# Chapter 5

# Inter-individual variability in the innate immune response

> ***Declaration***
>
> *QTL analysis was conducted using a pipeline developed by Marc Jan Bonder (Stegle Group, EMBL-EBI). The pipeline was run with the support of Ni Huang (Teichmann Group, WSI).*

## 5.1 Introduction

Alongside characterising the innate immune response across the scRNA-seq dataset as a whole, as seen in the previous chapter, it is possible to consider the variability between donors, illuminating differences in response to infections within the healthy human population. One method to do so is by fitting a linear mixed model to each gene, partitioning variation in gene expression into components. This can give insight into the source of variation within a dataset globally, but doesn't pinpoint any effects that may derive from specific genetic differences.

To elucidate the effect of genetic variation, an alternative approach that is commonly used is the eQTL approach, as described in Chapter 1.1. In the case of scRNA-sequencing data, it is possible to generate a mean expression value in two ways: either averaging across all cells to create a single 'pseudobulk' expression level per donor, or treating each cell from a donor as an independent replicate. While the latter approach increases sample size, it comes at the price of increased noise and computational cost. For this reason, the 'pseudobulk' expression value was used in this work.

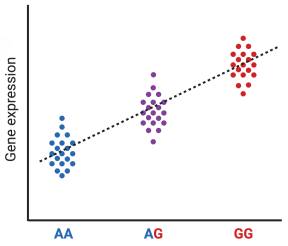
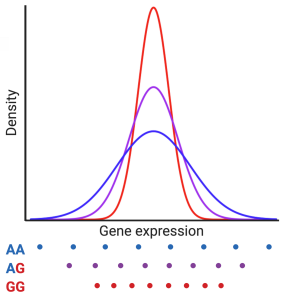
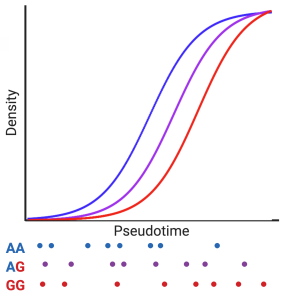
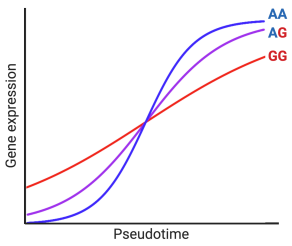
Alongside mean expression level, scRNA-seq also opens up the possibility of identifying variation in the heterogeneity of expression within each donor. For the current study, the simple metric of variance of each gene per donor per response pathway was calculated. As discussed in Chapter 1.2, however, there are alternate methods to reflect variability in expression, such as BASiCS [88] and DM [25].

Beyond metrics around the average and variance of expression between cells, it is possible to define alternative phenotypes capturing the difference in response between individuals (Table 5.1). One example is the proportion of cells expressing a gene, particularly of importance for cytokines and other signalling molecules that show stochastic expression across cells (Chapter 1.3).

Another element of variability is not just the level of expression, but the temporality of this. It may be the case that certain donors respond 'earlier' than others - a phenotype that is not captured by considering expression alone. However, this only provides one phenotype per donor, rather than a per-gene value allowing testing against the specific gene in question. The dynamics of expression may be inferred on a individual gene basis, for example using the SwitchDE package [97]. This infers parameters for the activation time (t0), expression level ($\mu$) and slope of activation (k) (Table 5.1).

In this chapter, I describe the application of these approaches to the IFN-$\beta$ and poly(I:C) stimulation dataset previously described, using both bulk and single cell RNA sequencing. By considering variability in gene expression within these data, I aim to identify a genetic basis for differences in innate immune response between individuals.

Table 5.1 Phenotypes derived from scRNA-seq data.

| Phenotype | Description | Schematic representation |
|---|---|---|
| *Mean* | The mean expression level per gene per sample. Cells from each donor and condition are averaged to produce a 'pseudobulk'. |  |
| *Variance* | The variance of expression per gene per sample. As above, cells from each donor and condition are combined together. |  |
| *Cell proportion* | The proportion of cells expressing a gene, per donor and condition |  |
| *Average pseudotime* | For each donor, the average of all cells for the IFN and poly(I:C) pseudotimes |  |
| *SwitchDE parameters* | Applying 'switchDE' [97] to each donor for the IFN and poly(I:C) pathway, inferring parameters t0, $\mu$ and k per gene |  |

## 5.2 | Variance partitioning of gene expression

To investigate variability within the scRNA-seq data, a variance partitioning approach, as described above, was taken. This was applied to the 5000 most highly variable genes, using the variancePartition package [197]]. The components of 'donor', 'condition' and 'log10(counts)' were included in the model, along with a residual noise component.

Figure 5.1a highlights the large proportion of variance explained by residual noise in this single cell dataset. Despite this, there are 674 genes for which the donor component explains more than 5% of variance in gene expression, and 362 genes for which condition explains more than 5% of variance. The latter threshold may be used to define a set of 'response' genes - those that vary most with stimulation conditions.

While many genes show large variance explained by donor or condition independently (Figure 5.1b), there are 139 genes for which both components explain more than 5% of gene expression variation. This shows promise for the ability to identify genes that vary between individuals in a stimulation-specific manner.

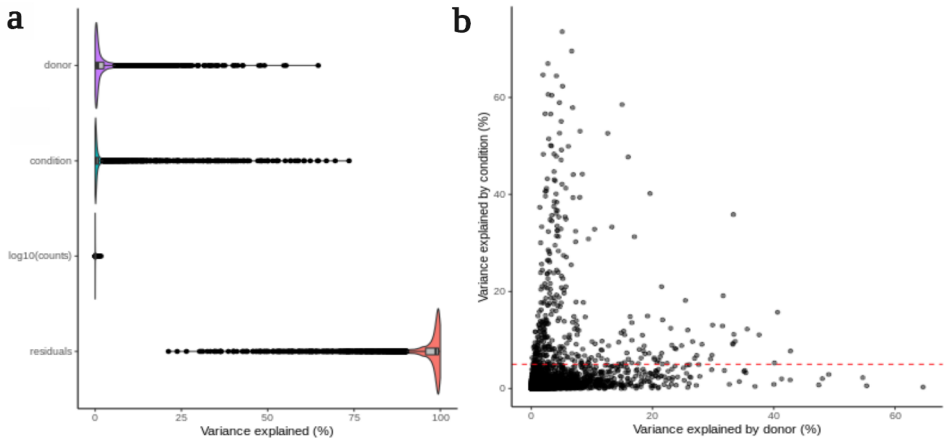



Fig. 5.1 Variance partitioning of gene expression across scRNA-seq data. a) A linear mixed model was fitted for each gene, and variance in expression was partitioned into the following components: 'donor', 'condition' (stimulation and time point), 'log10(total counts)', and a residual noise component. b) For each gene, the variance explained by 'donor' (x-axis) and 'condition' (y-axis) is shown. The red line indicates a 5% threshold for variance explained by condition, used to define a set of 'response' genes.

## 5.3 | eQTL analysis on bulk RNA-seq data

In order to identify the effect of common variants on the innate immune response, differences in the expression of response genes were examined using an eQTL approach, described further below. Gene sets were defined for the IFN-$\beta$ response and poly(I:C) response independently.

To identify genes with a change in expression in response to the two stimuli, the generalised linear model quasi-likelihood F test (glmQLF) in the edgeR package [90] was applied to the bulk RNA-seq data generated in parallel to the scRNA-seq described above. The test was conducted in a pairwise manner between each stimulation condition (for example the IFN-$\beta$ 2 hour time point) and the unstimulated sample, and genes were labelled as 'response genes' using an FDR threshold of 0.05. The union for the two time points per stimulus were taken, yielding an overall set of 'IFN-$\beta$ response' and 'poly(I:C) response' genes. These were supplemented with genes determined as having a high condition-dependent variance in the single cell data: the 362 genes for which the 'condition' component explained more than 5% of variance in expression.

A consistent eQTL mapping strategy was applied to bulk RNA-seq expression and expression traits derived from scRNA-seq. We considered common variants (minor allele frequency $> 5\%$) within a cis-region spanning 100kb up- and downstream of the gene body for cis QTL analysis. Association tests were performed using a linear mixed model (LMM), accounting for population structure and sample repeat structure as random effects (using a kinship matrix estimated using PLINK [198]). All models were fitted using LIMIX [199]. The significance was tested using a likelihood ratio test (LRT). To adjust for global differences in expression across samples, we included the first 10 principal components, calculated on the 500 mostly highly variable genes,

as covariates. To control for multiple testing, we then applied Benjamini-Hochberg correction [200].

The results of eQTL testing on the bulk RNA-seq data for each condition are shown in Figure 5.2: panel a shows testing of the set of IFN-$\beta$ response genes, while panel b shows the equivalent for poly(I:C) response genes. The values refer to the number of significant genes using a multiple testing-corrected p-value threshold of 0.1, with the lower part of each panel showing the condition (or overlap of conditions) in which this set was significant. The total number of significant hits in each condition is shown in the bottom left.

For both the IFN-$\beta$ and poly(I:C) response, the eQTL effects identified are largely context specific, with low overlap seen between the different conditions. Unfortunately the poly(I:C) 6 hour timepoint had a lower number of samples, reducing the power to detect QTL genes. However, the remaining conditions show detection of many response genes. Within these sets, several identified genes are within the list of known innate immune genes (IIGs) described in the previous chapter. These are listed in Table 5.2.

The identification of IIGs with a genetically-determined variation in the unstimulated state raises the intriguing possibility of differences between individuals in their ability to respond based upon expression prior to infection. This is highlighted by QTLs in DDX1 and UNC93B1, both of which are involved in sensing the presence of viral dsRNA. In the case of DDX1, this is in a complex with DDX21, DDx36 and TRIF [201], while UNC93B1 is involved in direct interaction with TLRs [202]. The expression differences across genotypes and conditions is shown for DDX1 as an example (Figure 5.3a). The presence of eQTLs in DDX1 has been found elsewhere; significant results can be seen in multiple tissues of the GTEx resource [203] (Figure 5.3b), including the SNP shown in panel a. Interestingly, this eQTL is most significant in transformed fibroblast cells, although significant effects can be seen in other tissues.

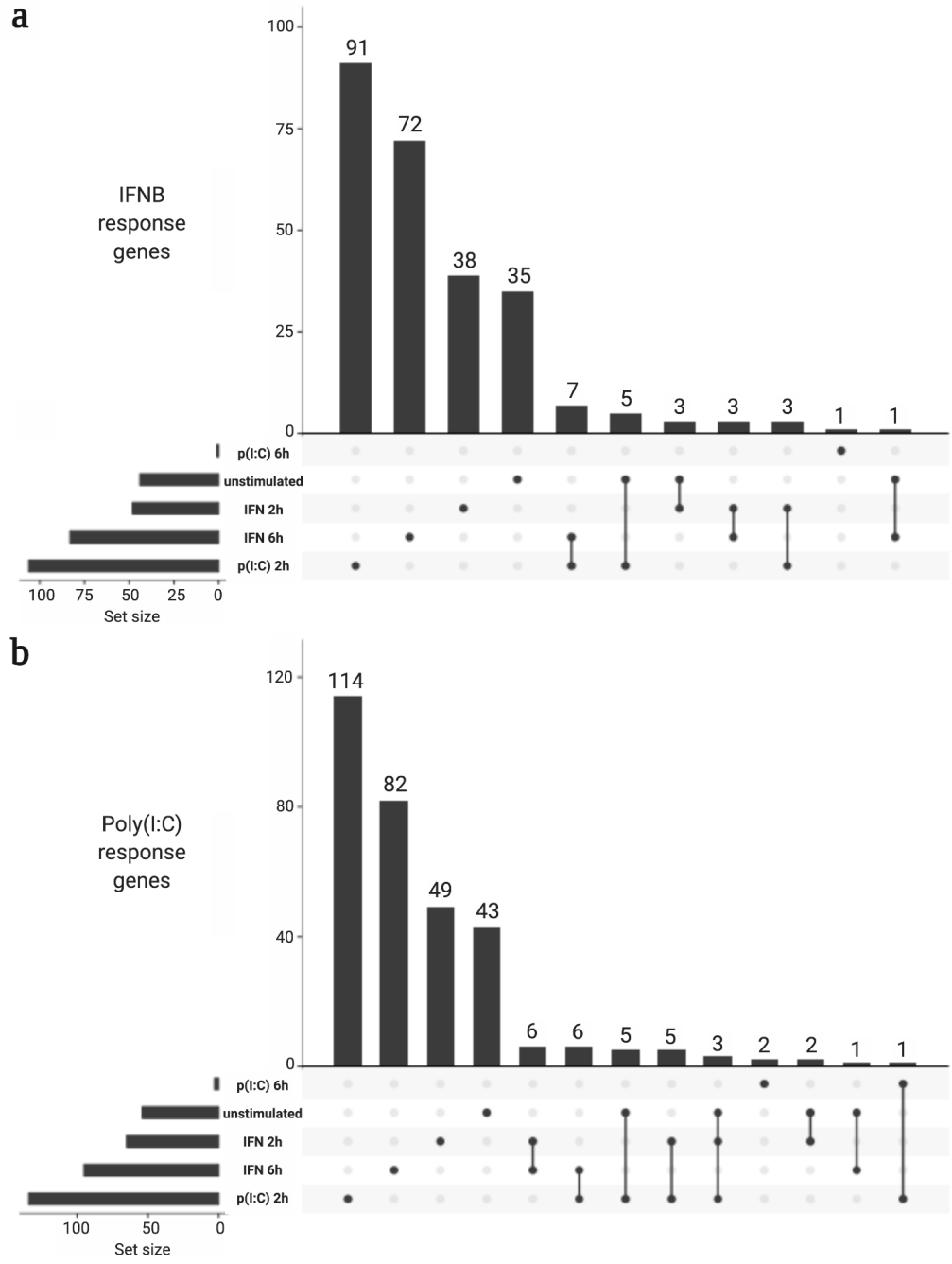



Fig. 5.2 Overlap of eQTLs across stimulation conditions in bulk RNA-seq data, for a) IFN-$\beta$ response genes, and b) poly(I:C) response genes. Values refer to the number of significant genes (multiple testing-corrected p-value $< 0.1$). The total number of significant hits in each condition is shown in the bottom left.

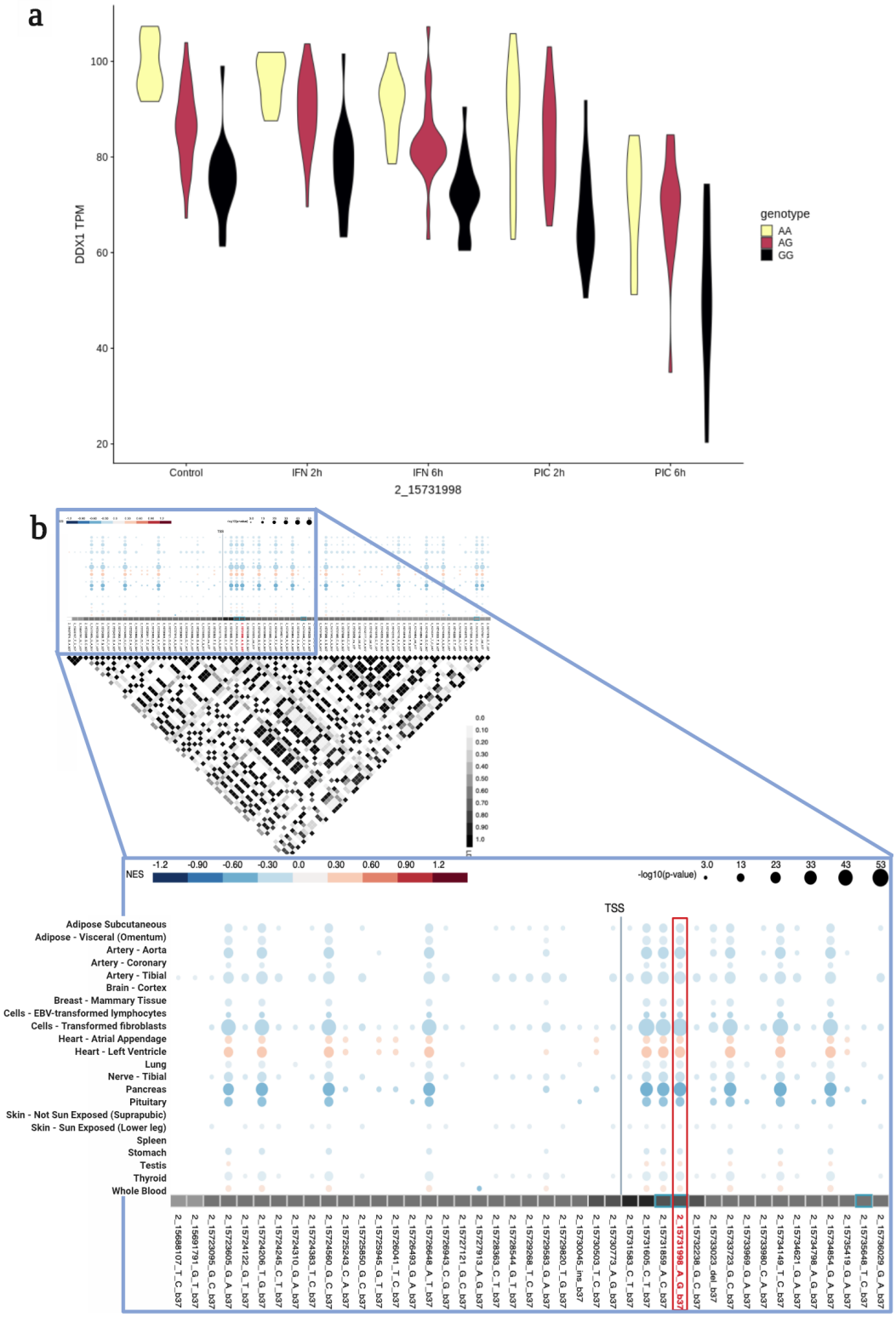



Fig. 5.3 Expression of DDX1 varies with genotype and conditions. a) Expression level (TPM) in bulk RNA-seq data of the DDX1 gene, grouped by stimulation condition and coloured by genotype. b) Presence of eQTLs in the DDX1 gene in the GTEx project. Upper panel shows a zoomed out view of the LD structure, while the lower panel shows the region around the TSS. The SNP identified in panel a is highlighted in red.

Looking at the remaining list of IIGs, several genes known to play a genetically-causative role in disease are identified. For example, mutations in TREX1 have been shown to play a role in both systemic lupus erythematosus and Aicardi–Goutieres syndrome [204–206]. As described in Chapter 1, heterozygous null mutations in IRF7 have been shown to lead to life-threatening influenza infection, and alter type I interferon signalling capacity of dermal fibroblasts [124]. The observation of variability in expression of these genes in healthy individuals could underpin differences in the response to infections within the phenotypically normal human population.

Table 5.2 Significant eQTL hits from bulk RNA-seq - known IIGs.

| Condition | Innate Immune Genes |
|---|---|
| **IFN-$\beta$ response genes** | |
| Unstimulated | *AMACR DDX1 UNC93B1* |
| IFN-$\beta$ 2h | *DNAJA3 TRIM69 TREX1 PRKAR2A UBA7* |
| IFN-$\beta$ 6h | *TREX1 BTN3A2 AMACR IRF7 TRIM69 TRIM4 APOBEC3F CALCOCO2 DUSP7 CCL2* |
| Poly(I:C) 2h | *AMACR PLEC LGALS9 IFIT5 BTN3A2 FES CTSS PRDX1 DDX1 IRAK1BP1 OAS3 CASP7 DUSP7* |
| Poly(I:C) 6h | - |
| **Poly(I:C) response genes** | |
| Unstimulated | *AMACR DDX1 UNC93B1* |
| IFN-$\beta$ 2h | *TRIM69 TREX1 PRKAR2A UBA7 PRKAR2A* |
| IFN-$\beta$ 6h | *TREX1 BTN3A2 AMACR IRF7 TRIM69 TRIM4 CASP12 APOBEC3F CALCOCO2 DUSP7 CCL2* |
| Poly(I:C) 2h | *ABCF1 AMACR PLEC LGALS9 IFIT5 BTN3A2 ULBP3 CTSS PRDX1 DDX1 IRAK1BP1 OAS3 CASP7 ACE* |
| Poly(I:C) 6h | - |

## 5.4 QTL analysis on single cell phenotypes

### 5.4.1 Mean expression

The results of eQTL testing on 'pseudobulk' expression values for each condition are shown in Figure 5.4, with panel a showing IFN-$\beta$ response genes, and poly(I:C) response genes in panel b, as before. While the overall number of genes identified is lower than bulk RNA-seq data, likely due to a slightly smaller sample size and increased noise within the dataset, it is still possible to detect significant QTLs at a multiple-testing corrected p-value threshold of 0.1. Once again, these effects are highly context specific.

Considering the innate immune genes within these sets (Table 5.3), several of the previously identified genes from bulk eQTL analysis appear, such as DDX1, IFIT5, OAS3 and BTN3A2. However, novel genes are identified through this analysis, such as ZC3HAV1, TRIM23 and TRIM25, highlighting the potential of scRNA-seq as an orthogonal data type in eQTL discovery. An example is shown for TRIM25 in Figure 5.5, in which expression in single cell (panel a) versus bulk (panel b) data is shown. The expression level in bulk RNA-seq is low (values between 1-3 TPM), which may be the cause of lack of ability to detect a significant effect in this dataset.

Table 5.3 Significant eQTL hits from scRNA-seq 'pseudobulk' values - known IIGs. The genes detected are all classified as both IFN-$\beta$ and Poly(I:C) response genes.

| Condition | Innate Immune Genes |
|---|---|
| **IFN-$\beta$ and poly(I:C) response genes** | |
| Unstimulated | *TRIM5 ZC3HAV1 DDX1 TRIM23 IFIT5* |
| IFN-$\beta$ 2h | *TRIM5 ZC3HAV1* |
| IFN-$\beta$ 6h | *BTN3A2 OAS3 ZC3HAV1 TRIM25* |
| Poly(I:C) 2h | *TRIM5* |
| Poly(I:C) 6h | *TRIM5 ZC3HAV1* |

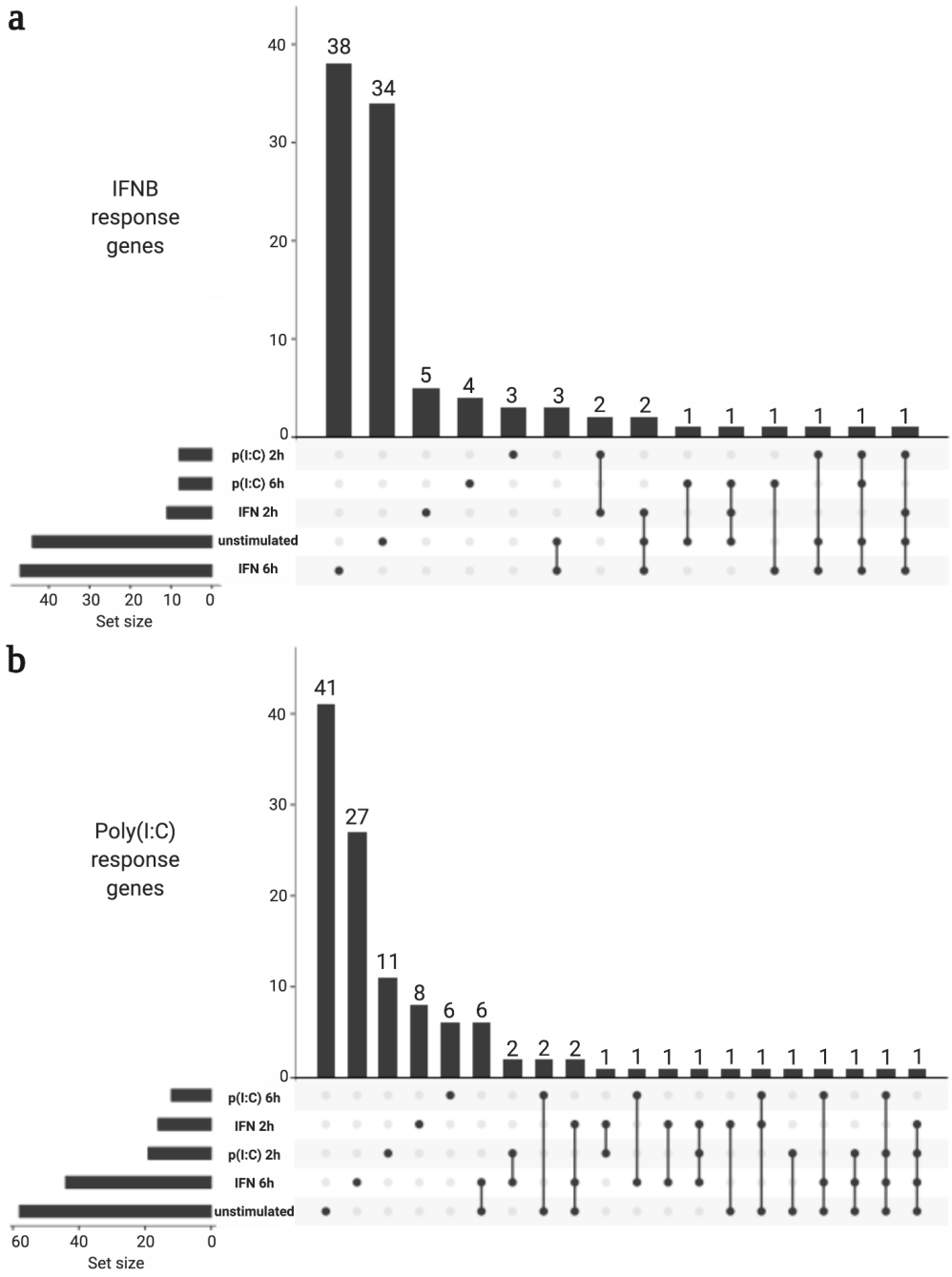



Fig. 5.4 Overlap of eQTLs across stimulation conditions in scRNA-seq derived 'pseudobulk' data, for a) IFN-$\beta$ response genes, and b) poly(I:C) response genes. Values refer to the number of significant genes (multiple testing-corrected p-value $< 0.1$). The total number of significant hits in each condition is shown in the bottom left.

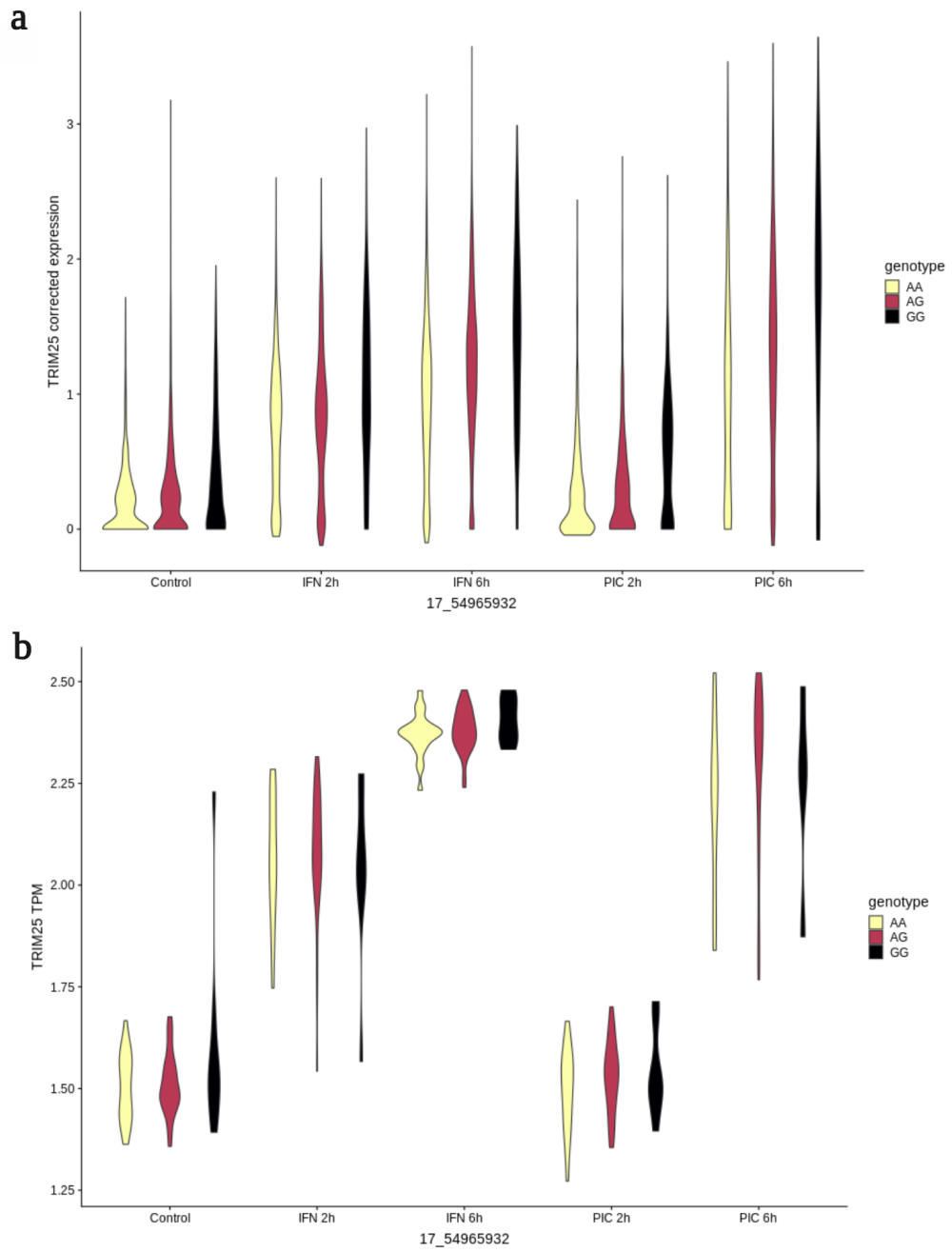



Fig. 5.5 Detection of a TRIM25 eQTL in scRNA-seq data. a) Expression level (Seurat-corrected value) in scRNA-seq data of the TRIM25 gene, grouped by stimulation condition and coloured by genotype. b) Expression level (TPM) in bulk RNA-seq data of the TRIM25 gene, grouped by stimulation condition and coloured by genotype.

### 5.4.2 Other response phenotypes

Using the IFN-$\beta$ and poly(I:C) response pseudotimes described in Chapter 4, the average position of cells for each donor was calculated, for each pseudotime independently. As the pseudotime was inferred based upon all donors, the pseudotime average for each donor should reflect the speed of response relative to other donors. Furthermore, the parameters reflecting dynamics of gene expression (t0, mu and k) were inferred for each donor across the two pseudotimes using the SwitchDE package [97]. The variance of gene expression, along with the 'cell proportion' (i.e. the number of cells expression each gene), across the two responses was also calculated.

In preliminary attempts at using these scRNA-seq derived phenotypes, only the variance of genes and proportion of expressing cells across the IFN-$\beta$ and poly(I:C) response pseudotimes showed significant QTL genes. While this did not result in many innate immune genes being identified, two novel hits were TECPR1, which had a significant cell proportion QTL in both the IFN-$\beta$ and poly(I:C) response, and SMARCE1, which has a significant cell proportion QTL in the poly(I:C) response. Furthermore, genes identified above - ZC3HAV1 and TRIM5 - were also detected as cell proportion QTLs. These results show the potential of scRNA-seq to identify variation in the proportion of cells expressing innate immune genes between individuals. ZC3HAV1 is shown as an example in Figure 5.6. For this gene, no bulk eQTL was detected (Figure 5.6a), however there appears to be a shift in the distribution of cells expressing the gene (and also a shift in expression level) between the genotypes.

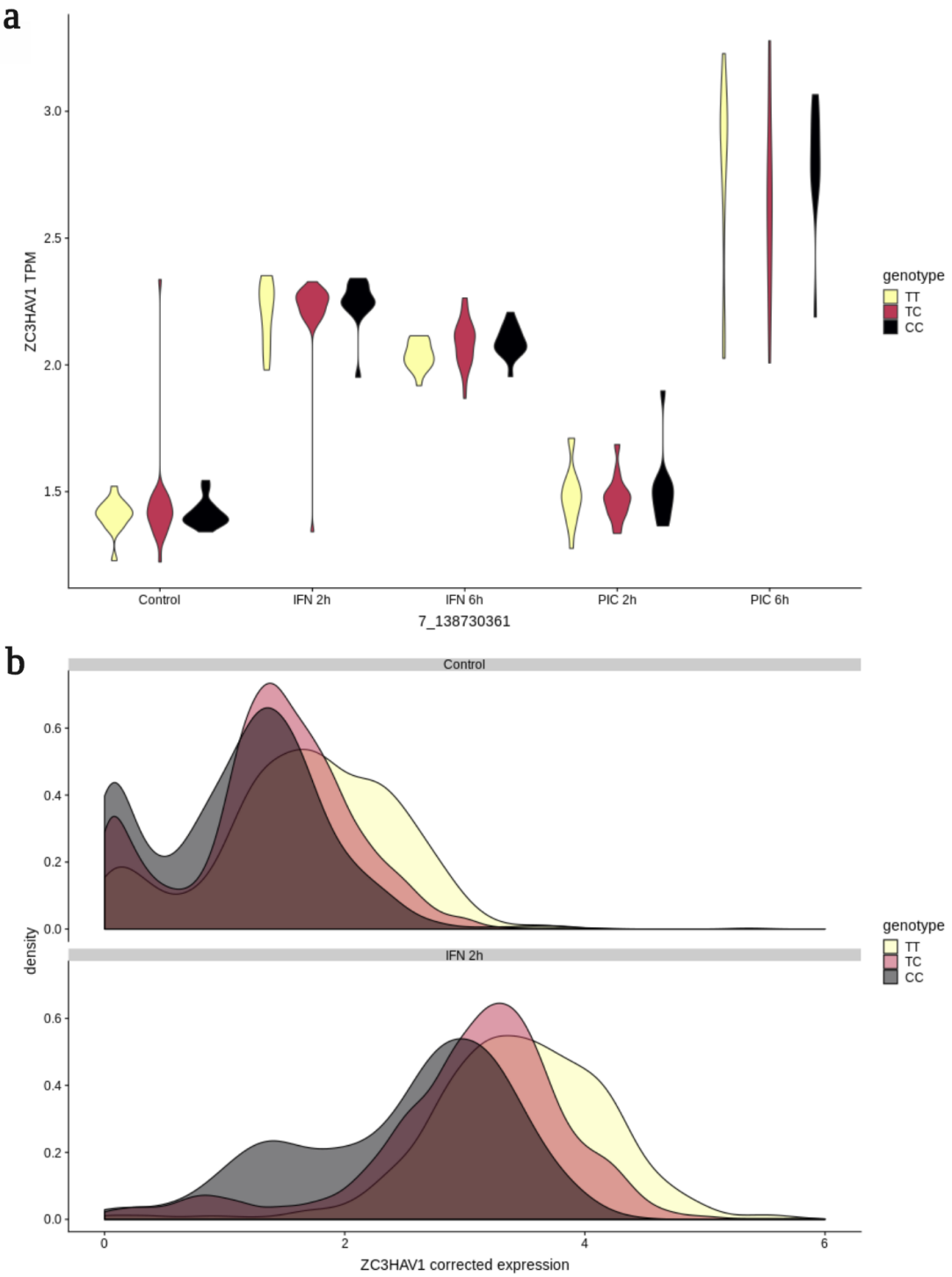



Fig. 5.6 Detection of a ZC3HAV1 cell proportion QTL in scRNA-seq data. a) Expression level (TPM) in bulk RNA-seq data of the ZC3HAV1 gene, grouped by stimulation condition and coloured by genotype. b) Expression level (Seurat-corrected value) in SCRNA-seq data of the ZC3HAV1 gene coloured by genotype. Distribution of expression is shown for the unstimulated cells (upper panel) and cells after 2 hours of IFN-$\beta$ treatment (lower panel).

## 5.5 Characterisation of QTL innate immune genes

Taking a combination of genes identified across QTL approaches yields a total set of 391 genes. While the majority were identified through bulk eQTL analysis (Figure 5.7a), use of the single cell data identified 89 additional genes.

In order to characterise these further, enrichment for particular functional categories was investigated for genes within the known IIG set (Figure 5.7b). Each functional class was compared against the background number in the entire scRNA-seq dataset. Sensors were found to be significantly enriched (hypergeometric test, pvalue = 0.021), which could suggest an interesting source of variability in response to infection through differences in detection of pathogens.

Having identified many response genes with a genetic basis for variation between individuals, it is interesting to consider whether these genes show co-regulated expression at a single cell level and across pseudotime. To this end, the expression of all genes with significant QTLs identified was plotted against the IFN-$\beta$ and poly(I:C) response pseudotimes defined in Chapter 4 (Figures 5.7c-d). From this analysis, it appears that there are modules of co-expressed genes, particularly in response to poly(I:C) treatment, however further work will be needed to elucidate whether there is a genetic mechanism underpinning co-regulation of these genes.

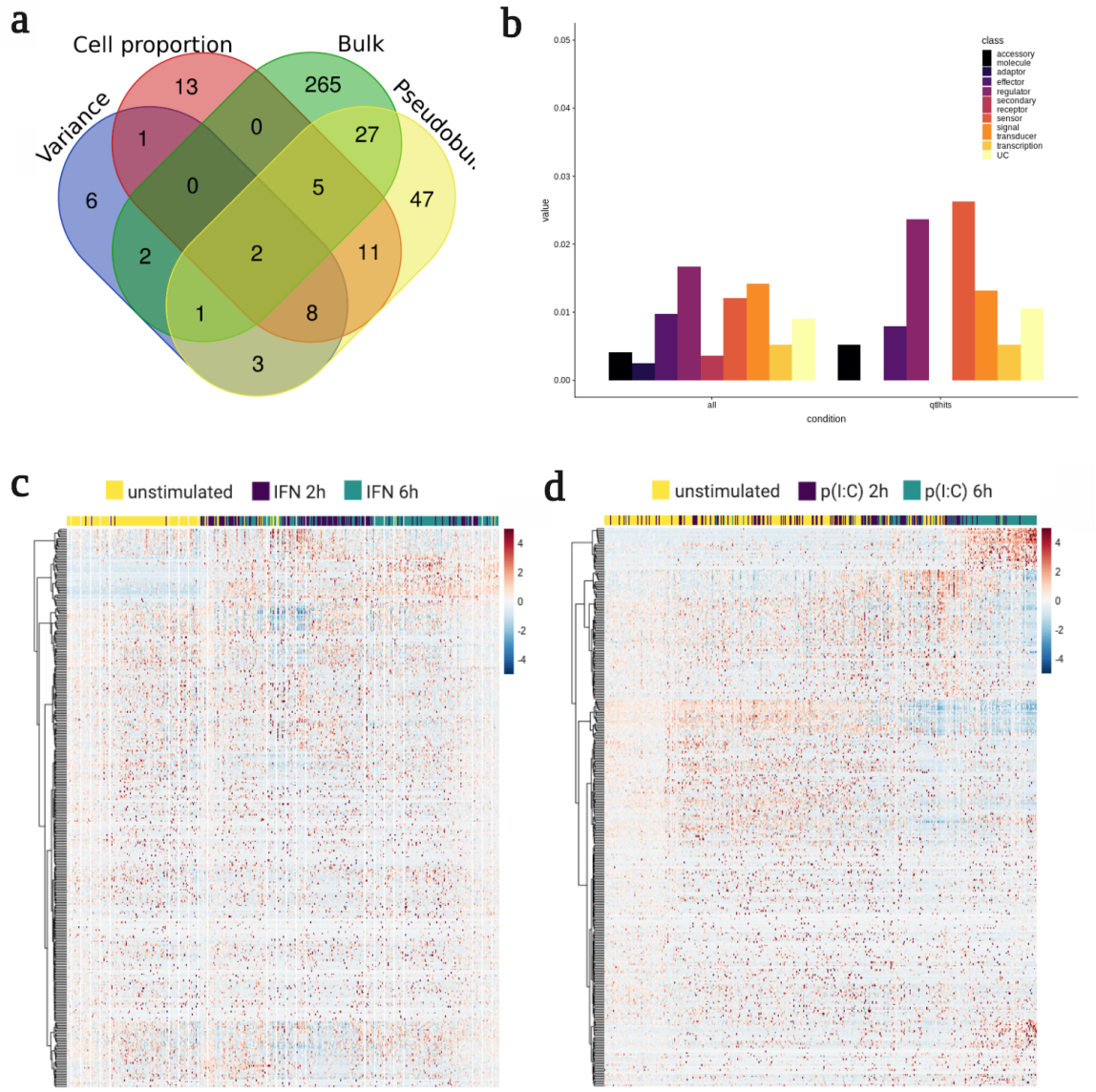



Fig. 5.7 Investigation of significant genes across QTL approaches. a) Overlap in genes identified from bulk RNA-seq data and scRNA-seq (pseudobulk, variance and cell proportion QTLs). b) Distribution of IIGs across functional categories - background on the left, vs significant QTL genes on the right. c-d) Expression of QTL genes across the IFN-$\beta$ and poly(I:C) response pseudotimes respectively.

## 5.6 Discussion

Variability in the response to viruses has long been studied, predominantly through investigation of individuals with susceptibility to particular infections, or with a genetic disorder in innate immune genes. In this work, I showed variability in this response across healthy individuals, first using a variance partitioning approach to highlight genes whose expression was explained by condition or donor.

Applying QTL approaches to bulk RNA-seq data revealed hundreds of genes with a genetic basis for inter-individual variation. Several of these have been previously implicated in disease, such as TREX1 [204–206] and IRF7 [124], validating the detection of biologically interesting hits. However, the identification of other innate immune genes in donors with a normal phenotype highlights the potential to understand variability in the response within the population as a whole. Furthermore, characterisation of QTL genes which are not annotated as known innate immune genes may yield novel insight into the type I interferon response.

Using scRNA-seq data, it was possible to expand the set of QTL genes, primarily through calculation of a 'pseudobulk' expression metric. In the case of phenotypes reflecting differences in temporal dynamics, such as average position in response pseudotime, or SwitchDE parameters, it is likely that the number of cells and donors in the current study is not large enough given the amount of noise within the data, and hence the difficulty in robustly inferring dynamic parameters. For such insights, it will be necessary to further develop the phenotypes used in such approaches. Furthermore, an increase in sample size will be required to improve power in single cell-based QTL studies. In a recent computational analysis, Sarker *et al.* estimated the sample size that would be required to detect dispersion QTLs in scRNA seq data (QTLs that affect the variability but not mean expression level) [207]. They showed that 4,015

individuals would be the lower bound to achieve 80% power to detect the strongest dQTLs in iPSCs. While this number will be lower for phenotypes that also affect mean expression level, an increase in number of individuals profiled will broaden the range of molecular phenotypes interrogatable with QTL approaches.

Moving forward, further analysis will shed light on the nature of the QTL hits identified. Through combination with ChIP-seq and ATAC-seq data, it will be possible to overlap identified genomic loci with regulatory regions. This will shed light on the mechanism of regulation, for example through transcription factor binding sites or enhancer regions. This will also allow detection of regions that may be 'primed' in an unstimulated state, as shown previously in human macrophages [208].