To decipher cell state transitions from single-cell transcriptomes it is crucial to quantify weak expression of lineage-determining factors, which requires computational methods that are sensitive to the variability of weakly expressed genes. Here, I introduce VarID, a computational method that identifies locally homogenous neighborhoods in cell state space, permitting the quantification of local variability in gene expression. VarID delineates neighborhoods with differential gene expression variability and reveals pseudo-temporal dynamics of variability during differentiation.

With the emergence of a growing number of single-cell sequencing technologies, the ability to decipher the cell type composition of complex tissues is rapidly improving. Single-cell transcriptomes can reveal manifolds in cell state space representing trajectories of cell state transitions. It is of core interest to understand the molecular control of these transitions, but the investigation of the transcription factor and signaling networks underpinning cell state transitions is frequently hindered by the low and highly variable expression of these classes of genes. Since differences in the expression of such genes are difficult to detect owing to technical and biological noise, I here introduce VarID, a method for the inference of local variability. Increased local variability could indicate the onset of expression in local neighborhoods or the response to fluctuating signaling inputs from the microenvironment. Available methods for the inference of noise parameters were not designed for complex mixtures of cell types and do not permit the local estimation of variability.

A fundamental challenge is the definition of local neighborhoods in cell state space, since admixtures of distinct cell types or states could inflate the variability estimates. Since k-nearest neighbor (KNN) networks have been used successfully for the inference of cell types and differentiation trajectories, I reasoned that the k-nearest neighborhood would be a useful starting point. I devised a statistical test to determine whether the expression levels of all genes for each neighbor are in accordance with the expected distribution of the ‘central’ cell. It was previously demonstrated that unique molecular identifier (UMI)-derived transcript counts are well described by a negative binomial distribution, which is uniquely determined by mean and variance. I thus computed a local mean by averaging expression across the central cell and its KNNs with weights determined by their similarity to the central cell (Methods). An additional parameter, α, can be varied to adjust the degree of locality. I next determined the variance associated with the local mean estimate from a global background distribution. As I showed previously, the mean–variance relation in logarithmic space is well described by a second order polynomial, robustly averaging across genes of similar mean expression (Supplementary Fig. 1a). Hence, a local mean allows the local background distributions to be defined for all genes, and links to any of the KNNs with expression levels not explained by this distribution are discarded (Fig. 1a). The resulting pruned KNN network thus connects only locally homogenous neighborhoods.

To identify distinct cell states and types I applied Louvain density clustering to the pruned network. To demonstrate increased sensitivity of cell type detection when using the pruned network, I analyzed murine hematopoietic progenitor single-cell transcriptomes (Fig. 1b,c). I recovered all lineages described in the original study, and resolved additional subpopulations such as Mpl<sup>lo</sup> versus P<sub>2</sub> megakaryocyte states, Ebf1<sup>lo</sup> pro-B cells and Dntt<sup>lo</sup> progenitors, and eosinophils (Fig. 1b,c). These subpopulations remain unresolved when clustering is performed on the full network (Supplementary Fig. 1b,c) or when Seurat<sup>lo</sup> analysis is performed (Fig. 1d and Supplementary Fig. 1d–f). As the clustering depends on the choice of parameters α and KNN, I evaluated the resolution of rare populations within this dataset (that is, lymphoid progenitors, B cells, basophils, eosinophils, dendritic cells, and megakaryocytes) based on the resolution of the expression domains of corresponding marker genes (Supplementary Fig. 1g). This analysis supports α = 10 and KNN = 10 as an optimal parameter choice. I observed similar clustering performance when determining KNNs with a supplied Pearson's correlation-based distance matrix and when using the default method based on Euclidean distances in principal component analysis (PCA) space (Supplementary Fig. 1h and Methods).

I next predicted transition probabilities between the inferred clusters on the pruned KNN network. Assuming a random starting cell within a given cluster, the probability to transition into another cluster within a single step on the network can be computed (Methods). These probabilities were in good agreement with known differentiation pathways (Fig. 1e): multipotent progenitors (cluster 16) were directly linked to megakaryocytes, dendritic cells, basophils, monocytes, and the major branches of erythrocytes and neutrophils.

To explore differences in differentially variable genes between cell states, I derived estimates of gene expression variability in local neighborhoods on the pruned KNN graph. To account for the convex variance-mean dependence in logarithmic space as a consequence of biological and technical noise<sup>2–4,11</sup> (Fig. 2a), I fitted a second order polynomial to the baseline level of the combined technical and biological variability (Methods). This allowed the systematic baseline mean-dependence to be regressed out and the corrected variability estimates to be directly compared between neighborhoods (Fig. 2b). As an alternative approach, I followed a recently published method based on a negative binomial generalized linear model with the total transcript count of each cell as an independent variable<sup>14</sup>. After averaging regression parameters across genes of similar mean expression (Methods and Supplementary Fig. 2), the

---

1Max-Planck-Institute of Immunobiology and Epigenetics, Freiburg, Germany. 2Centre for Integrative Biological Signaling Studies, University of Freiburg, Freiburg, Germany. *e-mail: gruen@ie-freiburg.mpg.de
variance of the Pearson residuals should, in theory, be independent of the mean expression.

To test the sensitivity and specificity of VarID for the detection of genes with enhanced variability, I performed a simulation experiment, which revealed that variability fold changes greater than 1.25 can be detected at a false positive rate of around 5% and a true positive rate of more than 50%, depending on the average expression (Supplementary Fig. 3).

To explore differences in the variability of gene expression across cell states, I inferred local estimates of the corrected variability for the murine hematopoietic progenitors using the first approach (that is, corrected variance) (Fig. 2a,b), since a residual mean-variance dependence remained for the second approach (that is, negative binomial generalized linear model) (Supplementary Fig. 2a,b). I noticed that increased local variability was frequently associated with the onset of lineage markers in multipotent progenitors (cluster 16)—for example, the early erythrocyte lineage transcription factor Gata1 or the neutrophil marker Mpo (Fig. 2c). However, although the corrected variability remains high in the case of Gata1 throughout erythrocyte differentiation, it becomes strongly suppressed for Mpo with increasing expression during neutrophil differentiation (Fig. 2c), indicating the gene-specific dynamics of expression variability.

I next extracted all genes with increased local variability within the multipotent progenitor population (cluster 16) in comparison to the remaining populations (one-sided Wilcoxon rank-sum test $P < 0.001$, Benjamini-Hochberg corrected, fold change $>1.25$) (Supplementary Fig. 2c). Differentially variable genes exhibited only limited overlap with differentially expressed genes ($P < 0.001$, Benjamini-Hochberg corrected, see Methods, fold change $>1.25$).
between the populations, Fig. 2d). Comparing the corrected variability of the top 50 variable genes with their expression across cell clusters revealed groups of genes with stochastic expression in cluster 16 and markedly increasing expression, for example, on the neutrophil branch (such as Mpo, Prtn3, or Elane), and classes of genes that are also most highly expressed in cluster 16 (such as Fli3, Cd27, Cd34, and Il12a). To investigate the variability of transcriptional regulators relevant for lineage decisions, I selected all transcriptional regulators from the list of significantly variable genes in cluster 16 (P<0.001, Benjamini–Hochberg corrected, see Methods, fold change >1.25 between cluster 16 and the remaining population) and predicted a regulatory network by running GENIE3 (ref. 14) (Methods, Fig. 2e, and Supplementary Fig. 2d). This network recovered modules associated with hematopoietic stem cells (HSCs) comprising Runx2 (ref. 15) and Hlf4, the megakaryocyte lineage (Pbx1, Fli1, Mef2c17, the lymphoid lineage (Sabb1 (ref. 18) Etv6 (ref. 19)), and monocyte differentiation (Sp110, Ifr8)20, indicating variable activity of lineage-associated transcription factors in multipotent progenitors, potentially indicating the onset of lineage bias.

To investigate the dynamics of variability during differentiation, I focused on the neutrophil branch and inferred a pseudo-temporal ordering of single-cell transcriptomes with StemID2 (ref. 20) (Fig. 3a). I then ordered the pseudo-temporal profiles of gene expression (Supplementary Fig. 4a) and of the corrected variability (Fig. 3b) into co-expressed and co-variable modules, respectively, using self-organizing maps as implemented in FateID (ref. 1). I observed modules with distinct variability profiles, such as genes with increased variability at naïve and mature states (for example, modules 1 and 8), or during intermediate stages (for example, module 11). Modules with similar dynamics of variability did not necessarily exhibit similar gene expression dynamics (Fig. 3c and Supplementary Fig. 4a). Of note, particular modules were enriched in specific functions (Methods and Supplementary Fig. 4b,c), suggesting that regulation of variability depends on functional pathways.

To investigate the impact of perturbation on gene expression variability, I co-analyzed hematopoietic cells sequenced from bone marrow after 48 h of erythropoietin (EPO) stimulation21 together with the cells sequenced from normal bone marrow. EPO stimulation leads to an expansion of the erythroid lineage at the expense of the other lineages19. My analysis confirmed that transcriptome changes following EPO stimulation affected only the erythroid lineage (Supplementary Fig. 5a–c), and revealed an enrichment of innate immunity pathways among genes with increased variability in EPO-stimulated versus normal erythrocyte progenitors (Supplementary Fig. 5d) (hypergeometric test P=0.002, Methods). This finding suggests that progenitors of other lineages could indeed be diverted towards the erythrocyte fate following EPO stimulation. Importantly, there was only marginal overlap with differentially expressed genes, and those differentially expressed genes do not exhibit a functional enrichment other than for ribosomal RNA processing (Supplementary Fig. 5e).

Finally, application of VarID to murine intestinal epithelial cells22 revealed the stochastic activity of secretory lineage transcription factors in Lgr5-positive intestinal stem cells, suggesting the existence of secretory fate-biased stem cells (Supplementary Results and Supplementary Figs. 6–8).

In conclusion, by quantifying the dynamics of gene expression variability, VarID reveals differential variability of lineage regulators involved in cell state transition and facilitates the investigation of the molecular control of fate decision by single-cell RNA sequencing (RNA-seq).
**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-019-0632-3.

Received: 11 June 2019; Accepted: 8 October 2019; Published online: 18 November 2019

**References**

1. Grün, D. Revealing routes of cellular differentiation by single-cell RNA-seq. *Curr. Opin. Syst. Biol.* **11**, 9–17 (2018).
2. Grün, D., Keister, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
3. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* **11**, e1004333 (2015).
4. Eling, N., Richard, A. C., Richardson, S., Marioni, J. C. & Vallejos, C. A. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst.* **7**, 284–294 (2018).
5. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
6. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
7. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
8. Herman, J. S. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
9. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
10. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).

---

**Fig. 3 | Exploring dynamics of gene expression variability during neutrophil differentiation.**

**a.** Separate RaceID3 and StemID2 (ref. 8) analysis of all cells from the original clusters 16, 5, 3, and 11. The link color indicates the link $P$ value, and the vertex color represents transcriptome entropy. The link $P$ value and transcriptome entropies were derived by StemID2 (ref. 8). MPP, multipotent progenitor population. **b.** Self-organizing map (SOM) of pseudo-temporal corrected variability profiles inferred by FateID (ref. 8) using the variability matrix as input. 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 a by StemID2. Original clusters (see Fig. 1b) are highlighted at the bottom. Modules were obtained by grouping SOM nodes based on correlation of averaged profiles (Pearson correlation $>0.85$). Only modules with more than ten genes are shown in the map. Genes with more than two transcripts in at least one cell were included. **c.** Pseudo-temporal variability (left) and corresponding gene expression (right) profiles averaged across all genes in a module. Pseudo-temporal profiles were normalized to the same scale by dividing transcript counts and corrected variabilities by the sum across all cells on the trajectory. **a–c.** Data from $n=2$ biologically independent experiments.
11. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10, 1093–1095 (2013).

12. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Preprint bioRxiv at https://doi.org/10.1101/576827 (2019).

13. Hu, H. et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* 47, D33–D38 (2019).

14. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 5, e12776 (2010).

15. Liting, X., Gerstein, R., Socolovsky, M. & Castilla, L. H. Deletion of core binding factors Runx1 and Runx2 leads to perturbed hematopoiesis in multiple lineages. *Blood* 122, 46 (2013).

16. Komorowska, K. et al. Hepatic leukemia factor maintains quiescence of hematopoietic stem cells and protects the stem cell pool during regeneration. *Cell Rep.* 21, 3514–3523 (2017).

17. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* 163, 1663–1677 (2015).

18. Doi, Y. et al. SATB1 expression marks lymphoid-lineage biased hematopoietic stem cells in mouse bone marrow. *Blood* 126, 2356 (2015).

19. Jones, C. L. et al. ETV6 regulates Pax5 expression in early B cell development. *Blood* 128, 2655 (2016).

20. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* 551, 333–339 (2017).

21. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv https://arxiv.org/abs/1802.03426v2 (2018).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

The VarID method. Inference of a pruned KNN network. The first step of VarID is the inference of a KNN network. This network can be constructed based on different metrics. As one alternative, a user-defined distance metric can be provided, or directly computed by VarID (for example, by using the Euclidean metric, or Spearman’s or Pearson’s correlation). Since, for datasets with tens of thousands of cells, computation and storage of a distance matrix is prohibitive owing to massive memory requirements, VarID provides an alternative approach. After an initial PCA to achieve dimensionality reduction, a fast KNN search was performed based on the Euclidean metric in PCA space. The number of principal components used can be specified and was set to 100 by default to ensure that the major variability was captured. I recommend keeping this default setting. Since the memory requirement for distance matrices of n cells scales with O(n^2) and the fast KNN search (using the FNN R-package v1.1.3) scales with O(n), the difference in memory requirement will be substantial for large datasets.

To eliminate the effect of cell-to-cell variability in total transcript counts, or sequencing depth, on the dimensional reduction and the downstream analysis, an optional regression with a negative binomial error model by a generalized linear model was computed, with the total transcript count of a cell as an independent variable, following a recently proposed method. If \( x_i \) is the transcript count of gene \( i \) in cell \( j \) (\( j = 1, \ldots, N \)), I computed a negative binomial generalized linear model:

\[
\log E(x_{ij}) = \beta_0 + \beta_1 \log m_{ij} + \eta_j = \sum_{k=1}^{N} x_{jk} \quad i = 1, \ldots, n \quad j = 1, \ldots, N
\]

with a log link function. The negative binomial distribution is over-dispersed and has been shown to be suitable for modeling technical and biological noise in single-cell RNA-seq data. The dispersion parameter \( \theta \) was estimated during the regression in addition to the intercept \( \beta_0 \) and the coefficient \( \beta_1 \). \( \eta_j \) determines the deviation of mean and variance \( \mu_j \) and \( \gamma_j \)

\[
\sigma_j^2 = \mu_j + \frac{\mu_j^2}{\theta_j}
\]

Following a similar procedure to that used by Hafemeister and Satija, information was shared between genes by a locally weighted scatter plot smoothing (loess):

\[
\beta_0 \to \beta_0(m) \quad \beta_1 \to \beta_1(m) \quad \eta \to \eta(m)
\]

resulting in the dependence of the parameters solely on the expression level \( m \).

The resulting KNN network was subject to pruning in the next step. For this purpose, a background model of the combined technical and biological variability was defined, using raw transcript counts as input. The variance \( \gamma_j \) and the mean \( m_j \) across the entire dataset were computed for each gene \( j \), and the variance-mean dependence across all genes was fitted by a second order polynomial after log-transformation to obtain a function \( v_j \) capturing the average dependence of the expression variability on the mean expression \( m_j \):

\[
v_j(m) \sim \exp\left(\alpha + \log(m) + \beta \log(m)^2\right)
\]

following a similar approach to that previously implemented in RaceID, to share information across genes with similar expression levels. The variance derived from this function fit for a fixed mean uniquely defines a negative binomial distribution, which serves as a background model:

\[
f(x_{ij}, \mu_j) = \text{NB}\left(x_{ij}; \mu_j, \gamma_j \right)\]

For every cell \( j \) a background model was inferred based on the local mean \( \mu_j \) for each gene \( i \). To account for the impact of sampling noise and to avoid skewing of the mean estimate by neighbors sampled from a distinct distribution which serves as a background model:

\[
\log E(x_{ij}) = \beta_0 + \beta_1 \log m_{ij} + \eta_j = \sum_{k=1}^{N} x_{jk} \quad i = 1, \ldots, n \quad j = 1, \ldots, N
\]

This function fit for a fixed mean uniquely defines a negative binomial distribution, on different metrics. As one alternative, a user-defined distance matrix can be provided, or directly computed by VarID (for example, by using the Euclidean metric, or Spearman’s or Pearson’s correlation).

Estimating local variability. The main goal of VarID is the quantification of local variability. As an alternative approach, the local variability \( \nu_{ij} \) of gene \( i \) in the neighborhood of cell \( j \) was computed as the variance of the transcript counts across all bins (unlike in equation (5)) to obtain a function \( v_j \) capturing the baseline variability as a function of mean expression \( m_j \):

\[
v_j(m) \sim \exp\left(\alpha + \log(m) + \beta \log(m)^2\right)
\]

The local mean \( \mu_j \) denotes the vector of mean expression values \( \mu_i \) for all genes \( i \) in cell \( j \) and uniquely defines a local transcript count distribution based on the inferred variance-mean relation (equation (5)). For each of the KNNs, the probability of the observed transcript count was computed for every gene from this local distribution. More precisely, for every gene the hypothesis that the observed expression is explained by the respective distribution was tested, and the \( P \) value for rejecting this hypothesis was computed as the probability of residing in one of the two tails of the distribution (that is, a two-sided test is performed). The total number of null hypotheses therefore corresponds to the number of tested genes. To control for the family-wise error rate at a given \( P \) value threshold, a Bonferroni correction was performed, resulting in link probabilities \( p_j \) for gene \( i \) between cell \( j \) and its KNNs \((l = j, \ldots, d_j)\). The minimum of these link probabilities, \( p_j = \min(p_{jl}) \), was compared to a probability threshold \( (P_l = 0.01 \text{ by default}) \) and all neighbors with \( p_l < P_l \) were pruned. This minimum was also assigned as link probability for further analysis.

Inferring cell type clusters and transition probabilities. The pruned KNN network connects only cells sampled from overlapping transcript count distributions across all genes. To accelerate the computation, the pruning procedure can be performed on a subset of selected genes (for example, based on expression or enhanced variability). The pruning procedure was implemented in VarID using the RaceID criterion for the optional selection of highly variable genes (that is, genes with expression variance exceeding the background level) (equation (5)).

Option 1: direct regression of the variance-mean dependence. A major problem is the dependence of the transcript count variance on the average transcript count. I observed that the baseline level of the variance as a function of the mean exhibits a convex behavior after log-transformation. This is mainly due to the presence of two sources of technical noise (that is, sampling noise and global cell-to-cell variability in sequencing efficiency) in addition to biological variability. To capture the baseline level of the noise, I split the gene variances into 100 equally populated bins after ordering by increasing mean expression. For each bin, I retained only the data points with variances below the 5% quantile of the variance distribution within this bin. I then applied a least square regression of a second order polynomial to the remaining data points across all bins (unlike in equation (5)) to obtain a function \( v_j \) capturing the baseline variability as a function of mean expression \( m_j \):

\[
v_j(m) \sim 2\nu_{ij} + \gamma_j \log(m) + \beta \log(m)^2
\]

The local variability \( \nu_{ij} \) of gene \( i \) in the neighborhood of cell \( j \), given by cell \( j \) and its nearest neighbors \((l = j, \ldots, d_j)\) that remained after pruning, was then estimated as the variance of the transcript counts \( x_{ij} \) across the neighborhood of cell \( j \), divided by \( \nu_j(m_j) \), where \( m_j \) was the mean of the transcript counts \( x_{ij} \) of gene \( i \) across the pruned neighborhood of cell \( j \):

\[
\nu_{ij} = \frac{\sum_{l=j}^{d_j} (x_{ij} - \mu_j)^2}{\bar{\nu}_j(m_j)}
\]

Option 2: eliminating the variance-mean dependence by regressing out total transcript counts from the expression data. As an alternative approach, the local variability \( \nu'_{ij} \) of gene \( i \) in the neighborhood of cell \( j \) was computed as the variance of the
Differentially expressed genes between two subgroups of cells were identified in a similar manner to a previously published method. First, negative binomial distributions reflecting the gene expression variability within each subgroup were inferred based on the background model for the expected transcript count variability computed by RaceID3. Based on these distributions, a $p$ value for the observed difference in transcript counts between the two subgroups was calculated and multiple testing corrected by the Benjamini-Hochberg method.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**
Primary data used in this manuscript were downloaded from GEO with accession code GSE89754 for the hematopoietic data, and GSE92332 for the intestinal data.

**Code availability**
VarID is integrated in the RaceID v0.1.4 package available from CRAN or github (https://github.com/dgrun/RaceID3_StemID2_package). Source code for reproducing the results of this manuscript is available on github (https://github.com/dgrun/VarID_analysis) and as Supplementary Software.

**References**
22. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* 12, 477–479 (2016).
23. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086 (2017).
24. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* 11, R106 (2010).

**Acknowledgements**
This study was supported by the Max Planck Society, the German Research Foundation (DFG) (grant numbers SPP1937 GR4980/1-1, GR4980/3-1, and GRK2344 MeInBio), by the DFG under Germany’s Excellence Strategy (CIBSS, EXC-2189, Project ID 39039984), by the ERC (818846, ImmuNiche, ERC-2018-COG), and by the Behrens-Weise Foundation.

**Author contributions**
D.G. conceived and implemented the method and performed the analysis.

**Competing interests**
The author declares no competing interests.

**Additional information**
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0632-3.

**Correspondence and requests for materials**
 should be addressed to D.G.

**Peer review information** Nicole Rusk and Nina Vogt were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Reprints and permissions information** is available at www.nature.com/reprints.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

| Data collection | Does not apply. Data were downloaded from GEO for published datasets. |
|-----------------|---------------------------------------------------------------------|
| Data analysis   | R package RaceID v0.1.4   |
|                 | R package FNN v.1.1.3    |
|                 | R package ReactomePA v1.22.0 |
|                 | R package Seurat v3.0.0  |
|                 | R package GENIE3 v.1.0.0 |
|                 | Source code for reproducing results is available on github: https://github.com/dgrun/VarID_analysis |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Primary data used in this manuscript was downloaded from GEO with accession code GSE89754 for the hematopoietic data, and GSE92332 for the intestinal data.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | Only published data were analyzed. The sample size was chosen based on the sample size in the publicly available datasets. |
|-------------|-----------------------------------------------------------------------------------------------------------------------|
| Data exclusions | Two replicates were excluded from the intestinal data (GSE92332) due to strong gender-related batch effects. This is described in the Methods section. The inclusion of these replicates would have confounded the cell type annotation as the batch effect could not be easily removed computationally. |
| Replication | In each dataset at least two independent biological replicates were included to demonstrate reproducibility of the findings. |
| Randomization | Only published data were analyzed. |
| Blinding | Only published data were analyzed. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|---------------------------------|---------|
| n/a | Involved in the study |
| ☑ | Antibodies |
| ☑ | Eukaryotic cell lines |
| ☑ | Palaeontology |
| ☑ | Animals and other organisms |
| ☑ | Human research participants |
| ☑ | Clinical data |
| ☑ | Involved in the study |
| ☑ | ChIP-seq |
| ☑ | Flow cytometry |
| ☑ | MRI-based neuroimaging |