What Is Claimed Is:

1. A method for deconvolving bulk or spatial RNA-sequencing data for a sample, the method comprising the following steps:

a) obtaining input from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations for the sample;

b) generating a UMI (unique molecular identifier) count matrix based on the bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations, wherein each column of the UMI count matrix corresponds to a cell within the sample or a spatial location within the sample, and each row of the UMI count matrix corresponds to a gene in the sample;

c) selecting a subset of the most variably expressed genes from the UMI count matrix based on a union of genes whose expression is enriched in each cell type in the UMI count matrix, or a union of genes that vary above a predetermined level across all the cells in the UMI count matrix, or both;

d) estimating the mean and dispersion parameters of the selected genes from the UMI count matrix;

e) computing the cross cell type specificity of each gene of the selected genes based on the estimated mean and dispersion parameters;

f) estimating cross-sample gene variability from the bulk or spatial RNA-seq data or single cell samples;

g) estimating, using a regression-based approach, gene-wise scaling factors using both the bulk or spatial RNA-seq data and single cell RNA-seq data; and

h) generating a regression model based on the cross-cell type specificity of each selected gene, the cross-sample gene variability from the bulk or spatial RNA-seq data or from the single cell samples, and the gene-wise scaling factors; and

i) estimating, using the regression model, cell type proportions in the bulk or spatial RNA-sequencing data, wherein the regression model assigns a weight to the selected genes with high cell type specificity and low cross-sample variability;

thereby determining the percentage of cell types in the bulk or spatial RNA-sequencing data.

2. The method according to claim 1, wherein the bulk data to be deconvolved is transcripts per kilobase million (TPM) or read counts.

3. The method according to claim 1, wherein the subset of the most variably expressed genes is selected based on the union of genes whose expression is enriched in each cell type in the UMI count matrix, and wherein a build-in tool takes as input the UMI count matrix and cell type annotations.

4. The method according to claim 3, wherein for each cell type, the build-in tool computes the fold change between the average UMI in that cell type and the average UMI in all other cell types, and ranks genes by descending fold changes.

5. The method according to claim 1, wherein the subset of the most variably expressed genes is selected based on the union of genes that vary above a predetermined level across all the cells in the UMI count matrix, and wherein variances for each gene after cell number balancing and VST normalisation are computed.

6. The method according to any one of claims 1 to 5, wherein the RNA-seq data is not normalized before estimating the mean.

7. The method according to any one of claims 1 to 6, wherein to compute the cell type specificity of a gene, the cell type in which the gene has: i) the highest expression, or ii) highest fold change compared to others, is identified, and the specificity of this gene is defined as the mean-to-variance ratio within the cell type.

8. The method according to any one of claims 1 to 7, wherein the cross-sample gene variability is computed using the variance-to-mean ratio (VMR) computed across samples.

9. The method according to claim 8, wherein the bulk or spatial RNA-seq data do not have multiple samples, and multiple compound samples are synthesized, each of which is an average

of all cells belonging to one of the samples in the single cell reference.

10. The method according to claim 8, wherein if multiple samples are unavailable for both bulk or spatial RNA-seq data and single cell data, the method comprises generating multiple synthetic samples for the single cell data by averaging the expression of a subset of cells.

11. The method according to any one of claims 1 to 10, wherein the gene-wise scaling factors are estimated using an adaptive learning strategy, and each gene is rescaled with its respective scaling factor.

12. The method according to any one of claims 1 to 11, wherein each compound sample is independently estimated by the regression model.

13. A computer readable medium storing processor-executable instructions adapted to cause one or more computing devices to deconvolve bulk or spatial RNA-sequencing data for a sample by a method comprising the following steps:

i) obtaining input from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations for the sample;

ii) generating a UMI (unique molecular identifier) count matrix based on the bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations, wherein each column of the UMI count matrix corresponds to a cell within the sample or a spatial location within the sample, and each row of the UMI count matrix corresponds to a gene in the sample;

iii) selecting a subset of the most variably expressed genes from the UMI count matrix based on a union of genes whose expression is enriched in each cell type in the UMI count matrix, or a union of genes that vary above a predetermined level across all the cells in the UMI count matrix, or both;

iv) estimating the mean and dispersion parameters of the selected genes from the UMI count matrix;

v) estimating cross-sample gene variability from compound data or single cell samples;

vi) estimating, using a regression-based approach, gene-wise scaling factors using both

the bulk or spatial RNA-seq data and single cell RNA-seq data; and

vii) generating a regression model based on the cross-cell type specificity of each selected gene, the cross-sample gene variability from the bulk or spatial RNA-seq data or from the single cell samples, and the gene-wise scaling factors; and

viii) estimating, using the regression model, cell type proportions in the bulk or spatial RNA-sequencing data, wherein the regression model assigns a weight to the selected genes with high cell type specificity and low cross-sample variability;

thereby determining the percentage of cell types in the bulk or spatial RNA-sequencing data.

14. A system comprising:

one or more processors; and

a memory having processor executable instructions that, when executed by the one or more processors, cause the apparatus to deconvolve bulk or spatial RNA-sequencing data for a sample by a method comprising the following steps:

i) obtaining input from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations for the sample;

ii) generating a UMI (unique molecular identifier) count matrix based on the bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations, wherein each column of the UMI count matrix corresponds to a cell within the sample or a spatial location within the sample, and each row of the UMI count matrix corresponds to a gene in the sample;

iii) selecting a subset of the most variably expressed genes from the UMI count matrix based on a union of genes whose expression is enriched in each cell type in the UMI count matrix, or a union of genes that vary above a predetermined level across all the cells in the UMI count matrix, or both;

iv) estimating the mean and dispersion parameters of the selected genes from the UMI count matrix;

v) estimating cross-sample gene variability from compound data or single cell samples;

vi) estimating, using a regression-based approach, gene-wise scaling factors using both the bulk or spatial RNA-seq data and single cell RNA-seq data; and

vii) generating a regression model based on the cross-cell type specificity of each selected gene, the cross-sample gene variability from the bulk or spatial RNA-seq data or from the single cell samples, and the gene-wise scaling factors; and

viii) estimating, using the regression model, cell type proportions in the bulk or spatial RNA-sequencing data, wherein the regression model assigns a weight to the selcted genes with high cell type specificity and low cross-sample variability;

thereby determining the percentage of cell types in the bulk or spatial RNA-sequencing data.