Chapter 4

Cell-to-cell variability in the innate immune response

> **Declaration**
>
> *The cross-mammalian dataset presented in Section 4.1 was produced by Tzachi Hagai. This work was published in Nature, 2018, and the full paper is included in Appendix D. Analysis of bulk RNA-seq data and calculation of response divergence were conducted by Tzachi, along with production of final figures for the manuscript.*

4.1 | Introduction

The innate immune system acts as a first line of defence across cell types and species, inhibiting pathogen replication and signalling pathogen presence to other cells. A key feature of this response is the rapid evolution that many of the genes have undergone along the vertebrate lineage, often attributed to pathogen-driven selection. As described in Chapter 1.3, another characteristic of the response is the high level of heterogeneity among responding cells, however the functional importance of this variability is unclear.

These two characteristics - rapid evolutionary divergence and high cell-to-cell variability — seem to be at odds with the strong regulatory constraints imposed on the host immune response: the need to execute a well-coordinated and carefully balanced programme to avoid tissue damage and pathological immune conditions. How this tight regulation is maintained despite rapid evolutionary divergence and high cell-to-cell variability remains unclear, but it is central to our understanding of the innate immune response and its evolution.

In this chapter, I present two angles of this question. Firstly, in a study led by Tzachi Hagai, we studied the evolution of the innate immune programme using two cells types — fibroblasts and mononuclear phagocytes — in different mammalian clades challenged with several immune stimuli. The results presented here focus on the fibroblast results; the experimental methods are described in Chapter 2.4. I then go on to use a larger human scRNA-seq dataset, described in Chapter 2.2, to define the dynamics of the response at a single cell resolution, characterising response gene modules.

4.2 | Innate immune variability: a cross-mammalian study

4.2.1 | Transcriptional divergence in immune response

First, we studied the transcriptional response of fibroblasts to stimulation with dsRNA (poly(I:C)) across the four species (human, macaque, rat and mouse). Bulk RNA-sequencing (RNA-seq) data was generated for each species after 4 h of stimulation, along with respective controls (Figure 4.1a).

In all species, dsRNA treatment induced rapid upregulation of genes that encode expected antiviral and inflammatory products, including IFN-β, TNF, IL1A and CCL5 (Figure 4.1b). A similar transcriptional response between species was observed when considering one-to-one orthologues (Spearman correlation, P < 10^{-10} in all comparisons), as reported in other immune contexts [183–185]. Furthermore, as seen in other expression programmes [186–188], the response tended to be more strongly correlated between closely related species than between more distantly related species (Appendix D; Extended Data Figure 1).

Using these cross-species bulk transcriptomics data, we characterized the differences in response to dsRNA between species for each gene. While some genes, such as those encoding the NF-B subunits RELB and NFKB2, respond similarly across species, other genes respond differently in the primate and rodent clades (Figure 4.1c). For example, Ifi27 (which encodes a restriction factor against numerous viruses) is strongly upregulated in primates but not in rodents, whereas Daxx (which encodes an antiviral transcriptional repressor) exhibits the opposite behaviour.

To quantify transcriptional divergence in immune responses between species, we focused on genes that were differentially expressed during the stimulation (see Appendix

D; Methods) referred to as 'responsive genes' (Figure 4.1d). In this analysis, we study the subset of these genes with one-to-one orthologues across the studied species, of which there are 955 such responsive genes in dsRNA-stimulated human fibroblasts. We define a measure of response divergence by calculating the differences between the fold-change estimates while taking the phylogenetic relationship into account (Appendix D; Methods).

For subsequent analyses, we split the 955 responsive genes into three groups on the basis of their level of response divergence: (1) high-divergence dsRNA-responsive genes (the top 25% of genes with the highest divergence values in response to dsRNA across the four studied species); (2) low-divergence dsRNA-responsive genes (the bottom 25%); and (3) genes with medium divergence across species (the middle 50%; Figure 4.1d).

4.2.2 | Cell-to-cell variability in immune response

As described in Chapter 1.3, previous studies have shown that the innate immune response displays high variability across responding cells. However, the relationship between cell-to-cell transcriptional variability and response divergence between species is not well understood. To study heterogeneity across individual cells, single cell RNA-seq was performed in all species in a stimulation time course (Figure 4.1a).

Cell-to-cell variability was quantitatively measured using an established measure for variability: distance to median (DM) [25]. We found a clear trend in which genes that were highly divergent in response between species were also more variable in expression across individual cells within a species (Figure 4.2); observed across the stimulation time points and in different species.

Next, we examined the relationship between the presence of promoter elements (CpG islands - CGIs - and TATA-boxes) and a gene's cell-to-cell variability. Genes that are predicted to have a TATA-box in their promoter had higher transcriptional
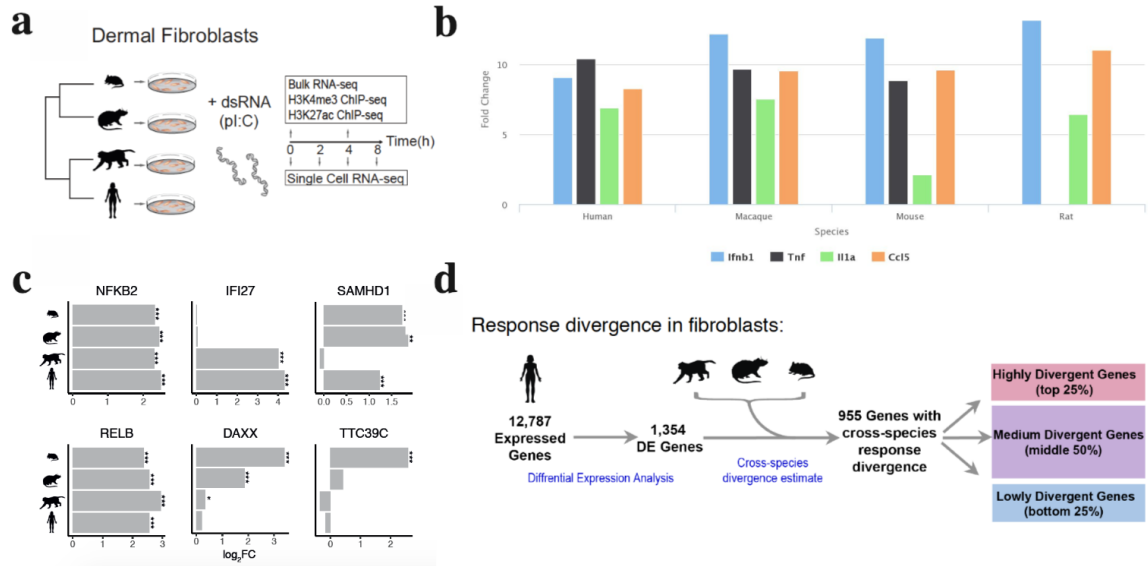
Fig. 4.1 Response divergence across species in innate immune response. a) Study design. Primary dermal fibroblasts from mouse, rat, human and macaque stimulated with dsRNA or controls. Samples were collected for bulk and single cell RNA-seq and ChIP–seq. b) Fold change of example genes (IFNB1, TNF, IL1A and CCL5) across the four species after 4h dsRNA stimulation. c) Fold-change (FC) after 4h dsRNA stimulation in fibroblasts for sample genes across species (edgeR exact test, based on n = 6, 5, 3 and 3 individuals from human, macaque, rat and mouse, respectively). False discovery rate (FDR)-corrected P values are shown (***P<0.001, **P<0.01, *P<0.05). d) Estimating each gene's level of cross-species divergence in transcriptional response to dsRNA stimulation. Using differential expression analysis, fold-change in dsRNA response was assessed for each gene in each species. We identified 1,358 human genes as differentially expressed (DE) (FDR-corrected q<0.01), of which 955 had one-to-one orthologues across the four studied species. For each gene with one-to-one orthologues across all species, a response divergence measure was estimated using: response divergence $= log[1/4 \times \Sigma_{i,j}(log[FCprimate_i] - log[FCrodent_j])^2]$. Genes were grouped into low, medium and high divergence according to their response divergence values for subsequent analysis.
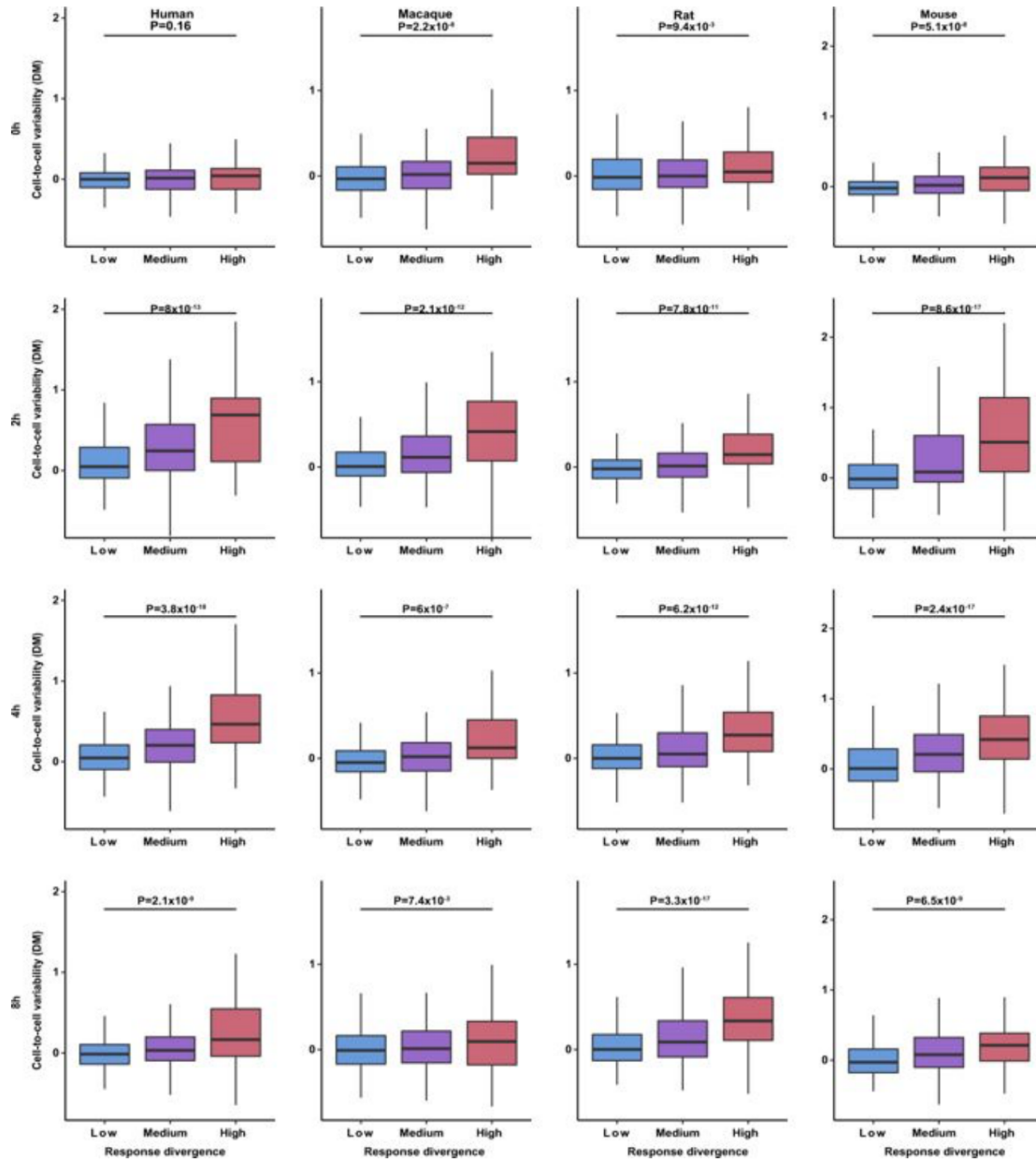
Fig. 4.2 Cell-to-cell variability versus response divergence across species and conditions. Cell-to-cell variability values, as measured with DM across individual cells, compared with response divergence between species (grouped into low, medium and high divergence). Variability values are based on n =29, 56, 55, 35 human cells, n = 20, 32, 29, 13 rhesus cells, n = 33, 70, 65, 40 rat cells, and n = 53, 81, 59, 30 mouse cells, stimulated with dsRNA for 0, 2, 4 and 8 h, respectively. Rows represent different time points (0, 2, 4 and 8 h), and columns represent different species. High-divergence genes were compared with low-divergence genes using a one-sided Mann–Whitney test. Data in boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the IQR.

variability, whereas CGI-containing genes tended to have lower variability (Figure 4.3a), in agreement with previous findings [189]. This finding also applied to transcriptional divergence between species (Figure 4.3b), showing that both these characteristics are associated with the presence of specific promoter elements.
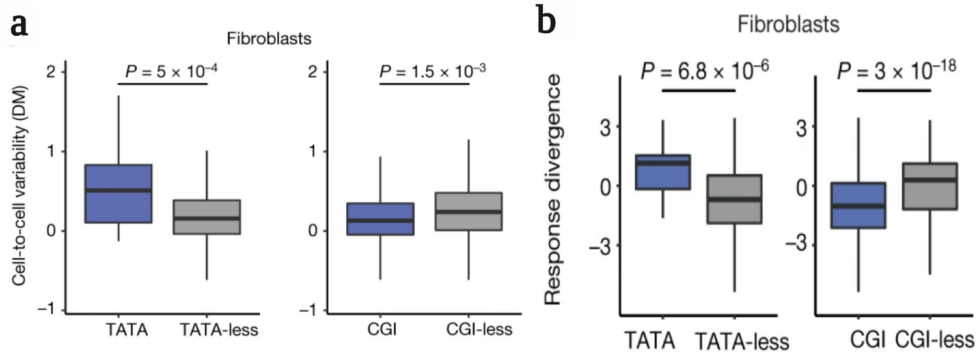

Fig. 4.3 Promoter architecture versus transcriptional divergence and variability. a) Comparison of cell-to-cell variability of genes with and without a TATA-box and a CGI (one-sided Mann–Whitney test). Cell-to-cell variability values are from DM estimations of human fibroblasts stimulated with dsRNA for 4 h (n = 55 cells). b) Comparison of divergence in response of genes with and without a TATA-box and a CGI in fibroblast dsRNA stimulation.

4.2.3 | Transcriptional divergence and variability of cytokines

We next investigated whether different functional classes among responsive genes are characterized by varying levels of transcriptional divergence. To this end, we divided responsive genes into categories according to function (such as cytokines, transcriptional factors and kinases) or the processes in which they are known to be involved (such as apoptosis or inflammation). Genes related to cellular defence and inflammation—most notably cytokines, chemokines and their receptors (hereafter 'cytokines')—tended to diverge in response significantly faster than genes involved in apoptosis or immune regulation (chromatin modulators, transcription factors, kinases and ligases) (Figure 4.4).
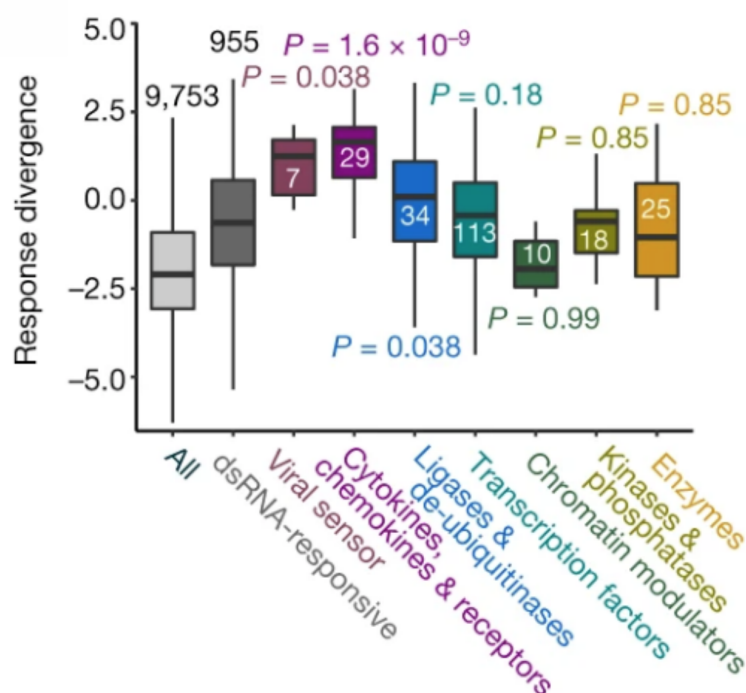
Fig. 4.4 Transcriptional divergence in genes of different functional categories. Distributions of divergence values of 9,753 expressed genes in fibroblasts, 955 dsRNA-responsive genes and different functional subsets of the dsRNA-responsive genes (each subset is compared with the set of 955 genes using a one-sided Mann–Whitney test and FDR-corrected P values are shown).

We subsequently compared the response divergence across species with the transcriptional cell-to-cell variability of three groups of responsive genes with different functions: cytokines, transcription factors, and kinases and phosphatases (referred to as 'kinases'). In contrast to kinases and transcription factors, many cytokines display relatively high levels of cell-to-cell variability across time points (Figure 4.5a). Furthermore, these are expressed only in a small subset of responding cells (Figure 4.5b).This has previously been reported for several cytokines, as described in Chapter 1.3. Here, we find that cells show high levels of variability in expression of cytokines from several families (for example, IFN-β, CXCL10 and CCL2).
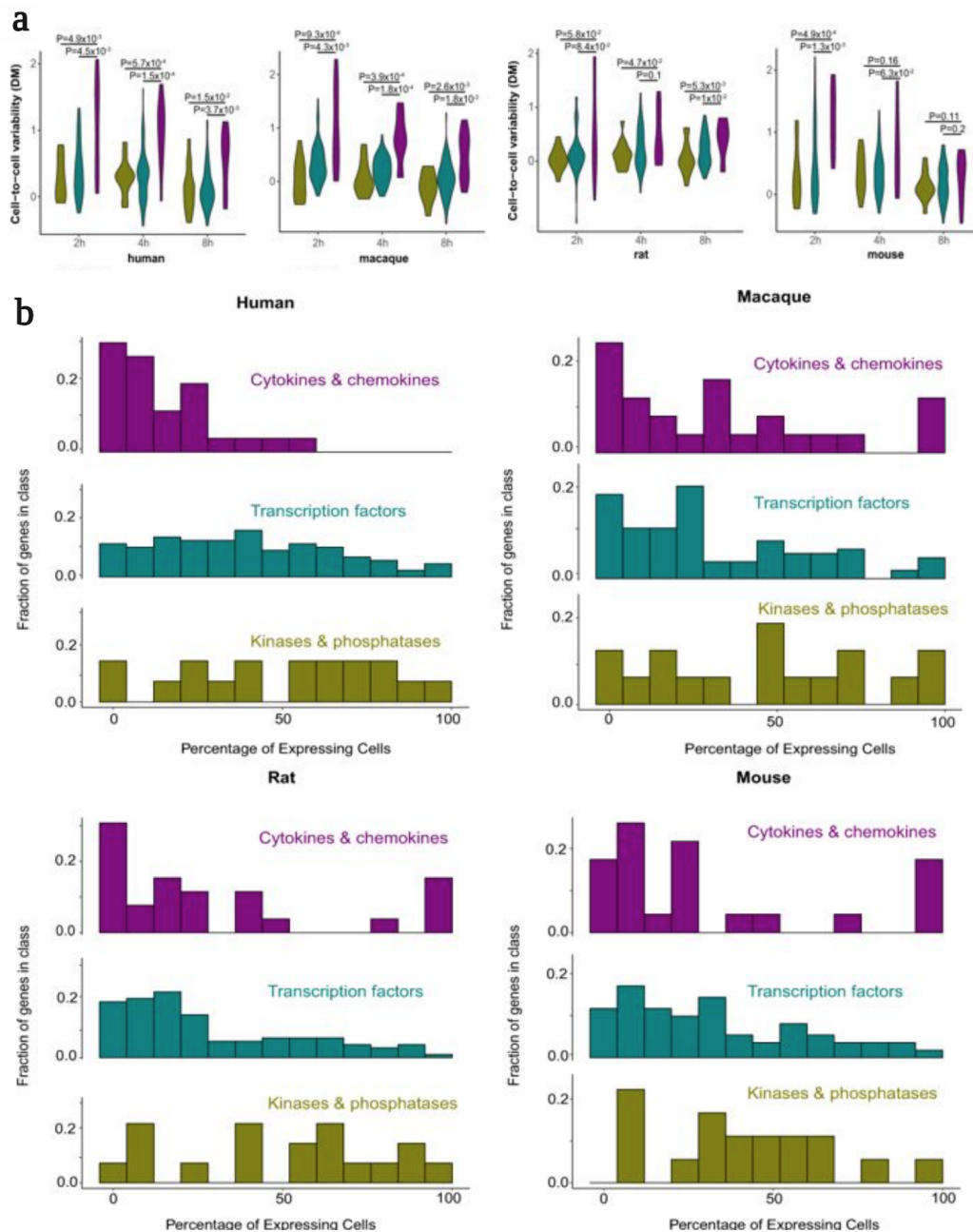
Fig. 4.5 Cell-to-cell variability levels in cytokines, transcription factors and kinases. a) Violin plots showing the distribution of cell-to-cell variability values (DM) of cytokines, transcription factors and kinases during a dsRNA stimulation time course in fibroblasts. Number of cells used in each species (at 2, 4, 8 h dsRNA, respectively): human, 56, 55, 35; macaque, 32, 29, 13; rat, 70, 65, 40; mouse, 81, 59, 30. Purple, cytokines; green, transcription factors; beige, kinases. Comparisons between groups of genes were performed using one-sided Mann–Whitney tests. Violin plots show the kernel probability density of the data. b) Histograms showing the percentage of fibroblasts expressing cytokines (top), transcription factors (middle) and kinases (bottom) following 4 h dsRNA stimulation, in human, macaque, rat and mouse cells. The percentage of expressing cells is divided into 13 bins (x-axis). The y-axis represents the fraction of genes from this gene class (for example, cytokines) that are expressed in each bin.

4.3 | Characterising the Type I interferon response in human fibroblasts

4.3.1 | Single-cell RNA-sequencing data

Having characterised variability in the innate immune response from an evolutionary perspective, the question of heterogeneity within the human population remains. In order to address this, a comprehensive dataset comprising both bulk and single cell RNA-sequencing data at two timepoints and with two stimulation conditions, along with a control, was generated - as described in Chapter 2.

The single cell dataset was filtered as described in Chapter 3.2, and UMAP dimensionality reduction was used to gain an oversight of the full dataset (Figure 4.6). It is clear to see that, as before, a major driver of variation is experimental batch effect, although cells also cluster by experimental condition. Once again, the 'integrate' function from the Seurat v3 package [167] was applied. This resulted in good mixing of the two batches in UMAP space, with experimental condition now being the major driver of variation in the dataset (Figure 4.6). The separation in unstimulated and interferon-treated cells seen in the 'condition' plot arises from cell cycle state, with cycling cells forming the cluster of mixed conditions on the left side of the plot.

4.3.2 | The temporal dynamics of the response

Harnessing the resolution available within the single-cell data generated, it is possible to comprehensively study the innate immune response over time. Although both poly(I:C) and IFN-β induce antiviral signalling within treated cells, the two elicit different responses, as can be seen in Figure 4.6.
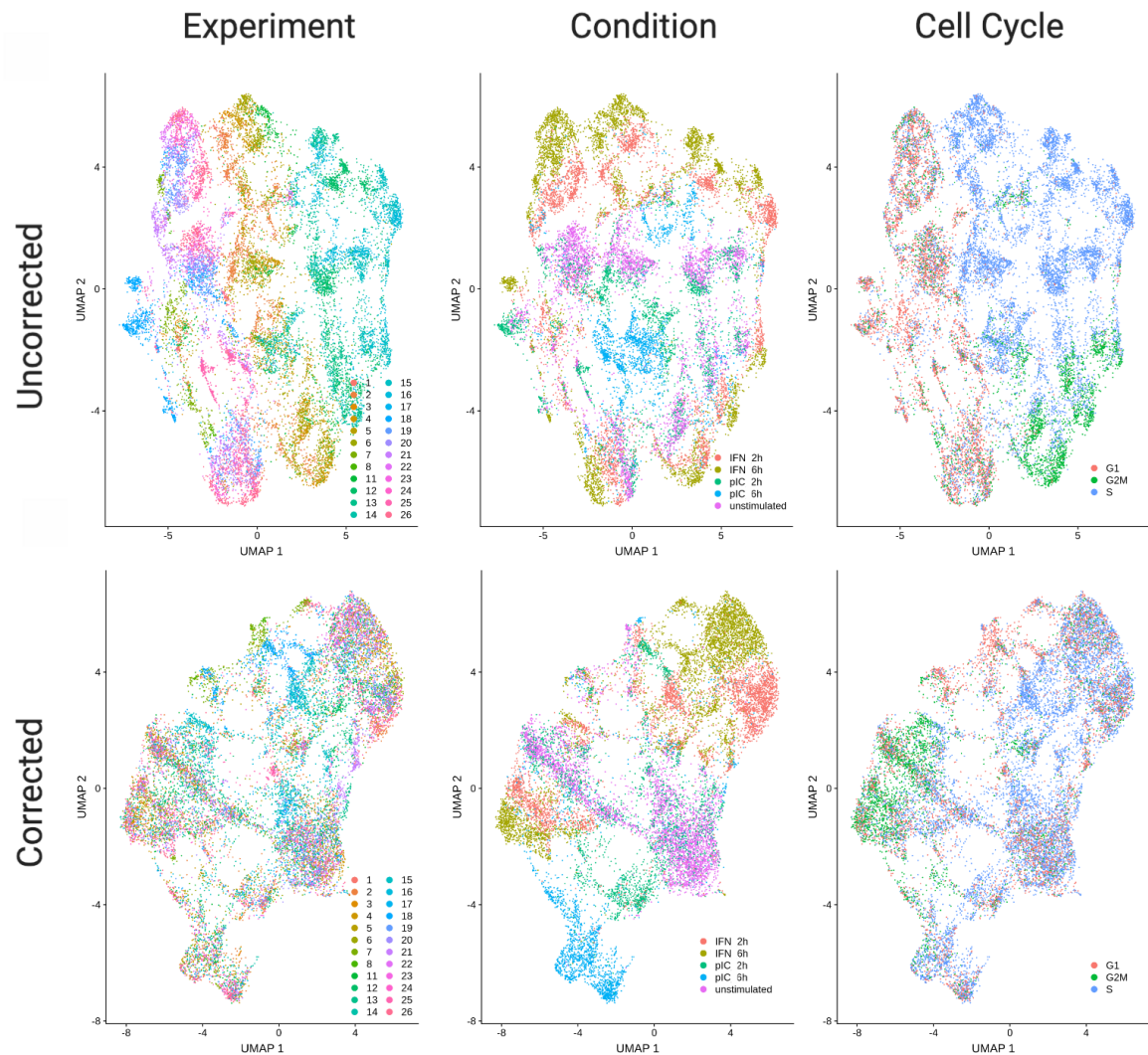
Fig. 4.6 Integration of scRNA-seq batches with Seurat. Dimensionality reduction using UMAP on uncorrected data (upper panel) and corrected data using Seurat v3's 'integrate' method (lower panel). Colours indicate, in order, experimental batch, stimulation condition, and cell cycle phase. The first two UMAP dimensions are shown.

In order to appropriately characterise the two response pathways, cells treated with poly(I:C) were separated from those treated with IFN-β. The two time points for each condition were considered together, along with control unstimulated cells. The benefit of combining all treated cells is particularly apparent after poly(I:C) stimulation, in which many cells after two hours of stimulation are transcriptionally similar to unstimulated cells, highlighting heterogeneity in this response.

To create a pseudotime, the destiny package was used, which employs a diffusion map approach [190]. This was applied to the 5000 most highly variable genes, calculated with Seurat's 'findVariableGenes' function, to the IFN-β and poly(I:C) pathways separately. Figure 4.7a shows Diffusion Components (DCs) 1 and 2 for each of these responses. This demonstrates that the largest source of variability, segregating along DC1, is stimulation condition. DC2 shows separation, particularly of unstimulated cells, representing cell cycle effects. This is confirmed by GO term enrichment analysis of the genes most highly correlated with DC2, along with visual inspection of the cell cycle phase distribution versus DC2 - shown in the inset plots in Figure 4.7a.

Given the correlation of DC1 with stimulation response in both treatment conditions, this is used as a 'response pseudotime'. In the case of poly(I:C) stimulation, the reverse of DC1 is taken as unstimulated cells lie on the right hand side. The distribution of cells in each stimulation condition across this response pseudotime highlights the heterogeneity in response (Figure 4.7b). This is particularly true in the response to poly(I:C) treatment, where the earlier timepoint shows a bimodality in the cells. Many of the cells show high similarity to the unstimulated state, while a subset are shifted to the right in response pseudotime, overlapping with the peak of the poly(I:C) 6 hour distribution (which itself has a broad distribution). In the IFN-β response pseudotime, the peak around the middle of DC1 corresponds to cell cycle state, however a breadth in distribution within responding cells can be seen.
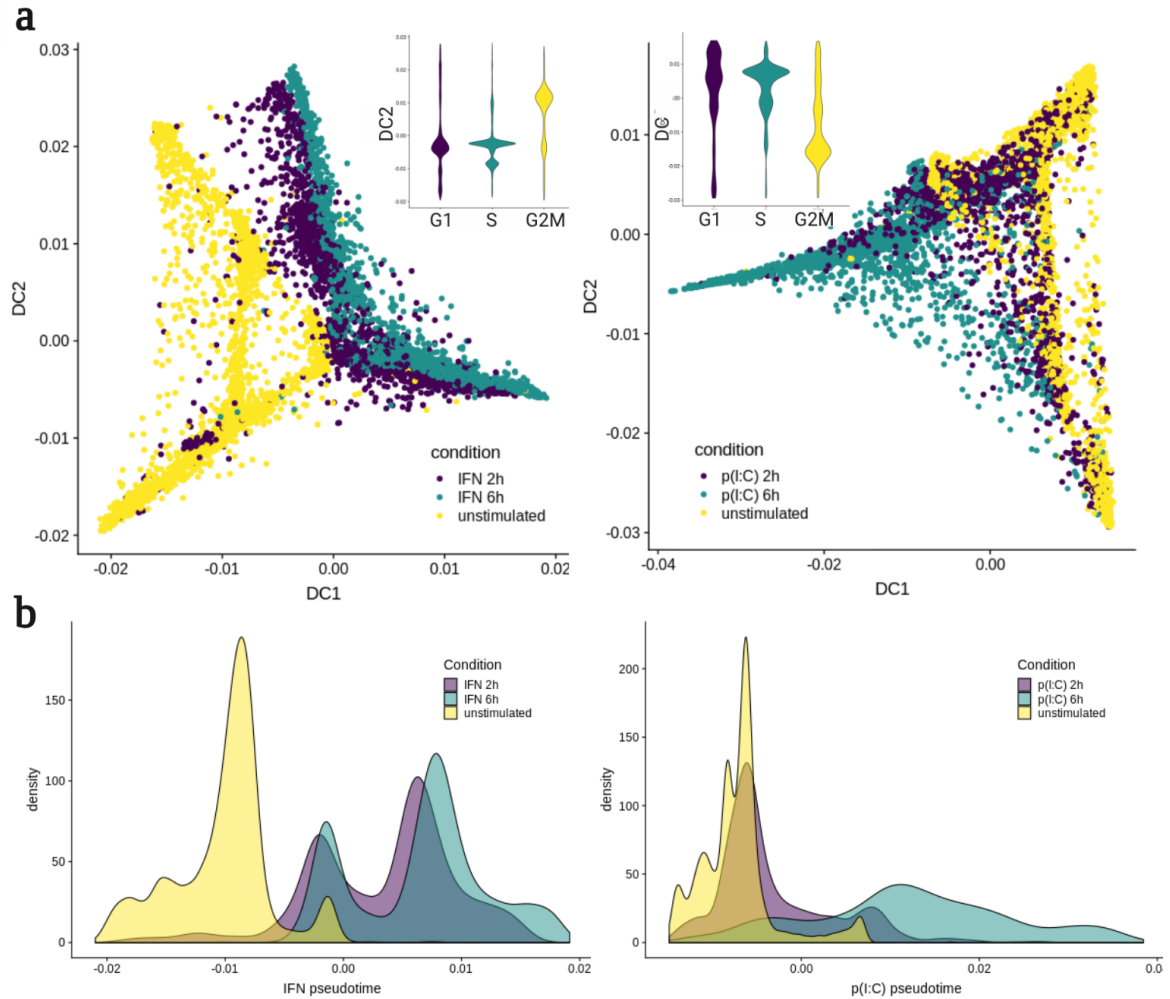
Fig. 4.7 A pseudotime of poly(I:C) and interferon response pathways. a) Cells plotted after dimensionality reduction with 'destiny' [190]; diffusion components (DCs) 1 and 2 shown for the 'IFN pathway', left, and 'poly(I:C) pathway', right, coloured by stimulation condition. Inset plots show the number of cells per cell cycle phase, assigned using 'cyclone' [174] against DC2. b) Density of cells from each stimulation condition across response pseudotimes, for the 'IFN pathway', left, and 'poly(I:C) pathway', right.

To confirm that the calculated pseudotimes capture the innate immune response, one can look at expression of known response genes, such as ISG15 and IFN-β (Figure 4.9a). It is worth noting that IFN-β treatment is not expected to induce IFN-β expression itself, and that in response to poly(I:C) treatment only a subset of cells produce IFN-β (rightmost panel), as discussed previously. Beyond example genes, it is possible to verify the expression of an entire innate immune response gene set. Deschamps *et al.* curated a set of 1553 innate immune genes (IIGs) from GO term annotation, InnateDB and manual addition [191]. These genes are classified into different functions, and examples of the genes and their annotated functions are shown in Figure 4.8. Looking at expression across the pseudotimes defined above, IIGs increase in the response to both IFN-β and poly(I:C) (Figure 4.9b), whereas the opposite is true for the remainder of genes (referred to as 'non-IIGs').
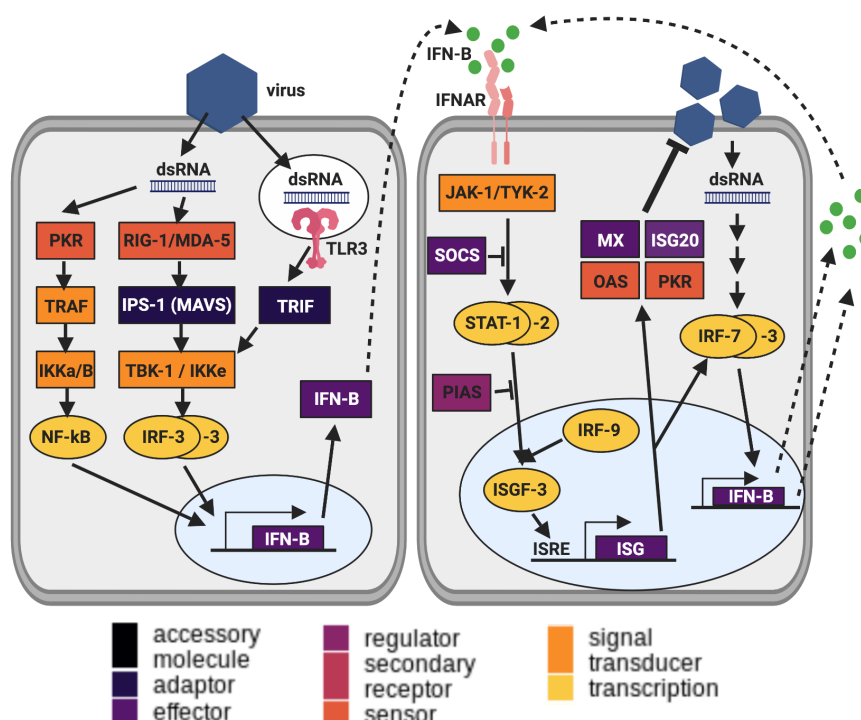

Fig. 4.8 Functional classification of innate immune genes. A curated list of innate immune genes (IIGs) was obtained from Deschamps *et al.* [191]. Examples involved in Type I interferon signalling, and their functional classification, are shown here.
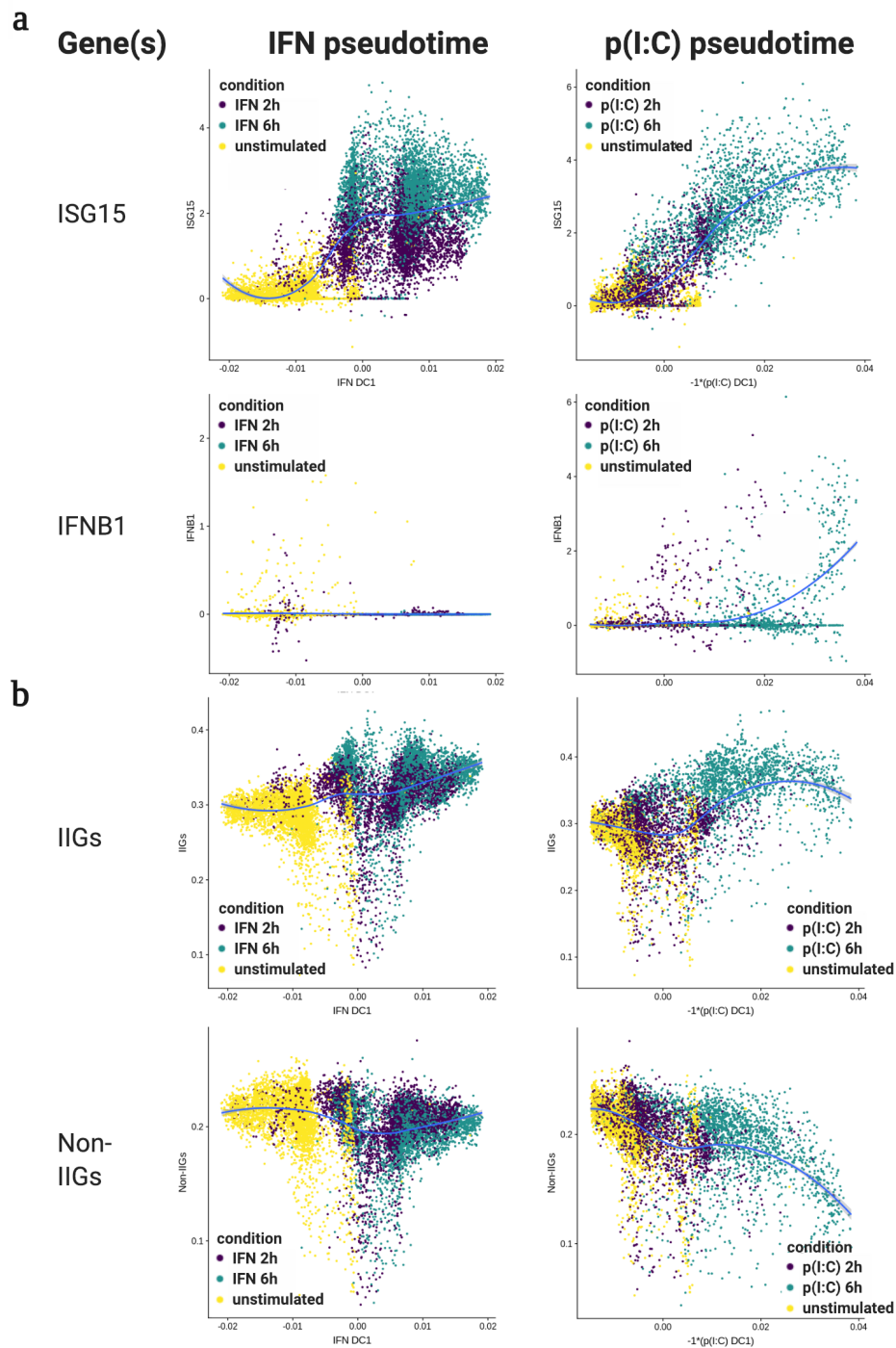
Fig. 4.9 Expression of innate immune genes across response pseudotime. a) The expression of ISG15 and IFNB1 against IFN pseudotime, left, and poly(I:C) pseudotime, right. b) Average expression of the set of innate immune genes (IIGs) [191] and non-IIGs over IFN pseudotime, left, and poly(I:C) pseudotime, right.

4.3.3 | **Defining gene modules in the innate immune response**

To define the dynamics of gene expression for each gene independently, it is possible to fit a model to the expression of each gene across response pseudotime. Using the SwitchDE package (Appendix E; Figure E.1a) [97], parameters for the activation time (t0), expression level (μ) and slope of activation (k) were inferred for the IFN-β and poly(I:C) responses. Using these inferred models, it is possible to look at the pattern of gene expression across time. Genes were categorised as 'on' or 'off' for each response pathway based upon whether they had a positive or negative k value, respectively. The top 500 most significant genes (taking the q-value from the SwitchDE model) in each direction were considered, and hierarchical clustering was used to define clusters of genes with shared temporal expression patterns. The number of clusters was determined visually based upon the dendrograms of gene similarity, for the IFN-β and poly(I:C) pseudotimes respectively. These clusters show good concordance with modules defined using an alternative approach: WGCNA (Weighted gene correlation network analysis)[100]; Appendix E, Figure E.1.

Each module was tested for enrichment of the IIGs described above, and results are shown in Tables 4.1 and 4.2 for the response to IFN-β and poly(I:C) respectively. The distribution of functional categories for these IIGs was considered, and is shown in the right hand plots of Figures 4.10 and 4.11. Furthermore, GO term enrichment analysis was conducted for each cluster, and the list of significantly enriched terms (pvalue < 0.05) is shown in Tables E.1 and E.2 (Appendix E).

In both the IFN-β and poly(I:C) response, there is one major cluster which represents the canonical Type I interferon response. In the case of IFN-β treatment, this is cluster coloured in black (Figure 4.10). This module of genes shows low expression in unstimulated cells (visible in the heatmap), a high enrichment of IIGs (hypergeometric test; pvalue = 1.98e-30), and inclusion of typical genes (such as DDX58, MYD88,

OAS3, ISG15, ISG20, IRF7, IFIT2, TRIM25, SAMHD1, IFI6, IFI35 and STAT1). The GO terms for this cluster reflect this signalling pathway, with the two most significant terms being "defense response to virus" and "type I interferon signaling pathway". This cluster has a particularly high representation of IIGs in the classes 'effector', 'regulator', and 'sensor'. highlighting functions across the pathway (Figure 4.10a; right panel). The same trend can be seen in the 'black' cluster in poly(I:C), which shows highest expression in later stages of the poly(I:C) response pseudotime. The most significant GO terms include "innate immune response", "defense response to virus" and "cytokine-mediated signaling pathway", and all of the example genes listed above fall within the cluster (with the exception of IRF7, however IRF9 is included). Again, IIGs are highly enriched (hypergeometric test; pvalue = 3.58e-42), and show functions across the pathway (Figure 4.11a; right panel).

Beyond these two major clusters, modules of genes with discrete innate immune response functions can be identified. For example, in response to IFN-β, there is a co-expressed set of genes (pink) which show involvement in signal transduction and regulation. This cluster includes genes such as DHX58, JAK2, STAT3 and TRADD. The third cluster, on the other hand, shows a higher level of effector function, with enrichment of GO terms relating to cytokine production, and genes such as CCl2, CXCL11 and CXCL16.

These alternative modules are less clear in the poly(I:C) response. Here, the second group of genes is a small cluster dominated by mitochondrial genes, which is reflected in the enriched GO terms. The two annotated IIGs within this cluster are IFITM2 and IFITM3, both of which are classified as 'effector' proteins. The third cluster, while enriched in IIGs, shows less ubiquitous expression in responding cells than cluster 1. There are no significant GO terms for this cluster, however members include genes

known to be involved in the type I interferon response, such as NFKBIA + NFKBID, SOCS3, CCL5, DDX3X and JUN, and particularly in signal transduction.

Along with categorising up-regulated gene sets, it is interesting to consider the set of genes down-regulated in response to the mock-viral stimulations. In response to IFN-β treatment, two major clusters of genes are down-regulated. One cluster, expressed in unstimulated cells but switched off in the response (Figure 4.10b; left panel) reflects processes around chromatin organisation and nucleic acid processing. Example genes in this cluster are HDAC2, SMARCA2, and ZNF287. The other cluster represents the cell cycle, with strongly enriched GO terms such as 'cell cycle' and 'DNA metabolic process'. Genes include CDK1, CCNA2, CDCA2, CCDC18, and several members of the CENP family.

In response to poly(I:C) stimulation, there are two major functions of down-regulated genes. The largest cluster of genes, which show decreased expression in responding cells, are involved in biological processes such as 'organelle organisation' and 'establishment of localisation in cell'. The second and third cluster are less clearly defined, however one cluster (pink) shows enrichment of GO terms highlighting metabolic processes, while the other centres on protein localisation and processing. Furthermore, these two modules show different temporal dynamics across the response pseudotime (Figure 4.11b; left panel).

The definition of these modules across response pseudotimes highlights a tightly regulated type I interferon response, with coordinated modules of genes showing discrete innate immune functions.

Table 4.1 Enrichment of IIGs in modules of co-expressed genes in the IFN-β response.

Gene module	Total group size	Number of IIGs	Enrichment p-value
Canonical Type I IFN	81	44	1.98e-30
Regulator/signal transduction	197	53	4.66e-18
Effector	222	40	2.65e-08
Cell cycle	244	19	0.33
Chromatin organisation	256	16	0.69

Table 4.2 Enrichment of IIGs in modules of co-expressed genes in the poly(I:C) response.

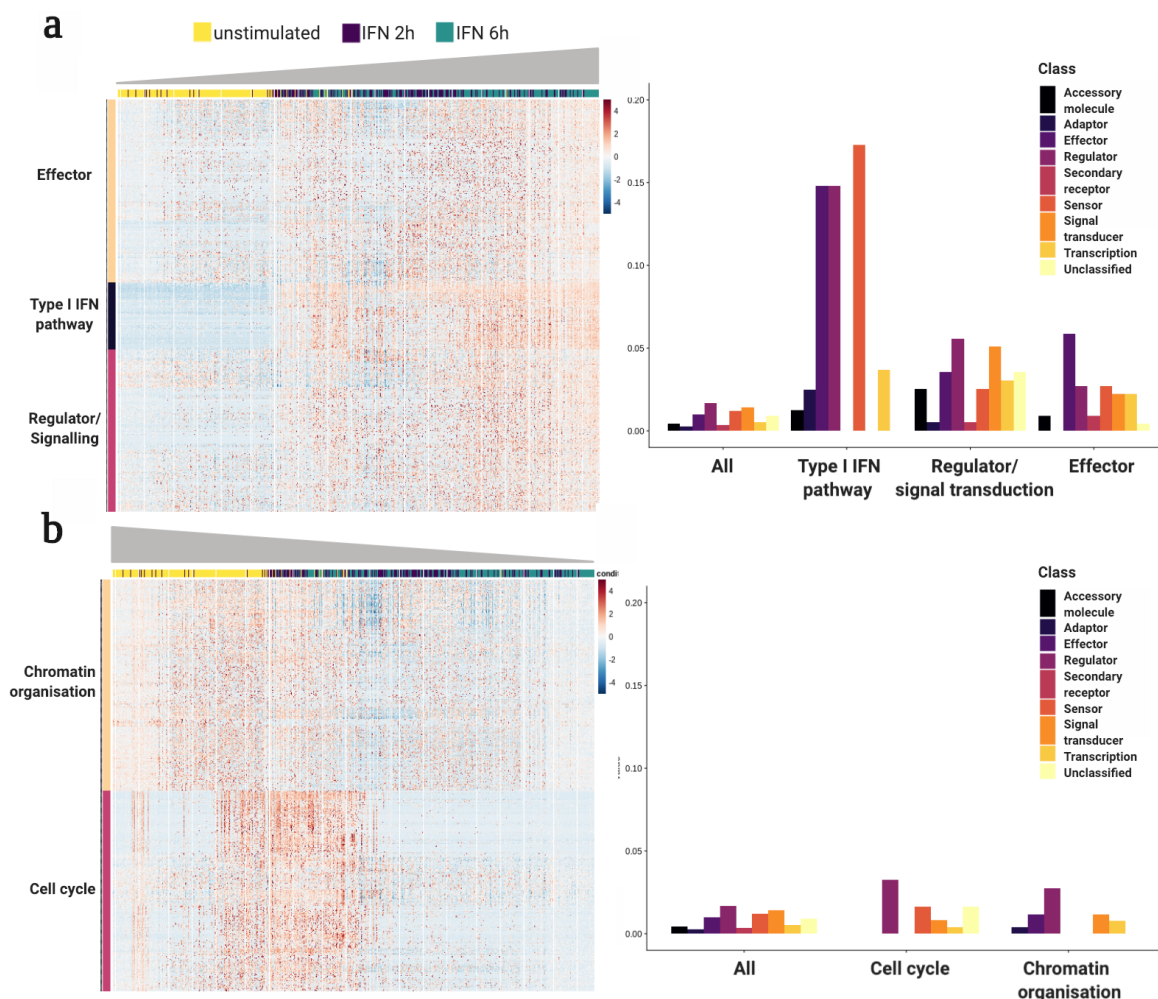| Group	Total group size	Number of IIGs	Enrichment p-value
Canonical Type I IFN	311	103	3.58e-42
Mitochondrial	25	2	0.27
Signal transduction	164	24	0.00032
Organelle localisation	298	26	0.15
Metabolic processes	127	21	0.00011
Protein regulation	75	10	0.018

Fig. 4.10 Modules of co-expressed genes in the response to IFN-β. The SwitchDE package [97] was used to infer a dynamic model of expression for each gene. Genes with a positive 'k' value were termed 'on' genes and those with a negative 'k' as 'off' genes, shown in panels a and b respectively. The 500 'on' and 'off' genes with the most significant qvalue were selected. Their z-score normalised expression across the pseudotime defined in Figure 4.7 is shown on the left; genes were clustered using hierarchical clustering with the ward method. Right: proportion of genes from each IIG functional category within the total cluster; the background set shows representation in the entire set of 15363 genes tested in SwitchDE.
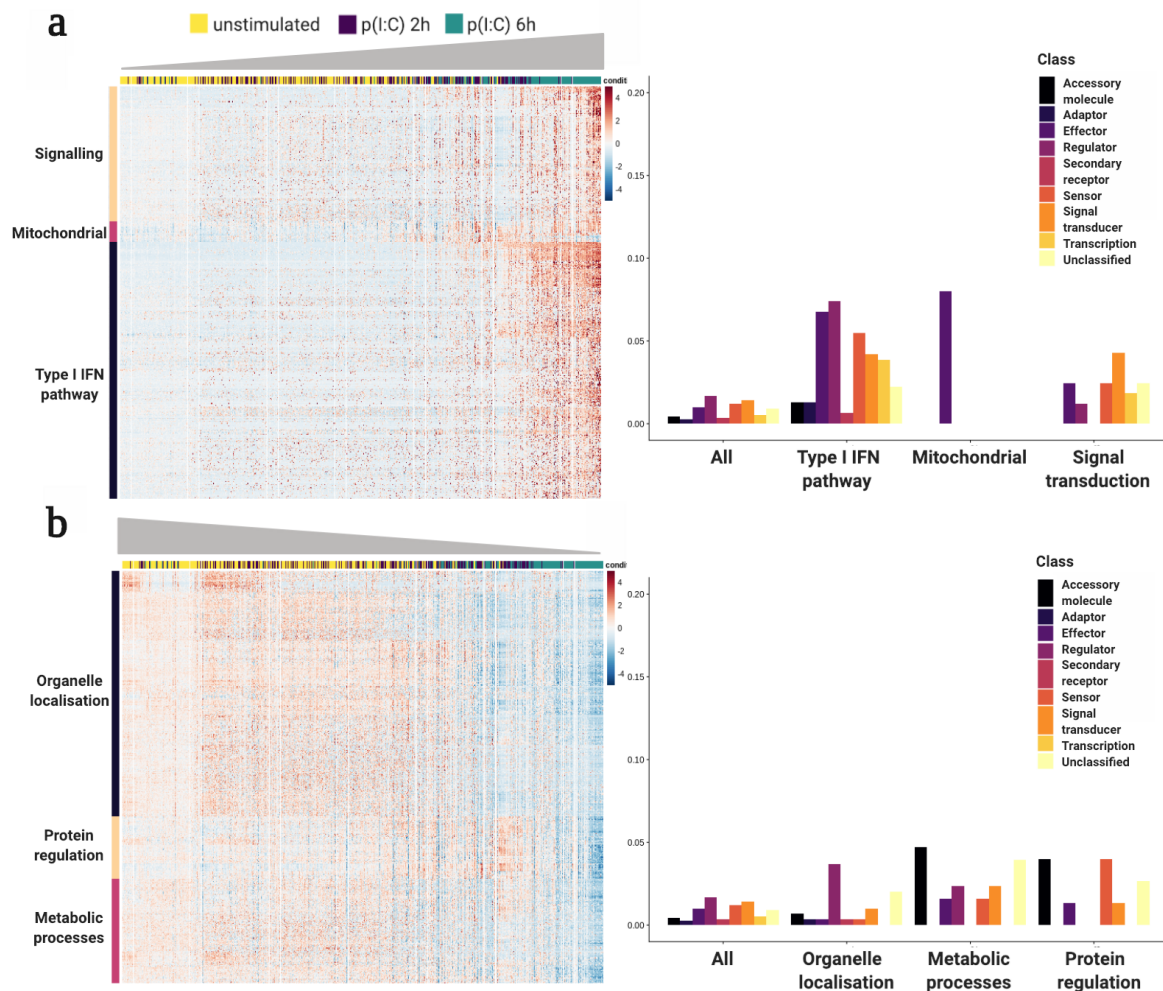
Fig. 4.11 Modules of co-expressed genes in the response to poly(I:C). The SwitchDE package [97] was used to infer a dynamic model of expression for each gene. Genes with a positive 'k' value were termed 'on' genes and those with a negative 'k' as 'off' genes, shown in panels a and b respectively. The 500 'on' and 'off' genes with the most significant qvalue were selected. Their z-score normalised expression across the pseudotime defined in Figure 4.7 is shown on the left; genes were clustered using hierarchical clustering with the ward method. Right: proportion of genes from each IIG functional category within the total cluster; the background set shows representation in the entire set of 15363 genes tested in SwitchDE.

4.4 | Discussion

In this chapter, I have described work charting the evolutionary architecture of the innate immune response. We showed that genes that diverge rapidly between species show higher levels of variability in their expression across individual cells than genes that diverge more slowly. Both of these characteristics are associated with a similar promoter architecture, enriched in TATA-boxes and depleted of CGIs. Notably, such promoter architecture is also associated with the high transcriptional range of genes during the immune response. Thus, transcriptional changes between conditions (stimulated versus unstimulated), species (transcriptional divergence), and individual cells (cell-to-cell variability) may all be mechanistically related to the same promoter characteristics. In yeast, TATA-boxes are enriched in promoters of stress-related genes, displaying rapid transcriptional divergence between species and high variability in expression [192, 193]. This finding suggests intriguing analogies between the mammalian immune and yeast stress responses—two systems that have been exposed to continuous changes in external stimuli during evolution.

We have also shown that genes involved in regulation of the immune response—such as transcription factors and kinases—are relatively conserved in their transcriptional responses. These genes might be under stronger functional and regulatory constraints, owing to their roles in multiple contexts and pathways, which would limit their ability to evolve. This limitation could represent an Achilles' heel that is used by pathogens to subvert the immune system. Cytokines, on the other hand, diverge rapidly between species, owing to their promoter architecture and because they have fewer constraints imposed by intracellular interactions or additional non-immune functions. Cytokines may therefore represent a successful host strategy to counteract rapidly evolving pathogens as part of the host–pathogen evolutionary arms race.

Cytokines also display high cell-to-cell variability and tend to be co-expressed with other cytokines and cytokine regulators in a small subset of cells, and this pattern is conserved across species. As prolonged or increased cytokine expression can result in tissue damage [194–196], restriction of cytokine production to only a few cells may enable a rapid, but controlled, response across the tissue to avoid long-lasting and potentially damaging effects. This cellular variability in response is also observed in the larger human scRNA-seq dataset, where cells in each stimulation condition show a wide distribution of positions across the IFN-β and poly(I:C) response pseudotimes. This further strengthens the notion that the response is heterogenous but highly regulated.

One mechanism to achieve a strongly coordinated response is the regulation of gene modules with discrete functions. By characterising genes whose expression changes across response stimulation, I showed that it is possible identify distinct modules. In both stimulation timepoints, a gene module representing the canonical type I interferon pathway was observed. Further discrete gene clusters were seen, such as those involved in signalling or effector functions. These modules showed differences in temporality and variability of expression. For example, the 'effector' module showed less ubiquitous expression across cells in response to IFN-β treatment compared to the type I interferon module, mirroring the cytokine heterogeneity seen in the cross-mammalian work. These features further suggest tight regulation of expression within each gene set, and across the response as a whole.