol.* **196**, 2923–2931 (2016).
53. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).
54. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
55. Panousis, N. I. et al. Gene expression QTL mapping in stimulated iPSC-derived macrophages provides insights into common complex diseases. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.29.542425 (2023).
56. Langenkamp, A., Messi, M., Lanzavecchia, A. & Sallusto, F. Kinetics of dendritic cell activation: impact on priming of TH1, TH2 and nonpolarized T cells. *Nat. Immunol.* **1**, 311–316 (2000).
57. Nagorsen, D. et al. Polarized monocyte response to cytokine stimulation. *Genome Biol.* **6**, R15 (2005).
58. Hamidzadeh, K., Christensen, S. M., Dalby, E., Chandrasekaran, P. & Mosser, D. M. Macrophages and the recovery from acute and chronic inflammation. *Annu. Rev. Physiol.* **79**, 567–592 (2017).
59. Hsu, A. T. et al. Epigenetic and transcriptional regulation of IL4-induced CCL17 production in human monocytes and murine macrophages. *J. Biol. Chem.* **293**, 11415–11423 (2018).
60. Salvador-Martínez, I., Coronado-Zamora, M., Castellano, D., Barbadilla, A. & Salazar-Ciudad, I. Mapping selection within Drosophila melanogaster embryo's anatomy. *Mol. Biol. Evol.* **35**, 66–79 (2018).
61. Moutinho, A. F., Eyre-Walker, A. & Dutheil, J. Y. Strong evidence for the adaptive walk model of gene evolution in Drosophila and Arabidopsis. *PLoS Biol.* **20**, e3001775 (2022).
62. Litman, T. & Stein, W. D. Obtaining estimates for the ages of all the protein-coding genes and most of the ontology-identified non-coding genes of the human genome, assigned to 19 phylostrata. *Semin. Oncol.* **46**, 3–9 (2019).
63. Cooper, M. D. & Alder, M. N. The evolution of adaptive immune systems. *Cell* **124**, 815–822 (2006).
64. Simon, A. K., Hollander, G. A. & McMichael, A. Evolution of the immune system in humans from infancy to old age. *Proc. R. Soc. B Biol. Sci.* **282**, 20143085 (2015).
65. McGovern, N. et al. Human fetal dendritic cells promote prenatal T-cell immune suppression through arginase-2. *Nature* **546**, 662–666 (2017).
66. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).
67. Copley, M. R. & Eaves, C. J. Developmental changes in hematopoietic stem cell properties. *Exp. Mol. Med.* **45**, e55 (2013).
68. López, D. A. et al. Prenatal inflammation perturbs murine fetal hematopoietic development and causes persistent changes to postnatal immunity. *Cell Rep.* https://doi.org/10.1016/j.celrep.2022.111677 (2022).
69. Bjornson-Hooper, Z. B. et al. A comprehensive atlas of immunological differences between humans, mice, and non-human primates. *Front. Immunol.* **13**, https://doi.org/10.3389/fimmu.2022.867015 (2022).
70. McMaster, S. R., Wilson, J. J., Wang, H. & Kohlmeier, J. E. Airway-resident memory CD8 T cells provide antigen-specific protection against respiratory virus challenge through rapid IFN-γ production. *J. Immunol.* **195**, 203–209 (2015).
71. Mosser, D. M. & Edwards, J. P. Exploring the full spectrum of macrophage activation. *Nat. Rev. Immunol.* **8**, 958–969 (2008).
72. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
73. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
74. Rainer, J., Gatto, L. & Weichenberger, C. X. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* **35**, 3151–3153 (2019).
75. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
76. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
77. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
78. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* https://doi.org/10.1016/j.xinn.2021.100141 (2021).
79. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
80. Wang, M. & Kong, L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinform.* **20**, 28 (2019).
81. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**, 2582–2584 (2018).
82. Kosakovsky Pond, S. L. et al. HyPhy 2. *Mol. Biol. Evol.* **37**, 295–299 (2020).
83. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
84. Kent, W. J. BLAT-the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
85. Booker, T. R. Inferring parameters of the distribution of fitness effects of new mutations when beneficial mutations are strongly advantageous and rare. *G3* **10**, 2317–2326 (2020).
86. Charlesworth, J. & Eyre-Walker, A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* **25**, 1007–1015 (2008).
87. Eyre-Walker, A. & and Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26**, 2097–2108 (2009).
88. Galtier, N. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet.* **12**, e1005774 (2016).
89. Tataru, P., Mollion, M., Glémin, S. & Bataillon, T. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* **207**, 1103–1119 (2017).
90. Boyko, A. R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLOS Genet.* **4**, e1000083 (2008).
91. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
92. Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* **24**, 1792–1800 (2007).
93. Soni, V., Vos, M. & Eyre-Walker, A. A new test suggests hundreds of amino acid polymorphisms in humans are subject to balancing selection. *PLoS Biol.* **20**, e3001645 (2022).

94. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
95. Luisi, P. et al. Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. *Genome Biol. Evol.* **7**, 1141–1154 (2015).
96. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
97. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, 1–16 (2009).
98. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
99. Salvador-Martínez, I. & Murga-Moreno, J. irepansalvador/Immune_adaptation_atlas_2023: revised manuscript version, 2024. *Zenodo* https://zenodo.org/doi/10.5281/zenodo.13638051 (2024).
100. Milacic, M. et al. The reactome pathway knowledgebase 2024. *Nucleic Acids Res.* **52**, D672–D678 (2024).

Acknowledgements

This publication is part of the Human Cell Atlas. We would like to thank the Human Cell consortium for providing a comprehensive transcriptomic characterisation of the human immune system that made this study possible. We also thank the members of the Single Cell Genomics lab in the CNAG and Antonio Barbadilla for their useful comments and feedback. Also, we thank Marta Coronado-Zamora and Aina Vaquer for suggesting using the consensus PS dataset. D.E. is funded by NIH NIGMS MIRA grant 5R35GM142677.

Author contributions

Conceived the study: I.S.-M. Designed the research: I.S.-M., J.M.-M., H.H. and D.E. Analysed the data: I.S.-M. and J.M.-M. Proposed and facilitated data for the iPSC analysis: C.A. Wrote the initial draft: I.S.-M., J.M.-M. and J.C.N. Edited and approved paper: I.S.-M., J.M.-M., C.A., J.C.N., D.E. and H.H.

Competing interests

H.H. is a co-founder and shareholder of Omniscope, a member of the Scientific Advisory Board of Nanostring and MiRXES, and a consultant to Moderna and Singularity. J.C.N. is a consultant to Omniscope. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-54603-5.

Correspondence and requests for materials should be addressed to David Enard or Holger Heyn.

Peer review information *Nature Communications* thanks Maxime Rotival and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024