SI Appendix for “Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system”

Authors: Alexis Vandenbon[a,1], Viet H. Dinh[a], Norihisa Mikami[b], Yohko Kitagawa[b], Shunsuke Teraguchi[c], Naganari Ohkura[b,d], Shimon Sakaguchi[b,1]

[a] Immuno-Genomics Research Unit, Immunology Frontier Research Center (IFReC), Osaka University, Suita, 565-0871, Japan

[b] Laboratory of Experimental Immunology, Immunology Frontier Research Center (IFReC), Osaka University, Suita, 565-0871, Japan

[c] Quantitative Immunology Research Unit, Immunology Frontier Research Center (IFReC), Osaka University, Suita, 565-0871, Japan

[d] Frontier Research in Tumor Immunology, Graduate School of Medicine, Osaka University, Suita, 565-0871, Japan

[1] To whom correspondence may be addressed. Email: alexisvdb@ifrec.osaka-u.ac.jp or shimon@ifrec.osaka-u.ac.jp

Table of Contents

SI Results

Practical example analysis using the Immuno-Navigator database

Analysis of single genes

Immuno-Navigator can be accessed here. Only a brief description of an example query is given here. For more information, we also refer to the relevant sections of the main text, and to the online documentation of Immuno-Navigator.

We will use *Foxp3* as an example query gene. On the top page, we can input the gene symbol “Foxp3” as a query (Fig. S1A). The gene page (Fig. S1B) includes basic information and links to external databases. Below this, there are several tabs with additional data. The “Probes” tab shows the available probe set identifiers for this gene, as well as a boxplot showing the distribution of values observed for this probe in each cell type. In this case, *Foxp3* has only one probe set, and its highest signals are observed in regulatory T cell (Treg) samples, which fits with the known function of Foxp3 as master regulator in the development and function of Tregs (1, 2). Hovering over the boxplots shows additional information, such as the cell types and median signals.

Clicking the probe set ID takes the user to a table showing PCCs between this probe and all other probes in the dataset for all combined data and for cell types of interest (Fig. S1C). Probes can be sorted in order of increasing or decreasing PCC values for each dataset. Under “cell type selection”, a selection can be made of cell types to display in the table, using a menu in which cell types are roughly ordered according to the hematopoietic lineage tree (Fig. S1D).

From the table, for any cell type, scatter plots can be shown for the probe of interest versus any probe in the table (Fig. S1E). In this case, the probe of Foxp3 is shown against a probe for *Il2ra* (also known as *Cd25*), another marker for Tregs, over all data (left, with samples coloured according to cell type) and within the Treg data only (right). The correlation between *Foxp3* and *Il2ra* within the Treg samples is relatively high (PCC: 0.45), but their correlation is exceptionally strong over the entire dataset (PCC: 0.82), with Treg samples being the only samples with high signals for both genes. This is also true for other Treg markers (such as *Tnfrsf4* (also called *Ox40*), *Gpr83*, *Ctla4*, *Ikzf4* (also called *Eos*)), but also genes which have so far not been reported as candidate markers. Pairwise comparisons can also be directly accessed from the tab “Gene pair comparison” on the top page (Fig. S1A).

The tab “Top correlated genes” shows the top positively and negatively correlated genes for the query gene in each cell type, and the PCC values of the query gene versus the genome-wide set of genes can also be downloaded, for all datasets. The tab “Correlation network” shows for each cell type a small network of the query gene, its 5 most strongly correlated genes, and in turn their 5 most correlated genes (Fig. S1F). Thick edges represent significantly correlated genes. In this case, within the Treg samples, *Foxp3* is significantly correlated with *Il2ra* (*Cd25*), *Dst*, and *Ikzf4* (*Eos*). These genes, in turn, are highly correlated with other Treg markers, such as *Nrp1*, *Ctla4*, and *Tnfrsf4* (*Ox40*). On the other hand, *Foxp3* has also high (though not significantly high) correlation with *Nfkb1* and *Bcl3*. *Nfkb1* encodes a subunit of NF-κB, a key regulator of the response to various immune stimuli, and *Bcl3* encodes a transcriptional co-activator of NF-κB. These two genes are in turn connected with *Stat3*, an important regulator of responses to cytokines and immune tolerance (3). Thus, the inspection of neighboring genes in the correlation network can suggest the function of the query gene and the presence of distinct regulatory modules. These correlation networks can also be downloaded in the Cytoscape.js (cyjs) format (4).

Correlation Gene Set Enrichment Analysis

It is often interesting to see if a gene of interest has any bias in its correlation with a set of genes that share some particular features. We implemented a tool, “correlation GSEA”, to detect such biases using a modification of the widely used Gene Set Enrichment Analysis (GSEA) approach (5) (see SI Appendix, section “Haemcode ChIP-seq analysis”). Correlations of the query gene *X* with the set *S* of input genes are compared with those with non-input genes. Biases between them can subsequently be quantified using “enrichment scores”, as defined in the original GSEA study. High positive enrichment scores indicate a bias towards positive correlation between the query gene and *S*, high negative enrichment scores indicate a bias towards negative correlation. These are not exclusive; as we will show below, a regulator can have both a bias towards positive as well as towards negative correlations with its target genes. A lack of a clear bias results in enrichment scores close to 0.

An example of features suitable for this methodology would be DNA binding af a regulator protein, which can be inferred from ChIP-seq data. The query gene *X* can be the gene encoding the protein for which a ChIP-seq experiment was conducted (hereafter referred to as the “ChIPed regulator”), and the set *S* could be genes that appear to be bound by that regulator. Although this approach is not limited to such inputs, below we will focus on the analysis of ChIP-seq data.

Here, we present the analysis of 104 ChIP-seq data sets provided in the Heamcode database (6) (see SI Materials and Methods section "Haemcode ChIP-seq analysis"). For each experiment, we calculated GSEA enrichment scores between the ChIPed regulator and its target genes, using the expression data of the same cell type as used for the ChIP-seq experiment. Fig. S14 shows an overview of the results (Fig. S14A), as well as a few example cases with distinct tendencies (Fig. S14B-E). A complete list of positive and negative enrichment scores are also shown in Table S5. Roughly, we can distinguish 3 broad classes.

A first class is regulators with a bias towards positive correlation with their target genes. This tendency was frequently observed: as a rough illustration, in 46 out of 102 Haemcode datasets a positive enrichment score > 0.2 was obtained. Factors Elf1, Foxo1, and Ets1 in Tregs (marked in red in Fig. S14A) show a high positive enrichment score, indicating a relatively strong bias towards positive correlation with target genes (see also cumulative distribution plots in Fig. 5 in the main text). Stat1 and Nfkb1 in cDC cells show a similar tendency towards positive correlation (marked in Fig. S14A; see also Fig. S13 for the cumulative distribution plots). As an illustration, the enrichment score plot for Stat1 in cDC cells is shown in Fig. S14B. The high positive enrichment score (0.52) reflects a strong tendency for Stat1-bound genes to have positive correlation with the *Stat1* gene expression in cDC cells.

In a second group of cases, no clear bias towards either positive or negative correlation was seen. As described in the main text of this paper, Foxp3 expression in Tregs follows this pattern (see also cumulative distribution plots in Fig. 5 in the main text). Fig. S14C shows the enrichment plot for Foxp3 in Treg cells. The positive (0.114) and negative (-0.012) scores are low, reflecting a lack of correlated expression between *Foxp3* and Foxp3-bound genes in Tregs. In contrast, as described in the main text (see also Fig. S13A), *Foxp3* does tend to have positive correlation with its target genes when seen over the combined data for all cells (enrichment score: 0.225). This correlation is caused by *Foxp3* and its target genes both having high expression in Treg cells, even though they lack correlation within the Treg-derived data.

Plotting the positive enrichment score of each ChIPed regulator in the cell type that was used for the ChIP experiment versus that over the combined data, we could identify several additional regulators following a similar pattern to that of Foxp3 in Tregs (Fig. S15). Examples include E2f1, Hif1a, Maff in cDCs, Junb, Sfpi1, C/EBPβ, and Atf3 in macrophages, Foxp3 in Tregs, and Stat6 in Th2 cells.

A third pattern is shown in Fig. S14D. Here, PU.1 (encoded by *Sfpi1*) shows a weak enrichment towards both positive correlation as well as towards negative correlation of expression with its target genes in macrophage cells. This pattern was observed for a limited number of TFs, including C/EBPβ in macrophages, and to a lesser degree in cDC cells (see also Fig. S13C). These factors are known to have genome-wide widespread binding in these cell types, and have been described to pre-bind regulatory regions of stimulus-induced and -repressed genes even before stimulation (7). One possible explanation for the bias towards both positive and negative correlation is therefore that the binding of these factors prepares a scaffold for stimulus-dependent activators and repressors to bind to after stimulation of the cells, resulting in both positive and negative correlation of expression with the query gene. In addition, there was a strong bias towards positive correlation of expression between PU.1 and its target genes when considering the combined expression data of all cell types (Fig. S14E). This correlation is caused by PU.1 having high expression in macrophages and cDC cells, a pattern which is also observed for PU.1 target genes. This is similar to our observations for Foxp3 and its targets in Tregs.

One pattern which was not present in our data is regulators with a strong shift towards negative correlation of expression with target genes. This might reflect the absence of regulators with a strong, exclusively repressive function in the Haemcode dataset. In addition, as also noted in the main text, strong negative correlations of expression appear to be in general rare compared to positive correlations.

Several studies have reported binding of TFs to sites which might not have a direct role in transcriptional regulation, or which might be "non-functional" (see (8) for a general review). However, such reports are often based on the analysis of only a few gene expression samples. Our data and tools, on the contrary, allow a more thorough analysis, based on large numbers of samples from the relevant cell type, covering a wide range of conditions. Our results partly confirm the apparent widespread "non-functional" binding: in most of the ChIP-seq datasets a substantial fraction of bound genes lack clear correlation of expression with the ChIPed regulator. At the same time, our results suggest an alternative interpretation for some of these reports: for a subset of regulators that lack correlation in the relevant cell type, we did observe correlation of expression over the combined data of all cell types. This reflects the regulator and its target genes being expressed in the same cell types together. Such regulators might be more relevant in the establishment and maintenance of cell type identity, rather than in regulation of expression following stimulation. Indeed, several of the regulators which showed this tendency are generally regarded as so-called "master regulators" or "pioneer factors" (Fig. S15), and play a key role in the differentiation and establishment of cell type identity. Further over-expression or knock-down using RNAi of such TFs in these cell types might have only little effect on target gene expression once the cell type has been fully developed. Although such binding events might appear to be "non-functional", they obviously are not.

In summary, in combination with ChIP-seq (or similar) data, our correlation GSEA analysis can be useful in interpreting different types of regulatory binding events.

This approach is made available on the Immuno-Navigator website ("correlation GSEA"). In our tool, the user can give as input one query gene *X*, and a set *S* of genes for which to extract the correlation with gene *X*. In addition, a cell type can be specified. The tool subsequently extracts the correlation values between query *X* and all genes in *S* in the expression data of the specific cell type. Graphs are generated visualizing the biases in correlation values between the input and non-input. Enrichment scores and associated p values (based on a Kolmogorov-Smirnov test; not discussed here) are also shown. Resulting output files are made available for download.

Analysis LPS-inducible genes in dendritic cells

Here we describe the application of Correlation Network Hub Prediction (CNHP) on 345 genes with induction of expression 4 hours after lipopolysaccharide (LPS) stimulation in mouse dendritic cells (DCs), which is a relatively well studied system for which several regulators of importance are known.

Fig. S12A shows the genes that are frequently highly correlated with the LPS-inducible input genes in conventional DC (cDC)-derived expression data. Here, only genes with the annotation term “nucleic acid binding transcription factor activity” are shown. Several known regulators of the response to LPS are highly correlated with the input genes, such as STAT and IRF family members, NF-κB subunits (*Nfkb2*, p: 1e-83; *Rel*, p: 1e-73; *Relb*, p:1e-52; *Nfkb1*, p: 1e-19), *Junb* (p: 1e-61), and *Cebpb* (p: 1e-71). The promoter regions of the input genes are strongly enriched for binding sites for several of these transcription factors, further supporting the CNHP result (Fig. S12B). Thus, the frequently correlated genes might reveal potential regulators (not restricted to only transcription factors) of the input genes. For many of the frequently correlated genes in the cDC data, similarly high correlations are found in the data obtained from macrophages and to a lesser extent from plasmacytoid dendritic cells (pDCs) and monocytes (Fig. S12A). This suggests that in these four closely related cell types, similar patterns of expression correlation are present. A similar pattern can be seen in mature B cells as well. Although B cells are part of the adaptive immunity, they are also antigen presenting cells, a function which they share with DCs and macrophages. This might explain a partly shared regulatory network between these cell types. In contrast, relatively unrelated cell types, such as the Megakaryocyte-Erythroid Progenitor (MEP) cells show little similarity.

A final observation is that some genes have a high degree of correlation in the data of many cell types, while the correlation of other genes is restricted to one or a few cell types. For example, *Stat1* and *Irf7* are highly correlated with the input set of LPS-inducible genes in macrophages, cDCs, mature B cells, and pDCs, but also in CD4+ T cells, CD8+ T cells, Tregs, hematopoietic stem cells (HSCs), Pre-B cells, and a number of other cell types. In contrast, correlation of *Batf2* with these LPS-inducible genes appears to be specific to macrophages and cDCs. This suggest that the role of Stat1 as regulator of the response to pathogens might be more general, while that of Baft2 is restricted to a few cell types. In relation with this, we also refer to Figure 4 and the modes of expression correlation that we described above.

Analysis of Foxp3-dependent and -independent genes

In this section, we present two additional analyses that can be easily performed using the data in Immuno-Navigator. As input for this analysis, we use sets of Foxp3-dependent, Foxp3-amplified, and Foxp3-independent genes, as defined in the work by Gavin and colleagues (9). Such sets of genes are typical input sets to analyse using our data. Note that these gene sets are independent from the ChIP-seq based Foxp3-bound target gene set described in the main text of this study.

In a study on the differentiation of Treg cells, Gavin and colleagues uncovered sets of genes with varying dependence on Foxp3. For this, they used gene expression data of CD25+ Foxp3- CD4+ T cells (referred to as “T_{25}”), Foxp3^{null}-expressing T cells (“T_{FN}”, which actively transcribe a non-functional *Foxp3^{null}* allele, yet lack Foxp3 protein), regulatory T cells (referred to as “T_R”), ad naïve T cells (“T_N”)

in thymus and in peripheral lymphoid organs. They used hierarchical clustering and manual curation to define 16 sets of genes (see Fig. 3 and Supplementary Fig. 5 in the paper by Gavin *et al.*).

For the sake of brevity, we limit the discussion here to the peripheral gene clusters containing Foxp3-dependent (cluster P3), Foxp3-amplified (cluster P4), and Foxp3-independent (cluster P7) genes. Although the clusters reported by Gavin *et al.* contain both genes with induced and repressed expression in presence in Treg cells (compared to naïve T cells), here we focussed only on the induced genes within each cluster. Set P3 contained 124 genes, set P4 72 genes, and set P7 63 genes.

In the first analysis, we used correlation GSEA to investigate the correlation of expression of these three sets of genes with the expression of *Foxp3* within Treg-derived samples. In the secondly analysis, we used CNHP to find genes that are highly correlated in Treg cells with each gene set.

Correlation of expression with Foxp3 *in Treg cells*

Using the correlation GSEA function of the Immuno-Navigator database, we obtained the correlation of expression data for *Foxp3* versus all genes in the mouse genome in Treg-derived samples. Using this data, we evaluated whether genes in clusters P3, P4, and P7 tend to have correlated expression with Foxp3 or not. Results are summarized in Fig. S16A.

We observed that, as expected, genes in clusters P3 and P4 tend to be positively correlated with *Foxp3* expression in Tregs (Fig. S16A). Intuitively more surprising is the observation that the Foxp3-independent genes in P7 too tend to have positive correlation with *Foxp3* expression. However, P7 might include genes whose induction during Treg cell differentiation precedes, or regulates, that of *Foxp3*. Alternatively, P7 might also include genes which are Foxp3-independent yet are regulated by the same mechanism that controls *Foxp3* induction. Both cases can explain the tendency towards positive correlation of expression. A third possibility is that the classification in Gavin *et al*. was not completely accurate, and P7 includes a considerable amount of Foxp3-dependent genes. Since the classification is based on only a small number of samples, we can not rule out this last alternative.

In combination with the results presented by Gavin *et al*., the above observations support the key role of Foxp3 in Treg cells. A relatively large number of genes were shown to be Foxp3-dependent or Foxp3-amplified by Gavin *et al*. Here we showed that these genes indeed have correlation of expression with Foxp3, in a collection of 240 samples obtained from Treg cells. However, on the other hand, correlation values between *Foxp3* and these genes are in general relatively low (typically PCC values < 0.4). In addition, as described in the main text, we observed that the expression of genes that are bound by Foxp3 in Tregs is not necessarily correlated with *Foxp3* expression. Together with the weak correlation observed even between Foxp3-dependent genes and *Foxp3*, these results support the existence of additional regulatory mechanisms that are independent of, or supplementary to, Foxp3-mediated regulation.

Correlation Network Hub Prediction of the gene clusters

We used the above three gene sets as input for our CNHP function. Fig. S16B-D shows the 10 top-scoring genes for each set. Below, we present and discuss some of the observations we could make.

In general, as in the results presented in Fig. 6 of the main text, high-scoring genes typically contained several known genes of importance, in addition to several genes with no known function in Treg cells. These genes might present valuable candidates for further analysis.

For the Foxp3-dependent genes (cluster P3), high-scoring genes include known genes of importance (*Icos*; rank 17, and *Nrp1*; rank 22), as well as genes which were also high-scoring in the analysis of Treg-specific genes (see Fig. 6; *Fam129a*, *Tiam1*, *Lclat1*, etc). *Il1rl1* (rank 10; encodes the Il33 receptor ST2) has recently been reported to be especially induced in effector Treg cells and in colonic Treg cells, and to be essential for the development and maintenance of Treg cells in visceral adipose tissue (10, 11).

For the Foxp3-amplified genes (cluster P4), high-scoring genes include several of the known genes of importance in Treg cells, including *Il2ra* (*Cd25*), *Ctla4*, *Tnfrsf4* (*Ox40*), *Irf4* (rank 17), *Prnp* (rank 21), *Cd83* (rank 22), *Dusp4* (rank 28), *Icos* (rank 37), *Socs2* (rank 41), and *Ikzf4* (rank 44).

Here too, *Tiam1* (rank 7) is found to have correlated expression with many of the P4 genes. As mentioned in the main text, Tiam1 has been shown to be important in the activation of LFA-1 through TCR-signaling. Vav2 (rank 8), too, is known to play a role in TCR-signaling (12, 13).

While high-scoring genes in the P3 and P7 cluster show correlation only in Treg-derived samples (Fig. S16B,D), in contrast, for P4 there is correlation in Treg-derived as well as in CD4 T cell-derived data (Fig. S16C). Since genes in the cluster P4 are Foxp3-amplified (see Gavin *et al*.), it might suggest that the differential expression of these genes is already partly established even in absence of Foxp3, and thus perhaps shared with Foxp3- CD4+ T cells.

In the Foxp3-independent genes (cluster P7), the top scoring gene is *Tiam1*, which was highly scoring also in P3 and P4. Again, we observe a certain overlap between high-scoring genes of other clusters, and for the Treg-specific gene set described in the main paper. Igf1r (rank 4), like Itgb8 (rank 8), plays a role in focal adhesion, and several integrins have been shown to directly interact with Igf1r.

General data analysis approach

Fig. S22 shows a summary of the main steps in the processing and treatment of data before the populating of the Immuno-Navigator database. Input data consists of biological data (here: gene expression data), supplemented with prior knowledge of the biological system and experimental platform(s) of interest. Publicly available biological samples are processed (removal of duplicated samples, etc) and normalized in a standard way. Sample annotations include biological and experimental variables which at this stage will be used for assessing data quality and for batch effect reduction, in addition to their ultimate use in the cell type-specific analysis of the gene expression data. Using prior knowledge and reasonable assumptions, a number of indicators of data quality are defined. The indicators used in our study are described in SI Appendix, section “Evaluation of batch effect reduction”, and include general as well as system- (consistency with hematopoietic lineage tree) and platform-specific (correlation between probe pairs representing the same gene) measures. Importantly, these measures are completely independent of batch annotations or batch effect reduction methods, and are defined over the entire dataset (e.g. based on the genome-wide data, not just a small subset of genes). Using these indicators, sample annotations, and quality indicators, an exploratory analysis of batch effects in the data is conducted, and its quality is assessed.

In a next step, batch effects in the data are reduced, guided by the provided biological (here: cell type) and experimental (here: studies as proxy for batches) variables for each sample. The quality of the obtained batch-treated data is again assessed, and compared with the original data’s quality. As described in this paper, for this study we found a general improvement of the gene expression correlation data after batch treatment. If quality is judged not to be sufficient or shows additional room for improvement, additional processing might be undertaken. Obviously, this step should not involve “tuning” of the data to the quality indicators. If batch effects appear to be weak, there might be cases in which the untreated data is sufficient for analysis.

Finally, the obtained data is processed for populating the database (or further downstream analysis). In the present study, this involved, among other, processing of probe-to-gene annotations and cell type-specific expression data for populating a SQLite database, as well as pre-computing lists of top correlated genes, enriched GO annotations, etc. The final data is made accessible using a three-tier database architecture, in which various tools and supporting data are integrated into a user-friendly interface.

This general workflow is applicable to various omics data, biological systems, or species, provided batch information is available.

Assessment of the presence of batch effects

Principal Component Analysis

There exist a number of exploratory methods for assessing the presence of batch effects in biological data (14). One exploratory analysis is principal component analysis (PCA) followed by the plotting of samples marked by batch identifiers or by cell type. We performed PCA on a random selection of 3000 probes over all 3,434 samples, for the data before and after treatment of batch effects. Figure 1A in the main text shows all samples of the untreated data plotted according to the first 2 principal components (PC), with color codes indicating cell types. Fig. S3 shows similar plots for PC1-PC3 (Fig. S3A), and PC2-PC3 (Fig. S3C). As also described in the main text, the association between PCs and biological variables is not so clear, especially for PC2. PC1 appears to be associated with cell types of the myeloid lineage, and PC3 appears to separate to some degree progenitor cells from non-progenitor cells. PC1, PC2, and PC3 explain 19.0%, 10.8% and 7.4% of variance in the untreated data, respectively.

For the batch-treated data, we refer to Fig. 1B in the main text (for PC1-PC2), to Fig. S3B (for PC1-PC3), and to Fig. S3D (PC2-PC3). PC1, PC2, and PC3 explain 34.0%, 14.1% and 8.0% of variance in the data after batch effect reduction, respectively. As described in the main text, the PC1 divides cell types of the myeloid lineage (negative values), of the lymphoid lineage (positive values), and progenitor cells (intermediate values). PC2 is roughly associated with the degree of maturation of cells, with progenitor cells (such as hematopoietic stem cells, common lymphoid progenitors, common myeloid progenitors, megakaryocyte-erythroid progenitors) having high values, and more specialized cell types having lower values. PC3 of the separates especially mature B cells (large negative values) and to a lesser degree also Pre-B cells from the other cell types.

Hierarchical clustering of samples

A second exploratory analysis is to cluster samples according to their similarity, and label them with their “batch” identifiers (in this case the study by which the samples were published), and by biological variables (in this case: cell types). In the ideal case (when no batch effects are present), clustering of samples should result in samples for the same cell type forming clusters. In the presence of significant batch effects, however, samples produced by the same study form clusters. Note however that the situation is complicated by the fact that in our case batches and biological variables are heavily confounded (e.g. many studies focus on one particular cell type).

Because of the high number of studies (261) used in this analysis, it is hard to give a comprehensive overview of the batch effects over the entire dataset. Here we therefore briefly focus on the data for regulatory T cells (Tregs) only. Figures S4A and S4B show the clustering of samples before and after batch effect reduction, respectively. In the original data, samples are clearly clustered according to the batch (study) in which they were published. After batch treatment, samples from different studies are distributed much more evenly over the dendrogram, although some clustering by study still remains. Similar observations were made for other cell types.

Evaluation of batch effect reduction

As discussed in the main text, we found in general a tendency for probe pairs with high correlation in the raw data to be also correlated in the batch-processed data. On the level of correlated gene pairs too, we observed a significantly high overlap between the untreated and batch-treated data (Table S2). For example, in the macrophage (MΦ) data, the untreated and treated expression data contain each significantly positively 2,575,478 correlated gene pairs, of which 240,041 are shared. Although this represents only 9.3% of the gene pairs of the treated data, this is about 9.1 times more than the overlap one would expect at random (26,448 pairs, 1.0%). In all cell types, similarly high overlap was observed. Nevertheless, the shared number of correlate gene pairs was typically just 10 to 25% (range 5.5 to 32.6%; mean: 18.0%) of the total correlated gene pairs in each data set. In other words, although there is a significant overlap, there is also a considerable discrepancy between the untreated and batch-treated correlated gene pairs.

Here we present a number of results that indirectly indicate that batch effect reduction improved the estimated gene expression correlation. In brief, we show that after batch effect reduction:

1. Correlation values were more consistent between cell types.
2. Related to the above, the number of gene pairs that were found to be significantly correlated in the data of multiple cell types was increased.
3. Clustering of cell types according to similarity of their correlation data resulted in a clustering that was more consistent with the known hematopoietic lineage tree.
4. Genes with shared functional annotations were more frequently found to be highly correlated.
5. Probe pairs representing the same gene were more frequently highly correlated, compared to probe pairs representing different genes.

Below, each of these results is described in more detail.

Overlap in significantly correlated gene pairs between cell types

Although each of the cell types included in our dataset has its own distinct features and role in the immune system, it is reasonable to assume that many biological pathways are shared between them. Under the assumption that correlation of gene expression reflects (directly and indirectly) biological pathways, we would therefore expect to observe a considerable amount of overlap in significantly correlated gene pairs between different cell types. We found that treatment of batch effects resulted in an increase in consistency between cell types: Firstly, gene pairs that have highly correlated expression in multiple cell types increased in number after batch-treatment (Table S4). On the other hand, gene pairs that were found to be correlated in the data of only a single cell type decreased in number. Note that PCC thresholds were set in such a way that the number of significantly correlation gene pairs would be the same in the untreated and treated data, and that the above observations can thus not be explained simply by a change in the number of significantly correlated gene pairs. Secondly, batch effect reduction increased the overlap in correlated gene pairs between pairs of cell types in 183 (72%) out of the 253 cell type combinations (Fig. S6). These results are likely to reflect a reduction of spurious correlations observed in only one cell type, caused by batch effects.

Clustering of cell types by similarity of co-expression between pairs of probes

Blood cells differentiate from hematopoietic stem cells through a number of progenitor states into cells of the lymphoid and myeloid lineage. It is reasonable to assume that neighboring cell types in this lineage tree are defined by more similar biological pathways than distal ones. When we performed hierarchical clustering of cell types by their similarity of PCC values (see Materials and Methods section), the clustering is improved in the batch-treated data (Fig. S7B) compared with the raw data (Fig. S7A); roughly, 3 big clusters are formed, dominated by progenitor cells, by lymphoid cell types, and by myeloid cell types, respectively. On the other hand, for the untreated data, the clustering of cell types fits less well with the known lineage tree.

Correlation between gene pairs with shared functional annotations

Under the assumption that that genes with shared functions are expected to have correlated expression more often than gene pairs with unrelated functions, we compared the correlation between genes with shared Gene Ontology (GO) terms with the correlation between unrelated genes. Similar approaches have been proposed, such as a “GO score”, comparing genes with shared GO annotations with genes lacking shared annotations (15).

We mapped all child annotations in the GO annotation to each of their parent nodes. Next, we made a selection of GO annotations that contained between 4 and 20 associated mouse genes. Gene pairs associated with each of these GO annotations we regarded as being functionally related.

On the other hand, we made a set of functionally unrelated gene pairs as follows: we selected all GO annotation terms with at most 500 associated genes. Randomly, a large amount of gene pairs were selected, rejecting any gene pairs associated with a shared GO term.

PCC values were calculated for all functionally related gene pairs, and all functionally unrelated gene pairs. Finally, we calculated the fraction of functionally related gene pairs having a PCC higher than 99% of the PCCs of the functionally unrelated gene pairs.

Results are summarized in Fig. S8, and show that after reduction of batch effects, the correlation of expression between functionally related gene pairs is increased relative to that between functionally unrelated gene pairs. We found an improvement in the batch-treated data in 22, 22, and 18 out of 24 cell types, for the Biological Process, Molecular Function, and Cellular Component GO annotations, respectively.

Correlation of same-gene probe sets

As a final measure, we used the correlation between pairs of probes representing the same gene. A considerable portion of mouse genes are represented by more than one probe set on the Affymetrix GeneChip Mouse Genome 430 2.0 platform. Although the interpretation of such probe sets mapping to the same gene is not always straightforward (16), it is reasonable to assume that, on average, probe sets representing the same gene should have a higher tendency to be positively correlated than probe sets representing different genes.

Comparing the distribution of PCC values of all same-gene probe pairs (35,164 probe pairs, representing 10,556 genes that have multiple probes) with that of different-gene probes (for probes representing 35,164 randomly selected gene pairs), we found that for all cell types there was a relative increase of correlation between same-gene probes in the batch-treated data (Fig. S9). It should be stressed that batch effect reduction was performed on the probe intensity data, completely independent of probe-to-gene mapping data.

As an example, Figure S10A shows the distribution of PCC values over all untreated samples obtained from macrophages, for both randomly selected pairs of probe representing different pairs of genes (black), and for all pairs of probes representing the same gene (red lines). Clearly, even in the untreated data, same-gene probe pairs tend to be more positively correlated, in general. Figure S10B shows the same histogram for the batch-treated macrophage data. Compared with the untreated data, the variance in PCC values for randomly selected pairs of probes has strongly decreased. For the same-gene probes too, the variance has decreased, but a subset of probe pairs continue to show high positive correlation. As a result, batch reduction results in a relative increase in correlation between same-gene probes as compared to probes representing different genes. We evaluated this relative increase using Receiver operating characteristic (ROC) curves. The ROC curves for the macrophages data, before (black) and after (red) batch treatment (Fig. S10C) show an increased distinction between same-gene probe PCCs and the PCCs of randomly selected probes. Similar improvements were seen in the data for all cell types (Fig. S9).

SI Materials and Methods

RNA-seq analysis of CD25^{pos} and CD25^{neg} T cells

High-throughput sequencing of RNA (RNA-seq) was conducted for CD25^{pos} T cells, unstimulated CD25^{neg} T cells, PMA-stimulated CD25^{neg} T cells, and anti-CD3-stimulated CD25^{neg} T cells. C57BL/6 mice (Female, from 5-6weeks) were purchased from CLEA Japan. CD4+ T cells were isolated from splenic and lymph nodes as previously described (1). CD8-B220-CD16/32-NK1.1-CD4+CD25+ T cells (Treg cells) and CD8-B220-CD16/32-NK1.1-CD4+CD25−CD44low T cells (Tconv cells) were purified by sorting with a cell sorter (MoFlo; Beckman Coulter). For in vitro TCR stimulation of cells, plates coated with anti-CD3 (1 µg/mL) and anti-CD28 (1 µg/mL) for 6 h or phorbol 12-myristate 13-acetate (20 ng/mL) and ionomycin (1 µM) for 2 h with recombinant IL-2 for Treg or without recombinant IL-2 for Tconv were used. Anti-Il2ra (PC61), anti-CD4 (RM4.5), anti-CD44 (IM7), anti-CD8a (53-6.7), anti-B220 (RA3-6B2), anti-CD16/32 (2.4G2), and anti-NK1.1 (PK136) were obtained from BD PharMingen, Biolegend, or eBioscience. Anti-CD3 (2C11) and anti-CD28 (37.51) were used for in vitro T-cell stimulation. Mouse recombinant IL-2 was a gift from Shionogi Co. Total RNAs were extracted from sorted cells using Trizol (Qiagen), and were subjected to TruSeq library prep kit (Illumina), and read by Hiseq2000 (Illumina). Sequencing data have been deposited in the DNA Data Bank of Japan (DDBJ) under accession number DRA004105. Obtained sequences were mapped to the mouse genome (mm9) by tophat2 (17).

Treg-specific genes were defined as follows. The number of reads aligned to each gene was counted, and normalized using DESeq (18). Genes with high preferential expression in Tregs defined as genes with a sufficiently high tag count in the CD25^{pos} sample (higher than the median non-zero tag count, 308.7), and the tag count in the CD25^{pos} sample should be at least 2-fold higher than that in any of the CD25^{neg} samples. From this set, genes induced (>2-fold enrichment) upon stimulation of CD25^{neg} cells by PMA or anti-CD3 were removed. This resulted in a set of 248 genes (Refseq IDs).

ChIP-seq analysis in Tregs and DCs

For the analysis of Foxp3 binding in Tregs, ChIP-seq data for Foxp3 binding in Treg cells was obtained from DDBJ accession number DRA003955. Obtained sequences were mapped to the mouse genome (mm9) by bowtie2 (19), and peak-called by FindPeaks (20). Peaks with at least a 7-fold stronger signal in the ChIP sample than in the input sample were retained, and from those the 25% with the highest score were selected. Finally, 1,300 genes were associated with at least one of these peaks (region -100kb to +100kb around transcription start site). Results were consistent when other thresholds were used.

For Elf1 and Ets1 (21), and Foxo1 (22) binding in Tregs, ChIP-seq reads were obtained from NCBI Genome Expression Omnibus (GEO), access numbers GSE40684 and GSE40657. Mapping, peak calling, and selection of target genes were performed as described above. For Elf1, Ets1, and Foxo1, 2252 1278, and 2868 bound genes were obtained, respectively.

For PU.1, C/EBPβ, Nfkb1, and Stat1 binding in DCs before and after stimulation with LPS, ChIP-seq data was obtained from GEO accession number GSE36104. Peak scores were used as reported in the original study (7), with scores above 26.9 regarded as significant. For each transcription factor, target genes were defined as genes with significant peaks in the region -5kb to +5kb around their transcription start site in at least one of the ChIP-seq samples for the transcription factor. Thus, 8758, 7143, 500, and 618 bound genes were obtained for PU.1, C/EBPβ, Nfkb1, and Stat1, respectively.

Haemcode ChIP-seq analysis

We obtained target genes for a collection of 104 TFs and DNA-binding proteins from the Haemcode database (6). Haemcode contains, among others, ChIP-seq peak data obtained from publicly deposited ChIP-seq data, processed using a consistent analysis pipeline. Haemcode also provides annotation files in which ChIP peaks are assigned to candidate target genes, based on the overlap between peaks and genes or the distance between them. We collected target genes for the 104 TFs and DNA-binding proteins for which data was available for cell types present in our database. This covered 61 different factors, and 14 different cell types. Table S5 shows an overview of the data.

For each factor, cell type, and study, we defined target genes using the Haemcode annotation data, as follows: target genes should overlap with a ChIP-seq region peak, or the distance between the gene and the peak should be at most 100 kbs. Genes meeting this condition were regarded as targets, and other genes as non-targets.

From our database, we collected the expression correlation data for the gene encoding the ChIPed regulator in the cell type used for the ChIP-seq experiment. As measure for the bias in correlation between the CHIPed regulator and its targets we calculated “enrichment scores”, similar to those used in Gene Set Enrichment Analysis (GSEA) (5). In brief, genes are sorted by their correlation with the ChIPed regulator. Enrichment scores are subsequently calculated by going through the sorted list of genes, increasing a running-sum score whenever a target gene is encountered, and decreasing it when a non-target gene is encountered. The maximum and minimum of this running-sum score are used as a measure for the bias in correlation between the ChIPed regulator and its targets as compared to the non-targets. For a more detailed description about GSEA and enrichment scores, we refer to (5).

Enrichment scores were analysed for all 104 datasets (see SI Appendix, section “Correlation Gene Set Enrichment Analysis”).

FACS sorting and CpG methylation analysis of Foxp3$^+$ Itgb8$^+$ T cells

FITC-conjugated anti-CD45RA (HI100) mAb and V500-conjugated anti-CD4 (RPA-T4) mAb were purchased from BD Biosciences. PE-conjugated anti-FoxP3 (236A/E7) mAb and purified anti-Itgb8 (416922) were purchased from eBioscience and R&D Systems respectively. Anti-Itgb8 mAb was biotinylated by Biotin Labeling Kit - NH2 (Dojindo).

Human CD4+ T cells were enriched from PBMCs of healthy donors by using BD IMag system. Enriched Th cells were stained with anti-Itgb8 mAb for 30 min on ice. After washing, cells were incubated with streptavidin-labeled APC (BD Biosciences) and other antibodies for 30 min. FoxP3 staining was performed after fixation by Foxp3 / Transcription Factor Staining Buffer Set (eBioscience) and FoxP3+Itgb8+ cells were sorted by FACSAriaII.

All donors provided written informed consent before sampling according to the Declaration of Helsinki. The present study was approved by the institutional ethics committees of Osaka University.

Methods and primers for CpG methylation analysis were previously described (23). Briefly, genomic DNA was subjected to bisulfite treatment using MethylEasy Xceed (Human Genetic Signatures), followed by PCR amplification of target regions and subcloning into pTAC-1 plasmid in DynaExpress TA PCR Cloning Kit (BioDynamics Laboratory Inc). 16 colonies per region were amplified with the Illustra TempliPhi Amplification Kit (GE Healthcare) and sequenced.

LPS-induced genes in mouse DCs

A set of genes induced in GM-CSF-induced bone marrow-derived DCs by LPS was defined as follows. RNA-seq data for mouse DCs before and 4 hours after LPS stimulation was obtained from the DDBJ Sequence Read Archive; accession number DRA001131 (24). Reads were mapped to the mouse genome (mm10) using Tophat and Bowtie (19, 25). Uniquely mapped tags with at most 2 mismatches were counted per mouse Refseq gene, and converted to reads per million reads per kilobase (RPKM). Genes with at least a 5-fold induction of expression 4 hours after stimulation and at least one sample with an RPKM value higher than the genome-wide median RPKM were defined to be significantly induced. This resulted in a set of 449 Refseq IDs, representing 345 unique genes.

Construction of Shuffled Data

It is not hard to see that batch effects increase absolute values of PCCs. Fig. S18 shows a toy example of two probes which are not correlated in reality, measured in 2 batched of 10 samples each (Fig. S18A). Adding batch effects to the batches (in this case a simple shift; increasing values for both probes in batch 1 and decreasing them in batch 2) results in an apparent strong (in this case positive) correlation between the two probes (PCC for Fig. S18A: 0.079; for Fig. S18B: 0.854). It is therefore not surprising to observe that absolute PCC values are decreased after treatment of batch effect (see also Fig. 2A in main text, and Fig. S5).

For our analysis, we defined “significant” correlations using a false discovery rate (FDR) measure, comparing correlation values obtained from the actual data, with those obtained from artificial data. However, it is reasonable that some batch effect remains even in the batch-treated data. Because of this, the PCC values in this batch-treated data are still likely to be to some degree biased towards extreme values. To prevent this from inflating the number of significantly correlated probe pairs, we constructed our artificial data in a way that would preserve batch effects that might still be present in the batch-treated data. We did this using “batch-guided” shuffling, as follows:

1. For each cell type, we collected all batch-treated samples, along with their corresponding batch index (identifier for the study they were published by).
2. For each probe, values were randomly shuffled one batch at the time:
    a. if a batch contained at least 5 samples for the cell type of interest: values for the probe were shuffled within the batch.
    b. if the batch contained less than 5 samples for the cell type of interest: a corresponding number of values were randomly sampled from all values of this probe over all batches.

This approach assures that if some particular batch contains extreme values, these extreme values will also be preserved in the artificial data. Note that in the absence of batch effects this approach is essential identical to random shuffling of data per probe, regardless of batch indices.

Using the same toy example as above, we illustrate the effect of this “batch-guided” shuffling. Fig. S18C shows the scatter plot of the samples in the toy example with batch effects (see Fig. S18B) after default random shuffling regardless of batch indices. Fig. S18D shows the scatter plot of the same samples after batch-guided shuffling. The PCCs are 0.152 and 0.810, respectively. In the latter case, batch effects have been largely preserved, leading to PCCs with higher absolute values, similar to those observed in the original batch-affected data (Fig. S18B).

Application of Correlation Network Hub Prediction on Random Sets of Genes

In order to evaluate the behavior of CNHP further, we ran it on sets of randomly selected genes of different sizes. We randomly selected sets of genes of size 10, 20, 30, 40, 50, 100, 150, 200, 300, 400, and 500 and used these as input for 23 cell types (not for multipotent progenitor cells, for which we could not find a suitable FDR-based PCC threshold; see main text) and for the combined data. We repeated this 10 times (2640 runs in total), and for each input size we recorded the minimum p value for the gene with the most significant frequency of correlation with the input set of genes. Fig. S19 shows for each input size and dataset the lowest p value (-$\log_{10}$ values) we observed (see also main text for the estimation of the p values). Overall the lowest p value observed was 3.9e-8, suggesting that in general a p value threshold of 1e-10 is a reasonable choice for genuine analyses. There were no clear tendencies in minimum p values with regard to the size of the input gene set or the number of samples in the dataset.

Robustness of Results of Correlation Network Hub Analysis

In order to evaluate the robustness of the CNHP results, we performed CNHP on the set of genes that are induced in mouse dendritic cells (DCs) after lipopolysaccharide (LPS) stimulation (see main text). We considered all genes with a high level of correlation with the input genes (p value < 1e-10) as a reference set for "highly correlated genes", and all genes with a low level of correlation with the input genes (p value > 0.01) as a reference set for "non-correlated genes". We then ran the same analysis on smaller input sets by removing randomly parts of the original input gene set (removing 5%, 10%, …, 95% of the genes; 20 runs each). For each run, we checked how many of the reference "highly correlated genes" and "non-correlated genes" were highly correlated (p value < 1e-10) with the genes in the smaller input gene sets.

Figure S20 shows box plots of the fraction of "highly correlated genes" that were retained for each input set size. The plot shows that the retention rate first drops slowly when only 5 to about 50% of the original input set is randomly removed. The retention rate then drops more rapidly as the input gene sets become smaller. Importantly, more than half of the "highly correlated genes" could be retained even when 75% of the original input was removed. On the other hand, not a single "non-correlated" gene was reported as "highly correlated" in any of the runs we performed.

In another analysis, we gradually added randomly selected (non-input) genes to the original input. We did this in several steps (5%, 10%, …, 100% of the original input set size; 20 runs for each level). For each run, we checked how many of the reference "highly correlated genes" and "non-correlated genes" were highly correlated (p value < 1e-10) with the genes in the noisy input sets.

Figure S21 shows box plots of the fraction of "highly correlated genes" that were retained for each level of randomly selected genes added. Retention rates drop linearly in function of the amount of noise added. However, even when 100% noise is added, retention rates are still above 85%. This suggest that CNHP is rather robust against noise, as long as a set of biologically meaningful genes is included in the input set. Here too, not a single "non-correlated" gene was reported as "highly correlated" in any of the runs we performed.

Transcription factor binding sites

From the Jaspar database (26) we prepared a set of 543 PWMs, including the DNA binding motifs for mammalian transcription factors, as well as core promoter motifs. For each PWM, a threshold score was set in a way that results in about one predicted binding site per 5kb of the mouse genome (mm10). For each Refseq gene, the region -500 to +200 relative to its transcription start site was scanned using the set of PWMs and their corresponding threshold scores. Predicted transcription factor binding sites (TFBSs) for each gene can be downloaded from Immuno-Navigator.

In addition, for each gene, the TFBSs that are enriched in the promoters of the top 100 most highly correlated genes in each cell type have been pre-calculated and are available in the database. Our approach for TFBS enrichment prediction takes into account GC content biases in the promoter regions (27). For a more detailed description we refer to the section "TFBS enrichment analysis" below.

TFBS enrichment analysis

Vertebrate promoters can be roughly classified into CpG island-associated promoters and non-CpG island promoters (28, 29). Obviously, the presence or absence of predicted transcription factor binding sites (TFBSs) in promoter sequences is affected by the overall GC content and CpG scores of a DNA sequence. A number of studies have reported better performance in the prediction of enrichment of TFBSs when taking into account the GC content and CpG scores of the input sequences (27, 30).

We extracted the genomic sequences from position -500 to +200 relative to all Refseq transcription start sites (obtained using the UCSC Table Browser) (31). Next, we divided these sequences into 7 bins of 100 bps, and in each bin calculated the GC content and the CpG score. These values were combined into a single matrix, on which we applied PCA. Using the two first principal components we classified promoters into two clusters using k-means clustering (k=2). These two clusters correspond to 17,847 promoters with high GC content and high CpG scores, and 15,225 promoters with low GC content and low CpG scores.

We used the above classification in order to reduce biases in the TFBS enrichment analysis caused by GC content and CpG scores of promoter sequences under investigation. For each position weight matrix (PWM) *p*, we calculated the fraction of sequences containing a hit for *p* among the high GC content promoters ($fr_{p,high}$) and the low GC content promoters ($fr_{p,low}$). For each set of promoter DNA sequences *D* in which to predict enriched TFBSs, we count the number of sequences that contain a hit for *p* ($h_{p,D}$). We also look up the number of sequences in *D* that were classified in the high GC content class (n_{high}) and in the low GC content class (n_{low}), respectively. Finally, using the binomial distribution, we calculated the probability of observing $h_{p,D}$ or more hits for *p* in a set of n_{high} high GC content and n_{low} low GC content sequences, given $fr_{p,high}$ and $fr_{p,low}$. This probability was corrected for multiple testing using the Bonferroni correction, and PWMs with a corrected p value < 0.01 were considered as significantly enriched in the input set *D*.

GO annotations and GO term enrichment.

For each gene, associated GO terms are included in Immuno-Navigator. In addition, enriched GO terms in the top 100 most correlated genes in each cell type are available. GO term enrichment was estimated using a hypergeometric distribution and Bonferroni correction for multiple testing. GOslim annotations can be used for filtering results of CNHP.

SI References

1. Hori S, Nomura T, Sakaguchi S (2003) Control of Regulatory T Cell Development by the Transcription Factor Foxp3. *Science (80-)* 299(5609):1057–1061.

2. Josefowicz SZ, Lu L-F, Rudensky AY (2012) Regulatory T cells: mechanisms of differentiation and function. *Annu Rev Immunol* 30:531–64.

3. Pallandre J-R, et al. (2007) Role of STAT3 in CD4+CD25+FOXP3+ Regulatory Lymphocyte Generation: Implications in Graft-versus-Host Disease and Antitumor Immunity. *J Immunol*

179(11):7593–7604.

4. Shannon P, et al. (2003) Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* (13):2498–2504.

5. Subramanian A, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545–50.

6. Sánchez-Castillo M, et al. (2015) CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res* 43(Database issue):D1117–23.

7. Garber M, et al. (2012) A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals. *Mol Cell* 47(5):810–22.

8. MacQuarrie KL, Fong AP, Morse RH, Tapscott SJ (2011) Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet* 27(4):141–148.

9. Gavin M a, et al. (2007) Foxp3-dependent programme of regulatory T-cell differentiation. *Nature* 445(7129):771–5.

10. Vasanthakumar A, et al. (2015) The transcriptional regulators IRF4, BATF and IL-33 orchestrate development and maintenance of adipose tissue-resident regulatory T cells. *Nat Immunol* 16(3):276–85.

11. Schiering C, et al. (2014) The alarmin IL-33 promotes regulatory T-cell function in the intestine. *Nature* 513(7519):564–8.

12. Tartare-Deckert S, et al. (2001) Vav2 Activates c-fos Serum Response Element and CD69 Expression but Negatively Regulates Nuclear Factor of Activated T Cells and Interleukin-2 Gene Activation in T Lymphocyte. *J Biol Chem* 276(24):20849–20857.

13. Zakaria S (2004) Differential Regulation of TCR-mediated Gene Transcription by Vav Family Members. *J Exp Med* 199(3):429–434.

14. Leek JT, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11(10):733–9.

15. Okamura Y, et al. (2014) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res* 43(Database issue):D82–6.

16. Stalteri M a, Harrison AP (2007) Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics* 8(13).

17. Kim D, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36.

18. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.

19. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–9.

20. Fejes AP, et al. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively

parallel short-read sequencing technology. *Bioinformatics* 24(15):1729–30.

21. Samstein RM, et al. (2012) Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* 151(1):153–66.

22. Ouyang W, et al. (2012) Novel Foxo1-dependent transcriptional programs control T(reg) cell function. *Nature* 491(7425):554–9.

23. Ohkura N, et al. (2012) T Cell Receptor Stimulation-Induced Epigenetic Changes and Foxp3 Expression Are Independent and Complementary Events Required for Treg Cell Development. *Immunity*.

24. Patil A, Kumagai Y, Liang K-C, Suzuki Y, Nakai K (2013) Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response. *PLoS Comput Biol* 9(11):e1003323.

25. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–11.

26. Mathelier A, et al. (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42(Database issue):D142–7.

27. Vandenbon A, et al. (2013) A Parzen window-based approach for the detection of locally enriched transcription factor binding sites. *BMC Bioinformatics* 14(1):26.

28. Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* 13(4):233–45.

29. Illingworth RS, Bird AP (2009) CpG islands--’a rough guide'. *FEBS Lett* 583(11):1713–20.

30. Roider HG, Lenhard B, Kanhere A, Haas S a, Vingron M (2009) CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses. *Nucleic Acids Res* 37(19):6305–15.

31. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493–6.

SI Tables

	Cell type	Abbreviation	sample count
1	CD4 T cells	CD4	634
2	Macrophages	MΦ	601
3	Mature B cells	-	384
4	CD8 T cells	CD8	323
5	Regulatory T cells	Treg	240
6	hematopoietic stem cells	HSC	236
7	conventional dendritic cells	cDC	216
8	Pre-B cells	-	94
9	Granulocyte-macrophage progenitors	GMP	82
10	Common myeloid progenitors	CMP	74
11	Mature NK cells	-	68
12	Double Positive cells	DP	65
13	Mast cells	-	61
14	Type 1 helper T cells	Th1	51
15	memory T cells	Tmem	47
16	Monocytes	-	38
17	Common lymphoid progenitors	CLP	36
18	Type 2 helper T cells	Th2	35
19	Plasmacytoid dendritic cells	pDC	28
20	Pro-B cells	-	28
21	Megakaryocyte-erythroid progenitors	MEP	27
22	Natural killer T cells	NKT	24
23	Common dendritic cell progenitors	CDP	22
24	Multipotent progenitor cells	MPP	20
	Total		**3,434**

Table S1: Final count of samples per cell type included in this study and in the database.

Data set or cell type	Gene pairs before treatment	Gene pairs after treatment	Gene pairs shared (observed)	Gene pairs shared (expected)	Fold enrichment	Percentage shared (%)
Combined	1,920,557	1,920,557	345,182	14,704	23.5	18.0
CD4	2,009,223	2,009,223	430,558	16,093	26.8	21.4
MΦ	2,575,478	2,575,478	240,041	26,443	9.1	9.3
Mature B	2,934,765	2,934,765	700,198	34,335	20.4	23.9
CD8	3,071,390	3,071,390	413,689	37,606	11.0	13.5
Treg	3,395,245	3,395,245	1,078,507	45,955	23.5	31.8
HSC	1,269,071	1,269,071	208,280	6,420	32.4	16.4
cDC	2,748,539	2,748,539	278,675	30,116	9.3	10.1
Pre-B	2,450,745	2,450,745	742,093	23,944	31.0	30.3
GMP	3,374,613	3,374,613	568,447	45,398	12.5	16.8
CMP	3,349,175	3,349,175	1,046,023	44,717	23.4	31.2
Mature NK	2,678,781	2,678,781	903,851	28,607	31.6	33.7
DP	3,031,810	3,031,810	760,249	36,643	20.7	25.1
Mast	3,216,178	3,216,178	422,503	41,236	10.2	13.1
Th1	2,260,888	2,260,888	332,115	20,377	16.3	14.7
Tmem	2,597,571	2,597,571	240,206	26,898	8.9	9.2
monocyte	2,715,329	2,715,329	585,837	29,393	19.9	21.6
CLP	1,886,671	1,886,671	204,328	14,190	14.4	10.8
Th2	574,163	574,163	51,349	1,314	39.1	8.9
pDC	2,259,353	2,259,353	254,739	20,350	12.5	11.3
Pro-B	552,209	552,209	30,314	1,216	24.9	5.5
MEP	575,317	575,317	50,042	1,319	37.9	8.7
NKT	679,041	679,041	101,665	1,838	55.3	15.0
CDP	806,882	806,882	263,131	2,595	101.4	32.6

Table S2: Table showing the overlap between significantly correlated gene pairs in the untreated and batch-treated gene expression data. For the combined data and for each cell type's data, the number of significantly positively correlated gene pairs are shown in the untreated data, as well as in the data from which batch effects have been reduced. Note that we selected PCC thresholds so that the number of gene pairs would be the same in the untreated data as in the treated data. In addition, the observed number of shared correlated gene pairs, the number expected at random, the fold increase (observed vs expected), and the percentage of shared gene pairs (shared pairs vs total pairs in treated data) is shown. Cell type abbreviations are as in Table S1.

Rank	PWM ID	Motif name	Hits in input	Hits expected	Fold enrichment	P-value
1	MA0105.1	NFKB1	52	25.9	2.01	8.8e-06
2	MA0107.1	RELA	42	18.4	2.28	2.1e-05

Table S3: Significantly enriched regulatory motifs in the promoter sequences of the 100 genes with the highest correlation with *Jmjd3* in macrophage samples. For significantly enriched motifs, the PWM ID, motif name, the observed and expected number of hits in the 100 promoter sequences is shown. There is a strong enrichment for NF-κB binding motifs (“NFKB1” and “RELA”).

Shared in x cell types	Untreated data	Batch-treated data	Fold enrichment
1	17,225,773	14,761,149	0.95
2	4,828,296	4,243,716	0.98
3	2,084,875	1,975,279	1.05
4	1,111,558	1,125,073	1.12
5	661,501	715,661	1.20
6	421,602	485,707	1.28
7	280,150	343,279	1.36
8	191,257	245,731	1.43
9	130,765	177,722	1.51
10	89,559	128,039	1.59
11	60,635	91,881	1.68
12	40,211	64,772	1.79
13	25,909	45,583	1.95
14	16,081	31,177	2.15
15	9,200	20,552	2.48
16	5,253	12,852	2.72
17	2,849	7,626	2.97
18	1,487	4,047	3.02
19	637	1,912	3.33
20	245	834	3.78
21	93	338	4.04
22	28	109	4.32

23	10	25	2.78
Total number of correlated gene pairs in cell type-specific data	27,187,974	24,483,064	1.00

Table S4: Table showing the number of significantly positively correlated gene pairs in 1, 2, …, 23 cell types, for untreated gene expression data, and data after treatment of batch effects. The fold enrichment column shows the relative enrichment after batch effect treatment, taking into account the total number of correlated gene pairs observed in both datasets (shown at the bottom).

ChIPed factor	GEO accession number	Cell type in Immuno-Navigator	Positive enrichment score	Negative enrichment score
Ascl2	GSE52840	CD4	0.050	-0.010
Atf3	GSE54414	Macrophage	0.132	-0.003
Batf	GSE39756	CD4	0.238	-0.005
Batf	GSE40918	CD4	0.340	-0.001
Batf	GSE52773	cDC	0.095	-0.030
Batf	GSE54191	CD8	0.218	-0.001
Cbx7	GSE36658	HSC	0.104	-0.003
Cbx8	GSE36658	HSC	0.081	-0.002
Cebpa	GSE21512	Macrophage	0.220	-0.052
Cebpa	GSE50565	Macrophage	0.280	-0.035
Cebpb	GSE21512	Macrophage	0.123	-0.092
Ctcf	GSE36099	cDC	0.149	0.000
Ctcf	GSE40918	CD4	0.248	0.000
Ctcf	GSE44637	Mature B	0.131	-0.001
Ctcf	GSE48086	Mast	0.212	0.000
E2f1	GSE36099	cDC	0.051	-0.063
E2f4	GSE36099	cDC	0.018	-0.062
Ebf1	GSE19971	Pro-B	0.136	0.000
Ebf1	GSE35857	Mature B	0.192	0.000
Ebf1	GSE35915	Mature B	0.289	-0.003
Egr1	GSE36099	cDC	0.171	-0.009
Egr2	GSE36099	cDC	0.176	-0.002
Egr2	GSE49366	CD4	0.262	-0.016
Elf1	GSE40684	Treg	0.293	-0.011

Ep300	GSE40463	Th1	0.163	-0.001
Ep300	GSE40463	Th2	0.152	-0.003
Ep300	GSE40918	CD4	0.268	0.000
Erg	GSE48086	Mast	0.160	-0.003
Ets1	GSE40684	Treg	0.455	0.000
Ets2	GSE36099	cDC	0.062	-0.126
Fli1	GSE20898	Th2	0.071	-0.001
Fli1	GSE48086	Mast	0.407	0.000
Fos	GSE48086	Mast	0.115	-0.017
Fosl2	GSE40918	CD4	0.069	-0.012
Foxo1	GSE40656	Treg	0.437	0.000
Foxo1	GSE46525	CD4	0.403	-0.001
Foxp3	GSE40684	Treg	0.114	-0.012
Gata2	GSE26031	HSC	0.213	-0.001
Gata2	GSE42518	Mast	0.381	0.000
Gata3	GSE20898	CD4	0.325	0.000
Gata3	GSE20898	CD8	0.301	-0.001
Gata3	GSE20898	DP	0.127	-0.021
Gata3	GSE20898	Treg	0.205	0.000
Gata3	GSE20898	NKT	0.313	-0.002
Gata3	GSE20898	Th1	0.146	-0.010
Gata3	GSE20898	Th2	0.168	-0.004
Gfi1	GSE42518	Mast	0.209	-0.001
Hif1a	GSE36099	cDC	0.136	-0.040
Hoxb4	GSE34014	HSC	0.050	-0.050
Ikzf1	GSE38200	Pre-B	0.170	-0.012
Irf1	GSE36099	cDC	0.137	-0.017
Irf4	GSE39756	Mature B	0.222	-0.018
Irf4	GSE39756	CD4	0.148	-0.020
Irf4	GSE40918	CD4	0.298	-0.005
Irf4	GSE54191	CD8	0.181	-0.031
Irf8	GSE53311	cDC	0.134	-0.052
Jun	GSE54191	CD8	0.253	0.000
Junb	GSE38377	Macrophage	0.090	-0.105

Junb	GSE52773	cDC	0.161	-0.004
Junb	GSE54191	CD8	0.150	0.000
Jund	GSE54191	CD8	0.082	0.000
Ldb1	GSE26031	HSC	0.130	0.000
Lmo2	GSE48086	Mast	0.369	0.000
Maf	GSE47528	CD4	0.160	0.000
Maff	GSE36099	cDC	0.107	-0.021
Med1	GSE44288	Pro-B	0.153	0.000
Meis1	GSE48086	Mast	0.177	-0.008
Men1	GSE53831	CD4	0.349	0.000
Mitf	GSE48086	Mast	0.274	-0.001
Pax5	GSE38046	Mature B	0.259	-0.002
Polr2a	GSE54414	Macrophage	0.057	-0.084
Pou2f2	GSE21512	Mature B	0.369	0.000
Rel	GSE36099	cDC	0.334	-0.003
Rela	GSE16723	Macrophage	0.196	-0.015
Rela	GSE36099	cDC	0.210	-0.005
Rela	GSE48759	Macrophage	0.150	-0.099
Relb	GSE36099	cDC	0.363	-0.010
Runx1	GSE29515	HSC	0.132	-0.001
Runx1	GSE48086	Mast	0.260	0.000
Runx3	GSE48591	cDC	0.109	-0.060
Runx3	GSE50131	CD8	0.283	-0.002
Sfpi1	GSE21512	Mature B	0.239	-0.021
Sfpi1	GSE21512	Macrophage	0.091	-0.143
Sfpi1	GSE21614	Pro-B	0.062	-0.010
Sfpi1	GSE38377	Macrophage	0.072	-0.154
Sfpi1	GSE48086	Mast	0.221	0.000
Sfpi1	GSE48759	Macrophage	0.096	-0.126
Sfpi1	GSE52773	cDC	0.209	-0.005
Smad3	GSE21614	Pro-B	0.197	-0.001
Stat1	GSE33913	Macrophage	0.119	-0.097
Stat1	GSE38377	Macrophage	0.097	-0.155
Stat1	GSE40463	Th1	0.209	-0.007

Stat3	GSE27161	cDC	0.169	0.000
Stat3	GSE36099	cDC	0.235	-0.002
Stat3	GSE39756	CD4	0.235	-0.001
Stat4	GSE22104	Th1	0.120	-0.019
Stat5b	GSE27161	cDC	0.159	-0.010
Stat6	GSE22104	Th2	0.093	-0.017
Stat6	GSE38377	Macrophage	0.294	-0.010
Tal1	GSE26031	HSC	0.263	0.000
Tal1	GSE48086	Mast	0.312	0.000
Tbx21	GSE33802	Th1	0.229	0.000
Tbx21	GSE40623	Th1	0.155	0.000
Tcf3	GSE48086	Mast	0.237	0.000

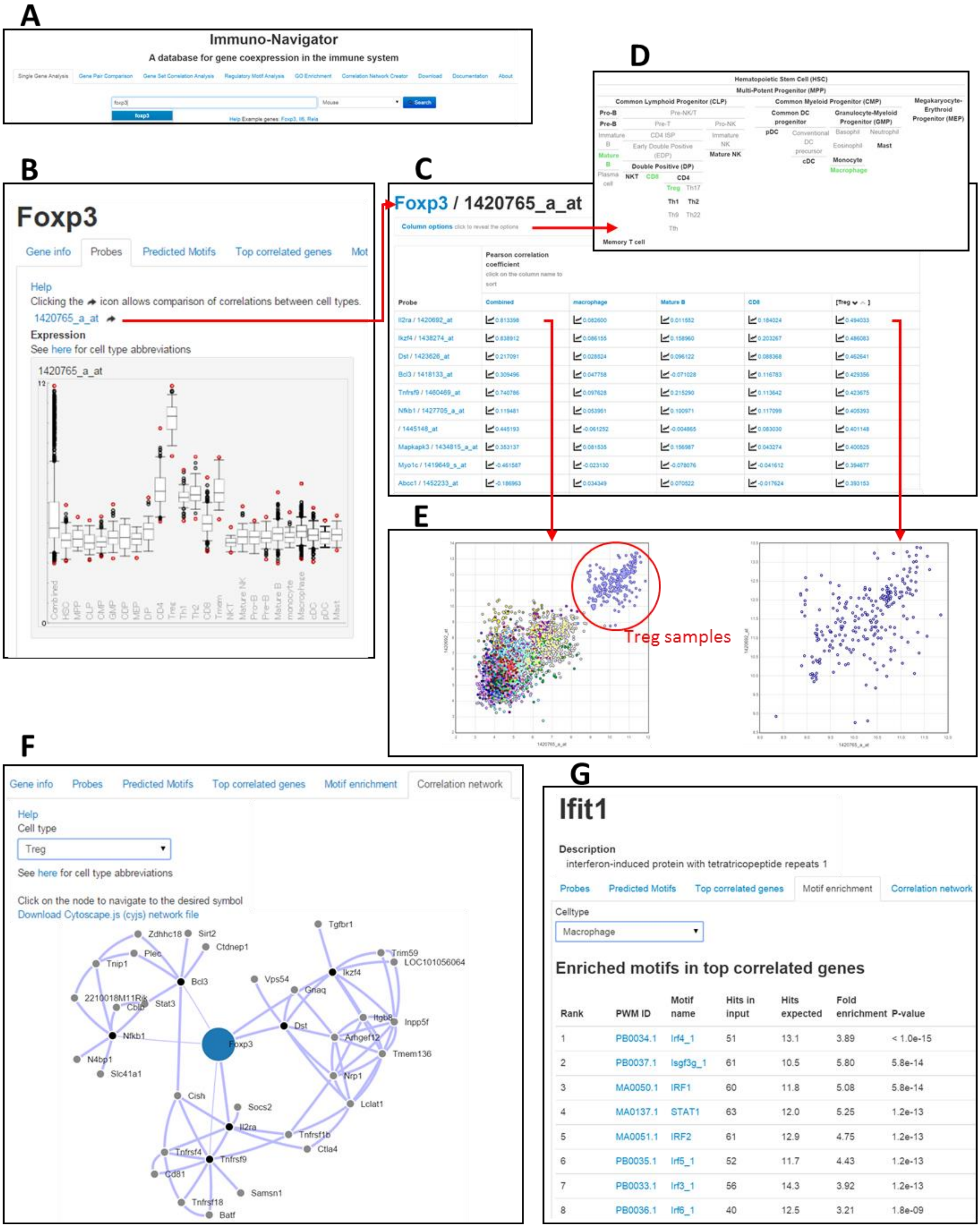
Table S5: Overview of the Haemcode-derived ChIP-seq data sets. For the 104 dataset included in this study, the table shows the ChIPed factor, the GEO accession number of the ChIP-seq data, the cell type used for the ChIP-seq experiment and for the correlation analysis, and the positive and negative enrichment scores observed in that cell type.

Cell type	**PCC threshold**	**Probe pairs**	**Gene pairs**	**Estimated FDR**
Combined data	0.620144	5999920	2924148	0*
CD4+ T cells	0.4	3498181	2032548	0
MΦ	0.4	5583890	3171592	0
mature B	0.416098	5999961	3429280	0
CD8+ T cells	0.403383	5999994	3316318	0
Treg	0.45861	5999972	3867165	8.28E-07
HSC	0.4	1888398	1411692	0
cDC	0.442626	5999973	3497173	0
Pre-B	0.479065	5999915	3321256	0.000541
GMP	0.518508	5999971	4394763	0.001781
mast	0.569727	5999908	4157054	0.009424
CMP	0.542823	5999954	4296330	0.000253
mature NK	0.543069	5999931	3595759	0.000541
DP	0.535053	5999909	3969803	0.002391
Th1	0.589866	5999903	3447620	0.003128
memory T cell	0.583928	5999934	3857087	0.005891

monocyte	0.680278	5999932	4228912	0.001371
CLP	0.639097	3859400	2870687	0.01
Th2	0.721279	1031368	761586	0.009999
pDC	0.706806	4647635	3349858	0.01
Pro-B	0.744162	1070121	818861	0.01
MEP	0.764744	1191676	881400	0.009999
NKT	0.778883	1320326	956253	0.01
CDP	0.792533	1458158	1092179	0.009999
MPP	NA	NA	NA	NA

Table S6: This table shows for each dataset the PCC threshold used, the number of probe pairs with higher PCC values than this threshold, the number of gene pairs with PCC values higher than this threshold, and the estimated false discovery rate (FDR). (*: for the combined dataset the threshold was decided using randomly shuffled data in which batch information was not used; see also section “Construction of Shuffled Data”).

SI Figures



Immuno-Navigator

A database for gene coexpression in the immune system

Help Example genes: Foxp3, Il6, Rela

Foxp3

Help

Clicking the ➜ icon allows comparison of correlations between cell types.

1420765_a_at ➜

Expression

See here for cell type abbreviations

Foxp3 / 1420765_a_at

Column options click to reveal the options

Probe	Pearson correlation coefficient click on the column name to sort Combined	macrophage	Mature B	CD8	[Treg ⌄ ⌃]
Il2ra / 1420692_at	0.813398	0.082600	0.011552	0.184024	0.494033
Ikzf4 / 1438274_at	0.838912	0.086155	0.158960	0.203267	0.486083
Dst / 1423626_at	0.217091	0.028524	0.096122	0.088368	0.462641
Bcl3 / 1418133_at	0.309496	0.047758	-0.071028	0.116783	0.429356
Tnfrsf9 / 1460469_at	0.740786	0.097628	0.215290	0.113642	0.423675
Nfkb1 / 1427705_a_at	0.119481	0.053951	0.100971	0.117099	0.405393
/ 1445148_at	0.445193	-0.061252	-0.004865	0.083030	0.401148
Mapkapk3 / 1434815_a_at	0.353137	0.081535	0.156987	0.043274	0.400525
Myo1c / 1419649_s_at	-0.461587	-0.023130	-0.078076	-0.041612	0.394677
Abcc1 / 1452233_at	-0.186963	0.034349	0.070522	-0.017624	0.393153

Hematopoietic Stem Cell (HSC)
Multi-Potent Progenitor (MPP)
Common Lymphoid Progenitor (CLP)
Common Myeloid Progenitor (CMP)
Megakaryocyte-Erythroid Progenitor (MEP)
Pro-B
Pre-NK/T
Common DC progenitor
Granulocyte-Myeloid Progenitor (GMP)
Pre-B
Pre-T
Pro-NK
Immature B
CD4 ISP
Immature NK
pDC
Conventional DC precursor
Basophil
Neutrophil
Early Double Positive (EDP)
Eosinophil
Mast
Mature B
Mature NK
cDC
Monocyte
Double Positive (DP)
Macrophage
Plasma cell
NKT
CD8
CD4
Treg
Th17
Th1
Th2
Th9
Th22
Tfh
Memory T cell
1420692_at
1420765_a_at

Help

Cell type

Treg ▾

See here for cell type abbreviations

Click on the node to navigate to the desired symbol

Download Cytoscape.js (cyjs) network file

Ifit1

Description

interferon-induced protein with tetratricopeptide repeats 1

Celltype

Macrophage ▾

Enriched motifs in top correlated genes

Rank	PWM ID	Motif name	Hits in input	Hits expected	Fold enrichment	P-value
1	PB0034.1	Irf4_1	51	13.1	3.89	< 1.0e-15
2	PB0037.1	Isgf3g_1	61	10.5	5.80	5.8e-14
3	MA0050.1	IRF1	60	11.8	5.08	5.8e-14
4	MA0137.1	STAT1	63	12.0	5.25	1.2e-13
5	MA0051.1	IRF2	61	12.9	4.75	1.2e-13
6	PB0035.1	Irf5_1	52	11.7	4.43	1.2e-13
7	PB0033.1	Irf3_1	56	14.3	3.92	1.2e-13
8	PB0036.1	Irf6_1	40	12.5	3.21	1.8e-09

Fig. S1. Example of single gene analysis using Immuno-Navigator. (A) Top page with several tabs and a gene search function. (B) Gene page for the gene *Foxp3*. A short description, IDs and external references are shown, as well as the “Probes” tab, showing the high expression of this gene in Treg samples. (C). Top correlated probes page. In this case, the probes are sorted by their correlation with the *Foxp3* probe in the Treg samples. (D) The “cell type selector” menu allows the user to select which cell types to show in the table. Black: cell

types that are currently included in Immuno-Navigator; Green: selected cell types; Grey: cell types which are not yet included in Immuno-Navigator. (E) Scatter plots of the probe of *Foxp3* (X axis) and the probe for *Il2ra* (Y axis) over all samples in the database (left) and the Treg samples only (right). For the scatter plot of all samples, colours reflect cell types. Moving the mouse over a sample displays the corresponding cell type. Treg samples are indicated. (F) Correlation network for *Foxp3* in the Treg samples. Thick edges represent significant correlations; thin edges represent high (top 5) but not significant correlations. The central blue node is the query gene, black nodes are the top 5 correlated genes of the query gene, and grey nodes represent their top 5 correlated genes, respectively. (G) Motif enrichment result page. In this case, enriched motifs are shown for the top 100 genes with highest correlation with the gene Ifit1, over all macrophage samples.

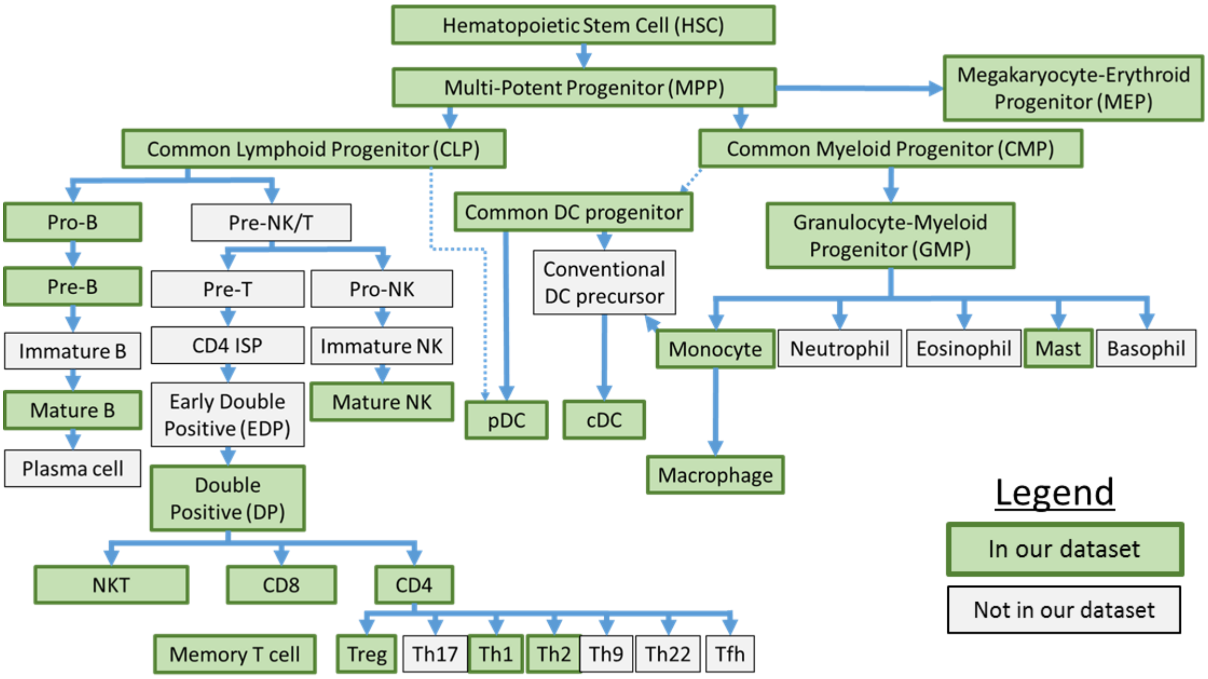


Fig. S2: Simplified representation of the hematopoietic lineage tree, indicating cell types which are currently included in our dataset (marked in green).

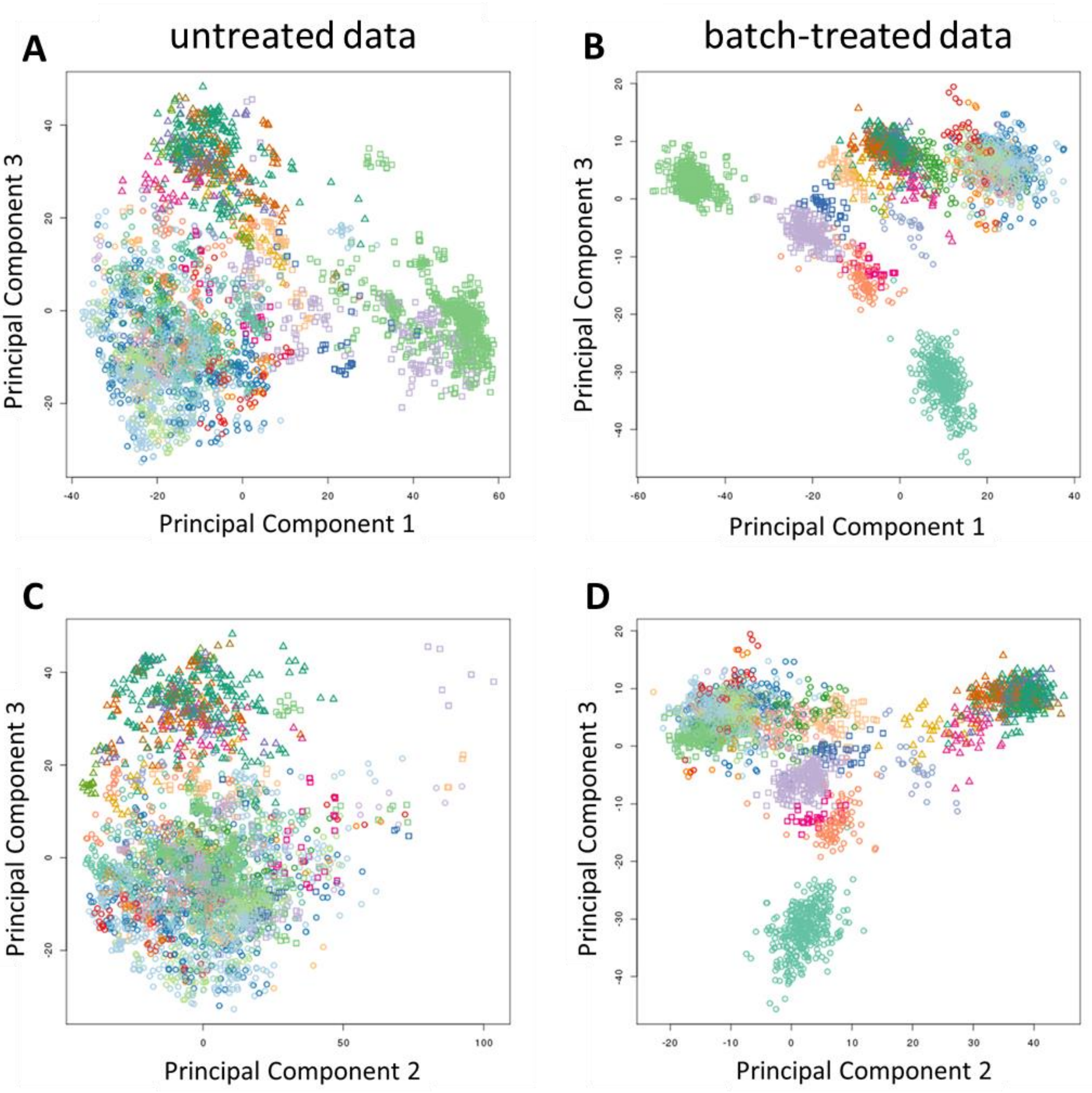


Fig. S3: Principal Component Analysis of gene expression data before and after batch effect reduction. Scatter plots are shown for all samples in the untreated data (**A**: PC1 vs PC3, **C**: PC2 vs PC3), and in the batch-treated data (**B**: PC1 vs PC3, **D**: PC2 vs PC3) . Shapes and colors reflect cell types (see legend of Fig. 1 in the main text of the paper).

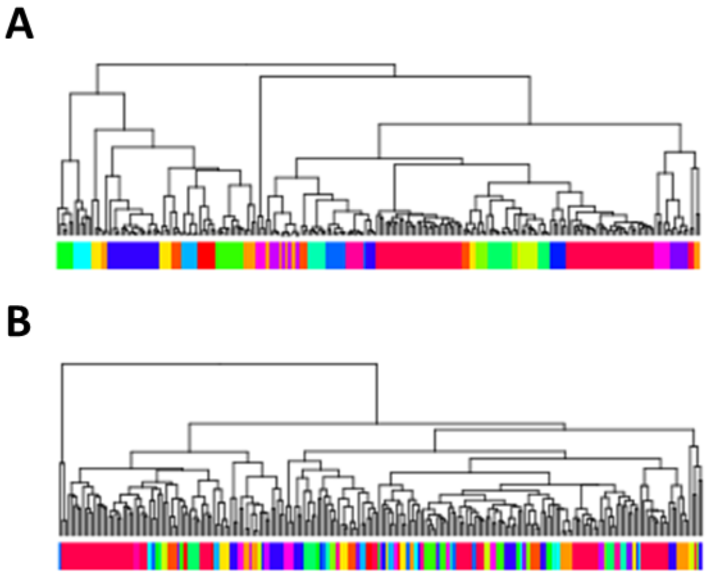


Fig. S4: Hierarchical clustering of Treg samples before (**A**) and after (**B**) batch effect reduction. Color codes below the dendrogram represent studies (different color for each study). Please note that because of the high number of studies some colors are hard to distinguish.

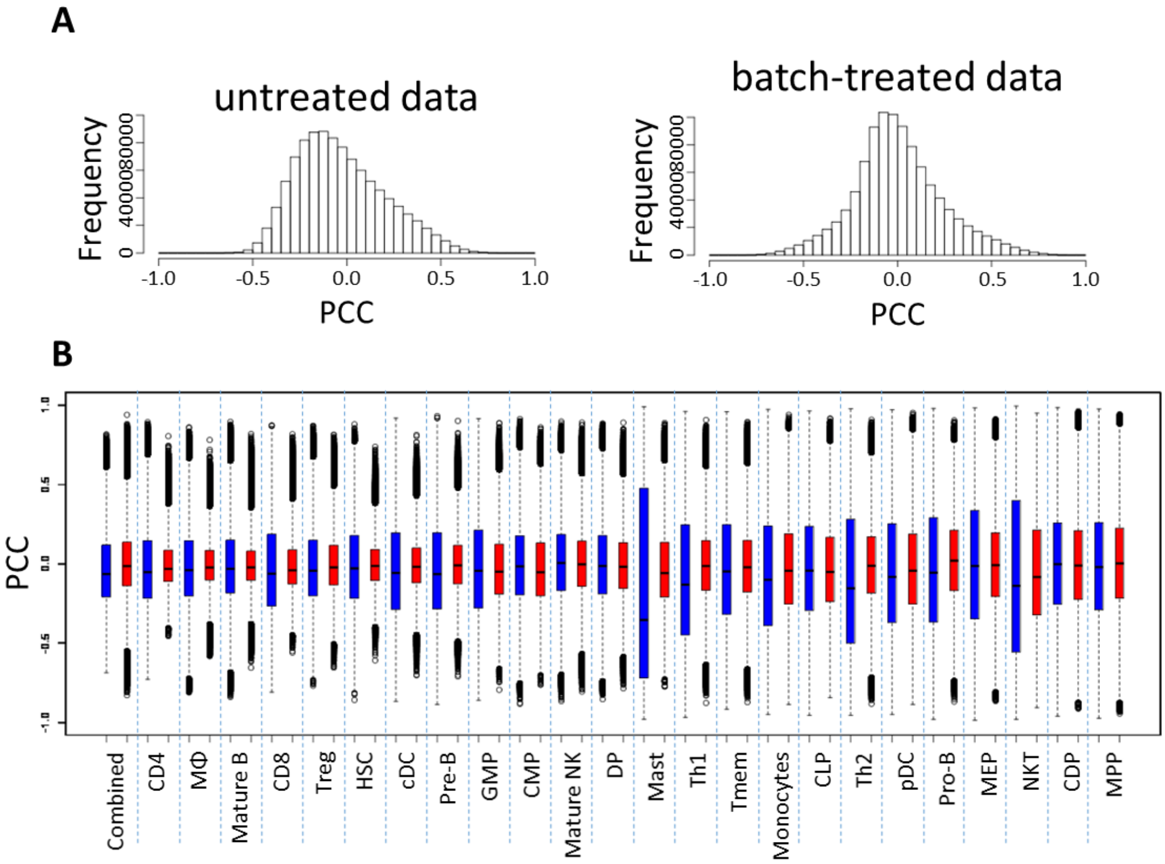
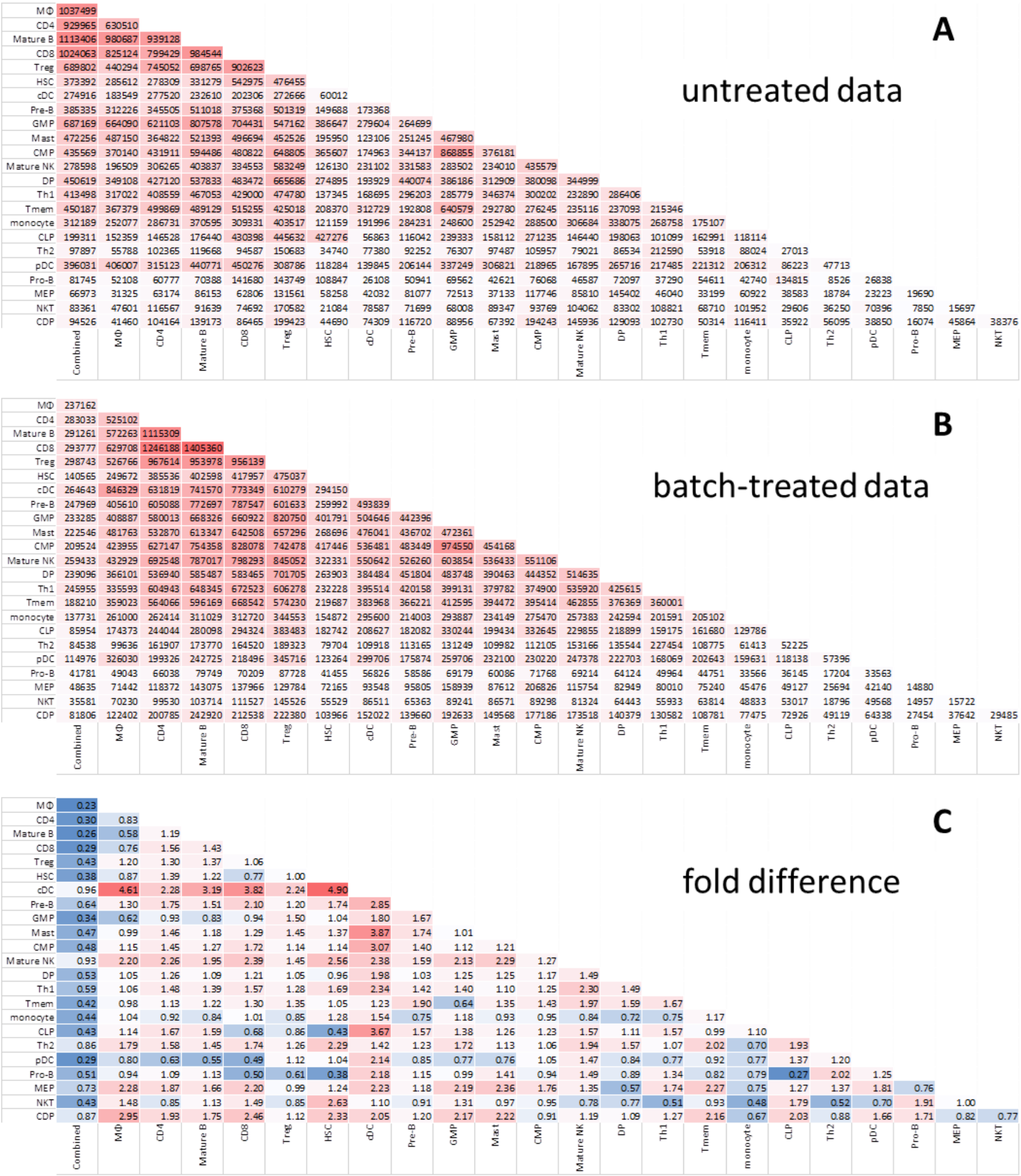


Fig. S5: Treatment of batch effects strongly changes gene expression correlation. (**A**) The distribution of PCC values in the entire set of 3,434 gene expression samples ("Combined" dataset) before (left) and after (right) batch effect reduction. (**B**) Boxplots show the distribution of PCC values observed in all data sets (combined data, and each cell type's data separately) before (blue) and after (red) batch effect treatment. In all datasets, batch effect treatment resulted in decrease of variance in PCC values, and a reduction of extremely high (positive and negative) PCC values.



A

untreated data

	Combined	MΦ	CD4	Mature B	CD8	Treg	HSC	cDC	Pre-B	GMP	Mast	CMP	Mature NK	DP	Th1	Tmem	monocyte	CLP	Th2	pDC	Pro-B	MEP	NKT
MΦ	1037499																						
CD4	929965	630510																					
Mature B	1113406	980687	939128																				
CD8	1024063	825124	799429	984544																			
Treg	689802	440294	745052	698765	902623																		
HSC	373392	285612	278309	331279	542975	476455																	
cDC	274916	183549	277520	232610	202306	272666	60012																
Pre-B	385335	312226	345505	511018	375368	501319	149688	173368															
GMP	687169	664090	621103	807578	704431	547162	386647	279604	264699														
Mast	472256	487150	364822	521393	496694	452526	195950	123106	251245	467980													
CMP	435569	370140	431911	594486	480822	648805	365607	174963	344137	868855	376181												
Mature NK	278598	196509	306265	403837	334553	583249	126130	231102	331583	283502	234010	435579											
DP	450619	349108	427120	537833	483472	665686	274895	193929	440074	386186	312909	380098	344999										
Th1	413498	317022	408559	467053	429000	474780	137345	168695	296203	285779	346374	300202	232890	286406									
Tmem	450187	367379	499869	489129	515255	425018	208370	312729	192808	640579	292780	276245	235116	237093	215346								
monocyte	312189	252077	286731	370595	309331	403517	121159	191996	284231	248600	252942	288500	306684	338075	268758	175107							
CLP	199311	152359	146528	176440	430398	445632	427276	56863	116042	239333	158112	271235	146440	198063	101099	162991	118114						
Th2	97897	55788	102365	119668	94587	150683	34740	77380	92252	76307	97487	105957	79021	86534	212590	53918	88024	27013					
pDC	396031	406007	315123	440771	450276	308786	118284	139845	206144	337249	306821	218965	167895	265716	217485	221312	206312	86223	47713				
Pro-B	81745	52108	60777	70388	141680	143749	108847	26108	50941	69562	42621	76068	46587	72097	37290	54611	42740	134815	8526	26838			
MEP	66973	31325	63174	86153	62806	131561	58258	42032	81077	72513	37133	117746	85810	145402	46040	33199	60922	38583	18784	23223	19690		
NKT	83361	47601	116567	91639	74692	170582	21084	78587	71699	68008	89347	93769	104062	83302	108821	68710	101952	29606	36250	70396	7850	15697	
CDP	94526	41460	104164	139173	86465	199423	44690	74309	116720	88956	67392	194243	145936	129093	102730	50314	116411	35922	56095	38850	16074	45864	38376

B

batch-treated data

	Combined	MΦ	CD4	Mature B	CD8	Treg	HSC	cDC	Pre-B	GMP	Mast	CMP	Mature NK	DP	Th1	Tmem	monocyte	CLP	Th2	pDC	Pro-B	MEP	NKT
MΦ	237162																						
CD4	283033	525102																					
Mature B	291261	572263	1115309																				
CD8	293777	629708	1246188	1405360																			
Treg	298743	526766	967614	953978	956139																		
HSC	140565	249672	385536	402598	417957	475037																	
cDC	264643	846329	631819	741570	773349	610279	294150																
Pre-B	247969	405610	605088	772697	787547	601633	259992	493839															
GMP	233285	408887	580013	668326	660922	820750	401791	504646	442396														
Mast	222546	481763	532870	613347	642508	657296	268696	476041	436702	472361													
CMP	209524	423955	627147	754358	828078	742478	417446	536481	483449	974550	454168												
Mature NK	259433	432929	692548	787017	798293	845052	322331	550642	526260	603854	536433	551106											
DP	239096	366101	536940	585487	583465	701705	263903	384484	451804	483748	390463	444352	514635										
Th1	245955	335593	604943	648345	672523	606278	232228	395514	420158	399131	379782	374900	535920	425615									
Tmem	188210	359023	564066	596169	668542	574230	219687	383968	366221	412595	394472	395414	462855	376369	360001								
monocyte	137731	261000	262414	311029	312720	344553	154872	295600	214003	293887	234149	275470	257383	242594	201591	205102							
CLP	85954	174373	244044	280098	294324	383483	182742	208627	182082	330244	199434	332645	229855	218899	159175	161680	129786						
Th2	84538	99636	161907	173770	164520	189323	79704	109918	113165	131249	109982	112105	153166	135544	227454	108775	61413	52225					
pDC	114976	326030	199326	242725	218496	345716	123264	299706	175874	259706	232100	230220	247378	222703	168069	202643	159631	118138	57396				
Pro-B	41781	49043	66038	79749	70209	87728	41455	56826	58586	69179	60086	71768	69214	64124	49964	44751	33566	36145	17204	33563			
MEP	48635	71442	118372	143075	137966	129784	72165	93548	95805	158939	87612	206826	115754	82949	80010	75240	45476	49127	25694	42140	14880		
NKT	35581	70230	99530	103714	111527	145526	55529	86511	65363	89241	86571	89298	81324	64443	55933	63814	48833	53017	18796	49568	14957	15722	
CDP	81806	122402	200785	242920	212538	222380	103966	152022	139660	192633	149568	177186	173518	140379	130582	108781	77475	72926	49119	64338	27454	37642	29485

C

fold difference

	Combined	MΦ	CD4	Mature B	CD8	Treg	HSC	cDC	Pre-B	GMP	Mast	CMP	Mature NK	DP	Th1	Tmem	monocyte	CLP	Th2	pDC	Pro-B	MEP	NKT
MΦ	0.23																						
CD4	0.30	0.83																					
Mature B	0.26	0.58	1.19																				
CD8	0.29	0.76	1.56	1.43																			
Treg	0.43	1.20	1.30	1.37	1.06																		
HSC	0.38	0.87	1.39	1.22	0.77	1.00																	
cDC	0.96	4.61	2.28	3.19	3.82	2.24	4.90																
Pre-B	0.64	1.30	1.75	1.51	2.10	1.20	1.74	2.85															
GMP	0.34	0.62	0.93	0.83	0.94	1.50	1.04	1.80	1.67														
Mast	0.47	0.99	1.46	1.18	1.29	1.45	1.37	3.87	1.74	1.01													
CMP	0.48	1.15	1.45	1.27	1.72	1.14	1.14	3.07	1.40	1.12	1.21												
Mature NK	0.93	2.20	2.26	1.95	2.39	1.45	2.56	2.38	1.59	2.13	2.29	1.27											
DP	0.53	1.05	1.26	1.09	1.21	1.05	0.96	1.98	1.03	1.25	1.25	1.17	1.49										
Th1	0.59	1.06	1.48	1.39	1.57	1.28	1.69	2.34	1.42	1.40	1.10	1.25	2.30	1.49									
Tmem	0.42	0.98	1.13	1.22	1.30	1.35	1.05	1.23	1.90	0.64	1.35	1.43	1.97	1.59	1.67								
monocyte	0.44	1.04	0.92	0.84	1.01	0.85	1.28	1.54	0.75	1.18	0.93	0.95	0.84	0.72	0.75	1.17							
CLP	0.43	1.14	1.67	1.59	0.68	0.86	0.43	3.67	1.57	1.38	1.26	1.23	1.57	1.11	1.57	0.99	1.10						
Th2	0.86	1.79	1.58	1.45	1.74	1.26	2.29	1.42	1.23	1.72	1.13	1.06	1.94	1.57	1.07	2.02	0.70	1.93					
pDC	0.29	0.80	0.63	0.55	0.49	1.12	1.04	2.14	0.85	0.77	0.76	1.05	1.47	0.84	0.77	0.92	0.77	1.37	1.20				
Pro-B	0.51	0.94	1.09	1.13	0.50	0.61	0.38	2.18	1.15	0.99	1.41	0.94	1.49	0.89	1.34	0.82	0.79	0.27	2.02	1.25			
MEP	0.73	2.28	1.87	1.66	2.20	0.99	1.24	2.23	1.18	2.19	2.36	1.76	1.35	0.57	1.74	2.27	0.75	1.27	1.37	1.81	0.76		
NKT	0.43	1.48	0.85	1.13	1.49	0.85	2.63	1.10	0.91	1.31	0.97	0.95	0.78	0.77	0.51	0.93	0.48	1.79	0.52	0.70	1.91	1.00	
CDP	0.87	2.95	1.93	1.75	2.46	1.12	2.33	2.05	1.20	2.17	2.22	0.91	1.19	1.09	1.27	2.16	0.67	2.03	0.88	1.66	1.71	0.82	0.77

Fig. S6: Overlap in positively correlated gene pairs in all datasets (combined data and cell type-specific data). (**A**) Table showing the number of shared significantly positively correlated gene pairs between all datasets in

the untreated data. A color code (white: low; red: high) is used to improve readability. (**B**) Same as (A) for the data after batch effect reduction. (**C**) Fold difference in the number of overlapping gene pairs between the batch-treated and untreated data. This shows the counts shown in (B) divided by those of (A). Here too a color code is used (red: higher overlap in the batch-treated data, blue: higher overlap in the untreated data). In 183 (72%) out of the 253 cell type combinations an increase in overlap was observed.

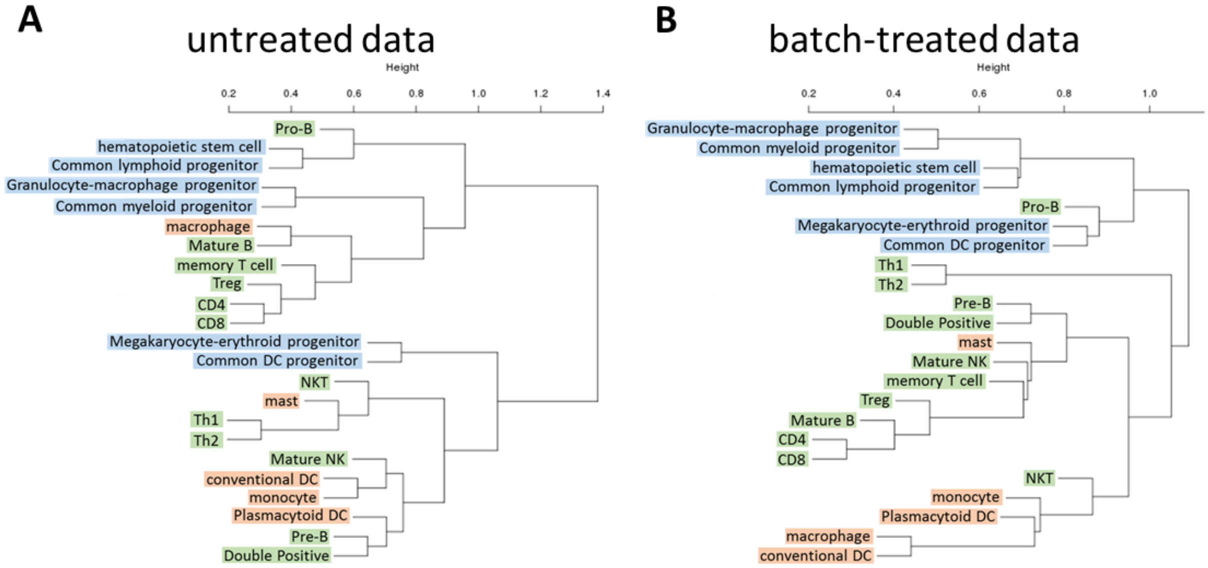


Fig. S7. Hierarchical clustering of cell types according to their similarity in gene pair correlation of expression values. Cluster dendrograms are shown for the untreated expression data (**A**) and for the data after batch effect reduction (**B**). Cell types are marked as follow: blue: progenitor cell types; green: lymphoid cell types; orange: myeloid cell types.

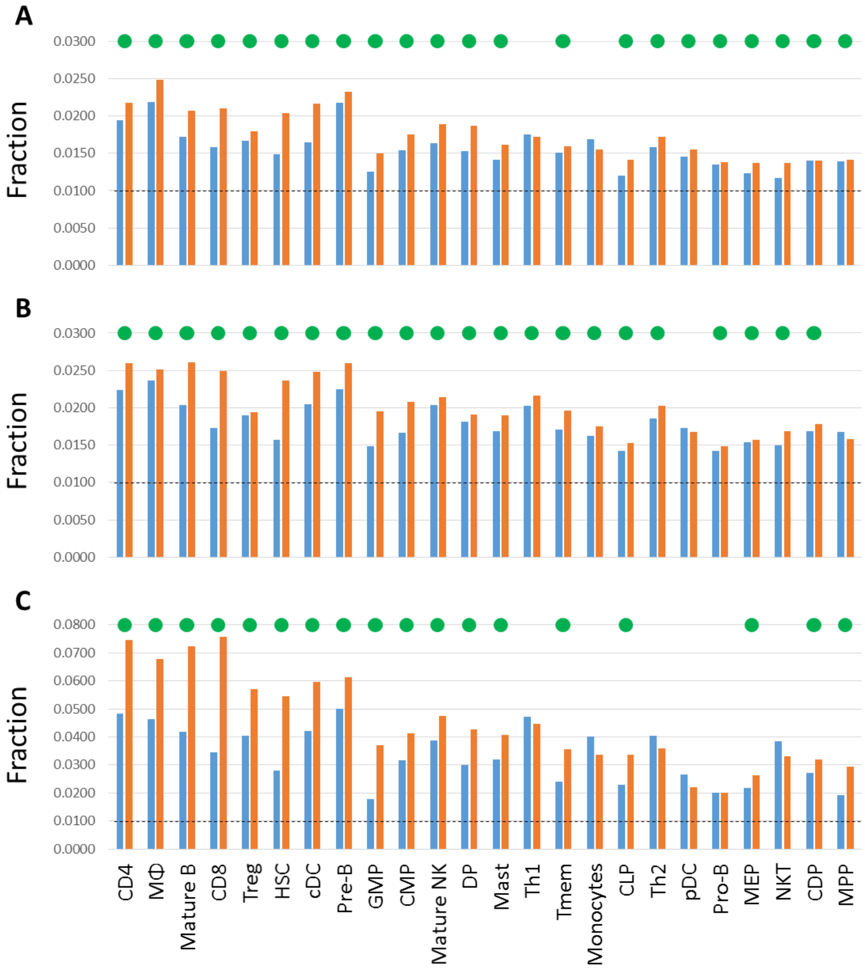


Fig. S8: Evaluation of batch effect reduction on correlation of expression between functionally similar genes. For each dataset (24 cell types), the fraction of functionally related gene pairs with high correlation is shown. High correlation was defined as correlation higher than that of 99% of functionally unrelated gene pairs. The fractions are shown for untreated (blue) and batch-treated (orange) data, for Biological Process (**A**), Molecular Function (**B**), and Cellular Component (**C**) GO annotations. Green dots indicate datasets in which an improvement was observed in the batch-treated data.

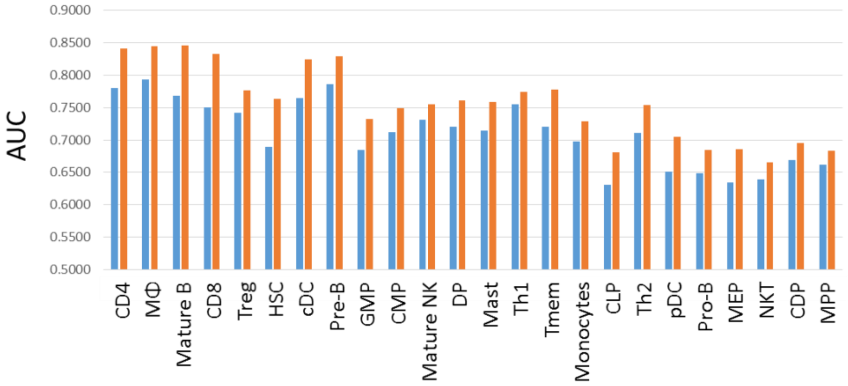


Fig. S9: Treatment of batch effects results in a relative increase of correlation between probe set pairs representing the same gene, in macrophage data. Differences in correlation between probe pairs representing the same gene, and probe pairs representing different genes were measured using Area Under the Curve (AUC)

values of ROC curves. The barplots in this figure show for all cell types the AUC values of these ROC curves, for untreated (blue) and batch-treated (orange) data. In all datasets an improvement was observed after batch effect reduction.

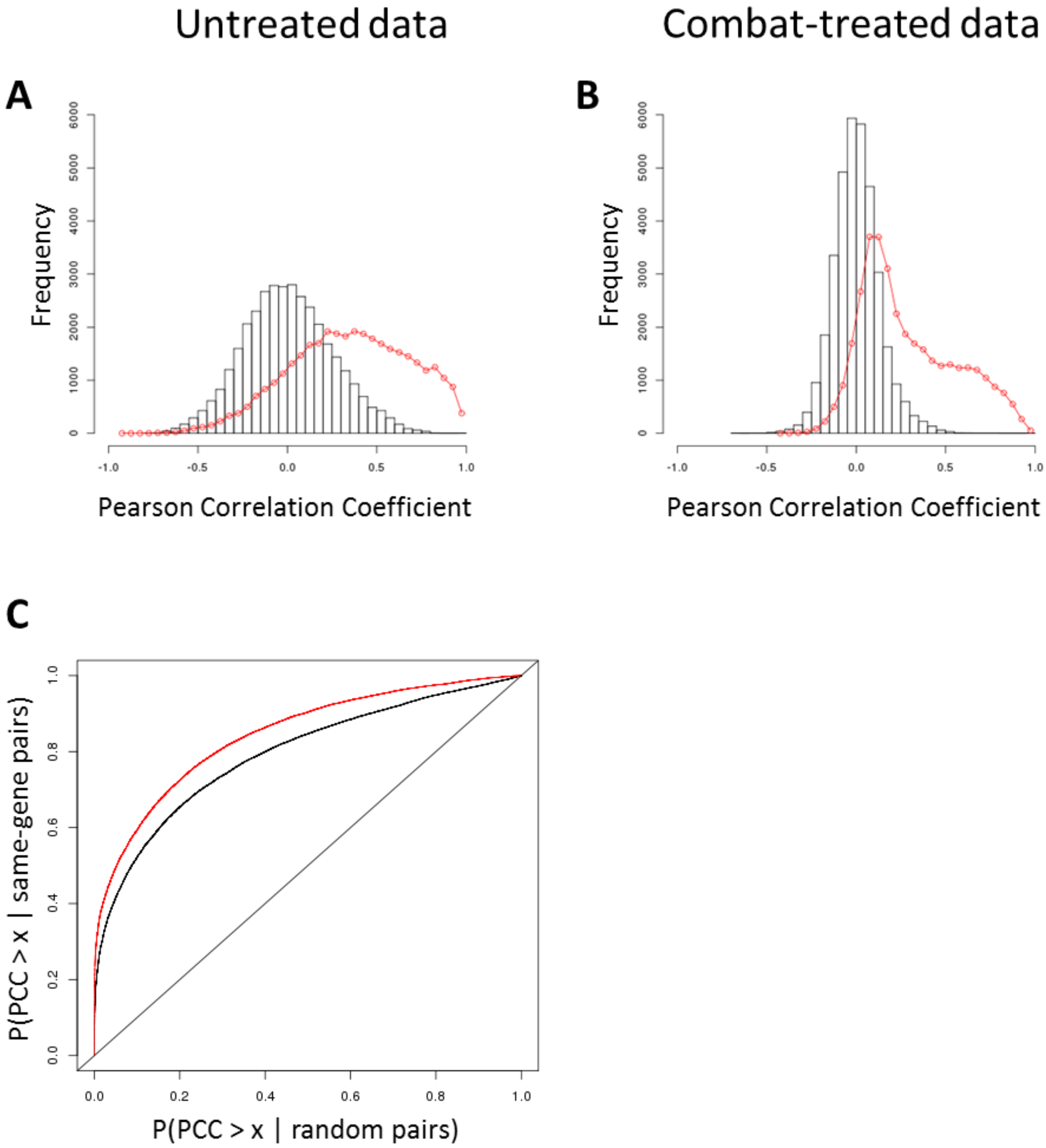


Fig. S10. Reduction of batch effects results in a relative increase of correlation between probe set pairs representing the same gene, in macrophage data. (**A**) Histograms for the distribution of PCC values in the raw, untreated, macrophages gene expression data. The histogram shows the distribution of PCC values for random probe pairs representing different genes (black), and for probe pairs representing the same gene (red). (**B**) Similar histogram for the batch-treated macrophage data. (**C**) ROC curve for PCC values in macrophage-derived expression data, between pairs of probes mapped to the same gene, and between randomly selected probes not mapped to the same gene, before (black line) and after (red line) bath reduction treatment. After batch reduction, probe pairs representing the same gene are relatively more positively correlated (see also Fig. S9). “x”: PCC threshold.

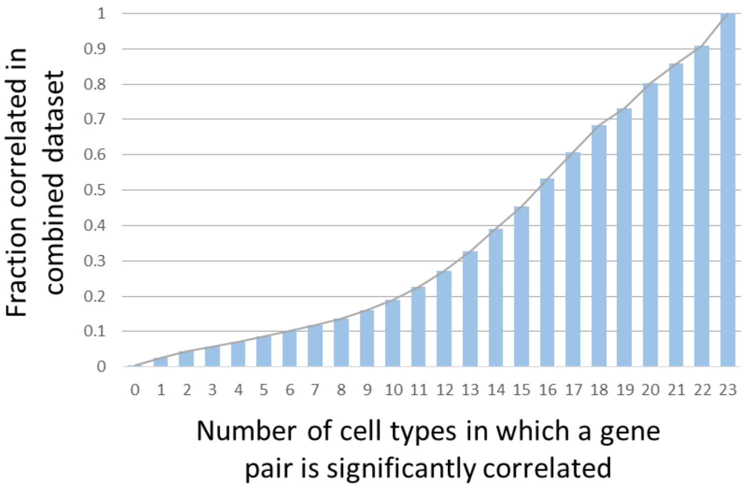


Fig. S11: The relationship between correlation in cell type-specific data and in the combined data. For gene pairs with significant positive correlation in 0, 1, 2,…, 23 cell types (X axis), the fraction of gene pairs that was also significantly positively correlated in the combined data (Y axis) is shown. For example, of gene pairs that are significantly correlated in 6 cell types, about 10% is also correlated in the data of all cell types combined together.

A

Rank	Gene Symbol	HSC	CLP	CMP	GMP	CDP	MEP	DP	CD4	Treg	Th1	Th2	CD8	Tmem	NKT	Mature NK	Pro-B	Pre-B	Mature B	monocyte	macrophage	cDC	pDC	mast	combined data
5	Mndal	29.3	0.4	27.4	15.7	1.2	7.8	4.2	16.7	1.6	0.6	9.1	14.5	2.7	18.6	2.8	7.1	3.6	14.4	42.1	204	179	82.2	6.6	21.7
14	Stat1	43.4	2.6	30.3	6.4	0.6	2.1	16.6	45.2	30.6	26.7	39.1	42.4	15.1	15.2	13.6	14.5	31.5	74.6	19.3	162	168	78.3	1.0	27.2
25	Batf2	0.5	0.0	1.0	0.7	0.6	0.0	0.0	0.0	0.0	5.3	0.0	0.0	0.7	0.0	2.5	0.0	0.0	1.8	3.6	193	163	3.8	0.0	45.7
42	Mxd1	1.5	0.0	0.0	3.0	1.7	0.0	0.8	1.7	1.2	22.7	2.8	0.0	3.5	0.7	7.0	0.4	0.1	1.6	0.7	199	156	53.1	2.3	0.7
50	Pyhin1	9.1	0.5	9.5	2.7	1.3	0.0	14.4	16.7	2.6	16.2	6.2	16.7	4.3	4.8	13.6	5.8	6.0	56.1	46.9	180	153	81.8	5.9	12.0
54	Irf7	33.2	2.9	9.7	8.1	1.5	0.0	18.9	46.8	26.5	38.5	21.6	41.8	14.1	7.3	16.8	2.7	30.5	73.6	37.6	145	152	49.4	0.0	32.9
62	Irf1	25.1	0.8	9.1	4.3	1.0	0.0	11.3	34.7	4.6	10.7	7.2	20.3	6.0	4.0	3.2	3.1	2.5	49.7	0.0	145	149	0.0	5.0	24.7
68	Pydc4	7.5	2.7	10.3	8.3	0.0	0.0	24.2	18.1	7.0	12.3	27.4	17.1	1.3	7.2	14.1	0.0	44.9	69.6	56.1	168	145	87.2	0.8	21.3
71	Stat2	9.0	0.7	0.0	0.0	0.8	0.0	0.5	32.6	7.2	15.1	17.2	19.8	5.9	1.5	5.2	0.7	9.9	36.4	50.5	159	145	59.9	0.2	40.9
83	Pou3f1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	3.2	0.0	0.6	0.0	0.0	0.0	0.3	114	141	12.5	0.0	44.0

B

Rank	PWM ID	Motif name	Hits in input	Hits expected	Fold enrichment	P-value
1	PB0037.1	Isgf3g_1	129	44.5	2.9	< 1.0e-15
2	MA0050.1	IRF1	132	50.6	2.61	< 1.0e-15
3	MA0137.1	STAT1	135	51.7	2.61	< 1.0e-15
4	MA0051.1	IRF2	134	57.1	2.35	< 1.0e-15
5	PB0035.1	Irf5_1	120	54.5	2.2	< 1.0e-15
6	PB0034.1	Irf4_1	130	61.7	2.11	< 1.0e-15
7	PB0033.1	Irf3_1	134	65.1	2.06	< 1.0e-15
8	MA0107.1	RELA	128	80.3	1.59	1.20E-05
9	MA0061.1	NF-kappaB	132	84.7	1.56	2.40E-05
10	MA0105.1	NFKB1	142	97.8	1.45	0.00031

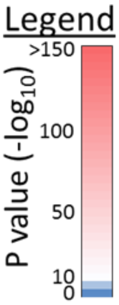


Fig. S12 Example analysis of genes induced after LPS stimulation of DCs. (**A**) Result of CNHP. Rows represent genes, and columns represent cell types. The first column shows the rank of genes as sorted by their enrichment score in cDC expression data. The second column shows gene symbols. In this case, only genes associated with the GOslim annotation GO:0001071 ("nucleic acid binding transcription factor activity") are shown. Genes are sorted by enrichment of high correlations with the input genes in the cDC data, and only the top 10 enriched genes are shown. A colour code is used to represent the degree of enrichment (-log$_{10}$ p value; blue: no enrichment; red: high enrichment). Values above 10 are considered to be significant. Cell type abbreviations are as in Table S1. (**B**) Motif enrichment analysis of the LPS induced genes. The top 10 regulatory motifs with significant enrichment in the input genes is shown. Several IRF and STAT family motifs are enriched, supporting the result shown in (A).

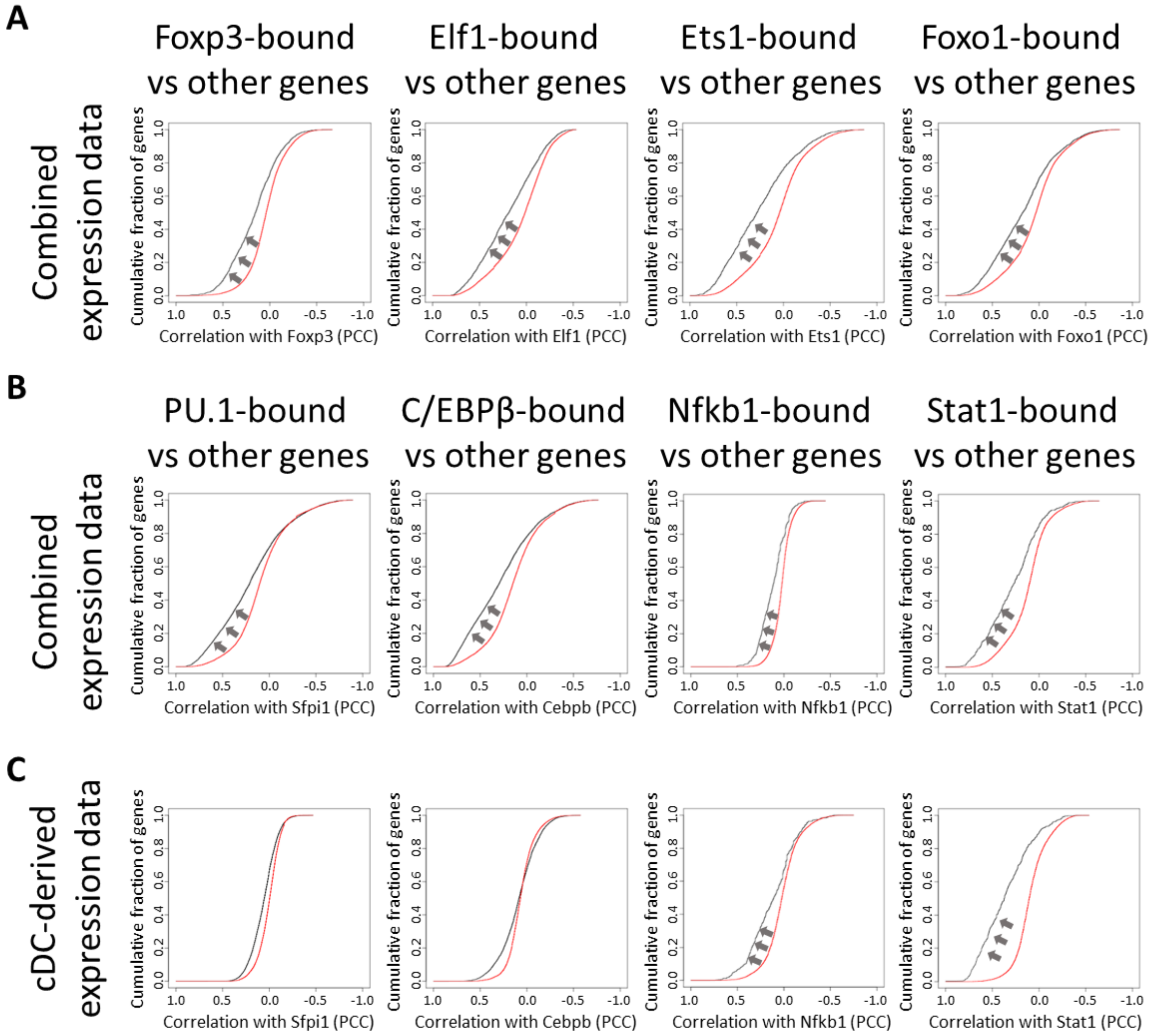


Fig. S13. Relationship between transcription factor binding and correlation of expression. (**A**) For four genes encoding transcription factors (Foxp3, Elf1, Ets1, and Foxo1) the cumulative distribution of correlation of expression is shown between the transcription factor and its target genes (black line) and non-target genes (red line) over the combined gene expression data (all cell types). The X axis represents the PCC. All four transcription factors showed higher correlation with their target genes than with their non-target genes. For the corresponding plots for correlation of expression in the Treg-derived samples only we refer to Fig. 5 in the main text. (**B**) and (**C**): Similar plots for four transcription factors (PU.1, C/EBPβ, Nfkb1, and Stat1) in dendritic cells. For all four regulators, higher correlation was observed with target genes than with non-target genes over the combined expression data (all cell types, B). In the cDC-derived data (C), however, higher correlation with target genes was observed only for Nfkb1 and Stat1, and not for PU.1 and C/EBPβ).

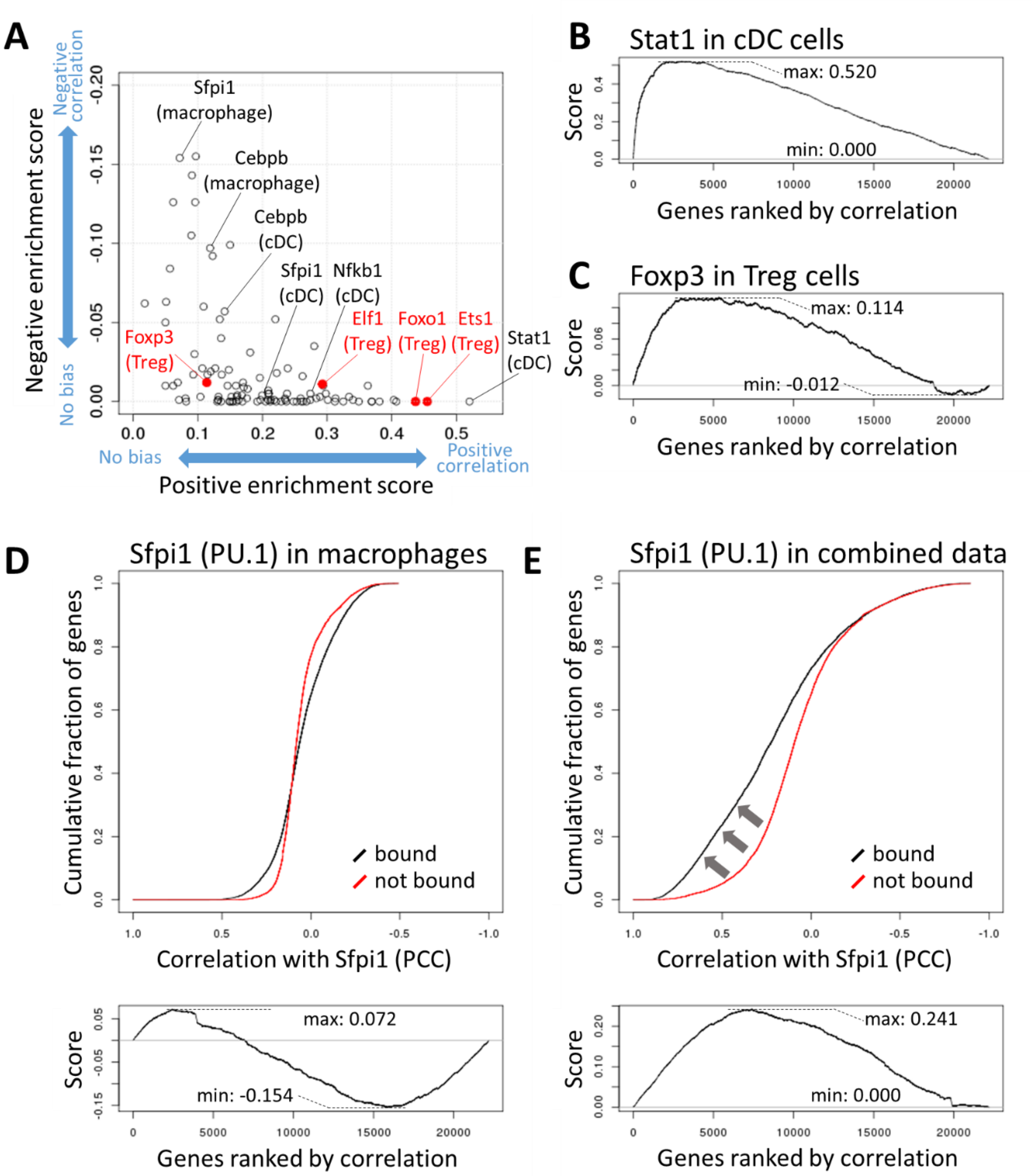
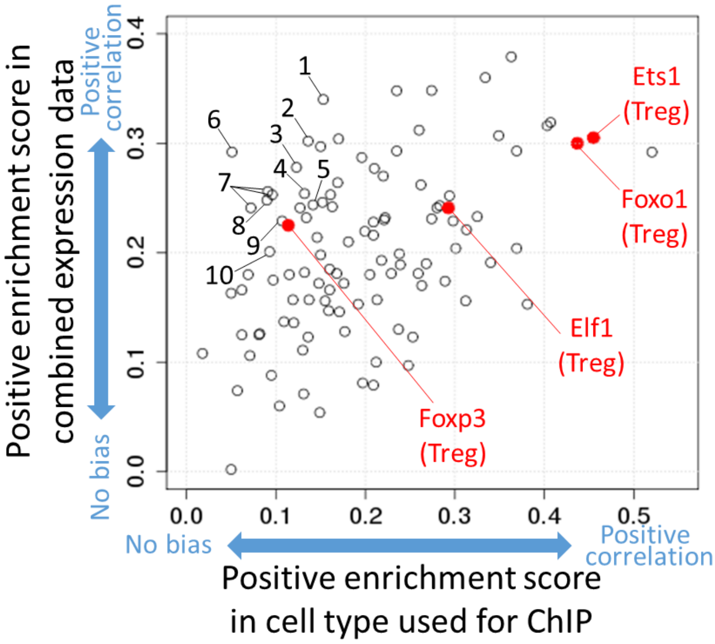


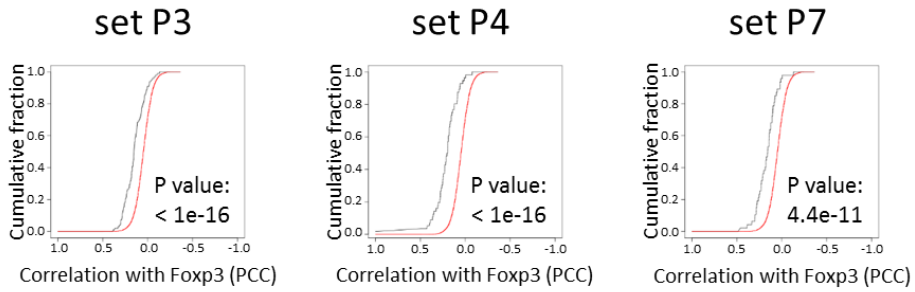
Figure S14: Large-scale analysis of expression correlation between regulators and their target genes. For 104 ChIP-seq dataset included in Haemcode and a few additional datasets included in this paper, we analysed the correlation of expression between the ChIPed factor and its target genes in the relevant cell types. **(A)** Scatter plot of positive and negative enrichment scores. Each point represents a ChIP-seq experiment for a specific factor in a specific cell type. High positive/negative scores reflect enrichment of target genes towards high positive/negative correlation of expression with the ChIPed factor. A number of datasets discussed in the paper are indicated. Note that the scales of the X and Y axes are different. **(B)** Enrichment plot of genes bound by Stat1 in cDC cells. Positive and negative enrichment scores are indicated. **(C)** As in (B), for Foxp3 target genes in Treg cells. **(D)** Cumulative distribution of PCC values for correlation with the *Sfpi1* gene (encoding PU.1) are shown for PU.1-bound and non-bound genes in macrophages. Below the corresponding enrichment plot is shown. **(E)** Same as in (D) for correlation with *Sfpi1* over the combined set of expression data for all cell types in our dataset.



Index in plot	**Regulator**	**Cell type**
1	Med1	Pro-B
2	Hif1a	cDC
3	Cebpb	Macrophage
4	Atf3	Macrophage
5	Cebpb	cDC
6	E2f1	cDC
7	Sfpi1	Macrophage
8	Junb	Macrophage
9	Maff	cDC
10	Stat6	Th2

Figure S15: Comparison of the correlation bias observed in cell type-specific data, and in the combined data for all cell types. Enrichment scores are shown for the same datasets as shown in Fig. S14. The X axis shows the bias in correlation between each regulator and its targets in the cell type used for the ChIP experiment. The Y axis shows the bias in the combined data for all cell types. Several Treg-derived datasets are indicated. In the upper left part of the plot, the indices 1 to 10 indicate regulators that have a similar pattern as Foxp3 in Tregs. Namely, these regulators lack correlation of expression with target genes in the cell type used for the ChIP-experiment, yet have correlation of expression when seen over the entire dataset. Details about these regulators are shown in the table at the right.

A



B

P3

Rank	Gene Symbol	HSC	CLP	CMP	GMP	CDP	MEP	DP	CD4	Treg	Th1	Th2	CD8	Tmem	NKT	Mature NK	Pro-B	Pre-B	Mature B	monocyte	macrophage	cDC	pDC	mast	combined data
1	Fam129a	2.4	1.6	7.9	1.2	0.8	0.0	2.1	10.7	25.2	0.4	1.0	3.2	0.3	0.0	0.7	1.5	5.6	0.8	1.8	0.0	0.6	2.4	3.7	5.2
2	Sdcbp2	0.0	0.0	0.0	0.5	0.0	0.0	0.0	5.3	19.8	0.8	0.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	4.0
3	Prdm1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.4	16.2	5.3	3.5	8.0	0.0	0.0	0.3	0.8	0.7	2.4	0.8	3.1	2.5	0.0	0.0	0.0
4	Pon3	0.1	1.6	0.0	0.2	0.0	0.0	0.8	3.8	15.7	0.6	1.2	0.0	0.0	0.0	0.0	0.4	0.0	3.4	0.0	0.0	0.7	0.0	0.6	4.0
5	Hacd3	0.0	0.0	0.2	0.1	0.0	0.5	0.5	1.8	14.7	0.0	0.0	0.1	0.3	0.0	0.0	0.0	0.4	0.0	0.2	0.4	0.1	2.1	0.6	0.0
6	Tiam1	0.0	0.2	0.6	1.9	2.5	0.0	0.0	1.2	13.6	0.5	3.3	0.6	0.2	1.3	0.0	0.0	0.3	0.0	1.5	0.0	0.3	0.0	0.2	0.0
7	Cass4	0.0	0.4	1.5	0.0	1.0	0.0	0.0	4.2	13.6	1.0	1.0	0.0	0.7	0.0	1.2	0.0	0.0	0.0	1.5	0.0	1.0	0.0	0.0	3.6
8	Tmbim1	1.5	0.0	0.2	2.9	0.7	0.0	0.0	3.2	13.6	1.1	0.5	0.5	1.7	0.0	0.2	0.0	1.0	0.1	0.1	1.0	0.2	0.0	0.9	3.9
9	Lclat1	0.0	0.6	0.2	0.3	0.0	0.0	0.4	0.0	13.4	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.9	0.6	0.5	0.1	0.0	0.1	0.0
10	Il1rl1	0.0	1.1	0.0	0.9	0.0	0.7	2.4	1.3	13.2	3.1	7.9	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.1	0.0

C

P4

Rank	Gene Symbol	HSC	CLP	CMP	GMP	CDP	MEP	DP	CD4	Treg	Th1	Th2	CD8	Tmem	NKT	Mature NK	Pro-B	Pre-B	Mature B	monocyte	macrophage	cDC	pDC	mast	combined data
1	Tnfrsf1b	0.0	0.6	5.4	1.1	0.0	0.0	3.2	20.3	29.9	0.0	0.0	4.2	0.7	2.2	0.4	1.6	0.7	1.1	0.6	4.1	1.1	4.1	1.6	4.4
2	Il2ra	0.0	1.0	0.0	0.0	0.0	0.0	1.1	12.0	24.3	0.4	2.5	1.7	4.1	0.0	2.7	1.1	0.0	1.3	0.8	0.0	5.2	0.7	0.0	13.6
3	Ctla4	0.0	0.0	0.9	0.0	0.0	0.0	0.0	29.5	23.5	2.3	4.6	8.5	0.0	0.0	2.6	0.0	0.0	3.8	0.0	0.0	0.0	3.0	0.0	13.4
4	Stx11	0.8	1.1	0.8	2.3	0.0	0.0	0.5	27.0	22.9	0.4	3.4	8.5	4.9	1.0	0.0	0.0	5.6	0.4	0.0	4.3	5.8	5.6	0.3	5.4
5	Fam129a	0.8	1.3	4.2	1.1	0.0	0.0	0.0	42.6	21.0	2.4	4.5	14.1	5.8	3.9	2.8	0.3	5.4	2.2	0.0	0.6	2.5	5.9	0.1	8.2
6	Tnfrsf4	0.0	0.0	0.0	0.0	0.0	0.0	3.2	27.2	20.8	1.6	9.1	9.3	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	11.9
7	Tiam1	0.7	0.0	0.0	0.8	0.0	0.0	0.4	15.3	19.9	2.8	4.5	5.9	3.9	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.5	1.0	1.6	0.0
8	Vav2	0.0	0.0	0.4	0.3	0.0	0.0	0.2	3.8	19.8	0.6	0.0	1.3	0.9	0.0	0.0	0.6	0.0	0.8	0.6	0.3	0.1	0.4	0.5	5.3
9	Slamf1	0.0	0.0	0.9	0.0	0.0	0.0	0.1	6.3	19.3	1.5	0.7	8.7	2.5	0.0	0.0	0.6	0.0	2.2	0.9	0.4	4.8	0.0	0.0	4.0
10	Dst	1.3	0.0	1.6	2.2	0.0	0.8	0.0	13.9	18.0	0.0	1.1	0.0	2.2	0.0	0.0	0.5	0.0	0.0	0.7	0.8	0.7	0.8	1.2	0.0

D

P7

Rank	Gene Symbol	HSC	CLP	CMP	GMP	CDP	MEP	DP	CD4	Treg	Th1	Th2	CD8	Tmem	NKT	Mature NK	Pro-B	Pre-B	Mature B	monocyte	macrophage	cDC	pDC	mast	combined data
1	Tiam1	0.8	0.0	0.0	2.1	0.4	0.0	0.4	1.8	18.2	0.0	1.4	0.0	0.5	1.6	0.0	0.0	0.5	0.0	1.3	0.0	0.6	0.0	1.9	0.0
2	Gpr83	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	14.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	3.5
3	Tmem158	0.0	0.0	0.0	0.0	0.9	1.0	0.2	0.0	12.8	0.0	0.8	0.0	0.3	0.0	1.1	0.0	0.0	0.0	0.0	0.5	0.0	0.3	0.5	0.0
4	Igf1r	1.1	0.6	0.7	0.9	0.6	0.7	0.0	0.0	11.4	0.3	0.0	2.0	0.0	0.0	0.5	2.1	0.0	2.2	0.0	0.0	0.0	0.1	1.0	3.1
5	Ikzf4	1.8	1.0	1.8	0.0	0.0	0.0	3.6	0.0	11.2	2.1	1.4	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	0.0	4.0	7.5
6	1110032F04Rik	0.0	0.0	2.2	0.0	1.9	0.0	2.0	0.9	10.6	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	0.0	0.0	2.2
7	Itih5	0.2	0.0	0.9	0.0	0.0	0.4	2.0	5.1	10.4	1.5	0.0	1.0	0.0	2.2	0.6	0.7	0.0	0.5	0.0	0.0	0.0	0.0	0.5	6.7
8	Itgb8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	1.5
9	Slamf1	2.4	0.9	0.9	0.0	0.0	3.3	0.3	0.7	10.0	0.0	0.0	0.6	0.8	0.7	0.0	0.0	0.0	0.0	1.0	0.4	3.3	0.0	0.3	0.8
10	Slc22a2	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	9.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0

Fig. S16: Analysis of correlation of expression of Foxp3-dependent, Foxp3-amplified, and Foxp3-independent gene sets. **(A)** Correlation of expression with *Foxp3* in Treg-derived samples. For Foxp3-dependent (P3), Foxp3-amplified (P4), and Foxp3-independent (P7) genes, the cumulative distribution of PCC values in the Treg-derived data is shown (black line). The red line represents genes not in each cluster. For all three sets, increased positive correlations with *Foxp3* were observed. P values for the difference in distribution is included in each graph (based on the Kolmogorov-Smirnov test). **(B-D)** Tables showing the top 10 genes with the highest correlation score (rank 1 to 10) for Foxp3-dependent (P3, **B**), Foxp3-amplified (P4, **C**), and Foxp3-independent (P7, **D**) gene sets. Scores are shown in 23 cell types, and in the combined dataset. Genes are sorted by their score in Treg-derived data. A color code represents the score (-log10 p value; blue: no enrichment; red: high enrichment). Cell type abbreviations are as in Table S1.

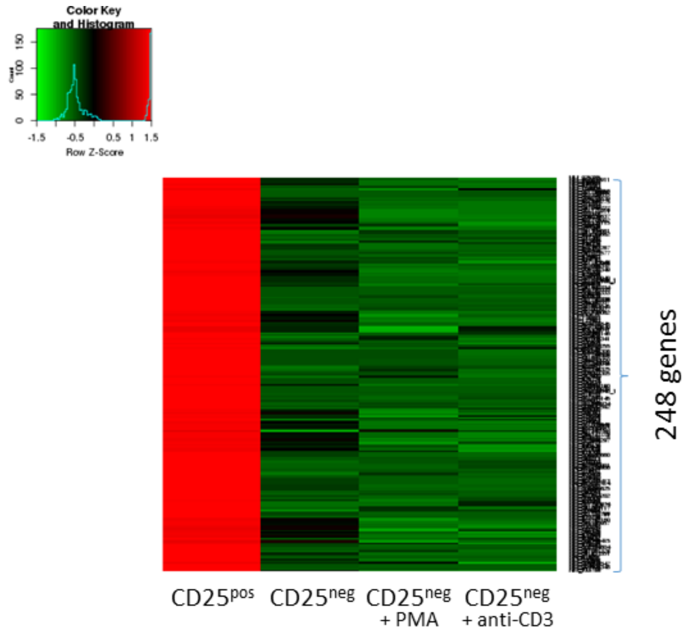


Fig. S17. Heatmap of 248 genes with Treg-specific expression. Columns represent RNA-seq-based expression levels of 248 genes in CD25^{pos}, and unstimulated and stimulated CD25^{neg} T cells.

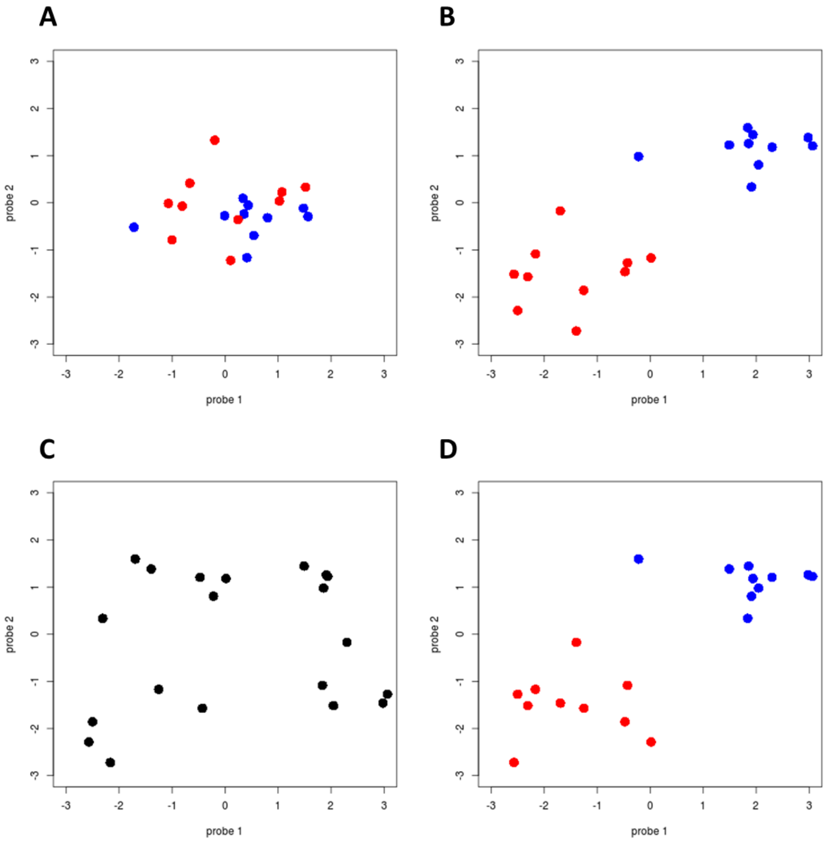


Fig. S18: Toy example of two probes measured in 2 batches (blue points: batch 1; red points: batch 2). (**A**) Without batch effects. (**B**) The same points with a simple batch effect added to both batches. (**C**) Toy example of (B) subjected to random shuffling, and (**D**) "batch-guided" shuffling. For (C), colors have been removed, as batch information is lost by the shuffling. In (D) colors are as in (A) and (B), and illustrate that batch effects have been preserved. The PCC values are 0.079 for (A), 0.854 for (B), 0.152 for (C) and 0.810 for (D), respectively.

Input size	combined	CD4+ T cells	MΦ	mature B	CD8+ T cells	Treg	HSC	cDC	Pre-B	GMP	CMP	mature NK	DP	mast	Th1	memory T cell	monocyte	CLP	Th2	pDC	Pro-B	MEP	NKT	CDP	MPC
10	3.35	3.57	3.67	4.50	3.84	3.59	4.38	3.87	4.22	4.33	5.36	4.04	4.57	4.59	5.26	4.64	5.12	5.65	3.52	4.49	3.40	4.70	4.59	4.41	NA
20	4.87	3.74	5.02	4.20	4.79	3.71	5.71	6.33	4.44	4.07	4.87	5.44	4.96	4.20	4.62	4.27	5.10	4.40	4.13	4.08	4.73	3.69	6.16	4.65	NA
30	5.43	4.11	5.11	4.51	4.52	4.51	4.59	4.51	4.35	5.10	4.62	4.75	4.20	5.11	4.51	4.51	4.65	4.35	4.63	4.37	4.20	4.34	4.67	5.02	NA
40	3.52	5.20	4.60	4.34	5.25	5.10	5.55	5.97	4.70	6.12	5.20	4.41	4.48	4.32	4.28	4.93	5.14	4.74	4.86	4.24	4.35	4.41	4.20	3.99	NA
50	3.51	5.75	3.97	5.12	4.20	5.33	4.81	7.39	4.22	4.50	4.30	4.14	5.97	4.78	4.56	4.88	4.35	4.75	5.01	4.15	4.45	4.69	4.83	5.43	NA
100	3.61	4.03	3.37	4.62	5.10	4.36	4.86	4.59	5.72	4.42	5.16	4.25	5.37	4.12	4.68	5.18	5.13	4.73	4.86	5.18	5.12	4.41	5.55	4.41	NA
150	4.25	4.27	4.94	4.24	4.81	4.54	5.92	4.48	4.79	4.20	4.54	4.64	4.60	6.28	4.81	4.69	5.55	5.09	4.43	4.80	5.79	4.31	4.31	4.09	NA
200	4.46	4.27	6.50	4.62	5.50	5.31	4.86	6.00	4.59	4.86	4.38	4.96	5.09	6.14	4.83	4.62	4.99	6.48	3.96	5.34	6.39	4.62	4.24	4.72	NA
300	4.39	4.47	5.23	4.90	4.21	4.09	4.73	3.92	5.27	7.41	4.06	4.61	4.25	4.85	4.81	4.59	5.20	5.60	4.23	4.22	5.34	4.31	4.31	4.61	NA
400	3.84	4.23	5.06	5.38	4.45	5.47	4.59	4.00	5.35	4.73	4.83	5.04	5.16	4.66	4.84	4.95	4.49	4.95	4.51	5.39	4.81	4.78	4.51	4.35	NA
500	4.60	4.92	3.91	4.17	4.21	5.13	4.41	4.69	4.71	5.78	4.03	5.39	4.39	6.70	4.65	4.61	5.90	4.76	4.88	4.69	5.26	5.14	4.88	4.41	NA

Fig. S19: Observed minimum p value (-$\log_{10}$ values) per dataset (columns; sorted by decreasing number of samples) per input set size (rows). Values are color coded to improve interpretability. We refer to Table S1 for cell type abbreviations.

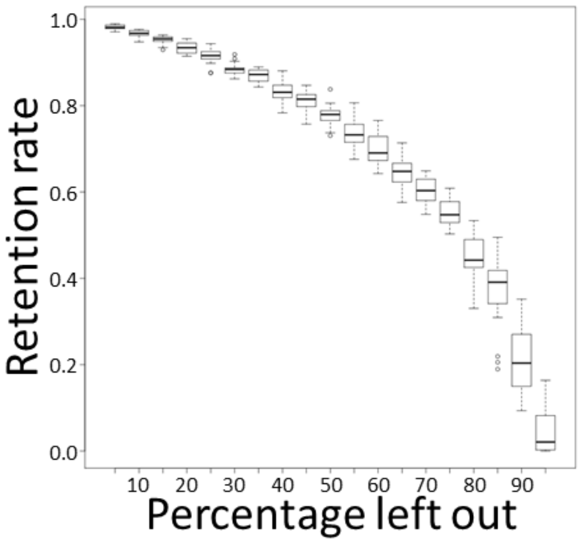


Fig. S20: Box plots showing the retention rates of "highly correlated genes" in function of the fraction of genes that were removed from the input (in percentages of the original set size).

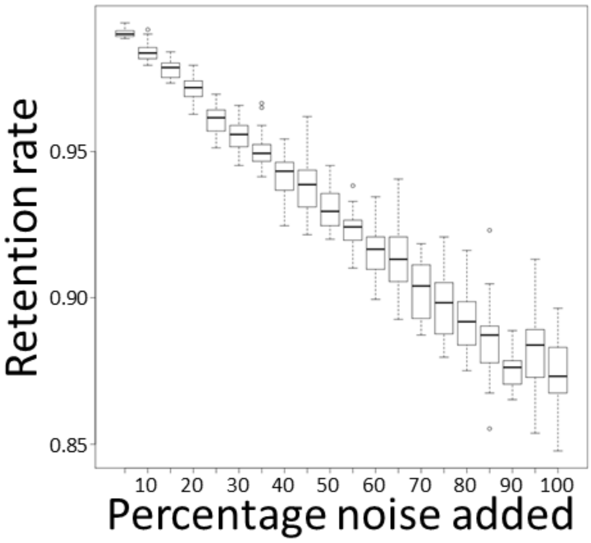


Fig. S21: Box plots showing the retention rates of “highly correlated genes” in function of the level of randomly selected genes added to the input gene set (in percentages of the original set size).

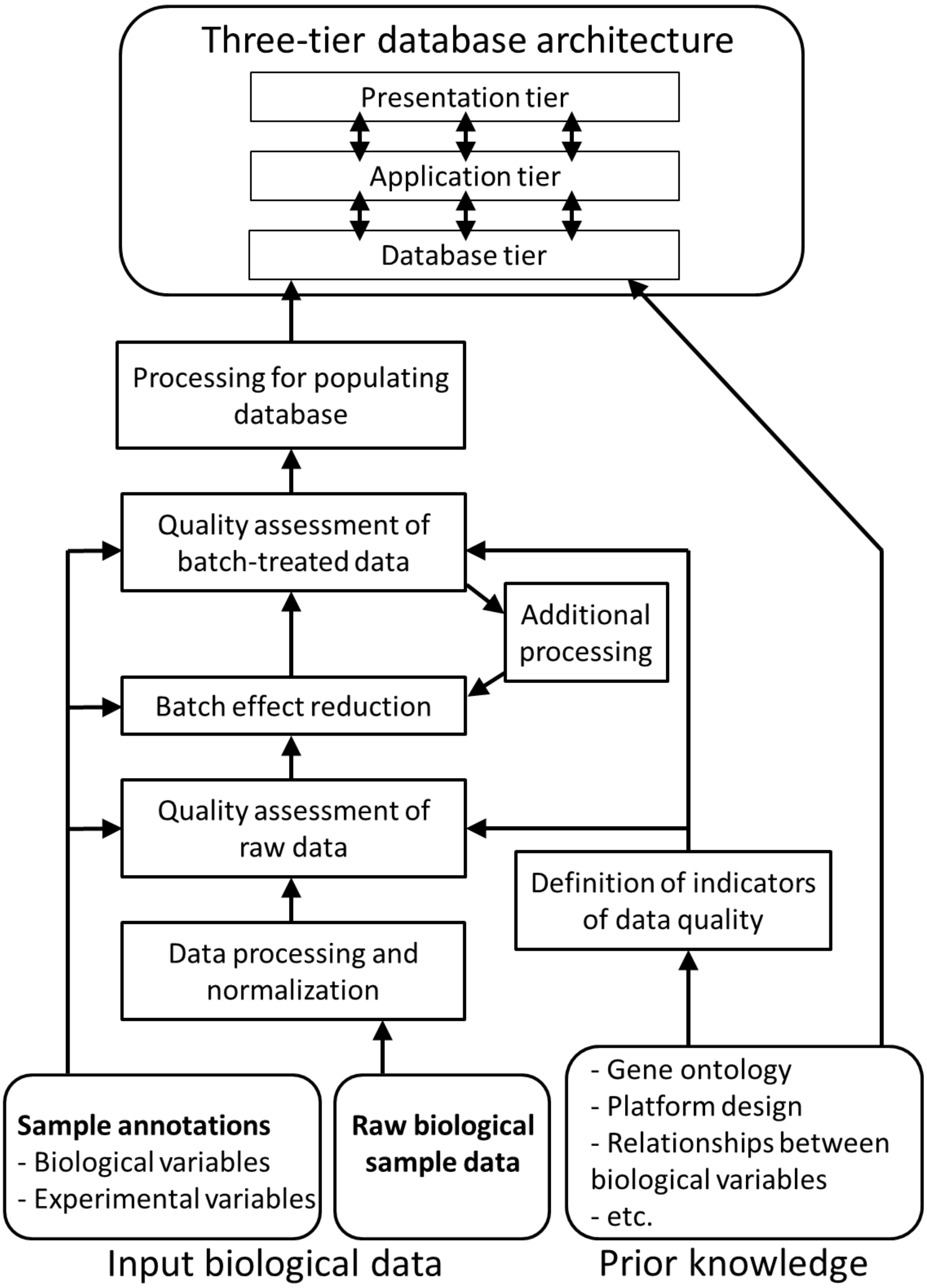

Three-tier database architecture
Presentation tier
Application tier
Database tier
Processing for populating database
Quality assessment of batch-treated data
Additional processing
Batch effect reduction
Quality assessment of raw data
Definition of indicators of data quality
Data processing and normalization
Sample annotations
- Biological variables
- Experimental variables
Raw biological sample data
- Gene ontology
- Platform design
- Relationships between biological variables
- etc.
Input biological data
Prior knowledge

Fig. S22: Flowchart summarizing the general strategy of our data processing approach, from the collection of input data to the population of a three-tier database. A description of the main steps is given in SI Appendix, section "General data analysis approach".