What Is Claimed Is:

1. A method for deconvolving bulk or spatial RNA-sequencing data, the method comprising any one or more of the following steps:

a) obtaining input from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations, and selecting a subset of the most variably expressed genes from a normalized matrix of counts-based sequencing data;

b) estimating the mean and dispersion parameters of the expression per gene per cell type;

c) computing the cross cell type specificity of each gene;

d) estimating cross-sample gene variability from the bulk or spatial RNA-seq data or single cell samples depending on multi-sample availability;

e) estimating gene-wise scaling factors using both the bulk or spatial RNA-seq data and single cell data; and

f) building a weighted and regularized regression model using all of the known quantities and using the model to estimate cell type proportions in the bulk or spatial RNA-sequencing data;

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

2. The method according to claim 1, wherein the matrix of counts-based sequencing data comprises counts-based sequencing counts against each gene within a plurality of genes for a fixed number of cells.

3. The method according to claim 1 or claim 2, wherein the input is a single cell UMI count matrix with cell type annotations associated with each cell.

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

5. The method according to claim 1 or claim 2, wherein the spatial data to be deconvolved

is a UMI count matrix.

6. The method according to any one of claims 1 to 5, wherein genes that contain important information to differentiate cell types, excluding non-informative genes that potentially introduce noise, are selected as the subset of the most variably expressed genes from a normalized matrix of counts-based sequencing data.

7. The method according to claim 6, wherein the subset of genes is selected from the union of marker genes whose expression is enriched in each cell type in the single cell UMI count matrix.

8. The method according to claim 7, wherein a build-in tool takes as input the single cell UMI count matrix and cell type annotations.

9. The method according to claim 8, wherein for each cell type, the 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.

~~10. The method according to claim 9, wherein about the top 200 genes from each cell type are selected.~~

~~11. The method according to claim 9, wherein selected marker genes presenting in no more than five cell types are selected.~~

~~12. The method according to claim 9, wherein selected marker genes presenting in a fixed number of cell types or a proportion of total number of cell types, whichever is smaller, are selected.~~

~~13. The method according to claim 9, wherein about 1,000 total unique genes are selected.~~

~~14~~10. The method according to claim 6, wherein the subset of genes is selected from the union of highly variable genes that vary the most across all the cells in the single cell UMI count

matrix.

~~15. The method according to claim 14, wherein variances for each gene after cell number balancing and VST normalization are computed.~~

~~16. The method according to claim 15, wherein genes with the highest variances are selected.~~

~~17. The method according to any one of claims 14 to 16, wherein the cell types in the single cell UMI count matrix are balanced by finding the median size of all the cell clusters, wherein cells from each cluster are sampled to make them equal to this size.~~

~~18. The method according to claim 17, wherein the variance of each gene across the cells in the balanced single cell UMI matrix is computed.~~

~~19. The method according to claim 18, wherein variances on the normalized data are computed by variance stabilization transformed (VST).~~

~~20. The method according to claim 19, wherein genes with the top 2,000 large variances are selected.~~

~~21~~11. The method according to any one of claims 1 to 20, wherein the RNA-seq data is not normalized before estimating the mean.

~~22. The method according to claim 21, wherein the mean is modeled using the raw UMI counts.~~

~~23. The method according to claim 21 or claim 22, wherein negative binomial distributions are fit to single cells of each cell type.~~

~~24. The method according to claim 23, wherein estimations are performed for each selected~~

~~gene in each cell type.~~

~~25~~12. The method according to any one of claims 1 to 24, 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.

~~26. The method according to claim 25, wherein the estimated mean and variance parameters from the negative binomial fitting is used to compute the cell type specificity weight for each gene in the set of selected genes.~~

~~27~~13. The method according to any one of claims 1 to 26, wherein the cross-sample gene variability is computed using the variance-to-mean ratio (VMR) computed across samples.

~~28. The method according to claim 27, wherein the cross-sample gene variability is computed from compound transcriptome data.~~

~~29~~14. The method according to claim 27, wherein the compound data do not have multi-samples whereas the single cell data 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.

~~30~~15. The method according to claim 27, wherein if multi-samples are unavailable for both compound data and single cell data, the method comprises generating multiple synthetic samples for the single cell data by averaging the expression of subset of cells.

~~31~~16. The method according to any one of claims 1 to 30, wherein the gene-wise scaling factors are estimated using an adaptive learning strategy, and each gene is rescaled with its respective scaling factor.

~~32~~17. The method according to any one of claims 1 to 31, wherein each compound sample is

independently estimated by the regression model.

~~33~~18. A method for deconvolving bulk or spatial RNA-sequencing data, the method comprising any one or more of the following steps:

a) computing the cross cell type specificity of each gene within a subset of the most variably expressed genes selected from a normalized matrix of counts-based sequencing data obtained from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations;

b) estimating cross-sample gene variability from the bulk or spatial RNA-seq data or single cell samples depending on multi-sample availability;

c) estimating gene-wise scaling factors using both the bulk or spatial RNA-seq data and single cell data; and

d) building a weighted and regularized regression model using all of the known quantities and using the model to estimate cell type proportions in the bulk or spatial RNA-sequencing data;

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

~~34~~19. A computer readable medium storing processor-executable instructions adapted to cause one or more computing devices to deconvolve bulk or spatial RNA-sequencing data by a method comprising any one or more of the following steps: i) obtaining input from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations, and selecting a subset of the most variably expressed genes from a normalized matrix of counts-based sequencing data, wherein the matrix of counts-based sequencing data comprises counts-based sequencing counts against each gene within a plurality of genes for a fixed number of cells; ii) estimating the mean and dispersion parameters of the expression per gene per cell type; iii) computing the cross cell type specificity of genes; iv) estimating cross-sample gene variability from compound data or single cell samples, depending on multi-sample availability; v) estimating gene-wise scaling factors using both compound data and single cell data; and vi) building a weighted and regularized regression model using all of the known quantities, and using the model to estimate cell type proportions in the bulk or spatial RNA-sequencing data;

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

~~35~~20. A computer readable medium storing processor-executable instructions adapted to cause one or more computing devices to deconvolve bulk or spatial RNA-sequencing data, by a method comprising any one or more of the following steps: i) computing the cross cell type specificity of genes within a subset of the most variably expressed genes selected from a normalized matrix of counts-based sequencing data obtained from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations; ii) estimating cross-sample gene variability from compound data or single cell samples, depending on multi-sample availability; iii) estimating gene-wise scaling factors using both compound data and single cell data; and iv) building a weighted and regularized regression model using all of the known quantities, and using the model to estimate cell type proportions in the bulk or spatial RNA-sequencing data; thereby inferring the percentage of cell types in the bulk or spatial RNA-sequencing data.

~~36~~21. 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 by a method comprising any one or more of the following steps: i) obtaining input from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations, and selecting a subset of the most variably expressed genes from a normalized matrix of counts-based sequencing data, wherein the matrix of counts-based sequencing data comprises counts-based sequencing counts against each gene within a plurality of genes for a fixed number of cells; ii) estimating the mean and dispersion parameters of the data per gene per cell type; iii) computing the cross cell type specificity of genes; iv) estimating cross-sample gene variability from compound data or single cell samples, depending on multi-sample availability; v) estimating gene-wise scaling factors using both compound data and single cell data; and vi) building a weighted and regularized regression model using all of the known quantities and using the model to estimate cell type proportions in the bulk or spatial RNA-sequencing data; thereby inferring the percentage of cell types in the bulk or spatial RNA-sequencing data.

~~37~~22. 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, by a method comprising: i) computing the cross cell type specificity of genes within a subset of the most variably expressed genes selected from a normalized matrix of counts-based sequencing data obtained from sources comprising bulk or spatial RNA-seq data, single cell RNA-seq data, and cell type annotations; ii) estimating cross-sample gene variability from compound data or single cell samples, depending on multi-sample availability; iii) estimating gene-wise scaling factors using both compound data and single cell data; and iv) building a weighted and regularized regression model using all of the known quantities and using the model to estimate cell type proportions in the bulk or spatial RNA-sequencing data; thereby inferring the percentage of cell types in the bulk or spatial RNA-sequencing data.