Supplementary Note

Contents

1 Linear mixed model

Consider expression levels of J genes for N cells are given. We model the vector of expression levels $y_j = (y_{1j}, \ldots, y_{Nj})^\top$ for gene j as a linear mixed model

$$y_j = X\beta + Z\alpha_j + \varepsilon_j; \quad j = 1, \ldots, J, \tag{1}$$

where β denotes a vector of fixed effects shared across all genes, α_j denotes a vector of gene specific random effects which is independently drawn from a multivariate normal distribution, and ε_j denotes a vector of residuals which follows a multivariate normal distribution $\mathcal{N}(0, \sigma_j^2 I_N)$ with the variance parameter σ_j^2, assuming homoscedasticity across all cells. Here I_n denotes $n \times n$ identity matrix. The design matrix X can be arbitrary, we set $X = Z$ in our model to capture the mean of α_j for all genes, around which variance parameters are estimated.

The design matrix Z is derived from a combination of known factors which account for transcriptional variation of y_j, so that the covariance matrix of α_j captures variance components of the given data. Here we assume $\alpha_j \sim \mathcal{N}(0, \sigma_j^2 D)$, suggesting the covariance matrix of α_j is shared across all genes except for the scaling by residual variance σ_j^2.

The model is general, thereby can be used for batch corrections in cell clustering and identification of cell populations as well as differential expression analysis of a target factor while treating all other factors as confounders. The subsequent sections describe three different utilisation of the model in detail.

1.1 Variance component analysis

Suppose we have K known factors $\{x_1, \ldots, x_K\}$ that could account for a part of transcriptional variation. Each factor is either a numerical or a categorical variable with N elements (*i.e.*, $x_k = (x_{1k}, \ldots, x_{Nk})^\top$). If the kth factor is a numerical variable, then we introduce $Z_k \in \mathbb{R}^{N \times 1}$ which is a column vector whose elements are scaled elements of x_k, so that the mean equal to 0 and variance equal to 1. If the kth factor is a categorical variable with m_k levels, we introduce a design matrix $Z_k \in \mathbb{R}^{N \times m_k}$ whose lth column is an indicator vector; the ith element is 1 if $x_{ik} = l$, otherwise 0, for $i = 1, \ldots, N$. We combine $\{Z_k; k = 1, \ldots, K\}$ as a design matrix

$$Z = (1, Z_1, \ldots, Z_K) \in \mathbb{R}^{N \times M}.$$

which is used in the model (Eq.1). Note that the first column of Z is the vector of all 1s which is introduced to estimate random intercepts of y_j. The corresponding random effect $\alpha_j = (\alpha_{0j}, \alpha_{j1}^\top, \ldots, \alpha_{jK}^\top)^\top \in \mathbb{R}^M$ is partitioned in accordance with Z, so that

$$Z\alpha_j = \sum_{k=0}^{K} Z_k \alpha_{jk}.$$

We assume

$$\alpha_{jk} \sim \mathcal{N}(0, \sigma_j^2 \varphi_k^2 I_{m_k}),$$

where φ_k^2 denotes the variance parameter for the factor k, which is shared across all genes.

The variance explained by each factor k is measured by the intraclass correlation using the maximum likelihood estimator $\hat{\varphi}_k^2$,

$$\text{(intraclass correlation)} = \frac{\sigma_j^2 \hat{\varphi}_k^2}{\hat{\sigma}_j^2 + \hat{\sigma}_j^2 \hat{\varphi}_k^2} = \frac{\hat{\varphi}_k^2}{1 + \hat{\varphi}_k^2},$$

where all other factors are kept constant. The standard error of the intraclass correlation is obtained by the delta method:

$$\text{SE(intraclass correlation)} \approx \frac{2|\hat{\varphi}_k|}{(1 + \hat{\varphi}_k^2)^2} \text{SE}(\hat{\varphi}_k)$$

Here the standard error $\text{SE}(\hat{\varphi}_k)$ is obtained by the inverse of Fisher information matrix for φ.

1.2 Latent factor estimation

A goal of a single cell experiment is to cluster cells by hidden cell types which are unknown a priori. We often use the principal component analysis (PCA) to find the most variable basis that splits single cells into biologically distinct clusters. However, PCA works only if the transcriptional variation among cell types is greater than the other factors, such as experimental batches. If, for example, two different experiments were performed in different laboratories or with completely different technologies (such as 10x and smart-seq), PCA is likely to captures the technical variation rather than the biological variation by cell types.

Although there are various batch correction methods previously proposed (MNN correction, Seurat v3 and scanorama), non of these can handle hundreds of batches or multiple factors, a combination of which explains a significant amount of variation in the data. The linear mixed model is a suitable approach to cope with such a complex situation. We introduce a handful number of latent factors in the variance component model in Eq.1, that could capture biologically meaningful principal components while adjusting effects of other known confounding factors. The linear mixed model with latent factors is given by

$$y_j = X\beta + Z\alpha_j + \Psi\gamma_j + \varepsilon_j; \quad j = 1, \ldots, J, \tag{2}$$

where $\Psi \in \mathbb{R}^{N \times L}$ denotes a matrix of L latent factors and γ_j denotes a random effect for each gene j independently following the normal distribution $N(0, \sigma_j^2 I)$. Note that, if there is no known factors in the model ($Z\alpha_j$), the maximum likelihood estimator of Ψ is identical to the L principal components from the standard PCA.

1.3 Differential expression analysis

We can also utilise the linear mixed model for the variance component analysis to adjust known confounding effects in the differential expression (DE) analysis. Because the factor of interest (*e.g.*, pathology) is confounded by other known factors (*e.g.*, patient or brain regions), the result of a standard DE analysis is likely to be biased by those confounding factors.

Consider the kth factor is of interest and we would like to test whether the gene j is differentially expressed by the factor k (between different levels of a categorical factor or by one-unit change of a numerical factor). Let $\alpha_{j,-k} = (\alpha_{j1}^\top, \ldots, \alpha_{j,k-1}^\top, \alpha_{j,k+1}^\top, \ldots, \alpha_{jK}^\top)^\top$ be the random effect without α_{jk} for gene

j, the Bayes factor that captures statistical significance of the factor k can be written as

$$BF_{jk} = \frac{\int p(y_j|\alpha_j, \hat{\beta}, \hat{\sigma}_j^2) p(\alpha_{j,-k}|\hat{\varphi}_{-k}^2, \hat{\sigma}_j^2) p(\alpha_{jk}|\varphi_k^2 = \tilde{\varphi}_k^2, \hat{\sigma}_j^2) d\alpha_j}{\int p(y_j|\alpha_{j,-k}, \hat{\beta}, \hat{\sigma}_j^2, \alpha_{jk} = 0) p(\alpha_{j,-k}|\hat{\varphi}_{-k}^2, \hat{\sigma}_j^2) d\alpha_{j,-k}}.$$

Here we set $\tilde{\varphi}_k^2 = 100\hat{\varphi}_k^2$ to make the prior distribution of α_{jk} non-informative.

For the categorical factor k with more than two levels, a high Bayes factor implies the gene is differentially expressed among those levels, but it does not provide any specific comparison in which a level (or a set of levels) is differentially expressed with others. Therefore, we introduced the contrast vector c_h which partitions all levels existing in the factor k into any of two groups. Using the contrast, the Bayes factor for the specific comparison h can be written as

$$BF_{jk}^{[h]} = \frac{\int p(y_j|\alpha_j, \hat{\beta}, \hat{\sigma}_j^2) p(\alpha_{j,-k}|\hat{\varphi}_{-k}^2, \hat{\sigma}_j^2) p(\alpha_{jk}|\hat{\sigma}_j^2 \tilde{\varphi}_k^2 D_k^{[h]}) d\alpha_j}{\int p(y_j|\alpha_{j,-k}, \hat{\beta}, \hat{\sigma}_j^2, \alpha_{jk} = 0) p(\alpha_{j,-k}|\hat{\varphi}_{-k}^2, \hat{\sigma}_j^2) d\alpha_{j,-k}},$$

where

$$D_k^{[h]} = (1 - c_h, c_h)(1 - c_h, c_h)^\top$$

is the covariance matrix for α_{jk} such that $\alpha_{jk} \sim \mathcal{N}(0, \hat{\sigma}_j^2 \tilde{\varphi}_k^2 D_k^{[h]})$. For example, if the categorical factor k has 5 levels, there are $2^5 - 1 = 15$ contrasts,

$$(c_1, \ldots, c_{15}) = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

which partition the 5 levels into any of two groups: *e.g.*, the comparison of the first two levels against the last three levels is given by c_3 and the covariance matrix is written as

$$D_k^{[3]} = (1 - c_3, c_3)(1 - c_3, c_3)^\top = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

The Bayes factors are then used to classify genes into one of the DE partitions. Assume there are H partitions, the likelihood is a product of finite mixture models

$$L(\pi) = \prod_{j=1}^{J} \left[\pi_0 + \sum_{h=1}^{H} \pi_h BF_{jk}^{[h]} \right]$$

with the prior probability $\pi = (\pi_0, \pi_1, \ldots, p_H)^\top$. We use a standard EM algorithm to maximise the likelihood. The posterior probability of the gene j being differentially expressed by partition h is given by

$$z_{jk}^{[h]} = \frac{\hat{\pi}_h BF_{jk}^{[h]}}{\hat{\pi}_0 + \sum_{i=1}^{H} \hat{\pi}_h BF_{jk}^{[i]}}.$$

The posterior distribution of α_j is useful to visualise the DE result. The mean of α_{jk} provides the averaged normalised expression levels for the levels of categorical factor k adjusted by other known confounding factors. The posterior probability is analytically obtained by

$$\tilde{\alpha}_j^{[h]} \sim \mathcal{N}(A^{-1}Z^\top(y_j - X\hat{\beta}), \hat{\sigma}_j^2 A^{-1})$$

where $A = Z^\top Z + D^{-1}$ and

$$D = \begin{pmatrix} \hat{\varphi}_1 I & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \hat{\varphi}_k D_k^{[h]} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \hat{\varphi}_K I \end{pmatrix}.$$

Therefore, the log fold change of expression levels between two groups can be written as

$$\delta_{jk}^{[h]} \sim \mathcal{N}\left((c_h^1 - c_h^0)^\top \alpha_{jk}^{[h]}, \hat{\sigma}_j^2 (c_h^1 - c_h^0)^\top (A^{-1})_{[k,k]} (c_h^1 - c_h^0)\right).$$

where

$$c_h^1 = \frac{c_h^\top}{|c_h|^2} \quad \text{and} \quad c_h^0 = \frac{1 - c_h}{|1 - c_h|^2}.$$

Here $M_{[k,k]}$ denotes a diagonal sub-matrix of any matrix M corresponding to the factor k (*e.g.*, $D_{[k,k]} = \hat{\varphi}_k D_k^{[h]}$). The posterior distribution also yield a more robust measure of significance for each partition h, the local true sign rate, or

$$ltsr_{jk}^{[h]} = z_{jk}^{[h]} \max\{p(\delta_{jk}^{[h]} > 0 | y_j), p(\delta_{jk}^{[h]} < 0 | y_j)\},$$

which is analogous to the local false sign rate proposed in [1]. The value is more stringent than the posterior probability $z_{jk}^{[h]}$ because it requires true discoveries to be not only nonzero, but also correctly singed in a consistent direction.

2 Three-way Bayesian hierarchical model

The three-way Bayesian hierarchical model is a simple extension of the hierarchical model first used in mapping eQTLs [2] or the pairwise fGWAS model to estimate a shared genetic architecture between paired GWAS traits [3]. To introduce the model, we begin by constructing the hierarchical model for mapping eQTLs in a single cell type. Then we increase the number of cell types (up to three) to sequentially build up the model. The prior probabilities of the model are empirically estimated from the data. We incorporate a three-stage optimisation which allows us to reduce the computational complexity and to increase the stability of model fitting process [4].

2.1 Hierarchical model for mapping eQTLs in one cell type

The hierarchical model for mapping eQTLs in single cell type is equivalent to the Bayesian hierarchical model proposed in [2]. We use genetic associations in cis window $\mathcal{W}_j$ of 1Mb centred at transcription start site (TSS) for each gene j. The association is measured by Bayes factor of a simple linear regression model, where $\log(TPM+1)$ is regressed on genotype of each genetic variant $l \in \mathcal{W}_j$. We use the asymptotic Bayes factor [5] which can be easily obtained from the estimated effect size $\hat{\beta}_{jl}$ and its standard error $\hat{\sigma}_{jl}$ of a variant l on expression of gene j, such that

$$BF_{jl} = \sqrt{1-r_{jl}} \exp\left\{\frac{z_{jl}^2}{2} r_{jl}\right\} \tag{3}$$

where

$$z_{jl} = \mathcal{Z}\left(\frac{\hat{\beta}_{jl}}{\hat{\sigma}_{jl}}\right)$$

and

$$r_{jl} = \frac{W}{W+\hat{\sigma}_{jl}^2}.$$

Here we use $\mathcal{Z}(\cdot)$ to convert student t statistic into normal z statistic to deal with small sample sizes (see Supplementary Note of [4] for details).

The model is a mixture of the following two hypotheses:

H_0 $(null)$: there are no genetic variants in $\mathcal{W}_j$ that associate with the expression of gene j;

H_1 $(eQTL)$: there is one causal variant in $\mathcal{W}_j$ that affects the expression of gene j.

We introduce the prior probability Π_1 with which a gene j is an eQTL. Assuming that there are J genes genome-wide and their expression levels are conditionally independent, the likelihood of the hierarchical model is then written as a product of mixture probability over $j = 1, \ldots, J$, such that

$$L(\Pi_1) = \prod_{j=1}^{J} \left[(1-\Pi_1) + \Pi_1 RBF_j\right], \tag{4}$$

where RBF_j denotes the regional Bayes factor which is the genetic association for gene j averaged over all variants $l \in \mathcal{W}_j$, defined as

$$RBF_j = \frac{1}{\#\mathcal{W}_j} \sum_{l \in \mathcal{W}_j} BF_{jl}. \tag{5}$$

Note that $\#\mathcal{W}_j$ denotes the number of variants in $\mathcal{W}_j$, assuming there is one variant causal to expression of gene j. The maximum likelihood estimator $\hat{\Pi}_1$ can be obtained by a standard EM algorithm.

2.2 Hierarchical model for jointly mapping eQTLs in two cell types

We then extend the hierarchical model for a pair of cell types. Again, we use genetic associations of variant $l \in \mathcal{W}_j$ that alter expression of gene j for two different cell types 1 and 2. We consider the following 5 different hypotheses:

H_0 $(null)$: there are no genetic variants in $\mathcal{W}_j$ that associate with expression of gene j in either cell types;

H_1 $(single)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j of cell type 1;

H_2 $(single)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j of cell type 2;

H_3 $(linkage)$: there are two independent causal variants in $\mathcal{W}_j$, one of which affects expression of gene j in cell type 1 and the other one affects expression of gene j in cell type 2, independently;

H_4 $(colocalisation)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in both two cell types simultaneously.

The likelihood of the model is given by a product of a finite mixture of the 5 different hypotheses,

$$L(\Psi_{12}, \Pi_{12}) = \prod_{j=1}^{J} \left[\Phi_0 + \sum_{h=1}^{4} \Phi_h RBF_j^{[h]} \right], \tag{6}$$

where

$$\Phi_h = \begin{cases} (1-\Psi_{12})(1-\Pi_1)(1-\Pi_2) + \Psi_{12}(1-\Pi_{12}) & h = H_0 \\ (1-\Psi_{12})\Pi_1(1-\Pi_2) & h = H_1 \\ (1-\Psi_{12})(1-\Pi_1)\Pi_2 & h = H_2 \\ (1-\Psi_{12})\Pi_1\Pi_2 & h = H_3 \\ \Psi_{12}\Pi_{12} & h = H_4 \end{cases} \tag{7}$$

denotes the prior probability that gene j belongs to one of the hypotheses $h = 0, \ldots, 4$. The prior probability Φ_h is a function of the probability Ψ_{12} that the gene j is pleiotropic in cell type 1 and 2 and the probability Π_{12} that the gene j is an eQTL driven by a same variant in $\mathcal{W}_j$. The probability Π_1 (or Π_2) is the probability that the gene j is an eQTL in cell type 1 (or cell type 2), independently from the other cell type. The maximum likelihood estimators, $\Pi_1 = \hat{\Pi}_1$ and $\Pi_2 = \hat{\Pi}_2$, are obtained by maximising Eq.4 for cell type 1 and 2 independently, and plugged into Eq.7, so that the likelihood (Eq.6) is a function of $\{\Psi_{12}, \Pi_{12}\}$. The regional Bayes factor for a hypothesis h, $RBF_j^{[h]}$, is defined by

$$RBF_j^{[h]} = \begin{cases} RBF_j^{(1)} & h = H_1 \\ RBF_j^{(2)} & h = H_2 \\ RBF_j^{(1)} RBF_j^{(2)} & h = H_3 \\ RBF_j^{(12)} & h = H_4 \end{cases}$$

where $RBF_j^{(1)}$ (or $RBF_j^{(2)}$) denotes the regional Bayes factor of gene j being an eQTL in cell type 1 (or cell

type 2) defined in Eq.5, and

$$RBF_j^{(12)} = \frac{1}{\#\mathcal{W}_j} \sum_{l \in \mathcal{W}_j} BF_{jl}^{(1)} BF_{jl}^{(2)} \tag{8}$$

denotes the joint association on gene expression j averaged over $l \in \mathcal{W}_j$ in cell type 1 and 2 under the conditional independence of gene expression in two cell types. We use a standard EM algorithm to maximise Eq.6 with respect to $\{\Psi_{12}, \Pi_{12}\}$.

2.3 Hierarchical model for jointly mapping eQTLs in three cell types

Finally, we extend the model to cope with three different cell types. We consider the following 15 hypotheses in the model to cover any potential shared genetic architecture:

H_0 $(null)$: there are no genetic variants in $\mathcal{W}_j$ that associate with expression of gene j in three cell types;

H_1 $(single)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in cell type 1;

H_2 $(single)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in cell type 2;

H_3 $(single)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in cell type 3;

H_4 $(linkage)$: there are two independent causal variants in $\mathcal{W}_j$, one of which affects expression of gene j in cell type 1 and the other one affects expression of gene j in cell type 2;

H_5 $(linkage)$: there are two independent causal variants in $\mathcal{W}_j$, one of which affects expression of gene j in cell type 1 and the other one affects expression of gene j in cell type 3;

H_6 $(linkage)$: there are two independent causal variants in $\mathcal{W}_j$, one of which affects expression of gene j in cell type 2 and the other one affects expression of gene j in cell type 3;

H_7 $(linkage)$: there are three independent causal variants in $\mathcal{W}_j$, each of which affects expression of gene j in each of the three cell types, independently;

H_8 $(colocalisation)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in cell type 1 and 2 simultaneously;

H_9. $(colocalisation \,\&\, linkage)$: there are two independent causal variants in $\mathcal{W}_j$, one of which affects expression of gene j in cell type 1 and 2 simultaneously and the other one affects expression of gene j in cell type 3, independently;

H_{10} $(colocalisation)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in cell type 1 and 3 simultaneously;

H_{11} $(colocalisation \,\&\, linkage)$: there are two independent causal variants in $\mathcal{W}_j$, one of which affects expression of gene j in cell type 1 and 3 simultaneously and the other one affects expression of gene j in cell type 2, independently;

H_{12} $(colocalisation)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in cell type 2 and 3 simultaneously;

H_{13} $(colocalisation \,\&\, linkage)$: there are two independent causal variants in $\mathcal{W}_j$, one of which affects expression of gene j in cell type 2 and 3 simultaneously and the other one affects expression of gene j in cell type 1, independently;

H_{14} $(colocalisation)$: there is one causal variant in $\mathcal{W}_j$ that affects expression of gene j in all three cell types simultaneously;

The likelihood of the model given expression data of J genes can be written as

$$L(\Psi_{123}, \Pi_{123}) = \prod_{j=1}^{J} \left[\Phi_0 + \sum_{h=1}^{14} \Phi_h RBF_j^{[h]} \right] \tag{9}$$

with the extended prior probabilities

$$\Phi_h = \begin{cases} \Psi_1(1-\Pi_1)(1-\Pi_2)(1-\Pi_3) + (\Psi_{12}-\Psi_{123})(1-\Pi_{12})(1-\Pi_3) \\ \quad +(\Psi_{13}-\Psi_{123})(1-\Pi_{13})(1-\Pi_2) + (\Psi_{23}-\Psi_{123})(1-\Pi_{23})(1-\Pi_1) \\ \quad\quad +\Psi_{123}(1-\Pi_{123}) & h = H_0 \\ \Psi_1\Pi_1(1-\Pi_2)(1-\Pi_3) + (\Psi_{23}-\Psi_{123})(1-\Pi_{23})\Pi_1 & h = H_1 \\ \Psi_1(1-\Pi_1)\Pi_2(1-\Pi_3) + (\Psi_{13}-\Psi_{123})(1-\Pi_{13})\Pi_2 & h = H_2 \\ \Psi_1(1-\Pi_1)(1-\Pi_2)\Pi_3 + (\Psi_{12}-\Psi_{123})(1-\Pi_{12})\Pi_3 & h = H_3 \\ \Psi_1\Pi_1\Pi_2(1-\Pi_3) & h = H_4 \\ \Psi_1\Pi_1(1-\Pi_2)\Pi_3 & h = H_5 \\ \Psi_1(1-\Pi_1)\Pi_2\Pi_3 & h = H_6 \\ \Psi_1\Pi_1\Pi_2\Pi_3 & h = H_7 \\ (\Psi_{12}-\Psi_{123})\Pi_{12}(1-\Pi_3) & h = H_8 \\ (\Psi_{12}-\Psi_{123})\Pi_{12}\Pi_3 & h = H_9 \\ (\Psi_{13}-\Psi_{123})\Pi_{13}(1-\Pi_2) & h = H_{10} \\ (\Psi_{13}-\Psi_{123})\Pi_{13}\Pi_2 & h = H_{11} \\ (\Psi_{23}-\Psi_{123})\Pi_{23}(1-\Pi_1) & h = H_{12} \\ (\Psi_{23}-\Psi_{123})\Pi_{23}\Pi_1 & h = H_{13} \\ \Psi_{123}\Pi_{123} & h = H_{14} \end{cases}$$

where $\Psi_1 = 1 - \Psi_{12} - \Psi_{13} - \Psi_{23} + 2\Psi_{123}$, so that $\sum_{h\in\mathcal{H}} \Phi_h = 1$. The prior probability Φ_h is a function of the probability Ψ_{123} that the gene j is pleiotropic in the three cell types and the probability Π_{123} that the gene j is an eQTL for the gene j in the three cell types. All other parameters are introduced in the pairwise or single hierarchical models in Section 2.1 and 2.2 (*e.g.*, Ψ_{23} is the probability that the gene j is pleiotropic in cell type 2 and 3). Those parameters are estimated a priori by maximising Eq. 4 and 6. Then they are plugged into the likelihood (Eq. 9), so that Eq. 9 is maximised only with respect to $\{\Psi_{123}, \Pi_{123}\}$.

The corresponding genetic association for each hypothesis is given by

$$RBF_j^{[h]} = \begin{cases} RBF_j^{(1)} & h = H_1 \\ RBF_j^{(2)} & h = H_2 \\ RBF_j^{(3)} & h = H_3 \\ RBF_j^{(1)} RBF_j^{(2)} & h = H_4 \\ RBF_j^{(1)} RBF_j^{(3)} & h = H_5 \\ RBF_j^{(2)} RBF_j^{(3)} & h = H_6 \\ RBF_j^{(1)} RBF_j^{(2)} RBF_j^{(3)} & h = H_7 \\ RBF_j^{(12)} & h = H_8 \\ RBF_j^{(12)} RBF_j^{(3)} & h = H_9 \\ RBF_j^{(13)} & h = H_{10} \\ RBF_j^{(13)} RBF_j^{(2)} & h = H_{11} \\ RBF_j^{(23)} & h = H_{12} \\ RBF_j^{(23)} RBF_j^{(1)} & h = H_{13} \\ RBF_j^{(123)} & h = H_{14} \end{cases}$$

where $RBF_j^{(i)}$ denotes the regional Bayes factor of gene j being an eQTL in cell type i (Eq.5), $RBF_j^{(ik)}$ denotes the regional Bayes factor of gene j being an eQTL in cell type i and k with a same causal variant (Eq.8), and

$$RBF_j^{(123)} = \frac{1}{\#\mathcal{W}_j} \sum_{l \in \mathcal{W}_j} BF_{jl}^{(1)} BF_{jl}^{(2)} BF_{jl}^{(3)} \tag{10}$$

is the joint association on gene expression j averaged over $l \in \mathcal{W}_j$ in cell type 1, 2 and 3 under conditional independence of gene expression in those cell types. We use a standard EM algorithm to maximise Eq.9 with respect to $\{\Psi_{123}, \Pi_{123}\}$.

2.4 Posterior probability calculation

Once the maximum likelihood estimator $\hat{\Phi}_h$ is estimated, the posterior probability that the gene j belongs to one of the 14 alternative hypotheses $H_1, \ldots, H_{14}$ is given by

$$z_j^{[h]} = \frac{\hat{\Phi}_h RBF_j^{[h]}}{\hat{\Phi}_0 + \sum_{i=1}^{14} \hat{\Phi}_i RBF_j^{[i]}}; \qquad h = 1, \ldots, 14.$$

Therefore the posterior probability that the gene j is an eQTL in cell type 1 is written as

$$\begin{aligned} &P(\text{gene } j \text{ is an eQTL in cell type 1}) \\ &\quad = z_j^{[1]} + z_j^{[4]} + z_j^{[5]} + z_j^{[7]} + z_j^{[8]} + z_j^{[9]} + z_j^{[10]} + z_j^{[11]} + z_j^{[13]} + z_j^{[14]}. \end{aligned}$$

Likewise,

$$\begin{aligned} &P(\text{gene } j \text{ is an eQTL in cell type 2}) \\ &\quad = z_j^{[2]} + z_j^{[4]} + z_j^{[6]} + z_j^{[7]} + z_j^{[8]} + z_j^{[9]} + z_j^{[11]} + z_j^{[12]} + z_j^{[13]} + z_j^{[14]} \end{aligned}$$

and

$$P(\text{gene } j \text{ is an eQTL in cell type 3}) = z_j^{[3]} + z_j^{[5]} + z_j^{[6]} + z_j^{[7]} + z_j^{[9]} + z_j^{[10]} + z_j^{[11]} + z_j^{[12]} + z_j^{[13]} + z_j^{[14]}.$$

3 Detailed experimental protocols

3.1 Total RNA extraction from primary microglia

Total RNA was extracted from primary microglia using the Qiagen AllPrep DNA/RNA Micro kit (Qiagen, 80284), according to the manufacturer's instructions. Total RNA quantity and quality was assessed using an Agilent Bioanalyzer with an RNA 6000 pico kit (Agilent Technologies, 5067-1513).

3.2 Low-input bulk RNA-seq library preparation for primary microglia

Between 0.3 ng and 10 ng of total RNA was diluted to a final volume of 25 μL with nuclease-free 10 mM Tris-HCl pH 7.5. To this, 25 μL of a 2x lysis/binding buffer (200 mM Tris-HCl pH 7.5, 1 M LiCl, 20 mM EDTA, 2 % w/v lithium dodecyl sulphate, and 10 mM 1,4-dithiothreitol) was added and mixed well. According to the manufacturer's instructions, mRNA was purified using the mRNA DIRECT kit (Thermo Fisher, 61012) from the total RNA using 20 μL of oligo dT Dynabeads, with a final elution volume of 7 μL of nuclease-free 10 mM Tris-HCl pH 7.5. The purified mRNA was processed using a modified Smart-seq2 protocol [6] as follows: 2 μL of oligo dT30VN (Integrated DNA Technologies) and 2.34 μL of 10 mM dNTPs (Thermo Fisher, R0193) were mixed with 7 μL of the purified mRNA and heated to 72 °C for 3 minutes to denature secondary structures, before rapidly chilling on ice for 5 minutes. 5 μL of 5x SMARTScribe first-strand buffer (Clontech Takara, 639538), 0.63 μL of SUPERase inhibitor (Thermo Fisher, AM2696), 1.25 μL of 100 mM 1,4-dithiothreitol, 5 μL of betaine (Sigma, B0300-5VL), 0.15 μL of 1 M MgCl2, 0.38 μL of template-switching LNA-oligo (TSO) (Qiagen) and 1.25 μL of SMARTScribe reverse transcriptase (Clontech Takara, 639538) were added to the denatured mRNA/dNTP/oligo dT30VN mix. Following a brief vortex mix, reverse transcription was performed at 42 °C for 90 minutes, followed by 10 cycles of 50 °C for 2 minutes, then 42 °C for 2 minutes. The reaction was stopped by incubating at 70 °C for 15 minutes. The first-strand cDNA was purified using 0.8 volumes of Ampure XP beads (Beckman Coulter, BCAG0006) to 1 volume of the reverse transcription reaction volume, according to the manufacturer's instructions, but leaving the eluted cDNA in 12 μL of 10 mM Tris-HCl pH7.5 with the beads in solution. This was done to maximise the amount of cDNA carried forward to the subsequent cDNA amplification reaction. The cDNA was amplified by adding 0.5 μL of 10 μM ISPCR primer (Integrated DNA Technologies) and 12.5 μL of 2x KAPA HiFi polymerase (Kapa Biosystems, KK2601) to the 12 μL of cDNA and mixed before heating at 98 °C for 3 minutes, followed by 11-18 cycles (depending on total RNA input quantity) of 98 °C for 20 seconds, 67 °C for 15 seconds and 72 °C for 6 minutes, followed by a final extension at 72 °C 5 minutes. The amplified double-stranded cDNA was purified as before, but this time the Ampure XP beads were removed from the 20 μL eluate. Amplified double-stranded cDNA was quantified with a Quant-iTTM dsDNA high sensitivity assay kit (Thermo Fisher, Q33120) in black v-bottom 96-well plates (Greiner Bio-One, 651209) on a FLUOstar Omega (BMG Labtech), according manufacturers' instructions. For cDNA tagmentation, 4 ng of cDNA was diluted with 10 mM Tris-HCl pH 7.5 to a volume of 9.5 μL. 5 μL of a 3x tagmentation buffer (99 mM Tris acetate, 198 mM potassium acetate, 30 mM magnesium acetate and 48 % v/v N,N-dimethylformamide) and 0.5 μL of TDE1 (Illumina, 20034197) were added, mixed and incubated at 55 °C for 5 minutes. The tagmentation reaction was stopped by the addition of 2.5 μL of a tagmentation stop buffer (220 mM EDTA and 1.1 % w/v sodium dodecyl sulphate) and mixed before incubating at room temperature for 10 minutes. The tagmented cDNA was diluted with 10 mM Tris-HCl pH 7.5 to a final volume of 50 μL,

before purifying with a 2:1 ratio of Ampure XP beads to sample volume, eluting the tagmented cDNA in 7 μL of 10 mM Tris-HCl pH 7.5. Tagmented cDNA samples were then amplified and sample-indexed by PCR as follows: 7 μL of tagmented cDNA was added to 2.5 μL of i5 index adapter and 2.5 μL of i7 index adapter from the Nextera®XT index kit v2 set A (Illumina, 15052163), 0.25 μL of 50 μM PC1 primer, 0.25 μL of 50 μM PC2 primer and 12.5 μL of 2x KAPA HiFi polymerase, before mixing and incubating at 72 °C for 3 minutes, 98 °C for 30 seconds, followed by 9 cycles at 98 °C for 15 seconds, 62 °C for 30 seconds and 72 °C for 30 seconds, followed by a final extension at 72 °C for 3 minutes. Individual libraries were purified and excess primers removed by performing 0.8:1 ratio of Ampure XP beads to PCR volume, eluting the finished library in 20 μL of 10 mM Tris-HCl pH 7.5. Libraries were quantified with a Quant-iTTM dsDNA high sensitivity assay kit, as mentioned above, before combining 96 libraries per pool in equimolar amounts. Library pools were assessed for fragment length and quantity on a Bioanalyser using a High Sensitivity DNA kit (Agilent Technologies, 5067-4626), according to the manufacturer's instructions. Each 96-library pool was sequenced over 8 lanes of a HiSeq SBS v4, collecting 75 bp paired-end reads.

3.3 iPS cell culture and macrophage differentiation

iPS cell culture and macrophage differentiation was carried as previously described [7] but with some minor modifications: embryoid bodies were harvested 3 days after formation and transferred onto gelatinised tissue-culture treated 10 cm dished in serum-free X-VIVO 15 (Lonza, BE02-060F) or Stem Pro-34 SFM (Thermo Fisher, 10640-019), with both mediums supplemented with 2 mM GlutaMAX (Thermo Fisher, 35050061), 50 IU/ml penicillin, 50 IU/ml streptomycin (Sigma, P4333), 100 ng/ml human macrophage colony stimulating factor (hM-CSF) (Peprotech, 300-25) and 25 ng/ml human interleukin-3 (hIL-3) (Peprotech, 200-03). Macrophage progenitor cells were counted and plated in RPMI 1640 (Thermo Fisher, 11875093) supplemented with 10% heat-inactivated FBS (Thermo Fisher, 10500-064), 2mM GlutaMAX (Thermo Fisher, 35050061) and 100 ng/ml hM-CSF (Peprotech, 300-25) at a cell density of 10,000 cells per well on a 96-well plate (for RNA-seq), 100,000 cells per well on a 6-well plate (for ATAC-seq) or 25,000 cells per well of black 96-well plate (VWR, 734-1661) (for the macrophage purity assay) and differentiated for another 7 days.

3.4 iPS-derived macrophage purity assay

iPS-derived macrophages progenitor cells were seeded and differentiated as above before fixing in 50 μL of 4 % formaldehyde (Applichem, A0823.2500) at 4°C for 20 minutes. After fixation, the formaldehyde was aspirated and the fixed cells were washed twice in 100 μL PBS with calcium and magnesium (Sigma, D8662). After the final wash, the PBS was removed and replaced with 100 μL of 10 % blocking solution (1 in 10 dilution of donkey serum (AbD Serotec, C06SBZ) with 0.1 % Triton X-100 (Sigma, 93420)) before incubating at room temperature for 1 hour. The blocking solution was removed and replaced with 50 μL of 1 % blocking solution with either a 200-fold dilution of mouse anti-CD14 (BioLegend, 301802) or an 800-fold dilution of rabbit anti-CD68 (Cell Signaling Technology, 76437S), before incubating at 4 °C overnight. Control wells were also set up with 50 μL of 1 % blocking solution only. After the overnight primary antibody incubation, the cells were washed three times with 100 μL of PBS, incubating at room temperature for 5 minutes for each wash. 50 μL of 1 % blocking solution and 10 μg/mL DAPI (AppliChem, A1001), with either a 1000-fold dilution of donkey anti-mouse Alexa Fluor

647 (Thermo Fisher, A31571) or a 1000-fold dilution of donkey anti-rabbit Alexa Fluor 488 (Thermo Fisher, A21206), was added to the anti-CD14-treated cells or the anti-CD68-treated cells, respectively, before incubating at room temperature for 1 hour. Control wells were treated with 50 μL of 1 % blocking solution and 10 μg/mL DAPI. After the secondary immunostaining was complete, the cells were washed three times with 100 μL of PBS. To determine the proportion of double-stained cells (CD14 and CD68), the cells were analysed on a Cellomics Arrayscan (Thermo Fisher), according to the manufacturer's instructions. Only cell lines with greater than 90 % of cells being double-stained for CD14 and CD68 were processed for RNA-seq and ATAC-seq.

3.5 ATAC-seq library preparation for iPS-derived macrophage cell lines and primary macrophages

ATAC-seq library creation was performed as previously described [7]. ATAC-seq libraries were quantified on a Bioanalyser using a High Sensitivity DNA kit (Agilent Technologies, 5067-4626), according to the manufacturer's instructions, before pooling libraries in equimolar amounts. Pools were sequenced at 4 libraries per lane of a HiSeq SBS v4, collecting 75 bp paired-end reads.

3.6 iPS-derived macrophage low-input bulk RNA-seq preparation

iPS-derived macrophages progenitor cells were seeded and differentiated as described above before 50 μL of a 1x lysis/binding buffer (100 mM Tris-HCl pH 7.5, 0.5 M LiCl, 10 mM EDTA, 1 % w/v lithium dodecyl sulphate, and 5 mM 1,4-dithiothreitol) was added and mixed well. Lysed cells were stored at -80 °C until needed. RNA-seq libraries were generated as with the primary microglia samples. All iPS-derived macrophage RNA-seq libraries were generated with 11 cycles of amplification during the ISPCR stage.

References

[1] Urbut SM, Wang G, Carbonetto P, Stephens M (2019) Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet 51: 187–195.

[2] Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-qtls yields insight into human gene regulation. PLoS Genet 4: e1000214.

[3] Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, et al. (2016) Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet 48: 709–717.

[4] Kumasaka N, Knights AJ, Gaffney DJ (2019) High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nat Genet 51: 128–137.

[5] Wakefield J (2010) Bayesian methods for examining hardy-weinberg equilibrium. Biometrics 66: 257-65.

[6] Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, et al. (2014) Full-length RNA-seq from single cells using smart-seq2. Nat Protoc 9: 171–181.

[7] Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, et al. (2018) Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet 50: 424–431.