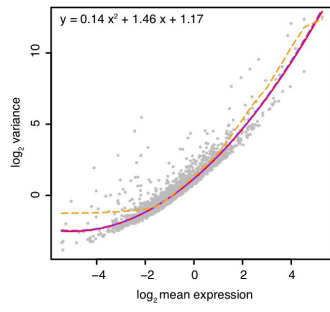
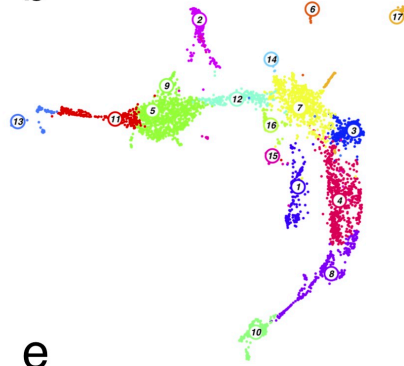
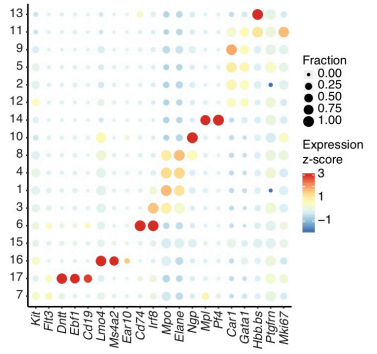
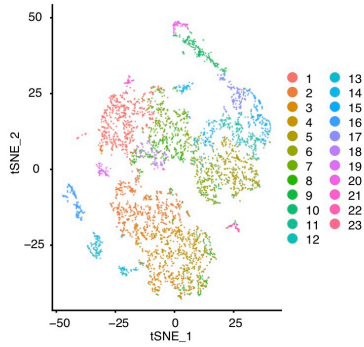
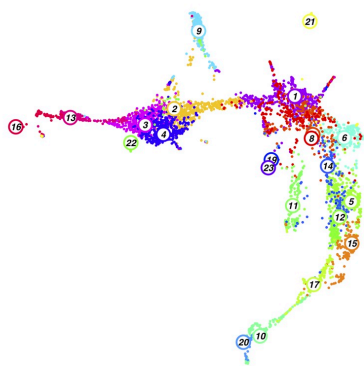
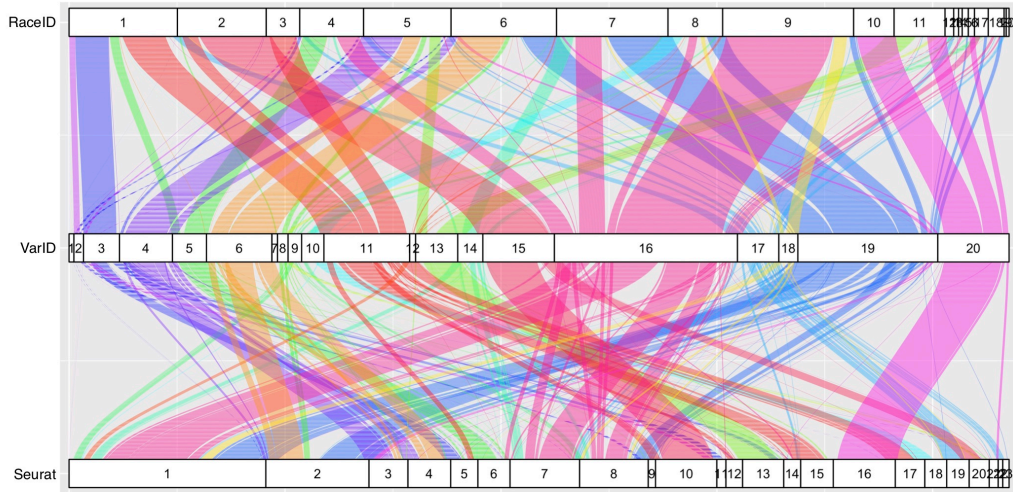
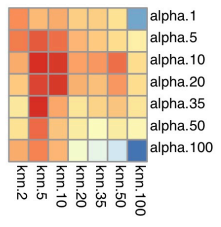
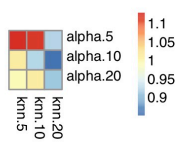
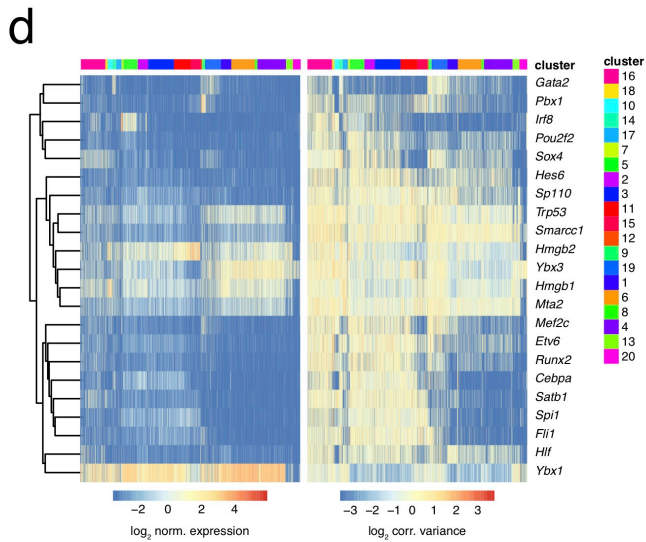
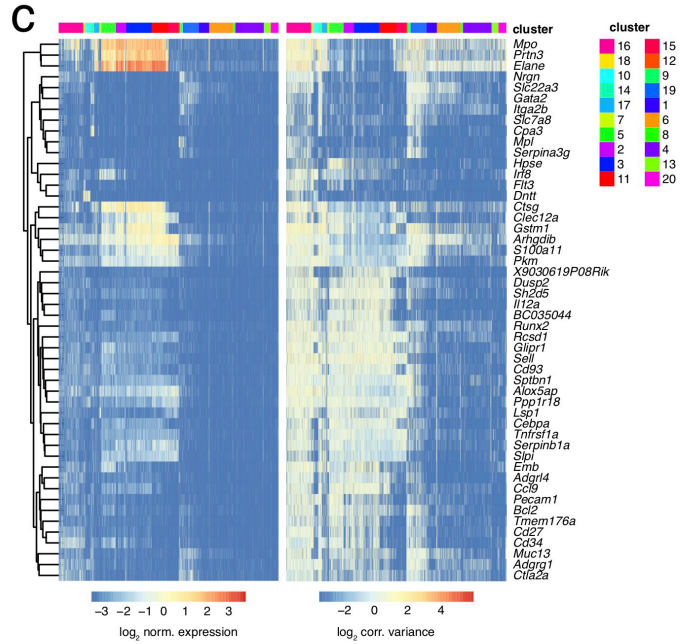
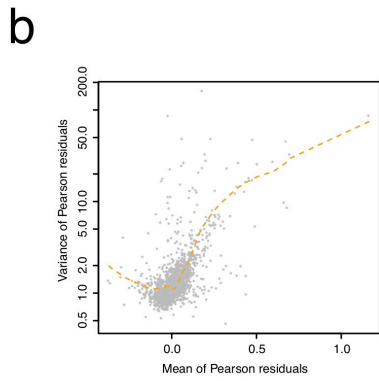
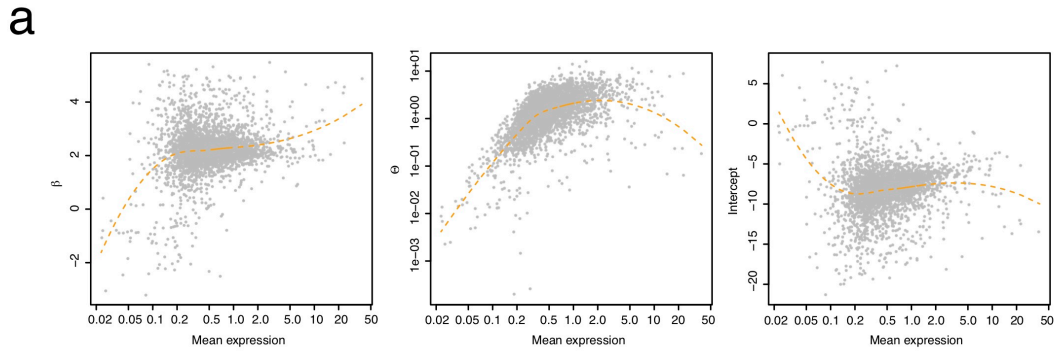


a**b****c****d****e****f****g****h**

Supplementary Figure 1

Hematopoietic cell type identification by Louvain clustering on knn networks.

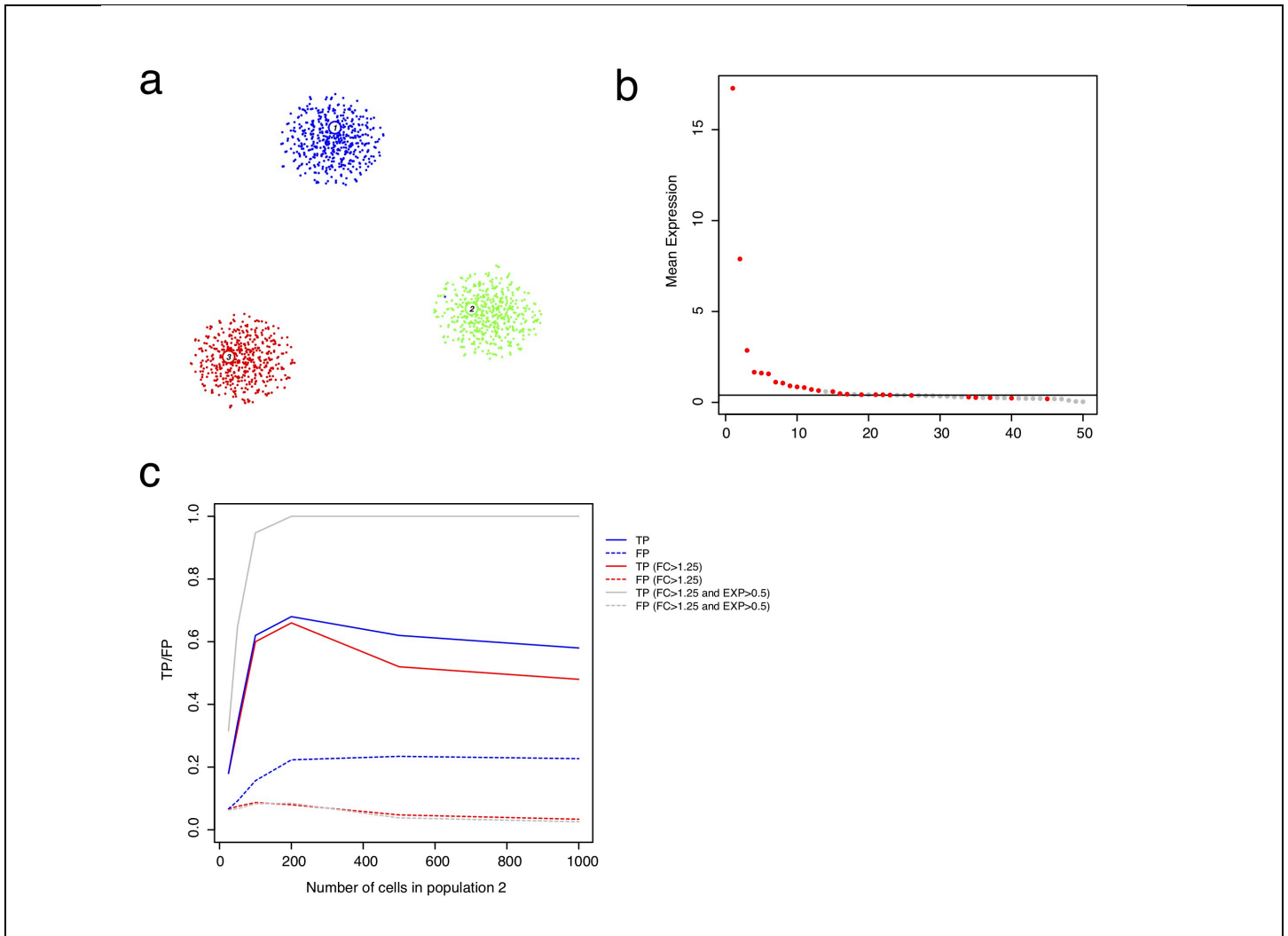
a, Scatterplot of variance and mean transcript count in logarithmic space for all genes across all cells in the mouse hematopoietic progenitor dataset. The red line indicates a second order polynomial fit to all genes. The blue line indicates the maximum deviation of the fit towards higher variability based on the error interval of the fitted coefficients. The broken orange line indicates a loess regression. The polynomial fit function is given at the top. **b**, UMAP representation highlighting clusters obtained by Louvain clustering on the full k-nearest neighbor network. **c**, Dot plot showing the expression z-score of lineage-specific marker genes across all clusters from (b). The dot size indicates the fraction of cells expressing a gene. **d**, t-SNE map of clustering output obtained from Seurat^{5,6} using high-resolution settings (Methods). **e**, UMAP representation highlighting clusters obtained with Seurat from (d). **f**, Alluvial diagram comparing the cluster composition obtained with VarID, RaceID3⁸, and Seurat. **g**, Evaluation of the resolution of rare populations as a function of α and knn within this dataset. I tested the overlap of inferred clusters with lymphoid progenitors (*Dntt*), B cells (*Ebfl*, *Cd19*), basophils (*Lmo4*, *Ms4a2*), eosinophils (*Ear10*), dendritic cells (*Cd74*), and megakaryocytes (*Mpl*, *Pf4*), based on expression of the corresponding marker genes (in parentheses). The fraction of cells in a cluster with positive transcript counts for the respective markers was computed for each cluster (termed enrichment), and the fraction of all marker-positive cells falling into that cluster (termed overlap). The clustering should maximize both the overlap and the enrichment. If a cluster perfectly recapitulates the marker expression domain, both values equal one. The heatmap shows the maximum of the product of overlap and enrichment across all clusters averaged across all marker genes as a function of the parameters, and supports $\alpha=10$ and knn=10 as an optimal parameter choice. Smaller values for knn would lead to higher variances of the variability estimates. **h**, The same analysis as in (g) was performed on a subset of parameters, either using a supplied distance matrix (1 – Pearson’s correlation coefficient) or the default method (Euclidean distance in PCA-space) for the knn search. The ratio of the overlap*enrichment product between the default and the correlation-based approach is shown in the heatmap and close to one for all parameter combinations. (a-h) Data from n=2 biologically independent experiments.



Supplementary Figure 2

Exploring local gene expression variability in hematopoietic progenitors.

a, Gene-specific parameter fit from the negative binomial generalized linear model with log link function and total transcript count of a cell as independent variable are shown in a scatterplot as a function of the mean expression. Robust parameter fits for the coefficient β , size factor θ , and intercept are obtained by a loess-regression of the parameter fits as a function of mean expression in order to share information between genes of similar expression (broken orange line). This method follows a recently published approach¹². **b**, Scatter plot of the variance of Pearson residuals from the generalized linear model fit as a function of the mean transcript expression in logarithmic space. The broken orange line represents a loess regression. **c**, Heatmap of normalized expression (left) and corrected variance (right) for the top 50 genes with enhanced variability from Figure 2d ordered by decreasing \log_2 -foldchange of variability between cluster 16 and the remaining cells. Clusters were manually grouped by lineage. Hierarchical clustering of rows was performed based on gene expression. **d**, Heatmap of normalized expression (left) and corrected variance (right) for all transcription factor genes with enhanced variability ordered by decreasing \log_2 -foldchange of variability between cluster 16 and the remaining cells (one-sided Wilcoxon rank sum-test $P < 0.001$, Benjamini Hochberg corrected, \log_2 -foldchange > 1.25). Clusters were manually grouped by lineage. Hierarchical clustering of rows was performed based on gene expression. (a-c) Data from $n=2$ biologically independent experiments.

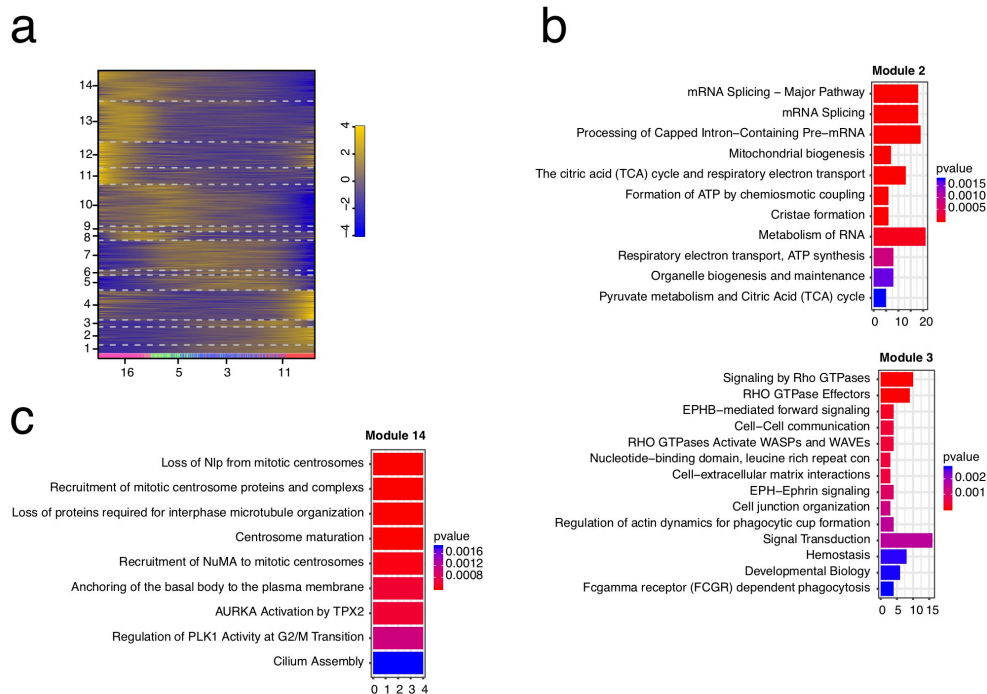


Supplementary Figure 3

Sensitivity and specificity for the identification of genes with enhanced variability.

a, Three populations were simulated with 500 cells each. The mean expression of population 1 was equal to all genes as measured in the hematopoietic dataset¹⁰. Gene expression was sampled from negative binomial distributions with size factors determined from the mean-variance relation in Figure 2a. Population 2 was generated from the same mean expression values after 5-fold up- or down-regulation of 100 genes each. Population 3 was simulated accordingly after 5-fold up- or down-regulation of 100 genes from population 2. For population 2 the variance of 50 genes taken from Fig. 2a was increased two-fold to simulate enhanced variability. The t-SNE map depicts the three populations resolved by VarID into three clusters. **b**, The plot shows the genes (n=50) with simulated noise differences ordered by average expression. If differentially variable genes are called by VarID in population 2 versus 1 and 3 with a fold change cut-off of >1.25 and one-sided Wilcoxon test $P<0.001$, the true positive rate is 52% at a false positive rate of 5%. The true positives are highlighted in red. Applying an average expression cut-off of >0.5 increases the true positive rate to 1 at a false positive rate of 4%. The solid black line indicates an average expression of 0.4. We note that a cut-off on the variability fold change is required to control for the false positive rate, since significant differences in variability can be induced by few tail events, i.e. cells with positive transcript counts for a lowly expressed genes, since these events affect a larger number of neighborhoods (determined by knn). The unconstrained false positive rate is $\sim 23\%$. I thus recommend applying a fold change cut-off of >1.25 , which I use throughout the manuscript. **c**, To test the dependence of sensitivity and specificity on the number of cells I varied the size of population 2 between 20 and 1,000 cells. The plot shows the true positive rate (solid lines) and false positive rate (broken line) as a function of the size of population 2. Rates were computed without filtering, after applying a variability fold change

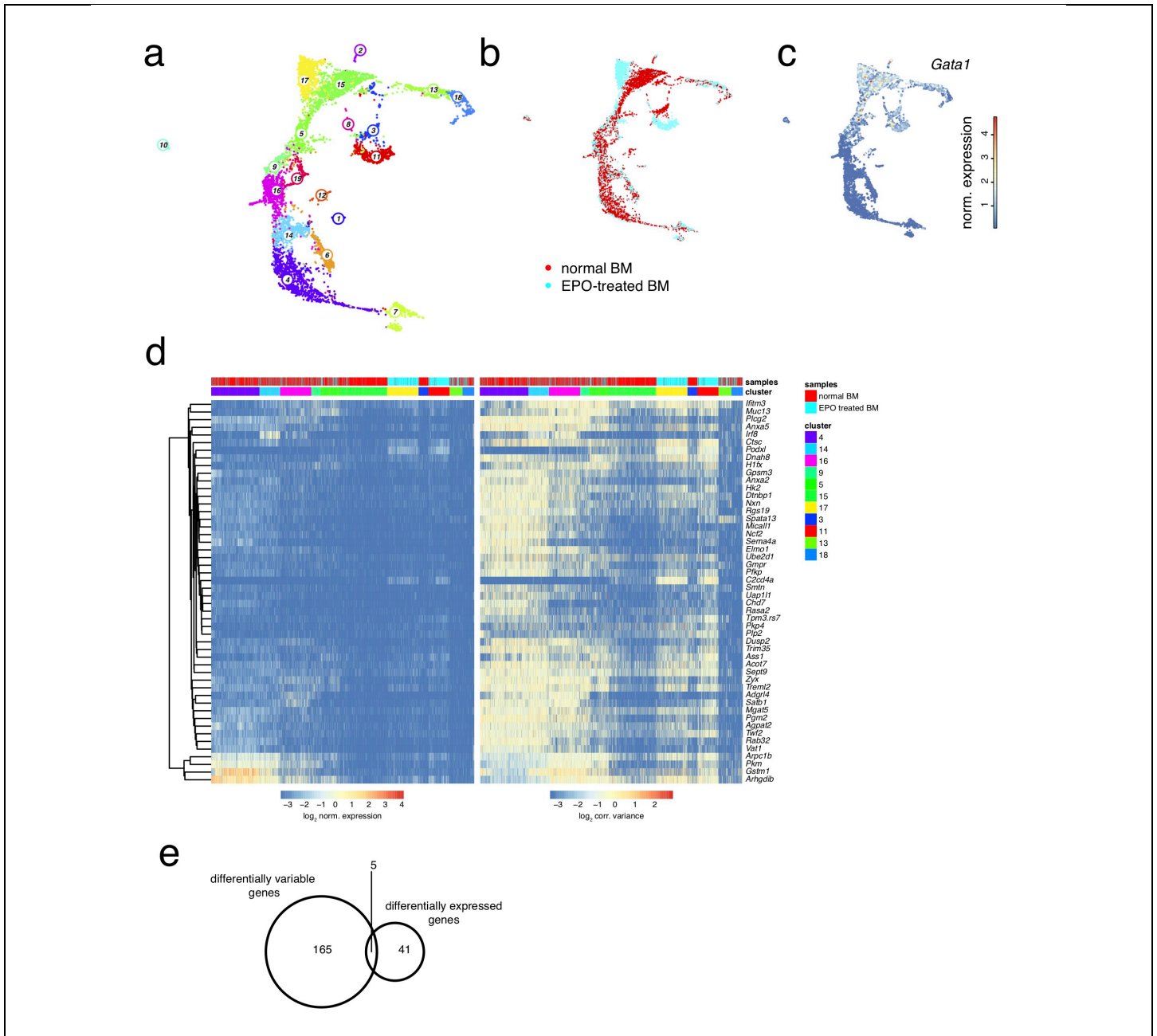
cut-off ($FC > 1.25$) and after applying an additional average expression cut-off ($EXP > 0.5$). While rates saturate beyond a population size of ~ 200 , sensitivity drops at small population sizes. For 50 cells, I observed a true positive rate of 32% at a false positive rate of 7% (64% and 7%, respectively, at an average expression cut-off of > 0.5) with a fold change cut-off > 1.25 .



Supplementary Figure 4

Characterization of co-expressed and co-varying genes during neutrophil differentiation.

a, Self-organizing map (SOM) of pseudo-temporal gene expression profiles inferred by FateID⁸. The color indicates the z-score of loess-smoothed profiles. Cells were ordered along the trajectory connecting clusters 5, 4, 3, 7, 1, and 2 in (Fig. 3a) by StemID2. Original clusters (cf. Fig. 1b) are highlighted at the bottom. Modules were obtained by grouping SOM nodes based on correlation (Pearson correlation > 0.85). Only modules with >10 genes are shown in the map. Genes with >2 transcripts in at least one cell were included. Data from n=2 biologically independent experiments. **b**, Reactome pathway analysis²² revealing enriched pathways in module 2 (n=112 genes) and module 3 (n=55 genes) (hypergeometric test $P < 0.05$, Methods) of SOM in Fig. 3b. **c**, Reactome pathway analysis revealing enriched pathways in module 14 (n=45 genes, hypergeometric test $P < 0.05$, Methods) of (a). (b,c) The x-axis shows the number of genes of a particular pathway present in the module. The gene universe comprised n=3,439 expressed genes.

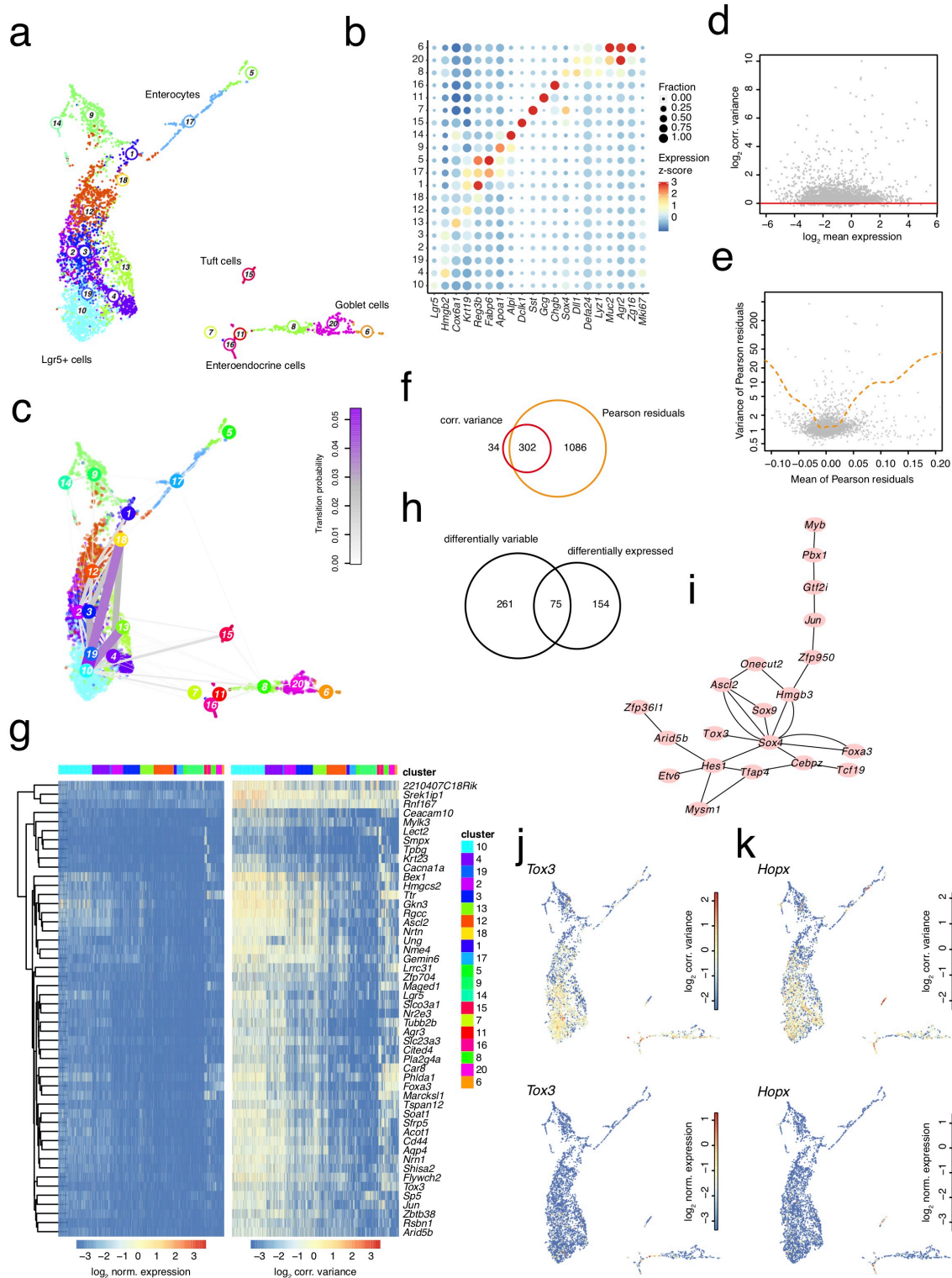


Supplementary Figure 5

EPO-stimulation of murine bone marrow cells leads to variable expression of innate immune genes in erythrocyte progenitors.

a, UMAP representation of combined EPO-stimulated and normal mouse hematopoietic progenitor single-cell RNA-seq data¹⁰ highlighting clusters inferred by Louvain clustering on the pruned knn network ($k=10$ and $\alpha=10$). **b**, UMAP representation indicating the sample of origin for each single-cell transcriptome. Only for the erythrocyte progenitor branch a separation of the samples is observed. **c**, UMAP highlighting expression of the erythrocyte progenitor marker gene *Gata1*. **d**, Heatmap of normalized expression (left) and corrected variance (right) for the top 50 genes with enhanced variability in cluster 17 ordered by decreasing \log_2 -foldchange of variability between cluster 17 and 15 (one-sided Wilcoxon rank sum-test $P < 0.001$, Benjamini Hochberg corrected, foldchange > 1.25). Clusters were manually grouped by lineage. Hierarchical

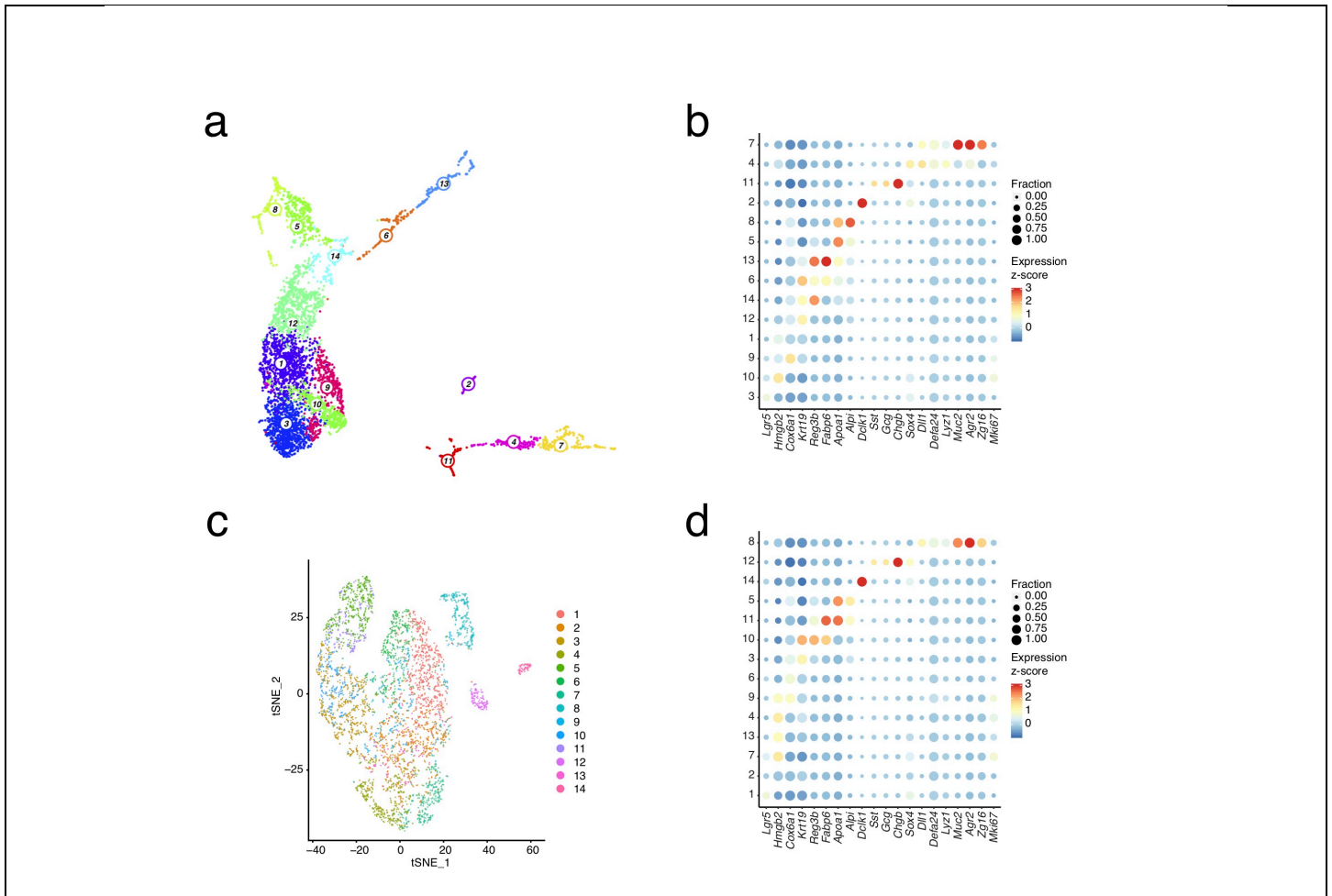
clustering of rows was performed based on gene expression. Pathway enrichment analysis revealed that 39 out of 170 differentially variable genes were annotated within the pathway “Innate Immune System”, adjusted (hypergeometric test $P=0.002$, Methods). **e**, Venn diagram showing the overlap of genes with enhanced local variability (one-sided Wilcoxon rank sum-test $P<0.001$, Benjamini Hochberg corrected, foldchange >1.25) and differentially expressed genes ($P<0.001$, Benjamini Hochberg corrected, see Methods, foldchange >1.25 between the populations) in cluster 17 versus 15. (a-e) Data from $n=2$ biologically independent experiments.



Supplementary Figure 6

Intestinal stem cells exhibit stochastic expression of secretory lineage transcription factors.

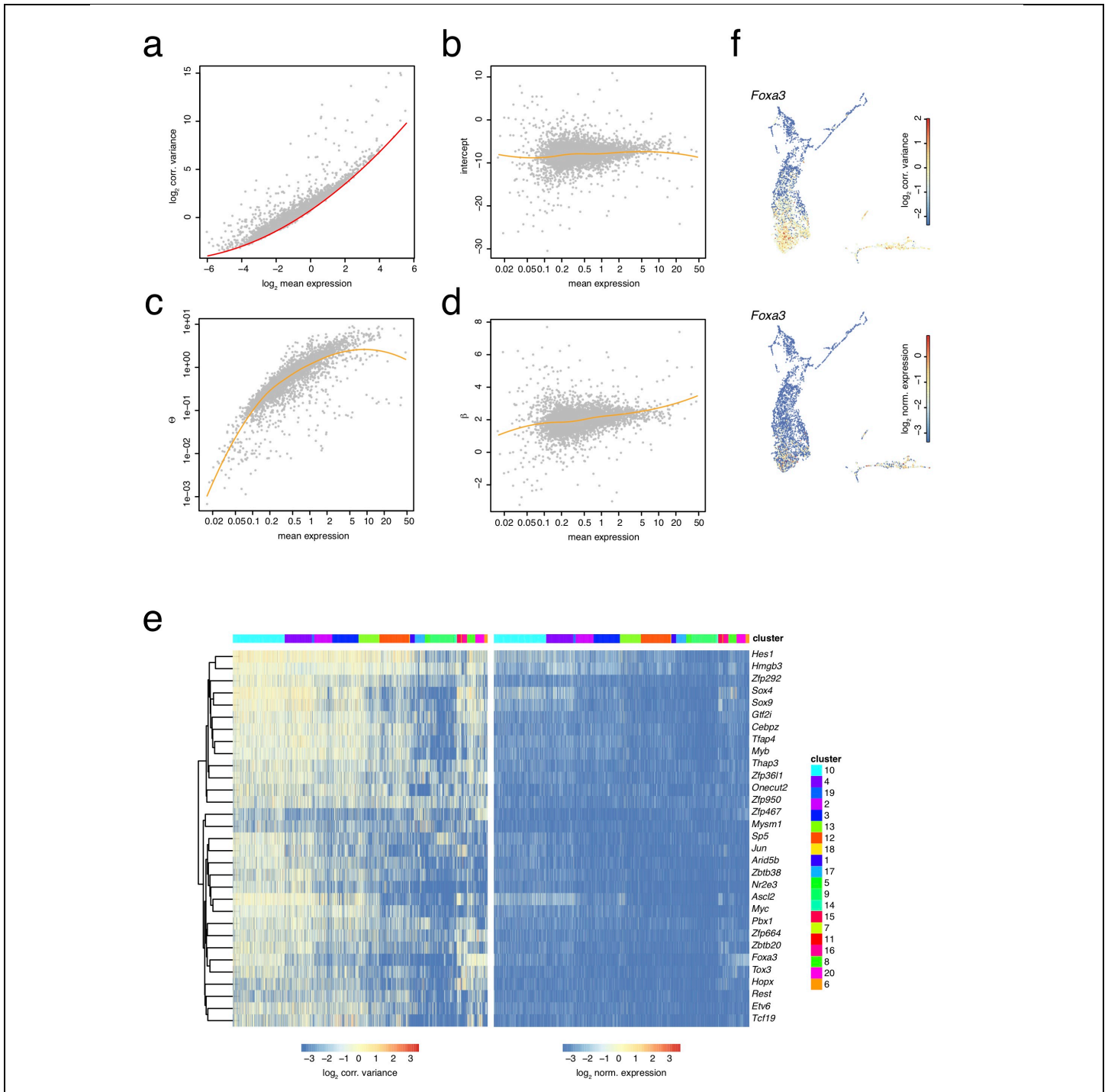
a, UMAP representation of mouse intestinal epithelial single-cell RNA-seq data²⁰ highlighting clusters inferred by Louvain clustering on the pruned knn network (k=10). Cell type labels are based on marker gene expression. **b**, Dot plot showing the expression z-score of lineage-specific marker genes across all clusters from (a). The dot size indicates the fraction of cells expressing a gene. **c**, UMAP representation with links connecting cluster medoids. The thickness and color of a link indicates the transition probability between the connected clusters. **d**, Scatterplot showing corrected variance of transcript counts as a function of the mean in logarithmic space after eliminating the mean-dependence by subtracting the baseline fit. The red line indicates the baseline level of the corrected variability. **e**, Scatter plot of the variance of Pearson residuals from a negative binomial generalized linear model with log link function and the total transcript count of a cell as independent variable as a function of the mean transcript expression in logarithmic space. The broken orange line represents a loess regression. Highly variable outliers at low and high expression are not visible, since the plot shows a zoom-in to increase visibility. **f**, Venn diagram showing the overlap of genes with enhanced local variability in cluster 10 versus the remaining cells as predicted after correcting the variance or computing the variance of Pearson residuals (one-sided Wilcoxon rank sum-test $P < 0.001$, Benjamini Hochberg corrected, foldchange > 1.25). **g**, Heatmap of normalized expression (left) and corrected variance (right) for the top 50 genes with enhanced variability ordered by decreasing \log_2 -foldchange of variability between cluster 10 and the remaining cells (one-sided Wilcoxon rank sum-test $P < 0.001$, Benjamini Hochberg corrected, see Methods, \log_2 -foldchange > 1.25). Clusters were manually grouped by lineage. Hierarchical clustering of rows was performed based on gene expression. **h**, Venn diagram showing the overlap of genes with enhanced local variability (one-sided Wilcoxon rank sum-test $P < 0.001$, Benjamini Hochberg corrected, foldchange > 1.25) and differentially expressed genes ($P < 0.001$, Benjamini Hochberg corrected, see Methods, foldchange > 1.25 between the populations) in cluster 10 versus the remaining cells. **i**, Gene regulatory network predicted by GENIE3 run on all transcription factors among the genes with enhanced variability, using the full dataset as input. **j**, UMAP representation highlighting corrected variability (upper panel) and normalized gene expression (lower panel) for *Tox3*. **k**, UMAP representation highlighting corrected variability (upper panel) and normalized gene expression (lower panel) for *Hopx*.



Supplementary Figure 7

Intestinal cell type identification by Louvain clustering on knn networks.

a, UMAP representation highlighting clusters obtained by Louvain clustering on the full knn network. **b**, Dot plot showing the expression z-score of lineage-specific marker genes across all clusters from (a). The dot size indicates the fraction of cells expressing the gene. **c**, t-SNE map of clustering output obtained from Seurat^{5,6} using high-resolution settings (Methods). **d**, Dot plot showing the expression z-score of lineage-specific marker genes across all Seurat clusters from (c). (a-d) Data from n=4 animals.



Supplementary Figure 8

Exploring local variability in intestinal epithelial stem cells.

a, Scatterplot showing variance and mean of the transcript count of all genes across all cells in the mouse intestinal dataset in logarithmic space. The red line indicates a second order polynomial fit to the baseline level of the variance comprising technical and biological variability. (b-d) Gene-specific parameter fits from the negative binomial generalized linear model with log link function and total transcript count of a cell as independent variable are shown in a scatterplot as a function of

the mean expression. Robust parameter fits for the intercept (**b**), the size factor θ (**c**), and coefficient β (**d**) are obtained by a loess-regression of the parameter fits as a function of mean expression in order to share information between genes of similar expression (broken orange line). This method follows a recently published approach¹². **e**, Heatmap of normalized expression (left) and corrected variance (right) for all transcription factor genes with enhanced variability ordered by decreasing \log_2 -foldchange of variability between cluster 10 and the remaining cells (one-sided Wilcoxon rank sum-test $P < 0.001$, Benjamini Hochberg corrected, \log_2 -foldchange > 1.25). Clusters were manually grouped by lineage. Hierarchical clustering of rows was performed based on gene expression. **f**, UMAP representation highlighting corrected variability (upper panel) and normalized gene expression (lower panel) for *Foxa3*. (a-f) Data from n=4 animals.