Computational deconvolution: extracting cell type-specific information from heterogeneous samples

Shai S Shen-Orr[1,2,3] and Renaud Gaujoux[2]

The quanta unit of the immune system is the cell, yet analyzed samples are often heterogeneous with respect to cell subsets which can mislead result interpretation. Experimentally, researchers face a difficult choice whether to profile heterogeneous samples with the ensuing confounding effects, or a priori focus on a few cell subsets of interest, potentially limiting new discoveries. An attractive alternative solution is to extract cell subset-specific information directly from heterogeneous samples via computational deconvolution techniques, thereby capturing both cell-centered and whole system level context. Such approaches are capable of unraveling novel biology, undetectable otherwise. Here we review the present state of available deconvolution techniques, their advantages and limitations, with a focus on blood expression data and immunological studies in general.

Addresses
[1] Rappaport Institute of Medical Research, Technion-Israel Institute of Technology, Haifa 31096, Israel
[2] Department of Immunology, Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 31096, Israel
[3] Faculty of Biology, Technion-Israel Institute of Technology, Haifa 31096, Israel

Corresponding author: Shen-Orr, Shai S (shenorr@technion.ac.il)

Introduction

The cellular composition of many biological samples is heterogeneous and varying, that is multiple cell-type subsets are present in each sample, and different samples show high variance between one another in relative cell subset proportions (from hereon, heterogeneous sample). Moreover, many physiological and pathological processes involve cell motility (e.g. infiltration) and differentiation, ultimately resulting in marked shifts in sample cell subset composition (Figure 1a). An example of importance is peripheral blood, which is composed of many different immune cell subsets, comprising the fundamental units of a complex system, whose state and interaction activity reflect the type of biological processes taking place whether in health or disease. The ability to measure

and interpret phenotypic changes between specific conditions at the cell subset level is therefore critical to obtain a detailed understanding of the role of each cell subset within the immune system.

While the problem of sample heterogeneity has long been acknowledged [1,2••,3], researchers have struggled between the choice of focusing on a single cell subset or ignoring the problem and assaying heterogeneous samples. Whereas cell-isolation entails a loss of a systems perspective (i.e. biologically meaningful changes happening in multiple cell subsets and between them), ignoring sample heterogeneity entails misleading and difficult to interpret results. This has been an especially sore point in genomic scale studies (predominantly whole genome gene expression), where it is easy to lose sight of the natural cellular-context of the data amongst thousands of measured features and draw incorrect conclusions.

An emerging solution to this dilemma comes in the form of computational deconvolution methodologies, capable of extracting cell type-specific information directly from data generated from heterogeneous samples. Research addressing this issue started with the pioneering work of Venet *et al.* [4], but received relatively little attention. More recently, the increased application of genomic tools to human samples, which exhibit high sample heterogeneity, has spurred further developments. In particular, the availability of data from large efforts that profiled the gene expression of multiple known cell subsets and highlighted distinct differences between them [5••,6••] both motivated such computational algorithm development and served as useful input for them.

This article reviews this active area of research and draws attention on the advantages and limitations of the different proposed approaches. For the sake of simplicity, we focus on peripheral blood, which is the primary source of samples in human immunology studies, and in particular on gene expression data and group differences analyses. However research has concomitantly focused on other tissue and data types. Eventually, such techniques have the potential to provide a valuable cell-centered view of the immune system, bringing new insights into inter and intra cellular dynamics, and cell subset states in health and disease.

The confounding effects of sample cell-type subset heterogeneity

Most genes are expressed to varying degrees across multiple cell subsets in a tissue or organism, implying

Figure 1
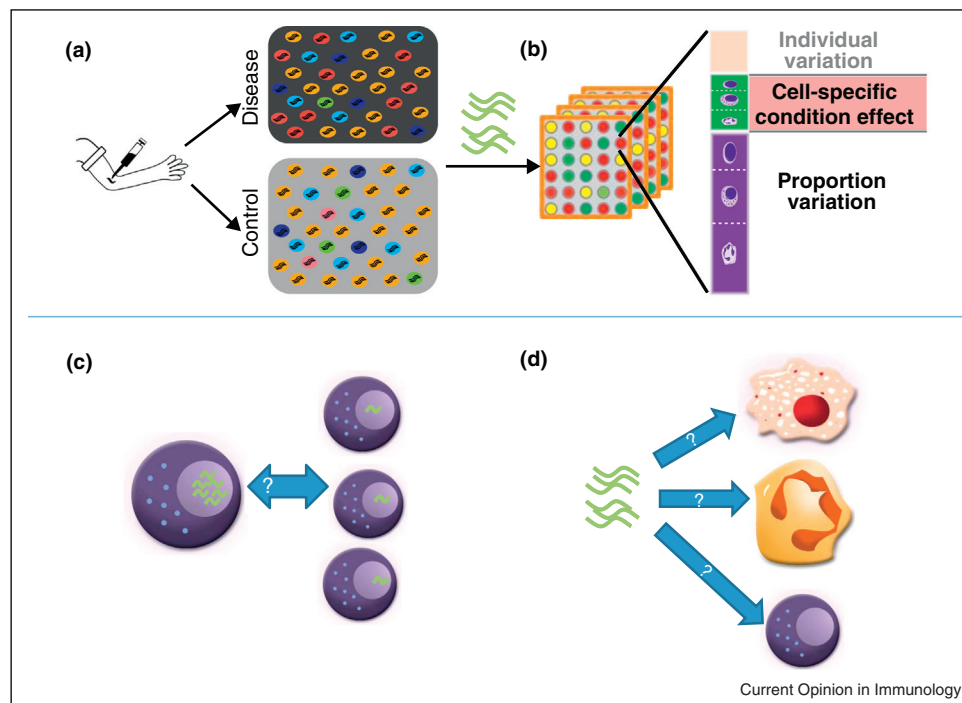
Biological samples are heterogeneous with respect to underlying cell subsets, with strong implications on downstream analysis. **(a)** Most tissue samples are composed of multiple cell subsets, and different samples show high variance between one another in relative cell subset proportions, especially under pathological conditions. **(b)** This implies that the total measured transcript abundance of a gene (as well as many other molecular species) is strongly affected by the cell subset composition of the sample and may be decoupled into three Abundance Components. Implications of sample heterogeneity include **(c)** an inability to identify whether increased total expression is due to the over-expression of a gene or to merely having more cells of a given subset in the sample, as well as **(d)** a difficulty in interpreting results and identifying the cell subset of origin of any detected differences.

that the measured abundance of any such transcript is confounded by the composition of the sample from which it is measured. More precisely, we may break down the total measured abundance of a gene in a sample into three *Abundance Components*: (1) that due to the characteristic condition of a sample (e.g. disease type, etc.), (2) that due to the individual variation, genotype-specific or technical measurement variation, and (3) that due to the average characteristic abundance of a gene as a function of the underlying cell subsets in a sample and their relative proportions (Figure 1b). The extent of contribution of each Abundance Component to total gene expression measured differs between cell subsets, conditions and individuals on a gene by gene basis. In general, however, we note that whereas Abundance Components (1) and (2) describe important biological differences in gene expression (excluding individual technical variation), Abundance Component (3) reflects gene expression variations that are more succinctly captured and interpreted by reporting differences in proportions of cell subsets.

As the composition of relative cell-type subset proportions varies between samples, analysis of sample differences by any methodology oblivious to the under-lying cell-type heterogeneity will suffer from several strong drawbacks:

Misleading information: if a cell subset's proportion difference is correlated with the phenotype of interest, classical analysis methods (i.e. cell subset agnostic) are prone to produce more false positive differentially expressed genes by detecting differences in total abundance stemming from this shared overall correlation (Figure 1c). Ideally, differences due to biological condition could be decoupled from those resulting from an artifact of sample heterogeneity (see next section).

Signal dilution: differences in genes expressed in cell subsets present at low proportion in a sample may be masked by the signal coming from the same gene expressed in a prevalent cell subset, or alternatively by a high sample-wise variance of the proportions of cell subsets in which the gene is expressed. A prime example of this problem is observed when analyzing gene expression samples derived from whole blood versus PBMCs of the same subjects, where tissue type is a greater determinant of similarity than subject condition [2••,7•] (primarily due to the presence of

neutrophils, which constitute anywhere from 50-70% of whole blood samples).

Result interpretation: without accounting for varying cell subset proportions, it is difficult to attribute the observed effect to any given cell subset (Figure 1d). Often a follow up experiment is required to understand from which cell subset a detected signal originated. Eventually, this makes it harder to build a detailed picture of the immune system and precisely delineate its main driving components — in the considered condition.

Given the above raised difficulties, it may be surprising that sample collection and analysis are still very commonly performed on highly heterogeneous tissues (e.g. peripheral blood in humans, spleen in the case of mice), despite the availability of experimental techniques to isolate and profile fairly homogeneous cell-population, and most recently to profile them at the single cell level [8]. From a research perspective standpoint, effectively choosing whether to profile a heterogeneous tissue or to isolate specific cell subsets should depend on whether one wishes to get a systems perspective by profiling all accessible cell subsets involved in a certain biological process, or conduct a focused study in which cell subsets of interest are a priori chosen. This is especially a problem in genomic level studies which often strive for a systems perspective but where the interpretation of data without cellular context is most difficult. In practice however, based on our own experiences, the decision as to how and what to profile is dictated by limitations in sample material, financial considerations, concerns with respect to biases introduced during isolation procedures [2••], or an under-appreciation to the inherent problems in using heterogeneous samples.

Extracting cell type-specific information from heterogeneous tissue

An attractive approach for gaining insight on cell-subset specific information is to estimate the proportion and/or gene expression profile of different cell subsets directly from the heterogeneous samples via computational methodologies, thereby preserving the whole-systems perspective, as well as obtaining an unbiased cell-based view (Figure 2).

Figure 2

Current Opinion in Immunology
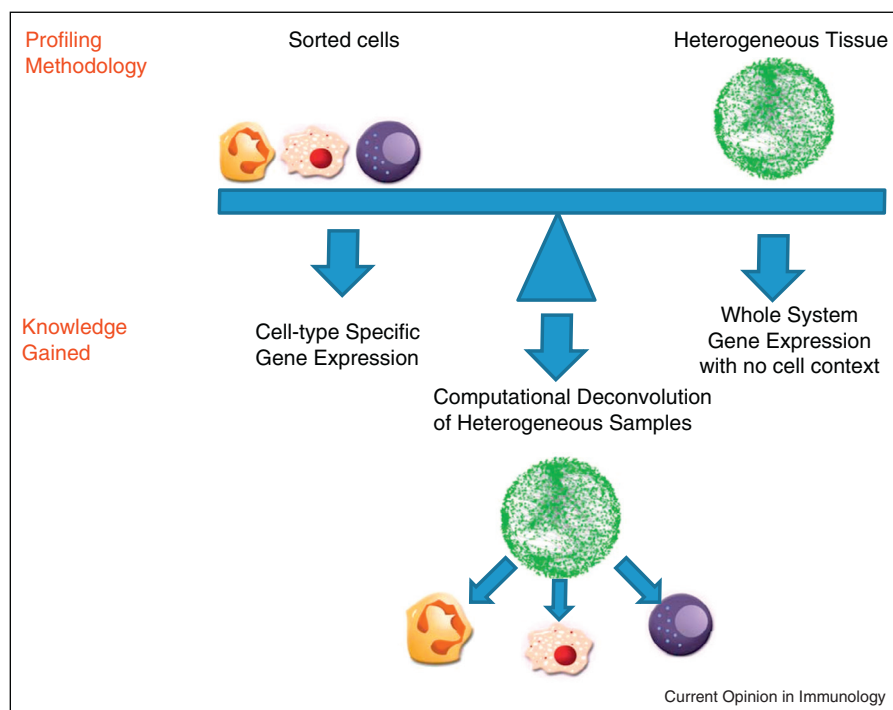
Computational deconvolution methodologies enable capturing both cell-centered and system wide information. Experimental methodologies for dealing with sample heterogeneity require either to isolate cells of interest, which perturbs the cells, entails a loss of perspective on the whole system, and is biased toward prior knowledge on which cells are of interest. The alternative, that of profiling the heterogeneous sample directly, provides a whole system view, which, however, lacks any cellular context. Computational deconvolution methodologies offer an intermediate alternative and allow to capture system level information in a cell-centered manner, a model proper to immunology, namely cells interacting with one another.

Present computational methodologies for extracting cell type-specific information from heterogeneous sample data may be divided into five main method classes (Figure 3, Table 1 for a detailed list of methodologies, and Supplementary Material for in depth review of requirements and limitations of each class), which we define based on the type of input they require, and the resolution of the output they offer: Quantifying the presence of different cell subsets in heterogeneous samples may be performed either at low-resolution, by (a) cell-type enrichment [9–14], or at high-resolution, by (b) cell proportion estimation methodologies [15••,16–19]. Both method classes rely on reference signature expression profiles of the different cell subsets or known cell subset-specific marker genes [20,12,21–24], and respectively provide detection of the presence/implication of a cell subset in an heterogeneous sample or a disease condition signature (Figure 3a), or actual numeric estimates of the relative proportions of each cell subset in the data (Figure 3b). Estimating a condition effect accounting for sample heterogeneity may be performed either at low-resolution, by (c) sample heterogeneity correction [25••,26,27••,28•,29], or at high-resolution, by (d) cell type-specific deconvolution methodologies [25••,26,30]. Both methodologies rely on the availability of the proportions of one or more cell subsets present in samples, and respectively provide a correction of the measured sample data for biases introduced due to cell subset sample heterogeneity (Figure 3c), or estimates of cell subset-specific gene expression profiles for each cell subset — for which proportion information was provided, which can be subsequently used to infer cell type-specific differential expression between groups of samples (Figure 3d). We note that the input cell-subset proportions in these two method classes, may come either from actual measurements or computationally estimated. In fact, a fifth category (e) consists of *complete deconvolution* methods which estimate both proportions and cell type-specific expression profiles, often using a combination of deconvolution methods (b and d), and require some limited prior knowledge on proportions [31] or expression profiles [19,30,32,33,34••,35,36].

## Gain in biological insights

Computational deconvolution methods aim at providing a cell-centered view of heterogeneous molecular data, by decoupling the effect of proportion from cell type-specific phenotype. In particular, they have the potential to mine high-throughput data in a way that even upcoming laboratory techniques may not yet or ever handle, for example, due to limitations of cell surface markers for cell-sorting or to the mere unavailability of biological material for past studies. Notably, they have already proved to be able to provide new insights in complex diseases such as autoimmune disease or cancer. [15••] deconvolve whole blood samples from Systemic Lupus Erythematosus (SLE)

patients, identifying specific changes in leukocyte proportions and activation (NK, T and monocytes, in particular), as well as correlation of proportions with treatment type and other clinical measures. Deconvolution based cell-type specific differential expression of acute rejection versus stable patients identified a previously unsuspected role for monocytes in both kidney [25••] and cardiac transplant [39], respectively undetectable or only mildly detectable from whole blood. [27••] showed that DNA methylation is a potential mediator of genetic risk in Rheumatoid Arthritis (RA), and highlighted the importance of correcting for sample heterogeneity in blood DNA methylation data. [34••] showed that significant improvements in outcome prediction of lung and prostate cancer can be achieved when building a classifier on computationally purified tumor data, as opposed to data from bulk biopsies. Given the large and increasing number of cell-types known, and the desire to capture their difference in behavior, we foresee deconvolution methodologies, which offer increased resolution and interpretability at a little or no extra costs, being increasingly utilized such that they become part of main stream analysis pipelines in human profiling studies.

Limitations of computational approaches

Despite the various successful application of computational deconvolution methodologies, we believe several open issues remain to be investigated before they become widely adopted. First, a better understanding of the accuracy lower bound for estimates of cell subsets proportions or differential gene expression detection must be developed. This general accuracy is difficult to assess because of the many factors to consider (proportion dependencies, individual variation, clinical condition, cell-cell interaction), a large scale evaluation on simulated and public data could provide much information with respect to their power. Second, and particularly relevant in the case of blood, is the development of algorithms capable of performing '*deep deconvolution*', that is, accurately estimating from a whole blood or PBMC sample, the proportions and expression patterns of a greater number of cell subsets, going further down into the hematopoietic tree (T-regs, naive, memory, and effector cell subsets). One challenge in achieving this is the minimum required sample size that increases together with the number of considered cell subsets, a restriction that may be lifted using regularization methodologies. Finally more research on how to efficiently perform single-sample deconvolution, where cell type-specific profiles are estimated for each individual sample rather than groups, as already proposed in cancer tissues [34••], would enable moving toward more personalized tests. Of note, one of us [RG] has recently published an R package, *CellMix* [40•], which compiles together many of the published computational gene expression deconvolution methodologies, and facilitates future algorithm development and benchmarking.

Figure 3
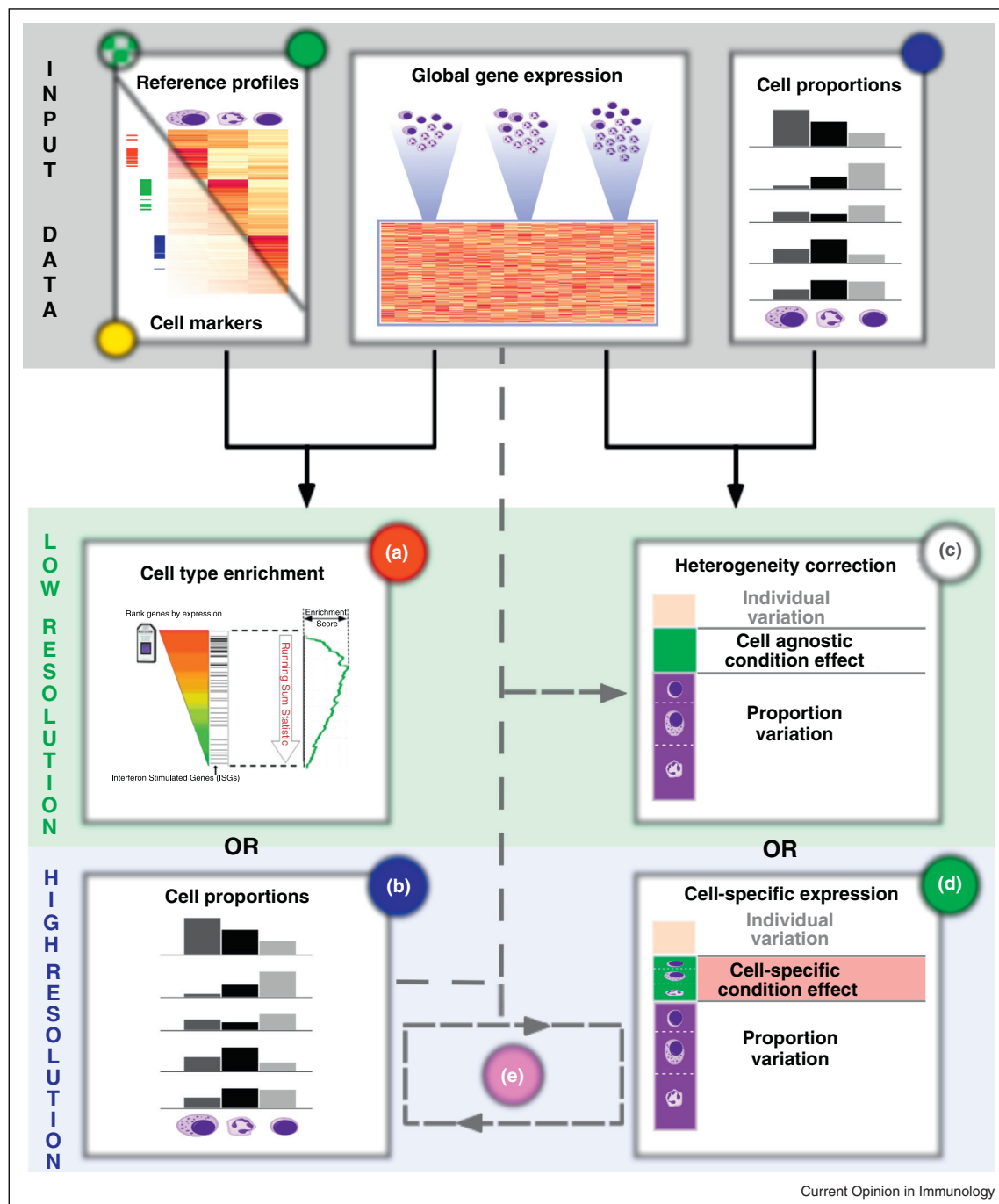
Five classes of computational approaches that extract cell type-specific information from heterogeneous sample data. Different classes of deconvolution methods defined according to the combination of the input data they require and the type and resolution of output they offer. All methods use data from heterogeneous samples, combined with either markers, signatures or proportions to **(a)** detect cell presence or implication of cell types, **(b)** estimate cell proportions, **(c)** correct for heterogeneity, or **(d)** estimate cell type-specific expression profiles. Dotted line indicates a possibility of using the output of one class of methods as input for another. Complete deconvolution methods **(e)** alternately estimate proportions from cell type-specific expression and vice-versa, starting with some limited prior knowledge on proportions or expression profiles (signatures, markers).

Table 1

Deconvolution methods with an available user interface

	In	Out	Name	Description	Availability	Tissue
A	○	A	SPEC★	Predicts the most likely cellular source of a given gene expression signature [11]	clip.med.yale.edu/SPEC R:SPEC	Blood
	○	A	CTEN★	Identifies enriched cell types in heterogeneous microarray data [12]	www.influenza-x.org/ jshoemaker/cten	-
	○	A	ssGSEA	Single sample GSEA [10]	ssGSEAProjection in GenePattern [9]	Cancer
B	○	B	collapseRows	Aggregates/selects a proportion proxy within cell type-specific co-expression modules [37]	R:WGCNA	-
	●	B	Abbas	Estimates proportions of 17 immune cell subsets using IRIS-based signatures [15]	CellMix	Blood
	●	B	DeconRNAseq	Similar to *Abbas* but uses quadratic programming instead of standard regression [38]	R + CellMix	-
	●	B (c)	PERT	Perturbation model that estimates proportions and a global condition effect the reflects deviance from reference pure profiles [26]	Octave code	Blood
	●	B	methylSpectrum	Estimate proportions from DNA methylation reference profiles [28]	R:methylSpectrum	Blood
	⊗	B	qpure	Estimate tumor cellularity SNP microarray data from paired (tumor and normal) samples [17]	R:qpure	Cancer
	⊗	B	ABSOLUTE	Infers tumor purity and malignant cell ploidy directly Copy-Number-Variation data and precomputed models of recurrent cancer karyotypes [18]	R	Cancer
D	●	D (c)	csSAM★	Estimates cell/tissue specific signatures from known proportions using SAM [25]	R:csSAM + XLS plugin + CellMix	-
	○	D	PSEA★	Population-Specific Expression Analysis, using a regression model selection schema [30]	R function(s)	-
E	⊗	B D	DeMix★	Estimate tumor fraction and individual purified tumor profiles using normal tissue profiles [35]	R function(s)	Cancer
	⊗	B D	ISOpure	Estimate tumor fraction and individual purified sample profiles using normal tissue profiles [34]	Matlab	Cancer
	●	B D	DSection★	Bayesian MCMC-based estimation from priors on cell proportions [31]	informatics.systemsbiology.net/DSection Matlab + CellMix	-
	○	B D	DSA	Digital Sorting Algorithm: complete deconvolution using a set of linear equations and quadratic programming [19]	R:dsa + CellMix	-
	○	B D	deconf	Alternate least-square NMF method, using heuristic constraints [33]	R:deconf + CellMix	Blood
	○	B D	ssNMF	Semi-supervised NMF algorithms, enforcing marker gene expression patterns [32]	CellMix	-

Methods are grouped by classes, which are identified using the labels from Figure 3. Methods are ordered by class. Methods matching more than one class are classified by the highest resolution they provide. For each method, the type of input required and output generated is listed. The color bullets match the labels of relevant blocks in Figure 3. ★ denotes methods that provide built-in capabilities for statistical testing or confidence interval estimation. For each method, the type of software implementation is indicated. For *R*, the name of the package implementing the method is listed, if available. *CellMix*, is an *R* package compiling together in a standardized interface many of the published computational deconvolution methodologies for gene expression data. The tissue for which a method was developed and may be expected to perform best is also listed.

Conclusion

Cell subset heterogeneity is inherent to most primary biological samples, which may confound downstream data analysis if it is not taken into account and strongly restricts result interpretability. From a systems immunology perspective to health and disease, it is critical to be able to assess each cell subset's state and interactions, over a range of condition and molecular environment. In this respect, computational deconvolution methodologies showed to be powerful tools capable of providing novel high-resolution system-wide insights.

In this review, we focused our discussion on gene expression studies, where the majority of research on the implications of sample cell subset heterogeneity for downstream analysis and on the benefit of addressing it has been done to date. Expressed transcripts are of course not the only molecular species whose interpretation of measured abundance is affected by sample cell subset heterogeneity. For example, computational techniques have already been applied to DNA methylation data, which revealed new biology otherwise masked behind sample heterogeneity [27••,28•]. These are not the

exception, and we would rather propose that the opposite is the case, namely the majority of molecular species assayed vary in their total abundance from one cell subset to another, and perhaps even for secreted molecular species. Thus their measurement and analysis would be affected by sample composition and benefit from computational techniques aimed at providing a cell-centered view of the system. We envision that different molecular species and measurement modalities would require the tailoring of their own particular methodology, but expect that any new methodology development will likely fall into one of the five method classes we defined here for obtaining cell type-specific information from cell subset heterogeneous gene expression samples.

Acknowledgements

Appendix A. Supplementary data
Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.coi.2013.09.015

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- •• of outstanding interest

1. Davey HM, Kell DB: **Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses**. *Microbiol Rev* 1996, **60**:641-696 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=239459&tool=pmcentrez&rendertype=abstract.

2. Whitney A, Diehn M, Popper S, Alizadeh A, Boldrick J, Relman D
•• *et al.*: **Individuality and variation in gene expression patterns in human blood**. *Proc Natl Acad Sci U S A* 2003, **100**:1896 http://www.pnas.org/content/100/4/1896.short.
The authors provide a comprehensive analysis of factors of variation in whole blood and PBMC gene expression, notably showing that a significant fraction of the observed variation refelects differences in cell type proportions.

3. De Ridder D, Van Der Linden CE, Schonewille T, Dik WA, Reinders MJT, Van Dongen JJM *et al.*: **Purity for clarity: the need for purification of tumor cells in DNA microarray studies**. *Leukemia: Off J Leukemia Soc Am Leukemia Res Fund UK* 2005, **19**:618-627 http://www.ncbi.nlm.nih.gov/pubmed/15744349.

4. Venet D, Pecasse F, Maenhaut C, Bersini H: **Separation of samples into their constituents using gene expression data**. *Bioinformatics* 2001, **17(Suppl. 1)**:S279 http://bioinformatics.oxfordjournals.org/content/17/suppl_1/S279.short.

5. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN,
•• McConkey ME *et al.*: **Densely interconnected transcriptional circuits control cell states in human hematopoiesis**. *Cell* 2011, **144**:296-309 http://dx.doi.org/10.1016/j.cell.2011.01.004 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3049864&tool=pmcentrez&rendertype=abstract.
D-MAP: the authors generated a dataset of 39 human cell types across the whole hematopoietic tree, showing, in particular, that gene expression programs get reused across lineages and cell subsets are generally more similar to one another in deep branches of the hematopoietic tree.

6. Shay T, Kang J: **Immunological genome project and systems
•• immunology**. *Trends Immunol* 2013 http://www.cell.com/trends/immunology/abstract/S1471-4906(13)00049-5.
The ImmGen project generated cell-specific profiles of 249 immune cells

in mice, and provide analytical tools to browse and mine the data for a variety of cell-specific information.

7. Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV,
• Xiao W *et al.*: **Application of genome-wide expression analysis to human health and disease**. *Proc Natl Acad Sci U S A* 2005, **102**:4801-4806 http://dx.doi.org/10.1073/pnas.0409768102 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=555033&tool=pmcentrez&rendertype=abstract.
Comparison of gene expression from whole blood and sorted cells (here T cells and Monocytes) suggests that differences in leukocyte proportions consititutes a major source of variation of the apparent global gene expression.

8. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR *et al.*: **Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells**. *Nat Biotechnol* 2012, **30**:777-782 http://dx.doi.org/10.1038/nbt.2282 http://www.nature.com/doifinder/10.1038/nbt.2282.

9. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF *et al.*: **Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1**. *Nature* 2009, **462**:108-112 http://www.ncbi.nlm.nih.gov/pubmed/19847166.

10. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD *et al.*: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1**. *Cancer Cell* 2010, **17**:98-110 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2818769&tool=pmcentrez&rendertype=abstract.

11. Bolen CR, Uduman M, Kleinstein SH: **Cell subset prediction for blood genomic studies**. *BMC Bioinformatics* 2011, **12**:258 http://dx.doi.org/10.1186/1471-2105-12-258 http://www.biomedcentral.com/1471-2105/12/258.

12. Shoemaker JE, Lopes TJ, Ghosh S, Matsuoka Y, Kawaoka Y, Kitano H: **CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data**. *BMC Genomics* 2012, **13**:460 http://dx.doi.org/10.1186/1471-2164-13-460 http://www.biomedcentral.com/1471-2164/13/460.

13. Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG: **Global prediction of tissue-specific gene expression and context-dependent gene networks in *Caenorhabditis elegans***. *PLoS Comput Biol* 2009, **5**:e1000417 http://dx.doi.org/10.1371/journal.pcbi.1000417 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2692103&tool=pmcentrez&rendertype=abstract.

14. Nakaya H, Wrammert J, Lee E: **Systems biology of seasonal influenza vaccination in humans**. *Nature* 2011, **12**:786-795 http://dx.doi.org/10.1038/ni.2067.Systems http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3140559/.

15. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF:
•• **Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus**. *PLoS ONE* 2009, **4**:e6098 http://dx.doi.org/10.1371/journal.pone.0006098 http://www.ncbi.nlm.nih.gov/pubmed/19568420.
The authors built an set of signature gene expression profiles for 17 immune resting/activated cell subsets, optimized for their ability to accurately estimate cell proportions in whole blood gene expression microarray data.

16. Gong T, Szustakowski JD: **DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data**. *Bioinformatics (Oxford, Engl)* 2013:1-2 http://dx.doi.org/10.1093/bioinformatics/btt090 http://www.ncbi.nlm.nih.gov/pubmed/23428642.

17. Song S, Nones K, Miller D, Harliwong I, Kassahn KS, Pinese M *et al.*: **qpure. A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles**. *PLoS ONE* 2012, **7**:e45835 http://dx.doi.org/10.1371/journal.pone.0045835 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3457972&tool=pmcentrez&rendertype=abstract.

18. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T *et al.*: **Absolute quantification of somatic DNA alterations in human cancer**. *Nat Biotechnol* 2012, **30**:413-421 http://

dx.doi.org/10.1038/nbt.2203 http://www.nature.com/doifinder/10.1038/nbt.2203.

19. Zhong Y, Wan YW, Pang K, Chow LM, Liu Z: **Digital sorting of complex tissues for cell type-specific gene expression profiles**. *BMC Bioinformatics* 2013, **14**:89 http://dx.doi.org/10.1186/1471-2105-14-89 http://www.biomedcentral.com/1471-2105/14/89.

20. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**:1739-1740 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3106198&tool=pmcentrez&rendertype=abstract.

21. Liu W, Yuan K, Ye D: **Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis**. *J Biomed Informatics* 2008, **41**:602-606 http://dx.doi.org/10.1016/j.jbi.2007.12.003 http://www.ncbi.nlm.nih.gov/pubmed/18234564.

22. Yang X, Ye Y, Wang G, Huang H, Yu D, Liang S: **VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery**. *Physiol Genomics* 2011, **43**:457-460 http://dx.doi.org/10.1152/physiolgenomics.00178.2010 http://physiolgenomics.physiology.org/cgi/content/abstract/43/8/457.

23. Xiao SJ, Zhang C, Zou Q, Ji ZL: **TiSGeD: a database for tissue-specific genes**. *Bioinformatics (Oxford, Engl)* 2010, **26**:1273-1275 http://dx.doi.org/10.1093/bioinformatics/btq109 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859128&tool=pmcentrez&rendertype=abstract.

24. Kogenaru S, Val C, Hotz-Wagenblatt A, Glatting KH: **TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles**. *Theor Chem Acc* 2009, **125**:651-658 http://dx.doi.org/10.1007/s00214-009-0670-5 http://www.springerlink.com/index/10.1007/s00214-009-0670-5.

25. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F,
•• Perry NM et al.: **Cell type-specific gene expression differences in complex tissues**. *Nat Methods* 2010, **7**:287-289 http://dx.doi.org/10.1038/nmeth.1439 http://www.ncbi.nlm.nih.gov/pubmed/20208531.
Computational method to estimate cell-specfic differential expression between two groups of samples (using SAM). A number of other authors used it to successfully identified novel cell-driven disease signatures.

26. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW: **PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions**. *PLoS Comput Biol* 2012, **8**:e1002838 http://dx.doi.org/10.1371/journal.pcbi.1002838 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3527275&tool=pmcentrez&rendertype=abstract.

27. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E,
•• Runarsson A et al.: **Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis**. *Nat Biotechnol* 2013, **31**:142-147 http://dx.doi.org/10.1038/nbt.2487 http://www.ncbi.nlm.nih.gov/pubmed/23334450.
The authors show that DNA methylation correction for heterogeneity significantly reduces the association signals of CpG methylation with disease (here rheumatoid arthritis (RA)), and highlight the importance of taking into account sample heterogeneity in the analysis of these data.

28. Houseman EA, Accomando WP, Koestler DC, Christensen BC,
• Marsit CJ, Nelson HH et al.: **DNA methylation arrays as surrogate measures of cell mixture distribution**. *BMC Bioinformatics* 2012, **13**:86 http://dx.doi.org/10.1186/1471-2105-13-86 http://www.biomedcentral.com/1471-2105/13/86.
The authors propose a deconvolution method for DNA metylation data, that accurately estimates the proportions of eight immune cell subsets from whole blood samples.

29. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3** http://dx.doi.org/10.2202/1544-6115.1027 http://www.ncbi.nlm.nih.gov/pubmed/16646809.

30. Kuhn A, Thu D, Waldvogel HJ, Faull RLM, Luthi-Carter R: **Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain**. *Nat Methods* 2011, **8**:945-947 http://dx.doi.org/10.1038/nmeth.1710 http://www.ncbi.nlm.nih.gov/pubmed/21983921.

31. a TE, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lähdesmäki H: **Probabilistic analysis of gene expression measurements from heterogeneous tissues**. *Bioinformatics (Oxford, Engl)* 2010, **26**:2571-2577 http://dx.doi.org/10.1093/bioinformatics/btq406 http://www.ncbi.nlm.nih.gov/pubmed/20631160.

32. Gaujoux R, Seoighe C: **Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study**. *Infect Genet Evol: J Mol Epidemiol Evol Genet Infect Dis* 2012, **12**:913-921 http://dx.doi.org/10.1016/j.meegid.2011.08.014 http://www.ncbi.nlm.nih.gov/pubmed/21930246.

33. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J et al.: **Biomarker discovery in heterogeneous tissue samples − taking the in-silico deconfounding approach**. *BMC Bioinformatics* 2010, **11**:27 http://dx.doi.org/10.1186/1471-2105-11-27 http://www.ncbi.nlm.nih.gov/pubmed/20070912.

34. Quon G, Haider S, Deshwar AG, Cui A, Boutros PC, Morris Q:
•• **Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction**. *Genome Med* 2013, **5**:29 http://dx.doi.org/10.1186/gm433 http://www.ncbi.nlm.nih.gov/pubmed/23537167.
Single sample deconvolution: given a set of separate profiles from tumor sample and their surrounding healthy tissue (not necessarily matched), the proposed method estimates, for each individual tumor sample, the fraction of normal tissue as well as the 'pure' expression profiles of each tissue type (normal/cancer).

35. Ahn J, Yuan Y, Parmigiani G, Suraokar MB, Diao L, Wistuba II et al.: **DeMix: deconvolution for mixed cancer transcriptomes using raw measured data**. *Bioinformatics (Oxford, Engl)* 2013, **29**:1865-1871 http://dx.doi.org/10.1093/bioinformatics/btt301 http://www.ncbi.nlm.nih.gov/pubmed/23712657.

36. Zuckerman NS, Noam Y, Goldsmith AJ, Lee PP: **A self-directed method for cell-type identification and separation of gene expression microarrays**. *PLoS Comput Biol* 2013, **9**:e1003189 http://dx.doi.org/10.1371/journal.pcbi.1003189 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3749952&tool=pmcentrez&rendertype=abstract.

37. Miller J, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR et al.: **Strategies for aggregating gene expression data: the collapseRows R function**. *BMC Bioinformatics* 2011, **12**:322 http://dx.doi.org/10.1186/1471-2105-12-322 http://www.biomedcentral.com/1471-2105/12/322.

38. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M et al.: **Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples**. *PLoS ONE* 2011, **6**:e27156 http://dx.doi.org/10.1371/journal.pone.0027156 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3217948&tool=pmcentrez&rendertype=abstract.

39. Shannon CP, Hollander Z, Wilson-McManus J, Balshaw R, Ng RT, McMaster R et al.: **White blood cell differentials enrich whole blood expression data in the context of acute cardiac allograft rejection**. *Bioinformatics Biol Insights* 2012, **6**:49-61 http://dx.doi.org/10.4137/BBI.S9197 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3329187&tool=pmcentrez&rendertype=abstract.

40. Gaujoux R, Seoighe C: **CellMix: a comprehensive toolbox for**
• **gene expression deconvolution**. *Bioinformatics (Oxford, Engl)* 2013:1-2 http://dx.doi.org/10.1093/bioinformatics/btt351 http://www.ncbi.nlm.nih.gov/pubmed/23825367.
R package that provides a standardized interface to many gene expression deconvolution methods, as well as related data like cell-specific marker gene sets or benchmark datasets, where cell proportions or sorted cell profiles are available.