Chapter 2

Establishing a system for studying innate immune responses in human fibroblasts

> ***Declaration***
>
> *Bulk RNA sequencing data for protocol optimisation was generated and processed by Tzachi Hagai.*
>
> *During the expansion and stimulation of HipSci lines, invaluable assistance was provided by the Cellular Genotyping and Phenotyping facility at the Wellcome Sanger Institute (WSI).*
>
> *Data processing was conducted with the help of the Cellular Genetics Informatics group, WSI, and Davis McCarthy, EMBL-EBI.*

2.1 Defining optimal stimulation conditions

It was first shown many years ago that synthetic dsRNA - polyinosinic:polycytidylic acid, also known as poly(I:C) - could induce an antiviral response and interferon production in treated cells [130, 131]. However, there are many factors in the stimulation procedure which may affect the response, such as concentration of poly(I:C) used, time after stimulation, and reagents used. Furthermore, the effect of these variables may differ between cell types[132]. It is also possible to induce interferon signalling more directly through administration of interferons, capturing the second cascade of signalling and removing the effect of upstream PAMP sensing.

In order to determine the effects of differing stimulation conditions in human fibroblasts, and enable optimisation of large-scale experiments, bulk RNA-sequencing data generated by Tzachi Hagai was analysed. In these experiments, primary human fibroblasts (HipSci resource) were either stimulated directly with poly(I:C) or interferons. Fibroblasts were seeded approximately 18 hours prior to stimulation, at a density of 100,000 cells per well (6 well plate) or equivalent numbers on smaller plates (12 well and 24 well plate). Cells were cultured either in specialised fibroblast medium (ATCC-PCS-201-041), or in alpha-MEM supplemented with 10% FBS, non-essential amino acids, vitamin C and L-glutamine. In order to achieve sufficient intracellular levels of poly(I:C), addition with lipofectamine 2000 (LF) in a transfection medium (opti-MEM) was used, at a ratio of 1 μg poly(I:C) : 2 μg LF : 100 μg opti-MEM. Interferon was added directly at a concentration of 1000 U/ml. At stated time points (1, 2, 3, 4, 6, 8, 12, 18 or 24 hours) post-stimulation, cells were lysed with RLT buffer containing 1% β-mercaptoethanol, and collected. Library preparation was performed according to Illumina Truseq/KAPA protocols and samples sequences using Illumina Hi-Seq (125bp paired-end sequencing).

The raw RNA-seq data was processed using two pipelines, either mapping with TopHat2 and using cuffLinks to quantify reads in order to identify differentially expressed genes with the cuffDiff tool, or mapping with Kallisto for analysis with Sleuth.

TopHat2 and CuffLinks

Reads were first mapped to the human reference genome (hg37) using TopHat2 [43], before calculating a normalized count for each gene (fragments per kilobase per million, FPKM) using cuffLinks [133]. Differentially expressed genes were identified using the cuffDiff command. These stages were performed by Tzachi Hagai, using default parameters. For analysis of patterns of expression, FPKM was first converted to TPM using the relationship derived by Lior Pachter [20, 134]. A threshold for TPM expression of 2 was chosen, and any transcripts which had expression below this across samples were discarded. The basis for this threshold was the finding by Wagner et al., [135], that RNA-seq data can be modelled as a mixture of two distributions: an exponential distribution for transcripts from inactive genes and a negative binomial distribution for actively transcribed genes. It is shown that the probability of TPM 2 for the exponential distribution (inactive genes) is $< 10^{-8}$, while the probability of a gene with TPM 2 belonging to the class of non-expressed genes is $< 1\%$, for all datasets considered. In order to compare the behaviour of genes across samples, in response to stimulation/time, it is also important to normalise within each gene. The normalised value for a gene in a given condition (z) was calculated using z-score normalisation: $z = (x - \mu)/\sigma$ where: x = raw TPM in the sample considered, μ = average TPM across all conditions, σ = standard deviation of the TPM values across all samples for the gene.

Kallisto and Sleuth

In order to model the time-course response to stimulants with the Sleuth program, reads were first pseudoaligned using kallisto [46]. This software is able to accurately quantify transcript abundances without the need for alignment. Using estimated counts from the kallisto output, a time-based model was fitted in sleuth, using natural splines with five degrees of freedom. Any transcripts for which a likelihood ratio test against a null model had a q-value < 0.01 were considered significant and included in further analysis.

2.1.1 Stimulation with Poly(I:C)

Effects of the transfection procedure

While the primary focus was the investigation of varying poly(I:C) concentration and duration of treatment, the effect of lipofecatmine transfection was first checked, as this is thought to be able to induce up-regulation of gene expression. To identify the effect of LF in fibroblasts, pairwise comparisons between poly(I:C) + LF samples and the media used and LF alone with media were conducted, with differentially expressed genes identified using the cuffLinks software.

Figure 2.1a shows the total number of differentially expressed genes for each sample. It is clear to see that lipofectamine alone does not cause up-regulation of many genes at any time point, while poly(I:C) causes an increase in the expression of hundreds/thousands of genes. Interestingly, at 12 hours there are more differentially expressed genes when the transfection medium is not added, however when only genes involved in innate immunity are considered (Figure 2.1b) this difference is reduced. This trend may be a result of less efficient transfection in the absence of the transfection medium, leading to a delayed but similar induction of the immune response.

To verify induction of the type 1 interferon response, expression of IFN-β across conditions was considered, Figure 2.1c. This confirms up-regulation of IFN-β gene expression when treated with poly(I:C) but not lipofectamine. Furthermore, there is a sustained and slightly higher expression when transfection enhancing medium is added.

Concentration of poly(I:C)

As poly(I:C) may have harmful effects at high concentrations, the effect of reducing the concentration on induction of the type I interferon response was examined. To see whether there was increased sensitivity in the detection of response genes at higher poly(I:C) concentration, the number of differentially expressed genes in the standard concentration (1 μg/ml) over reduced concentrations (either 0.5 μg/ml or 0.1 μg/ml), samples were compared directly using cuffLinks. While there are limited differences between different concentrations of poly(I:C) at 4 and 12 hours, there is a clear increase in differentially expressed genes at 8 hours (Figure 2.1d), particularly compared to 0.1 μg/ml.

The dynamic response to poly(I:C) in two individuals

In order to investigate the response over time after poly(I:C) stimulation, a model using null splines to capture dynamics over the time-course was fitted as described above. Any transcripts for which this model explained behaviour significantly better than a null model (likelihood ratio test, q-value <0.01) were selected. This process was carried out for experimental data from two individuals separately. Many of the most significant genes (smallest q-value) are known to be involved in the antiviral response, such as the IFIT genes, IRF1, CCL2 and OAS1. Plotting the expression profiles of several of these genes (Figure 2.2a) highlights different types of expression patterns and variability between individuals. For example, some transcripts show fairly rapid up-regulation, with a peak at around 8 hours, before a decrease in expression level (IRF1) while

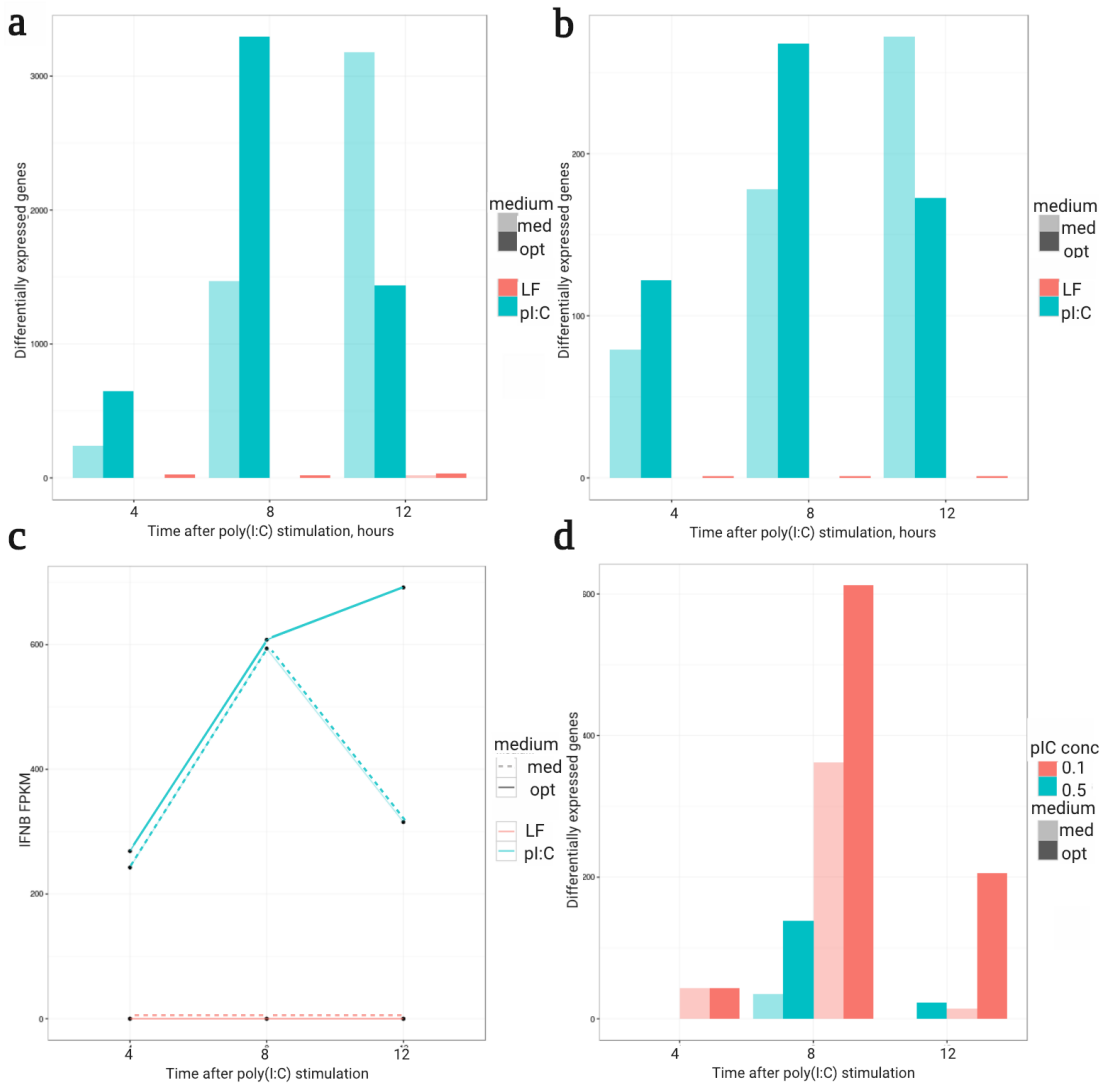

Fig. 2.1 Effects of the poly(I:C) transfection procedure. a-b) Number of genes up-regulated under different stimulant conditions. Colour indicates stimulant: red = lipofectamine alone (LF), blue = p(I:C) + LF (pI:C), while shade denotes transfection medium: light = medium alone ('med'), dark = medium + opti-MEM ('opt') a) Total differentially expressed genes identified. b) Number of differentially expressed innate immune genes. c) Expression of IFN-β (FPKM) across poly(I:C) and transfection control conditions. d) Comparison of response induced by different concentrations of poly(I:C): number of differentially expressed genes identified between 1 μg poly(I:C) and 0.5 or 0.1 μg.

others show up-regulation followed by sustained expression (IFIT3). In several cases, increase in expression is slower in individual 2 (IRF1, CCL2, OAS1) and there are large differences in transcript levels across time points between the individuals (IRF1, CCL2, OAS1). Figure 2.2b shows the behaviour of the entire set of significant transcripts (z-score normalised across conditions for each transcript) in the two individuals. Three groups of genes with distinct expression patterns can be seen, and similar groups are present in both individuals. The first cluster (orange) appears to be 'slow-response' genes, in which expression increases after 8-12 hours and reaches a maximum at 24 hours. In contrast, the cluster highlighted in purple are 'quick-response' genes, peaking at 8 hours. Finally, there are a group of transcripts (blue) which are expressed in the control and earliest time point, indicating genes which are down-regulated in response to poly(I:C). As cells show morphological changes signifying higher levels of apoptosis in later time points, the slow-response genes may be involved in this process. The enrichment (hypergeometric test) of genes marked as 'apoptotic' or 'interferon response' (GO term annotation) was investigated for each cluster, shown in Table 2.1. While all clusters showed enrichment of apoptotic genes, which may suggest that this is a more general feature of the transcripts selected as significant, only clusters 1 and 2 show enrichment of interferon response genes, suggesting that these specific patterns of expression reflect dynamic antiviral responses.

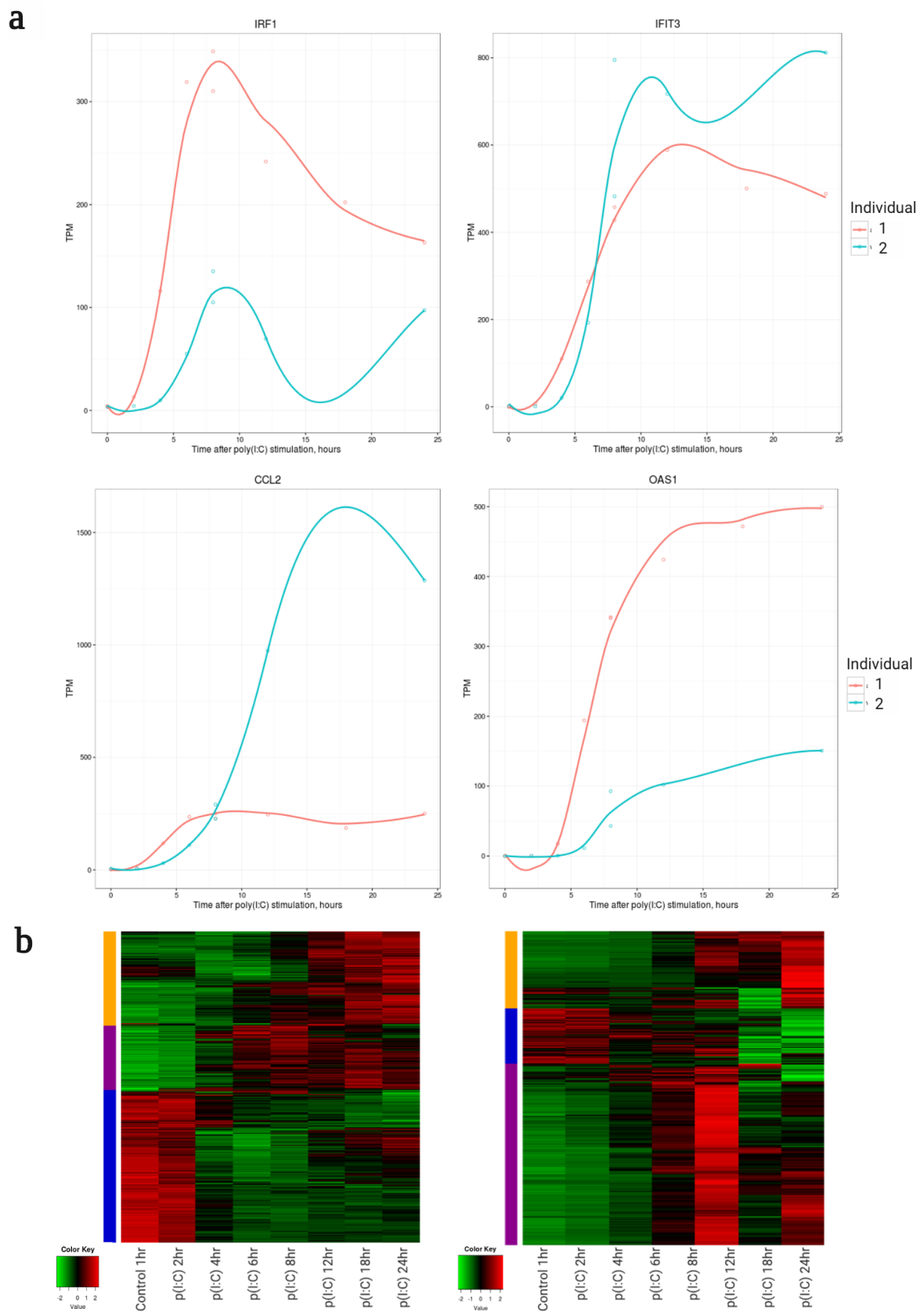

Fig. 2.2 Response to poly(I:C) stimulation over time in two individuals. a) TPM profile of IRF1, IFIT3, CCL2 amd OAS1, respectively, at timepoints of 0-24 hours after poly(I:C) treatment. b) Z-score normalised TPM of all transcripts for which the spline-based model was significant (likelihood ratio test, q value < 0.01) in two individuals.

Table 2.1 Enrichment of apoptotic v.s. IFN response genes in response to poly(I:C).

Group	Total	Apoptosis genes	Enrichment p-value	IFN response genes	Enrichment p-value
Background	*37978*	*1988*	-	*122*	-
Slow-response (orange)	2035	302	0	51	0
Quick-response (purple)	1904	285	0	36	0
Down-regulated (blue)	3024	387	0	6	0.87

2.1.2 Stimulation with interferons

While the results above show that poly(I:C) is capable of inducing an antiviral state in transfected cells, it is also possible to induce interferon signaling in a direct fashion through administration of interferons. As cell types respond differently to distinct interferons, an initial investigation into the response in fibroblasts was conducted, before looking at a more comprehensive time course of interferon-induced changes.

Type I vs Type II interferons

The response to IFN-α and IFN-β (type I) and IFN-γ (type II) at 1 and 4 hours, along with combined IFN-β and IFN-γ stimulation for 4 hours, was studied in two individuals. The heatmaps in Figure 2.3a show two distinct sets of genes in both individuals: one group which responds to interferons at 4 hours (blue), while another which shows higher expression in the controls and at 1 hour after stimulation (orange). As expected, the former group is very strongly enriched for interferon response genes (18/230, background proportion = 122/37978, p = 0), while the latter is not (1/279, p = 0.23). From these heatmaps, it appears that IFN-α and IFN-β elicit similar changes, while IFN-γ shows a distinct response (although up-regulation of type 1 response genes is seen in the sample stimulated with both IFN-β and IFN-γ).

To consider directly the similarity between response to the different interferons, Pearson correlation of gene expression (TPM) between samples was calculated, shown in Figure 2.3b. As would be expected, there is a high level of correlation between control and 1 hour time points, and between 4 hour time points of IFN-α and IFN-β stimulation. As seen above, IFN-γ treatment alone yields a more distinct response.

Interferon β time course dynamics

In order to elucidate the dynamics of response to interferon β, the same modelling approach as discussed for poly(I:C) above was utilised. Similarly, any transcripts for which the spline-based model significantly explained the data (compared to a null model, q-value <0.01) were selected. Again, many of these are known to function in the innate immune response, and the expression of example transcripts after IFN-β stimulation is shown in Figure 2.4a. The difference in dynamics is highlighted in these plots, in which some transcripts are more quickly up-regulated before decreasing (TAP1) or plateauing (DTX3L), while others steadily increase over time (ISG15, STAT1). Interestingly, in several of the transcripts, increase in expression in the second individual begins at later time points compared to the first individual. This may signify a broader delay in response to type I interferons. Although there is a lack of replicates for each individual, suggesting that further investigation may be needed in order to conclude differences between the individuals, the presence of several close time points in each time course deriving from different experimental wells adds reliability to the findings. Considering all significant transcripts (Figure 2.4b), there appear to be four main patterns of expression displayed in both individuals. In order of the timing of peak expression, there is first a group in which expression is highest in the control and at 1 and 2 hours of expression, but down-regulated after this (blue). Another group has similar expression but with a later peak (3-4 hours, brown). Neither of these groups are significantly enriched for genes involved in the interferon response (Table 2.2). The

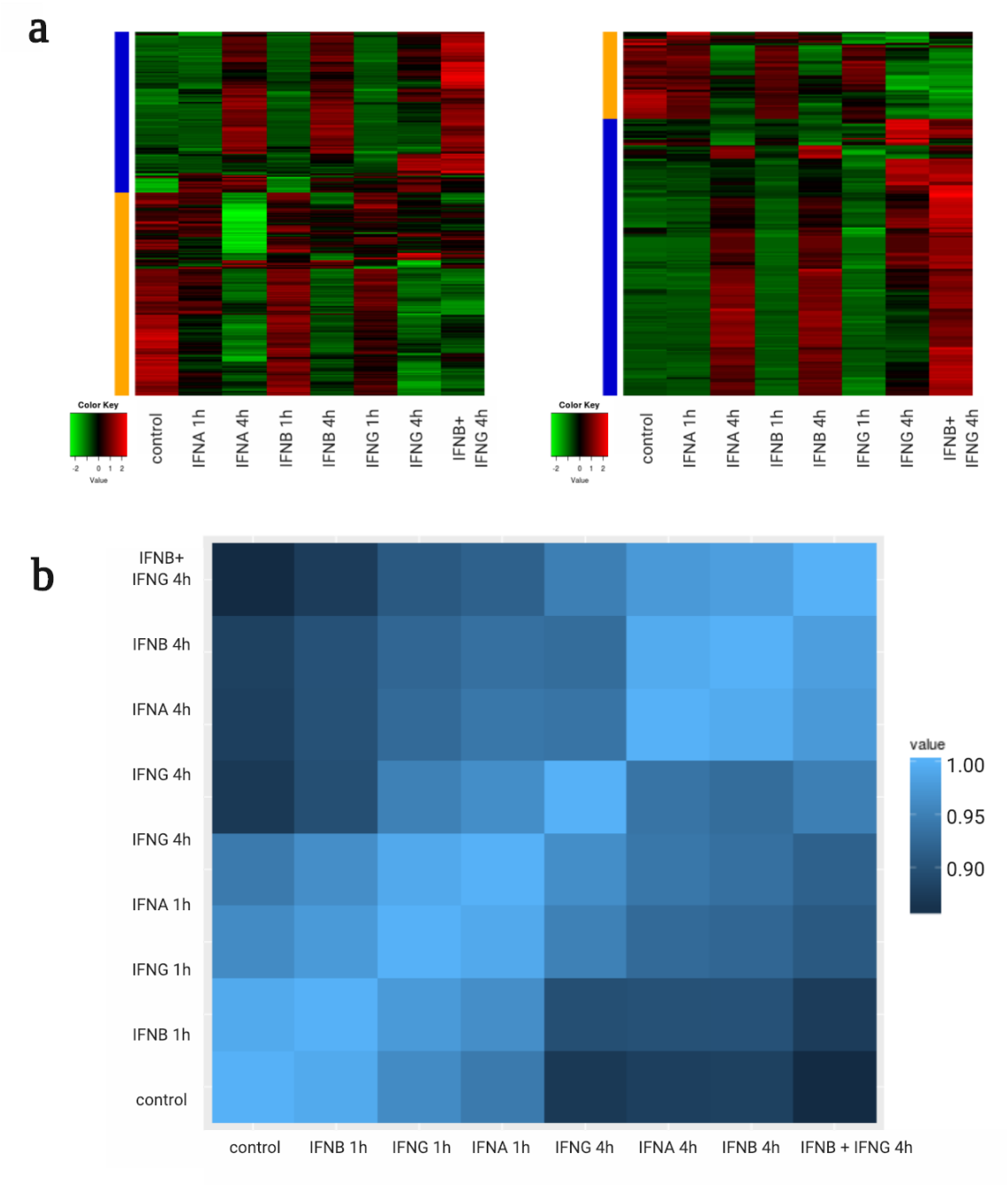

Fig. 2.3 Response to different interferon stimulations. a) Z-score normalised TPM across different interferon treatments (IFN-α, IFN-β, and IFN-γ) at 1 and 4 hours in two individuals; b) Correlation (Pearson coefficient) of TPM between different IFN treatments.

remaining two groups both show up-regulation in response to IFN-β, although in the group highlighted purple this is focused at 6-12 hours post-stimulation, while in the final group (orange) expression is highest at the latest time points. In both cases, there is high enrichment of genes known to play a role in the interferon response.

Table 2.2 Enrichment of apoptotic v.s. IFN response genes in response to IFN-β.

Group	Total	Apoptosis genes	Enrichment p-value	IFN response genes	Enrichment p-value
Background	*37978*	*1988*	-	*122*	-
Early (brown)	288	44	3.1 x $10^{-}13$	2	0.066
Down-regulated (blue)	448	68	8.9 x $10^{-}16$	2	0.17
Slow-response (orange)	480	81	0	26	0
Intermediate-response (purple)	366	68	0	27	0

2.1.3 Innate immunity vs. apoptotic genes across conditions

Thus far, the response to poly(I:C) and IFN-β stimulation has been considered separately, and presence of genes known to be involved in the innate immune response only seen through enrichment values. To further investigate this, alongside the presence of genes known to be involved in apoptosis (a factor in deciding optimal experimental conditions), the expression across control, poly(I:C) and IFN-β treated cells at many time points were considered for the set of innate immune and apoptotic genes. As there is overlap in these two sets of genes, only those which are annotated with one but not the other term were considered. Figure 2.5a) shows the normalised TPM across samples for innate immune and apoptotic genes respectively. While there are similar expression patterns, for example genes expressed most highly in poly(I:C) treatment (highlighted in orange), genes expressed across control and IFN-β samples (blue) and

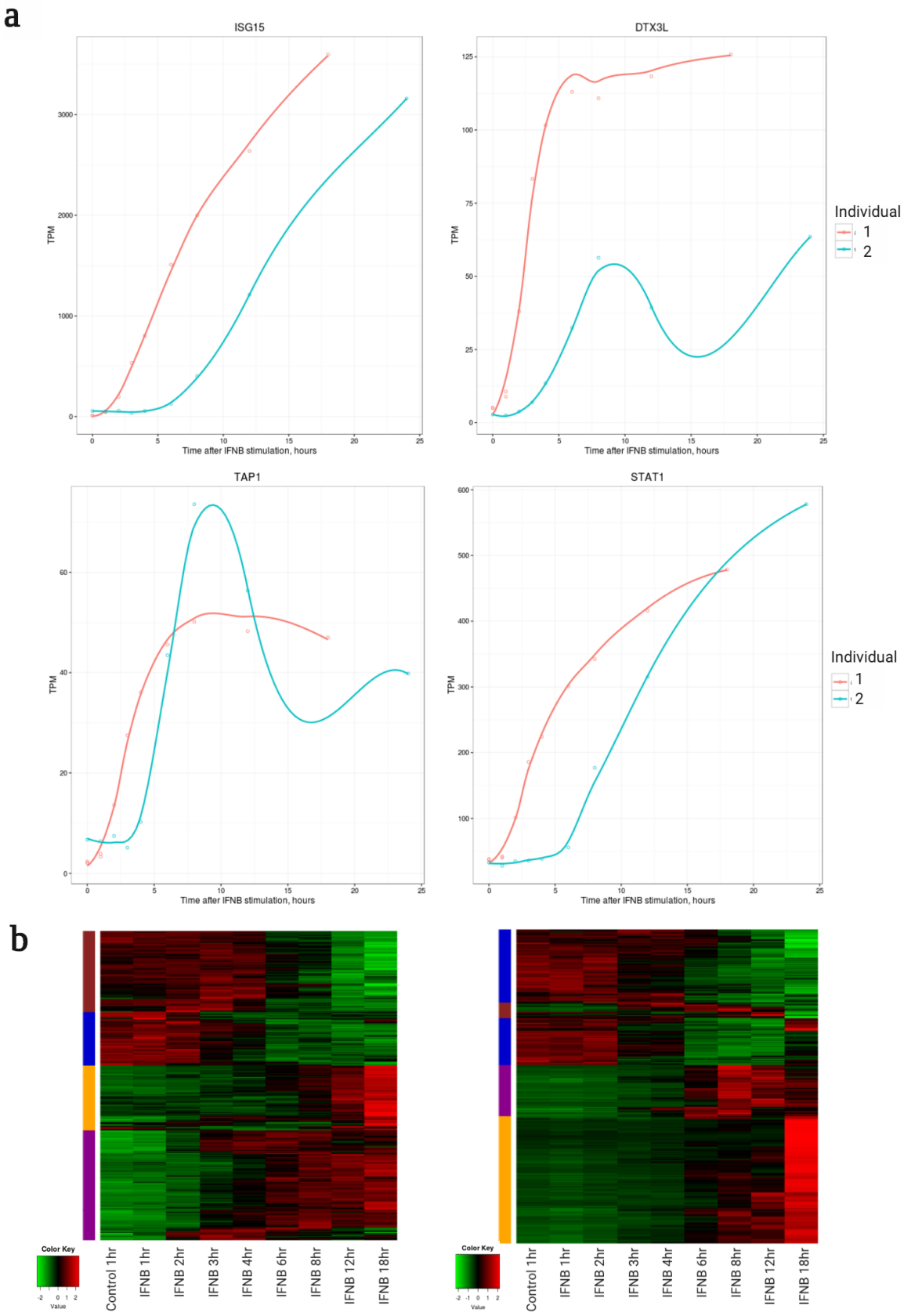

Fig. 2.4 Response to IFN-β stimulation over time in two individuals. a) TPM profile of ISG15, DTX3L, TAP1 amd STAT1, respectively, at timepoints of 0-24 hours after IFN-β treatment. e) Z-score normalised TPM of all transcripts for which the spline-based model was significant (likelihood ratio test, q value < 0.01) in two individuals.

those expressed mostly in control samples and at the latest time points after poly(I:C) stimulation (purple), there are some key differences between the innate immunity and apoptosis set of genes. In the heat map on the left (innate immunity), there are more distinct waves of expression through the poly(I:C) time-course, and there is a set of genes in the lower part of the top cluster which are expressed in both IFN-β and poly(I:C) stimulated cells – a feature missing from the right-hand heatmap.

In Figure 2.5b), the correlation between all samples is considered. While the similarity between samples of the same treatment type is to be expected, a clear difference in the similarity between IFN-β samples and controls can be seen – the correlation is much lower in innate immune genes than apoptotic genes.

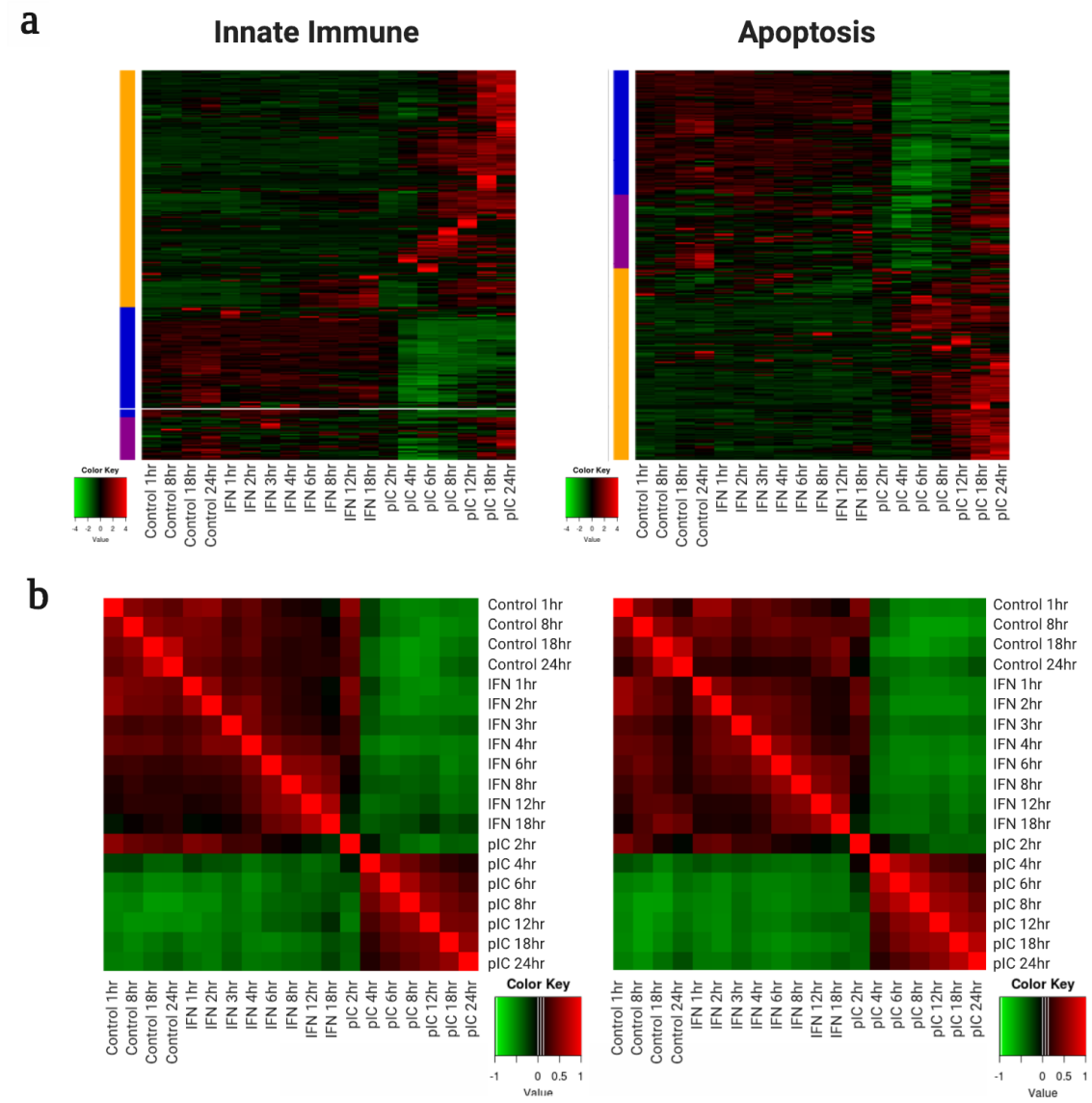

Fig. 2.5 Comparison of control, poly(I:C) and IFN-β treated cells with time. a) Z-score normalised TPM, and b) Correlation between samples, for genes with 'innate immune' (left) and 'apoptosis' (right) functions.

2.2 Large-scale stimulation experiments

In order to study the effect of genetic variation on the innate immune response, a large number of individuals was required. To this end, primary fibroblast cells from the Human Induced Pluripotent Stem Cell Initiative (HipSci; http://www.hipsci.org/) were used. These samples were collected initially for reprogramming into induced Pluripotent Stem Cells (iPSCs), however they provided a ideal resource for stimulation experiments, especially given the genetic profiling carried out through the intiative. The cells derived from healthy individuals spanning a range of ages and both genders (Appendix A).

2.2.1 Expansion of lines

As the initial sample from each line was one vial of 1 million cells, expansion of cells was required to ensure there were enough for stimulation experiments and further studies. Cells were cultured in supplemented DMEM (high glucose, pyruvate, GlutaMAX - Life Technologies), with 10% FBS and 1% penicillin-streptomycin added, until they had expanded at least three-fold. The passage numbers of fibroblasts ranged, as did the apparent quality of the cells, leading to the introduction of a 'grading' system. Not all lines were graded, however this qualitative score - based upon morphology under the microscope - was recorded for the majority of cultures, and cell viability scores were recorded for all lines. Only lines with a grade 3 or above were used in further experimental work; grades, where available, are shown in Appendix A.

2.2.2 Stimulation experiments

The aim of the stimulation experiments is to mimic a viral infection, inducing an effective type I interferon response while minimizing apoptosis. On the basis of the data

presented in section 2.1, it was determined that the following experimental conditions would be used:

- 0.5 μg/ml p(I:C), at 2 and 6 hours
- 1000 U/ml IFN-β, at 2 and 6 hours
- Unstimulated, medium-only (control) cells.

The concentration of p(I:C) was chosen based upon the observation that 1 μg/ml p(I:C) induced a similar response, and 0.1 μg/ml was significantly less effective (Figure 2.1d). The time points were chosen in order to capture the early induction of response at 2 hours, followed by later response at 6 hours, while minimising observation of the apoptotic effect seen at later times (Figure 2.5). To capture the secondary wave of type I interferon signalling, IFN-β was applied directly to cells. Both IFN-α and IFN-β induced a type I interferon response (Figure 2.3), however IFN-β was chosen due to its physiological relevance in fibroblast cells.

A schematic of the experimental setup is shown in Figure 2.6. As can be seen, this was carried out for fibroblasts from many individuals, with three donors being profiled in each experiment. Using the same experimental protocol as above, fibroblasts were stimulated directly with either rhodamine-conjugated poly(I:C) or human recombinant IFN-β. Poly(I:C) was mixed with 1 μl lipofectamine 2000 in 50 μl optiMEM, per well (6 well plate), for 5 minutes prior to transfection. IFN-β was diluted in the media immediately prior to addition. After the relevant period of time, cells were trypsinised and mixed (for example, 'unstimulated' cells from the three donors would be pooled together). The primary aim of this mixing step is to reduce downstream experimental variability between donors, while simultaneously streamlining the collection stage. However, this consequently necessitates the *in silico* deduction of the donor of origin for each cell, as described below.

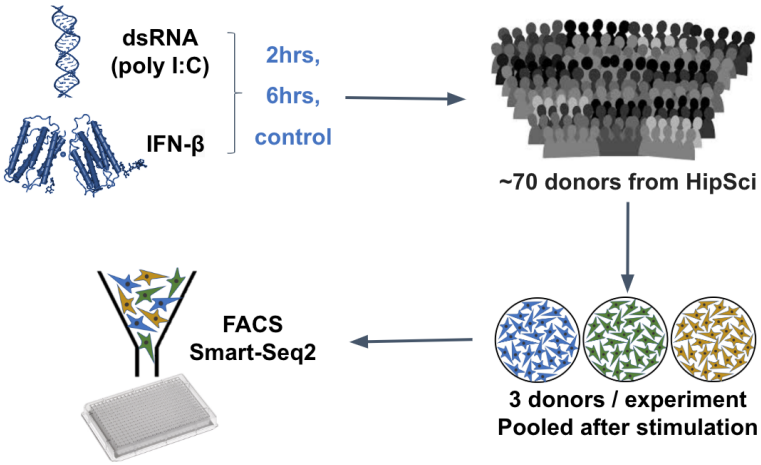

Fig. 2.6 An overview of stimulation experiments on HipSci fibroblast lines. Cells were stimulated with either 0.5 μg/ml p(I:C) or 1000 U/ml IFN-β for 2 or 6 hours, or left unstimulated as a control. Three donors per experiment were stimulated, and pooled together prior to FACS and consequent processing.

In a pilot study of three lines, cells were captured using both a droplet capture method (10X Genomics) and a flow cytometry plate based method. In the droplet capture protocol, control and stimulated cells were (separately) washed with PBS, trypsinised, and resuspended in PBS + 4% BSA. Cells were captured in droplet suspension on 10x Genomics' Chromium machine, and processed to sequencing libraries following the supplier's protocol. Multiplexed libraries were sequenced on an Illumina MiSeq instrument.

For the plate-based method used throughout, cells were washed with PBS, trypsinised, and resuspended in PBS + 0.1% DAPI. Cells were sorted on a Becton Dickinson IN-FLUX into plates containing 2 μl/well lysis buffer. Single cells were sorted individually (using FSC-W vs FSC-H), and apoptotic cells were excluded using DAPI. Rhodamine-positive cells were selected in the poly(I:C) treatments. Reverse transcription and cDNA amplification was performed according to the SmartSeq2 protocol (Picelli et al., 2014), and library preparation was performed using an Illumina Nextera kit. Samples were sequenced using paired-end 75bp reads on an Illumina HiSeq 2500 machine.

2.3 Data processing

For both bulk and single cell data the reads were submitted to fastqQC (version 0.11.7), mapped using Salmon (version 0.9.1) on genome build GRCh37, quantified by featureCounts from the Subread package (version 1.6.2). These programs were run using the Nextflow pipeline [1]. All programs, excepting the Subread package, were installed in the conda environment specified by the file 'env-rnaseq1.6.yml' in the same repository. The Subread package was installed separately. The index for salmon was built on reference v29lift37.

[1] https://github.com/cellgeni/rnaseq

2.4 Additional datasets

2.4.1 Primary skin data

To study the similarity of *in vitro* cultured fibroblasts to *ex vivo* skin cells, a primary skin tissue sample was used. This derived from a collaboration with Professor Muzlifah Haniffa (Newcastle University), and Roser Vento, Felipe Vieira Braga and Gozde Kar.

A skin sample taken from a human female was digested overnight in RPMI, 10% FCS, 100 U/ml penicillin, 100 μg/ml streptomycin, 1% L-Glutamine and 1.6 mg/ml collagenase. Dead cells were removed using beads from Miltenyi Biotec, followed by use of CD45+ beads (Miltenyi Biotec) to remove immune cells according to standard manufacturer protocol. To profile non-immune cells, the CD45- fraction was processed in a 10X Chromium machine (10X Genomics). Libraries were prepared according to the manufacturer's protocol. The resulting libraries were sequenced on two lanes of Illumina Hiseq 2,500 (rapid run mode). Droplet-based sequencing data was aligned, filtered and quantified using the Cell Ranger Single-Cell Software Suite, against the GRCh38

human reference genome provided by Cell Ranger. The output of this procedure (filtered matrix files) was used with the Seurat package. Low-quality cells (cells with less than 500 expressed genes and above 10% mitochondrial reads) were removed prior to further analysis.

2.4.2 Cross-mammalian data

These data were generated by Tzachi Hagai, and involved stimulation of primary dermal fibroblasts from sexually-mature females of four different species (human (European ancestry), rhesus macaque, C57BL/6 (black 6) mouse and brown Norway rat). All skin samples were taken from shoulders. Human cells were obtained from the Hipsci project, as described above. Rhesus macaque cells were extracted from skin tissues that were incubated for 2 h with 0.5% collagenase B after mechanical processing, and then filtered through 100 µm strainers before being plated and passaged before cryo-banking. Rodent cells were obtained from PeloBiotech where they were extracted using a similar protocol.

Prior to stimulation, cells were thawed and grown for several days in ATCC fibroblast growth medium with Fibroblast Growth Kit-Low serum (supplemented with Primocin and penicillin/streptomycin) - a controlled medium that has proven to provide good growing conditions for fibroblasts from all species, with slightly less than 24 h doubling times. About 18 h before stimulation, cells were trypsinized, counted and seeded into 6-well plates (100,000 cells per well). Cells were stimulated as follows: (1) stimulated with 1 μg/ml high-molecular mass poly(I:C) transfected with 2 μg/ml Lipofectamin 2,000; (2) mock transfected with Lipofectamin 2,000; (3) stimulated with 1,000 IU of IFNB for 8 h (human IFN-β for human and macaque cells, rat IFN-β for rat cells, mouse IFN-β for mouse cells; all IFNs were obtained from PBL, and had activity units based on similar virological assays); or (4) left untreated.