Bioinformatics, 35(12), 2019, 2093–2099
doi: 10.1093/bioinformatics/bty926
Advance Access Publication Date: 8 November 2018
Original Paper

OXFORD

Gene expression

dtangle: accurate and robust cell type deconvolution

Gregory J. Hunt [1,*], Saskia Freytag[2,3], Melanie Bahlo[2,3] and Johann A. Gagnon-Bartsch[1,*]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA, [2]Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia and [3]Department of Medical Biology, University of Melbourne, Parkville, VIC 3010, Australia

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on December 21, 2017; revised on October 20, 2018; editorial decision on November 4, 2018; accepted on November 6, 2018

Abstract

Motivation: Cell type composition of tissues is important in many biological processes. To help understand cell type composition using gene expression data, methods of estimating (deconvolving) cell type proportions have been developed. Such estimates are often used to adjust for confounding effects of cell type in differential expression analysis (DEA).

Results: We propose dtangle, a new cell type deconvolution method. dtangle works on a range of DNA microarray and bulk RNA-seq platforms. It estimates cell type proportions using publicly available, often cross-platform, reference data. We evaluate dtangle on 11 benchmark datasets showing that dtangle is competitive with published deconvolution methods, is robust to outliers and selection of tuning parameters, and is fast. As a case study, we investigate the human immune response to Lyme disease. dtangle's estimates reveal a temporal trend consistent with previous findings and are important covariates for DEA across disease status.

Availability and implementation: dtangle is on CRAN (cran.r-project.org/package=dtangle) or github (dtangle.github.io).

Contact: gjhunt@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Complex organisms have a vast collection of specialized cell types. The presence and interaction of these cell types is important to understanding many biological processes. For example, shifts in the relative composition of cell types is important to developmental processes of organisms including embryogenesis, morphogensis, cell differentiation and growth (Lu *et al.*, 2003). Likewise, understanding the presence or absence of cell types is of direct etiological interest for many diseases and dysfunctions (Abbas *et al.*, 2009; Altboum *et al.*, 2014; Lu *et al.*, 2003; Newman *et al.*, 2015). For example, changes in glial populations in brain tissue are characteristic of Alzheimer's disease (Mohammadi *et al.*, 2015). Similarly, white blood cell composition can be indicative of acute cellular rejection of transplanted kidneys (Shen-Orr *et al.*, 2010). Cell type composition is also important in tumorigenic processes. It has been shown that heterogeneity of tumors cells is implicated in the metastatic potential of cancer (Lu *et al.*, 2003; Marusyk and Polyak, 2011).

Given the importance of understanding cell type composition, several methods to estimate cell type proportions using high-throughput gene profiling experiments have been developed. Known as 'cell type deconvolution', these methods have been successfully employed in a variety of applications. Deconvolution algorithms have been used to study cell type compositional changes in patients in clinical studies (Abbas *et al.*, 2009; Altboum *et al.*, 2014; Bowling *et al.*, 2017; Gong *et al.*, 2011; Newman *et al.*, 2015). In these studies, estimating constituent cell types of carefully selected tissues reveals important cell type compositional dynamics of diseases.

Similarly, such gene expression deconvolution has been posited as useful for clinical cell type monitoring, for example, by tracking patients' leukocytes (Newman *et al.*, 2015). Finally, estimating cell type proportions is important for deconfounding differential expression analysis. In differential expression studies detecting gene expression differences within each cell type is confounded by changes in the cell type composition across the factor of interest. For example, diseases will simultaneously affect changes in gene expression within each cell type and through compositional changes in the tissues. Including estimated proportions of cell types to account for this confounding has been shown to improve differential expression analysis (Capurro *et al.*, 2015; Hagenauer *et al.*, 2016).

We present dtangle, a new deconvolution method that is accurate, robust and simple to compute. It estimates cell type proportions using biologically plausible models of high throughput profiling technology. We compare dtangle to other methods on 11 benchmark datasets. These datasets include many different cell types, profiling technologies and cover realistic scenarios like batch effects, mixed technologies and third party references. Analysis of this data shows that dtangle outcompetes existing methods in a broad range of applications.

2 Materials and methods

dtangle requires two pieces of external knowledge: (i) reference data and (ii) marker genes. First, dtangle requires auxiliary gene expression reference data for each cell type [e.g. from GEO (Edgar, 2002)]. Second, dtangle requires marker genes for each cell type. A gene is defined as marker of a cell type if it is predominantly expressed by that type. dtangle can determine marker genes using the reference data or they may be specified by the user.

dtangle's approach is built on a biologically appropriate linear mixing model of linear-scale expressions but robustly fitting the model using log-transformed data and thus sets it apart from other deconvolution methods.

2.1 The dtangle estimator

In this section we describe the mathematical form of dtangle's estimator. Intuition for the estimator follows in subsequent sections. Assume we have a mixture sample of K cell types. Let $Y \in \mathbb{R}^N$ be the (base-2) log-scale expression measurements of this mixture sample and $p_1, \ldots, p_K$ be the mixing proportions of the cell types. For $k = 1, \ldots, K$ assume that there are ν_k reference samples of cell type k and let $Z_{kr} \in \mathbb{R}^N$ be the log-scale expressions of the rth type k reference. Furthermore, let $G_k \subset \{1, \ldots, N\}$ be the set of type k marker genes. These marker gene sets are mutually disjoint.

Let $g_k = |G_k|$ and define $\overline{Y_{G_k}} = \frac{1}{g_k}\sum_{n \in G_k} Y_n$ and $\overline{Z_{G_k}} = \frac{1}{g_k \nu_k}\sum_{n \in G_k}\sum_{r=1}^{\nu_k} Z_{krn}$ to be the average of all type k marker genes across the mixture and reference samples, respectively. Define $D_{kt} = \frac{1}{\gamma}((\overline{Y_{G_k}} - \overline{Y_{G_t}}) - (\overline{Z_{G_k}} - \overline{Z_{G_t}}))$ and $D_k = (D_{k1}, \ldots, D_{kK})$. The value D_{kt} is a normalized measure of the type k marker genes' expression over the type t markers' expressions in the mixture. Precisely, D_{kt} is the average difference of marker expressions, $\overline{Y_{G_k}} - \overline{Y_{G_t}}$, baseline normalized by their average difference across the references, $\overline{Z_{G_k}} - \overline{Z_{G_t}}$, and adjusted by γ, a term we discuss in detail later. We estimate p_k by mapping $D_k \in \mathbb{R}^K$ into the unit interval $[0, 1]$ by a multivariate logistic function $L_k : \mathbb{R}^K \to [0, 1]$. Precisely, for $x \in \mathbb{R}^K$ let $L_k(x) = 1/(1 + \sum_{t \neq k} 2^{-x_t})$ and estimate p_k as

$$\widehat{p_k} = L_k(D_k) \tag{1}$$

(see Supplementary Section S5 for details). This definition ensures that $\widehat{p_k} \geq 0$ and $\sum_{k=1}^K \widehat{p_k} = 1$.

2.2 Motivation and model

Let us first define some terminology. Measured expressions are determined by a gene expression profiling (GEP) technology by measuring the amount of mRNA transcribed from each gene. Typically these measured expressions are further summarized, e.g. by MAS or RMA, and normalized, e.g. quantile or TPM normalization. We call these processed measurements the 'measured gene expressions.' Often, they are transformed by a logarithm to produce 'log-scale' measured expressions, otherwise, they are 'linear-scale'. We call the true, yet unobserved, amount of mRNA transcribed from each gene the 'actual expression' of the gene. This actual gene expression can also be considered on the linear-scale or the log-scale. (See Supplementary Fig. S1 for a graphical representation of these relationships.) Given these definitions, the statistical modeling that yields the dtangle estimator [Equation (1)] is as follows.

First we posit that actual expressions mix linearly on the linear-scale. If η_{kn} is the actual linear-scale expression of the nth gene in a sample of type k cells and η_n is the actual linear-scale expression in the mixture, then dtangle assumes

$$\eta_n = \sum_{k=1}^K p_k \eta_{kn}. \tag{2}$$

This assumption is simply that the total amount of mRNA in a mixture is the sum amount from each cell type.

Second, dtangle assumes that log-scale *measured* expressions are well modeled as linear in log-scale *actual* expressions. Statistically,

$$\begin{aligned} Y_n &= \mu + \theta_n + \gamma \log_2(\eta_n) + \varepsilon_n \\ Z_{krn} &= \alpha + \theta_n + \gamma \log_2(\eta_{kn}) + \varepsilon_{krn} \end{aligned} \tag{3}$$

for $n = 1, \ldots, N, r = 1, \ldots, \nu_k$, $k = 1, \ldots, K$. (Recall the $Y's$ and $Z's$ are on the log-scale and the $\eta's$ are not.) We assume uncorrelated errors ε with zero mean and finite variance.

Equation (3) models several important features of the transformation from actual to measured expressions by the GEP technology. First, μ and α model the samples' and references' mean measured expressions. This accounts for experimental features like quantity of mRNA or sequencing depth (for RNA-seq). We assume the references have been normalized (e.g. quantile normalized or mean centered) so that they share an intercept α. Second, θ_n accounts for gene-specific effects like length biases in RNA-seq or probe affinities in microarrays. Intuitively, γ is a factor to account for imperfect mRNA quantification. Ideally, $\gamma = 1$ meaning, on the linear-scale, increasing actual expression always leads to a proportional increase in measured expression. For RNA-seq we find $\gamma \approx 1$, however for microarray technology a γ slightly smaller than 1 helps account for saturation and attenuation of the intensity measurements for lowly and highly expressed genes (see Supplementary Section S5). While such measuring imperfections are well-known, dtangle is the only existing method to account for them.

Finally, dtangle assumes marker genes are (approximately) expressed by only one cell type. If n is a marker gene for cell type k ($n \in G_k$), this implies

$$\eta_{\ell n} = 0 \text{ for all } \ell \neq \mathrm{k} \tag{4}$$

(This is an approximation. See Supplementary Section S6.3 for further discussion.)

Combining Equation (2) with Equation (3) and Equation (4) we have

$$\begin{aligned} D_{kt} &= \frac{1}{\gamma}\left((\overline{Y_{G_k}} - \overline{Y_{G_t}}) - (\overline{Z_{G_k}} - \overline{Z_{G_t}})\right) \\ &= \log_2(p_k/p_t) + \delta \\ &\approx \log_2(p_k/p_t) \end{aligned} \tag{5}$$

where δ is a function of the ε's and $\delta \to 0$ as $g_k, g_t \to \infty$ (for details

see Supplementary Section S7). Thus assuming the approximation in Equation (5) holds for all t then

$$D_k \approx \left(\log_2(p_k/p_1), \ldots, \log_2(p_k/p_K)\right)$$

and so $L_k(D_k) \approx p_k$.

2.3 Relationship of dtangle to other deconvolution methods

The area of 'cell type deconvolution' encompasses several related inference problems. Hence, every deconvolution problem includes three main components: (1) measured expressions from mixture samples, (2) measured expressions from reference samples of each cell type and (3) the proportion each mixture sample is comprised of each cell type. Typically it is always assumed that (1) is known. The deconvolution problem is then estimating either: (a) the mixing proportions, given the reference expressions, (b) the reference expressions given the mixing proportions, or (c) the proportions and the references jointly. All three problems are considered instances of deconvolution. dtangle most closely resembles problem (a), of estimating unknown mixing proportions given measured expressions from the mixture and references. In Section 3 we compare dtangle to methods solving both (a) and (c) since they both estimate the proportions. Problem (a), called 'partial deconvolution' (Gaujoux, 2013), is typically solved as a regression or penalized regression problem (Abbas *et al.*, 2009; Altboum *et al.*, 2014; Gong *et al.*, 2011; Lu *et al.*, 2003; Newman *et al.*, 2015; Qiao *et al.*, 2012; Racle *et al.*, 2017; Wang *et al.*, 2006), problem (c), called 'full deconvolution', is usually accomplished by non-negative matrix factorization (Gaujoux and Seoighe, 2012; Repsilber *et al.*, 2010; Venet *et al.*, 2001; Zhong *et al.*, 2013).

2.3.1 Scale: Interpretability, robustness and efficiency

Existing methods to solve problems (a), (b) or (c) are based on a common linear mixing model. Let $X \in \mathbb{R}^{S\times N}$ be the S mixture samples' N linear measured expressions, $M \in \mathbb{R}^{S\times K}$ so that M_{sk} is the percentage of type k cells in sample s, and $U \in \mathbb{R}^{K\times N}$ so that the K rows of U are reference expressions of the K cell types. Existing methods presume a linear mixing model on either the linear scale,

$$X \approx MU, \tag{6a}$$

or the logarithmic scale,

$$\log(X) \approx M\log(U). \tag{6b}$$

They then solve for (a) M, (b) U (equiv. $\log(U)$) or (c) both, presuming the other components are known.

Both Equation (6a) and Equation (6b) have advantages and drawbacks. Equation (6a) is a physically plausible linear mixing model of linear measured expressions. It posits that mRNA from a sample of cells is the sum of the mRNA from each cell. While plausible, fitting this model on the linear-scale is non-robust and statistically inefficient. The highly skewed data means the fit is unduly influenced by data in the tail of the distribution (Li *et al.*, 2016). Furthermore, since the variance of gene expressions typically scales with their mean, regression approaches are sub-optimal (Li *et al.*, 2016). In contrast, Equation (6b) models a linear mixture of log expressions. This approach is more robust since the log transformation ameliorates the skewness and heteroskedasticity. However Equation (6b) is not physically plausible. It implicitly assumes that the mRNA in a mixture sample is the product (not sum) of the mRNA from each cell.

dtangle's approach is to take advantage of the beneficial aspects of each scale while avoiding their problems. Firstly, dtangle is based on a biologically plausible linear mixing model of linear-scale actual expressions [Equation (2)]. Second, dtangle's linear model between actual and measured expression [Equation (3)] and definition of D_{kt} [Equation (5)] are on the log-scale. This makes dtangle robust and statistically efficient. dtangle only transforms into the linear-scale in its final step robustly exponentiating after averaging, not before.

Similar to Equation (6a) dtangle uses a plausible and interpretable physical model of mixing [Equation (2)]. However dtangle robustly averages log-scale expressions [Equation (5)] and thus has robust character similar to fitting using Equation (6b). Supplementary Section S6.2 uses simulations to explore these points in more depth.

3 Results

3.1 Benchmarking

To evaluate dtangle we compare it to eight other deconvolution algorithms (Supplementary Table S1). Six methods are accessed through the CellMix R package (Gaujoux, 2013). We also compare to CIBERSORT and EPIC as they are recent and powerful methods (Newman *et al.*, 2015; Racle *et al.*, 2017). We only compare dtangle against methods that estimate cell type proportions from gene expression data for arbitrary cell types. We do not compare to methods like xCell (Aran *et al.*, 2017) which produce enrichment scores and not percentages. We also do not compare against the many deconvolution methods for methylation data or fully unsupervised methods whose cell types have to be inferred with further post-hoc analysis e.g. CAM (Wang *et al.*, 2016). Furthermore, we do not compare against methods that only estimate cell type proportions from a very specific subset of cells or only in the context of a specific problem, for example, immune cell infiltration of tumors by methods like TIMER (Li *et al.*, 2016).

Like dtangle, all methods require marker genes. However four 'full' deconvolution methods we analyze require only marker genes and do not explicitly require reference data. Nonetheless, we find marker genes through DEA on the reference data and so, in one way or another, all methods use reference data. There are several 'completely unsupervised' deconvolution methods in the literature (e.g. Wang *et al.*, 2016) that require neither markers nor references. However their estimates are difficult to interpret biologically unless reference data is used post-hoc to map proportions to cell types. For this reason we do not compare to such methods. Finally, while full deconvolution algorithms also estimate type-specific expressions profiles, we only compare dtangle to their estimated mixing proportions as this is what dtangle estimates.

We choose marker genes for deconvolution following Abbas *et al.* (2009). First we restrict analysis to genes in the highest quartile of variance. We then rank genes by P-value using a t-test between the reference expressions of the two most highly expressed cell types. For each cell type, the 10% of genes with lowest P-values are designated markers.

Note that many genes selected as markers using this approach do not exactly satisfy (4). Further filtering the set of marker genes to attempt to ensure they satisfy (4) could potentially improve the performance of dtangle. However, in our analysis we nonetheless follow the method of Abbas *et al.* (2009) without any further filtering to ensure that the method of marker selection is not biased in favor of dtangle. The exact same set of marker genes are used for each algorithm.

3.2 Datasets compared

We compare dtangle to the eight other algorithms across 11 benchmarking datasets (Supplementary Table S2). The true mixing proportions are known for each dataset either because the experiment was conducted by mixing each cell type in known proportions or because an independent physical sorting technique, like flow cytometry, was used to estimate the proportions. Most datasets include their own cell type references.

For the RNA-seq data we TPM normalize, transform as one plus the read count. For the microarray data we quantile normalize on a logarithmic scale. All data is re-exponentiated so it is on the linear-scale for algorithms that require it. Pre-processing code is available in the dtangle.data R package available at dtangle.github.io.

3.3 Microarray data

3.3.1 Mixture experiments with references

We consider five microarray mixture experiments: datasets Abbas, Kuhn, Gong, Shi and Shen-Orr (Supplementary Table S2). For each algorithm we estimate the mixing proportions in each dataset. We evaluate the algorithms' accuracy in terms of absolute error of estimated proportions from true proportions and by Pearson correlation and R^2 of the estimates against the truth for each cell type. dtangle has the lowest median error, second lowest mean error (behind CIBERSORT), and the highest mean and median correlation and R^2 across the datasets (Supplementary Fig. S4). This meta-analysis shows that for the microarray mixture experiments dtangle is the most accurate algorithm as measured by most metrics and also one of the most consistently accurate. For each dataset Supplementary Boxplots of error, correlation, R^2, as well as scatter plots may be found in Supplementary Figures S13, S15, S16, S22 and S23.

We highlight comparisons between dtangle, CIBERSORT and EPIC on two datasets where dtangle performs worst and best relative to other algorithms (Fig. 1a and b). For the Gong data blood and breast tissue were mixed in known proportions. While dtangle does not perform as well as other algorithms, it still performs quite well. The estimated mixing proportions are still highly correlated with the truth (see Supplementary Fig. S15). Conversely, the Shen-Orr data is from a microarray mixture experiment where rat liver, brain and lung cDNA were mixed in known proportions. Here, dtangle performs as well or better than the other algorithms (Fig. 1b, Supplementary Fig. S22). It has the comparable error and the highest correlation and R^2. dtangle performs on par with CIBERSORT and out-performs EPIC.

3.3.2 Mixtures without references

In practice pure reference samples of each cell type are not typically generated along with the mixed samples to be deconvolved. In this case existing reference data for each of the cell types to deconvolved must be procured. Typically these pure reference samples are collected from repositories like GEO.

The Becht dataset is a mixture experiment where cDNA from the HCT116 colorectal carcinoma line and various leukocytes (NK, B, neutrophils, T and monocytes) were mixed in known quantities and analyzed with an Affymetrix microarray. Unlike previous datasets no reference data was produced as part of the mixture experiment. Like the authors we use publicly available expression data from GEO as references for each cell type. In total there are 776 samples gathered from GEO which we use to create reference profiles for the six cell types. On this data dtangle performs as well or better than CIBERSORT and EPIC (Fig. 1c, Supplementary Fig. S14). dtangle has commensurate mean/median error, correlation and R^2 as these methods.

3.3.3 Performance evaluation with flow cytometry based cell sorting

Mixture experiments are only a surrogate for cell mixtures found in organisms. Realistically, deconvolution methodology is applied to complex tissue extracted from an organism. Such tissue will be a mixture of many cell types (more types than in a typical mixture experiment) and the cell types will have complex inter-cellular interactions modifying their gene expressions. The difficulties are estimating cell type proportions from such complex tissue is likely only partially explored by a mixture experiment.

The Newman follicular lymphoma (FL) data was generated by taking lymph node biopsy samples and enumerating immune cell sub-types using flow cytometry (Newman *et al.*, 2015). This process identified 3 leukocyte types (B, CD4 T and CD8 T) in various proportions across samples from 14 patients. As cell type expression reference data we use the same reference data used to create the LM22 reference by Newman *et al.* (2015). It contains gene expressions of 22 white blood cell types as references. Similar to Newman *et al.* (2015) we group these 22 types into 12.

The Newman peripheral blood mononuclear cells (PBMC) data was generated from blood samples from twenty adults where the proportions of nine types of leukocytes were determined by flow cytometry. We again use the same data to create references as used to create the LM22 dataset (Newman *et al.*, 2015). dtangle compares well with other deconvolution methods on these two datasets (Supplementary Figs S19 and S20). For the Newman PBMC dataset

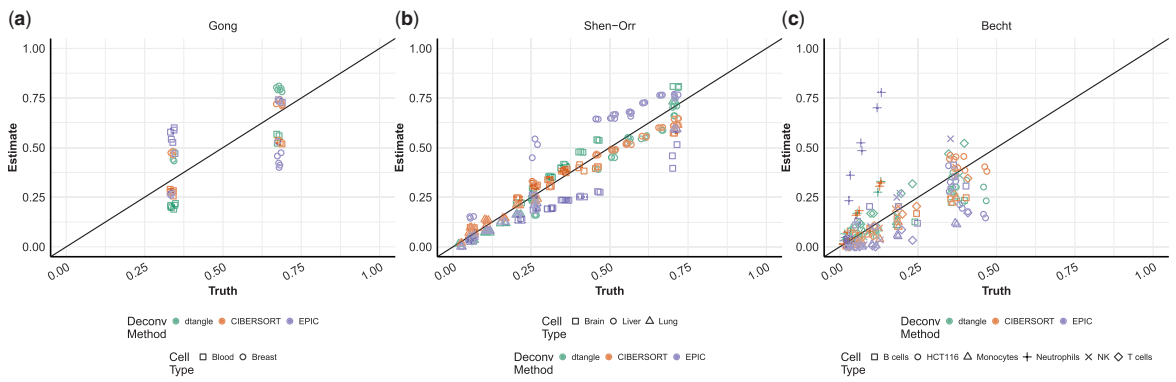

Fig. 1. Scatter plots of dtangle, CIBERSORT and EPIC on the Gong, Shen-Orr and Becht datasets. Each point is a particular cell type in a sample

dtangle has the highest average correlation and lowest average error. For the Newman FL data dtangle has the highest average correlation however the overall accuracy suffers somewhat because of biases in the CD4T and B cell types. This may be due to the large number of cell types making it difficult for our markers to distinguish among them.

To investigate the effect of marker gene selection we re-analyze both the Newman FL and PBMC datasets using the exact LM22 signature matrix used in Newman *et al.* (2015) (see Supplementary Figs S24 and S25). The LM22 signature matrix is a highly curated set of marker genes for 22 PBMCs developed by Newman *et al.* (2015). The results largely remain the same however the biases largely disappear for dtangle. In particular, dtangle is across the board the best performing method on the Newman PBMC data and dtangle has the highest average correlation and R^2 for the Newman FL data. This further underlines the fact that choosing references and markers is an important component of deconvolution and needs to be considered carefully.

3.4 RNA-seq

We also investigate the performance of deconvolution methods on RNA-seq mixture experiments (Supplementary Fig. S5). The Liu and Parsons datasets are RNA-seq mixture experiments with internal reference data. The Linsley dataset is a realistic dataset of leukocytes extracted from patients where the true proportions are determined by flow-cytometry and external references are used. dtangle, CIBERSORT, EPIC and LS Fit seem to be the best algorithms across the RNA-seq datasets. For each dataset Supplementary Boxplots of error, correlation, R^2, as well as scatter plots may be found in Supplementary Figures S17, S18 and S21.

3.5 Meta-analysis

We compare dtangle to the other algorithms in a meta-analysis (Fig. 2). We see that dtangle has the lowest median error, second lowest mean error, and highest mean and median correlation and R^2 across datasets. Existing methods in the literature require linear scale expressions. For example, CIBERSORT hard-codes this requirement and EPIC explicitly requires TPM normalized read counts. Nonetheless, we modify these other methods and fit them using log expressions. While dtangle is unchanged, switching to the log scale helps some of the other methods on some metrics but hurts other methods on other metrics. However after transformation dtangle performs the best on all metrics and so we conclude that these other methods do relatively worse after a log-transformation (Supplementary Fig. S3).

3.6 Robustness to marker selection

Thus far we have been selecting marker genes by, among the top 25% most variable genes in the references, ranking marker genes following Abbas *et al.* (2009) with a t-test *P*-value between the top two most expressed cell types for each gene and selecting the top 10% of differentially expressed genes. To analyze the sensitivity of dtangle to how the markers are ranked we consider another way of ranking marker genes. This second method looks at the ratio of the mean expression for each cell type to the sum of the mean expressions by all other cell types. We call the Abbas method '*P*-value' and call this latter approach 'Ratio.' In Figure 3 we look at the grand error median of each algorithm across all datasets for a range of marker tuning parameters. We compare partial deconvolution algorithms as they are the most competitive with dtangle. dtangle is robust to the way markers are ranked (*P*-value or Ratio) and ranking threshold determining the number of markers. CIBERSORT, EPIC, Q Prog. and LS Fit appear to have a strong dependence on the quantile cutoff. Indeed, the performance of CIBERSORT degrades much more quickly than other methods. As the number of marker genes approaches one per cell type its grand median error grows large very quickly. This can be seen as the sharply increasing blue line in Figure 3 for both the *P*-value and Ratio methods. In Supplementary Figures S6–S8 we include similar plots looking at the mean and median error, and mean and median correlation and R^2 for all datasets, microarray datasets and RNA-seq datasets.

Marker gene selection also influences computational time. For each dataset we timed all algorithms across a range of quantile cutoffs using *P*-value ranking (Supplementary Fig. S9). dtangle is consistently the fastest algorithm. It is between one and four orders of magnitude faster than other algorithms regardless of what quantile cutoff is used.

3.7 Application to Lyme disease

To demonstrate dtangle on a biological problem we consider RNA-seq data of PBMCs from Lyme disease patients (Bouquet *et al.*, 2016). To better understand persistent Lyme symptoms (e.g. fatigue or arthritis) it is of interest to understand the progression of the human immune response to Lyme (Bouquet *et al.*, 2016). To this

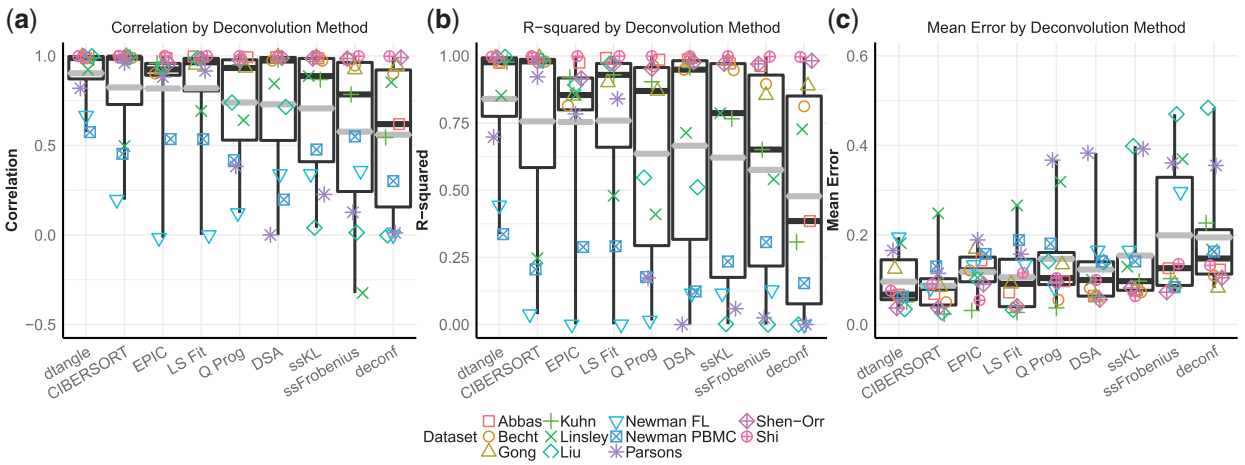

Fig. 2. Meta-analysis of deconvolution algorithms. Side-by-side box plots of the mean errors, correlations and R^2 across the algorithms. The bold black line is median, the grey line is mean. Overlapping are jittered points of the metric for each dataset

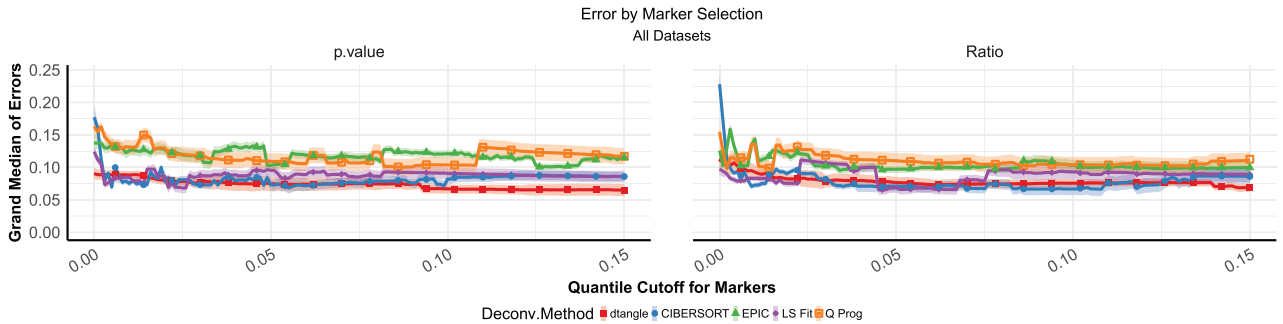

Fig. 3. Grand error medians across marker ranking methods (*P*-value and Ratio) and number of markers. 95% confidence bands included

end Bouquet *et al.* measure gene expression in a subset of white blood cells (PBMCs). PBMCs of 28 patients were collected at the point of diagnosis (V1), after a 3-week course of doxycycline (V2) and 6 months later (V5). PBMCs from 13 matched controls were also collected (C).

We use dtangle to estimate, for each sample, the cell type proportions of nine types of PBMCs (B, dendritic, macrophages, mast, monocytes, NK, CD4 T, CD8 T and gamma-delta T). We use as reference the LM22 dataset from Newman *et al.* (2015), choosing the top 10% of differentially expressed genes for each cell type as markers. We find that the phagocytes (dendritic, macrophages, mast and monocytes) make up a larger percentage of the patients' PBMCs earlier, rather than later, in the infection (Supplementary Fig. S26). We see a large difference between the control group and V1 and decreasing differences between the controls and V2 and V5. Natural killer (NK) cells follow this same pattern.

The estimated cell type percentages agree with the current understanding of Lyme. The initial infection induces an immune response where fast-acting phagocytes are recruited to attack the foreign bacteria (Dame *et al.*, 2007). This agrees with dtangle's estimates of a relatively large percentage phagocytes early in the infection that decreases with time. Phagocytes decrease in numbers once the bacteria has been cleared and they are no longer needed. Furthermore, NK cells follow the same pattern. This agrees with work from Horowitz *et al.* (2012) showing NK cells are rapidly activated by cytokines after a bacterial infection.

In Bouquet *et al.* (2016) the authors seek to find genes that are differentially expressed among the groups (V1, V2, V5 and C). Following Bouquet *et al.* (2016) we compare the control group to V1, V2 and V5 and find that there are 399 genes that are differentially expressed in the intersection of each of the three comparisons. This was done controlling for a FDR of 0.05 by the Benjamini-Hochberg procedure.

As this previous differential expression analysis was not corrected for cell type proportions we expect to find genes that are correlated with cell type. We add in covariates to account for composition of fast-acting cell types (phagocyte and NK). After doing so we only find 158 genes differentially expressed in the same comparison. Thus the cell type composition changes the results of the analysis greatly. dtangle is one tool practitioners can use to help tease apart histological changes in cell composition from changes in gene expression within particular cell types.

4 Discussion

dtangle is a simple and robust deconvolution estimator. It is a closed-form estimator deriving from plausible biological modeling. Our meta-analyses show that dtangle is a robust and accurate, typically performing better than eight of the best existing methods across eleven diverse datasets. It can accurately deconvolve cell types using microarray and RNA-seq technology and is very fast to compute where other methods are not. Furthermore it is consistent with standard physical sorting methods like flow cytometry on realistic complex clinical tissue. Finally, dtangle has competitive accuracy when dealing with realistic datasets where the reference samples are obtained from publicly available repositories. dtangle works well even when these reference datasets were created using a different profiling technology. This points to scRNA-seq data as a promising source for references.

dtangle has some of the same limitations as other algorithms. Primarily, it is necessary that the cell types comprising each sample be known in advance and that reference data is available. Furthermore dtangle needs to find marker genes for each cell type. This can be potentially difficult if there are many cell types or the cell types are closely related. Nonetheless, dtangle seems to perform well in many situations. We will continue to develop dtangle to overcome some of these challenges to broaden its utility.

Funding

Work by S.F. and M.B. was supported by the Victorian Government's Operational Infrastructure Support Program and Australian Government NHMRC IRIIS. M.B. is funded by NHMRC Senior Research Fellowship 110297 and NHMRC Program Grant 1054618. G.H. and J.G. were supported by the National Science Foundation grant no. DMS-1646108.

Conflict of Interest: none declared.

References

Abbas,A.R. *et al.* (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One*, **4**, e6098. Doi: 10.1371/journal.pone.0006098.

Altboum,Z. *et al.* (2014) Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.*, **10**, 1–14.

Aran,D. *et al.* (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.*, **18**, 220.

Bouquet,J. *et al.* (2016) Longitudinal transcriptome analysis reveals a sustained differential gene expression signature in patients treated for acute lyme disease. *mBio.*, **7**, 1–11.

Bowling,K.M. *et al.* (2017) Genomic diagnosis for children with intellectual disability and/or developmental delay. *bioRxiv*, **9**, 43.

Capurro,A. *et al.* (2015) Computational deconvolution of genome wide expression data from Parkinson's and Huntington's disease brain tissues using population-specific expression analysis. *Front. Neurosci.*, **9**, 1–12.

Dame,T.M. *et al.* (2007) Endothelium to favor chronic inflammation 1. *J. Immunol.* **178**, 1172–1179.

Edgar,R. *et al*. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.

Gaujoux,R. (2013) An introduction to gene expression deconvolution and the CellMix package. 1–45.

Gaujoux,R. and Seoighe,C. (2012) Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol*., **12**, 913–921.

Gong,T. *et al*. (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One*, **13**, e0200003.

Hagenauer,M.H. *et al*. (2016) Inference of cell type composition from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. *bioRxiv*, https://doi.org/10.1101/089391.

Horowitz,A. *et al*. (2012) Activation of natural killer cells during microbial infections. *Front. Immunol*., **2**, 1–13.

Li,B. *et al*. (2016) Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol*., **17**, 174.

Lu,P. *et al*. (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci. USA*, **100**, 10370–10375.

Marusyk,A. and Polyak,K. (2011) Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta*, **105**, 1–28.

Mohammadi,S. *et al*. (2015) A critical survey of deconvolution methods for separating cell-types in complex tissues. *arXiv*, **105**, 1–20.

Newman,A.M. *et al*. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 193–201.

Qiao,W. *et al*. (2012) PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol*., **8**.

Racle,J. *et al*. (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, **6**.

Repsilber,D. *et al*. (2010) Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 27.

Shen-Orr,S.S. *et al*. (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

Venet,D. *et al*. (2001) Separation of samples into their constituents using gene expression data. *Bioinformatics*, **17**, S279–S287.

Wang,M. *et al*. (2006) Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics*, **7**, 328.

Wang,N. *et al*. (2016) Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep*., **6**, 18909.

Zhong,Y. *et al*. (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, **14**, 89.