dtangle

a quick and accurate cell-type deconvolution estimator

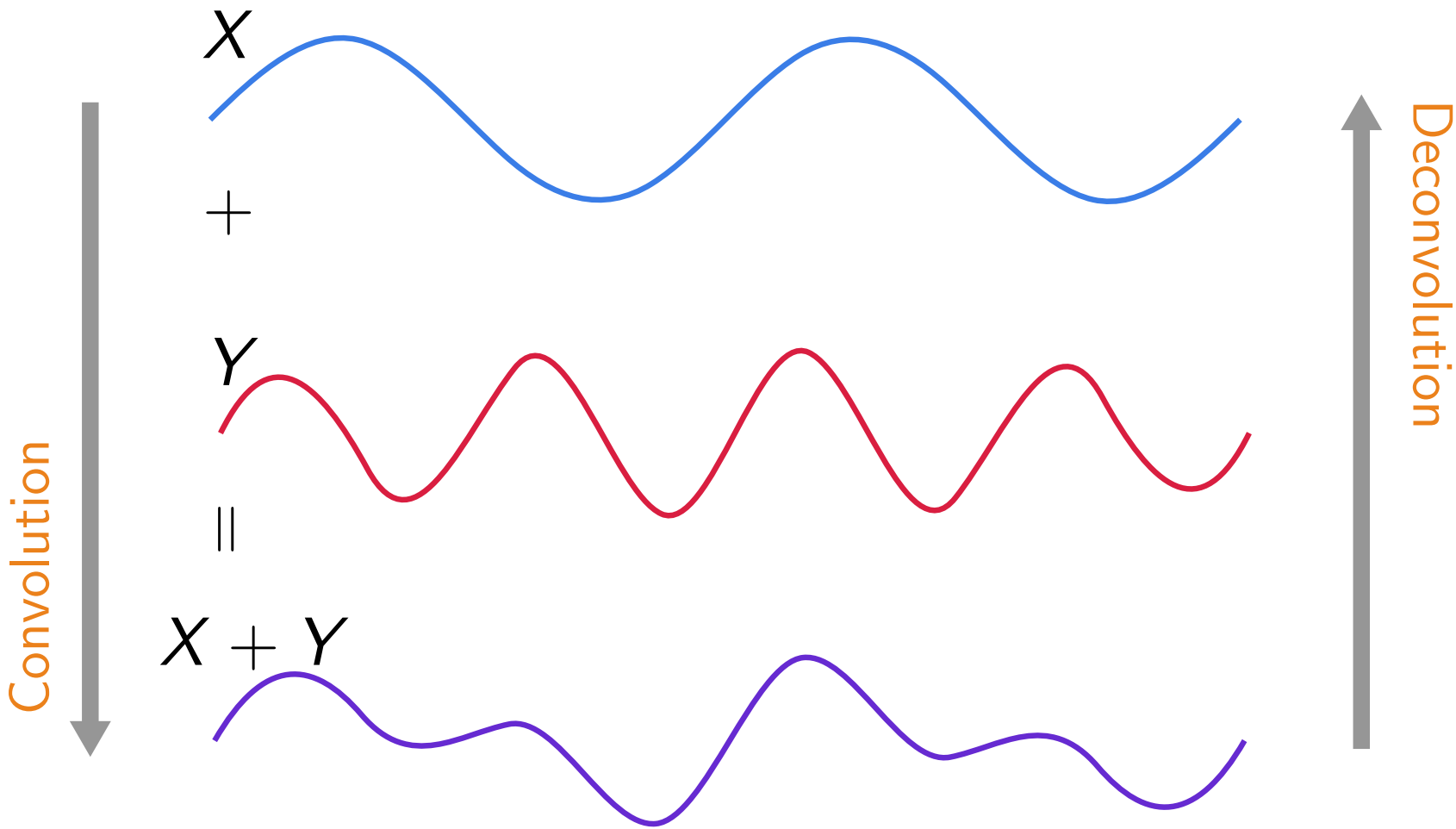
Greg Hunt[1], Saskia Freytag[2], Melanie Bahlo[2] and Johann Gagnon-Bartsch[1]

August 2, 2017

[1]Statistics at the University of Michigan
[2]Bioinformatics at the Walter and Eliza Hall Institute

Deconvolution = Decomposing Mixtures

Convolution*: $X, Y \to X + Y$, Deconvolution*: $X + Y \to X, Y$



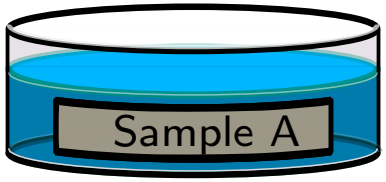
*probably should be called "mixing/de-mixing"

Sample Deconvolution = Decomposing Expression Data

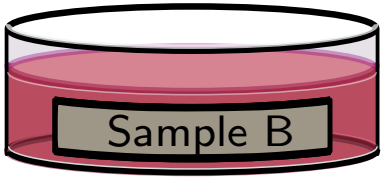
Tissue Samples

Gene Expressions

Reference Samples



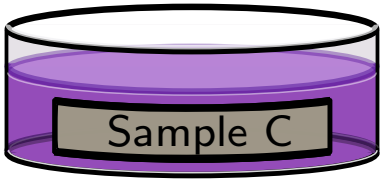
Gene Name	Gene 1	Gene 2	Gene 3	···
Expression	6.39	12.305	7.129	···



Gene Name	Gene 1	Gene 2	Gene 3	···
Expression	8.221	9.234	3.123	···

Sample C is a mixture of A and B.

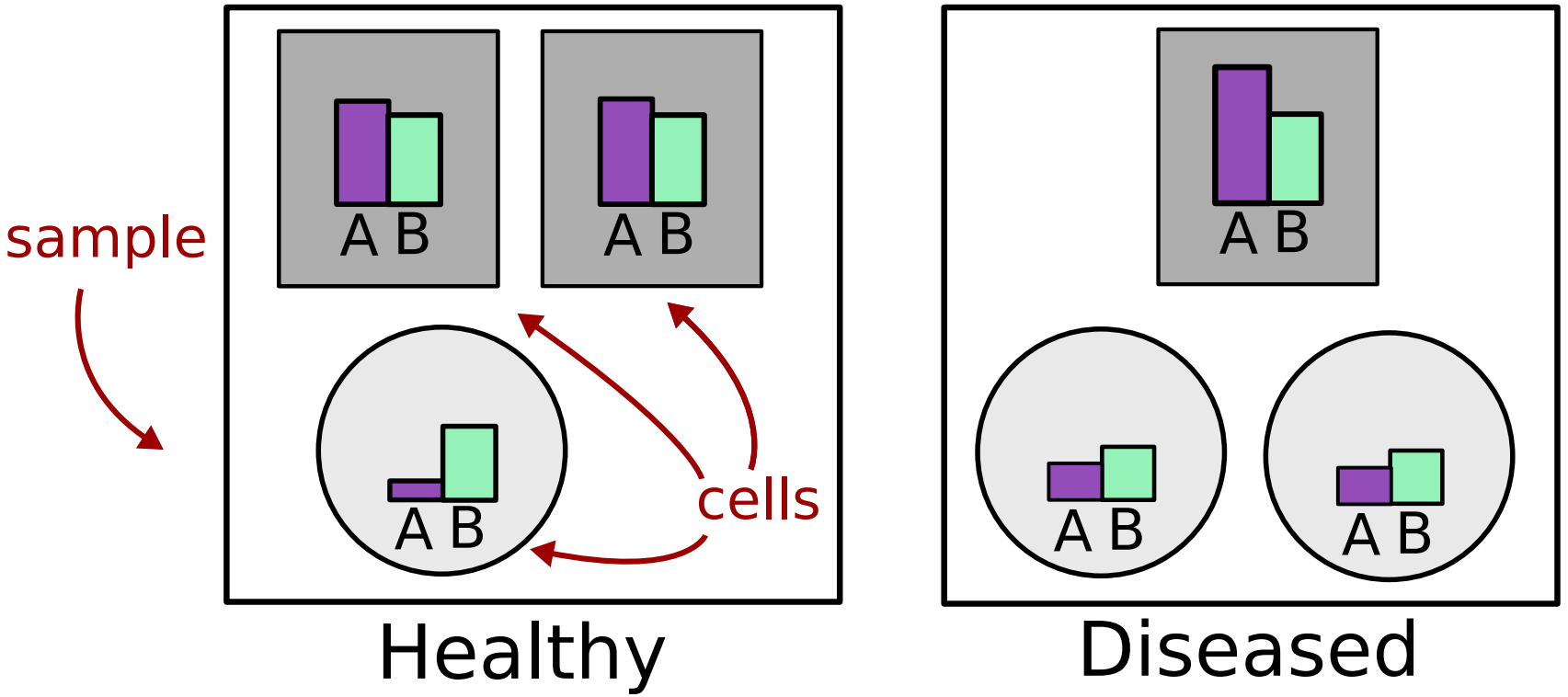
Mixture Sample



Gene Name	Gene 1	Gene 2	Gene 3	···
Expression	7.223	10.332	5.238	···

Deconvolution Controls for Confounding

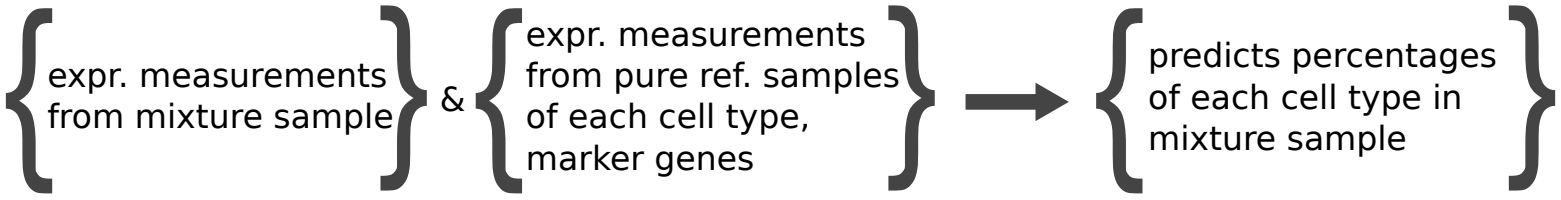
Expression measuring technology measure "mean" expression of a sample. This is affected by by changes within cell types and changes in cell-type composition.



dtangle

dtangle departs from existing methods

dtangle is a new deconvolution estimator.



1. it does not assume a linear mixing model of the expressions
2. it is based on a linear mixing model of mRNA transcripts and and affine relationship between transcripts and expressions
3. it is fast and simple to compute
4. it performs consistently well across many datasets and technologies
5. it performs well deconvolving closely related cell types, many cell types and across mixed technology

dtangle maps relative abundances to percentage by a logistic

The general deconvolution setting is

1. Expression measurements from a **mixture sample**
2. unknown **mixture proportions** $p_1, \ldots, p_K$ $(p_k \geq 0, \sum_{k=1}^{K} p_k = 1)$
3. **reference samples** for each cell types
4. **marker genes** for each cell type

The dtangled estimator of p_k is

$$\widehat{p_k} = \mathit{logistic}\left(\begin{array}{l}\text{baseline normalized}\\ \text{relative abundance of}\\ \text{type } k \text{ over type } 1\end{array}, \begin{array}{l}\text{baseline normalized}\\ \text{relative abundance of}\\ \text{type } k \text{ over type } 2\end{array}, \ldots\right)$$

dtangle can also be thought of as a linear model

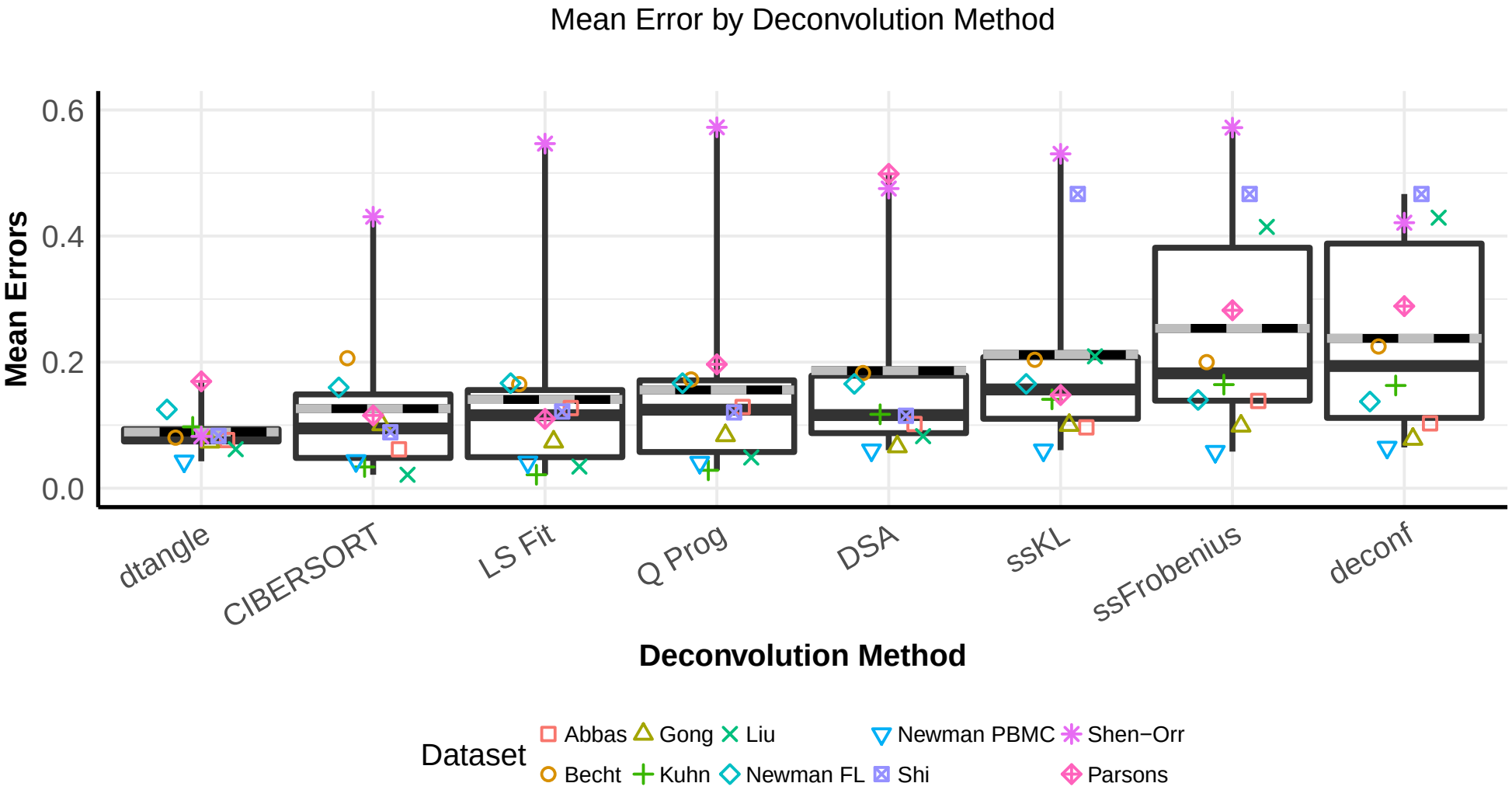
dtangle is based on a linear model (on the log – log level) between measured and actual transcript concentrations.

The assumed linear model (plus some straight-forward mathematical assumptions) imply

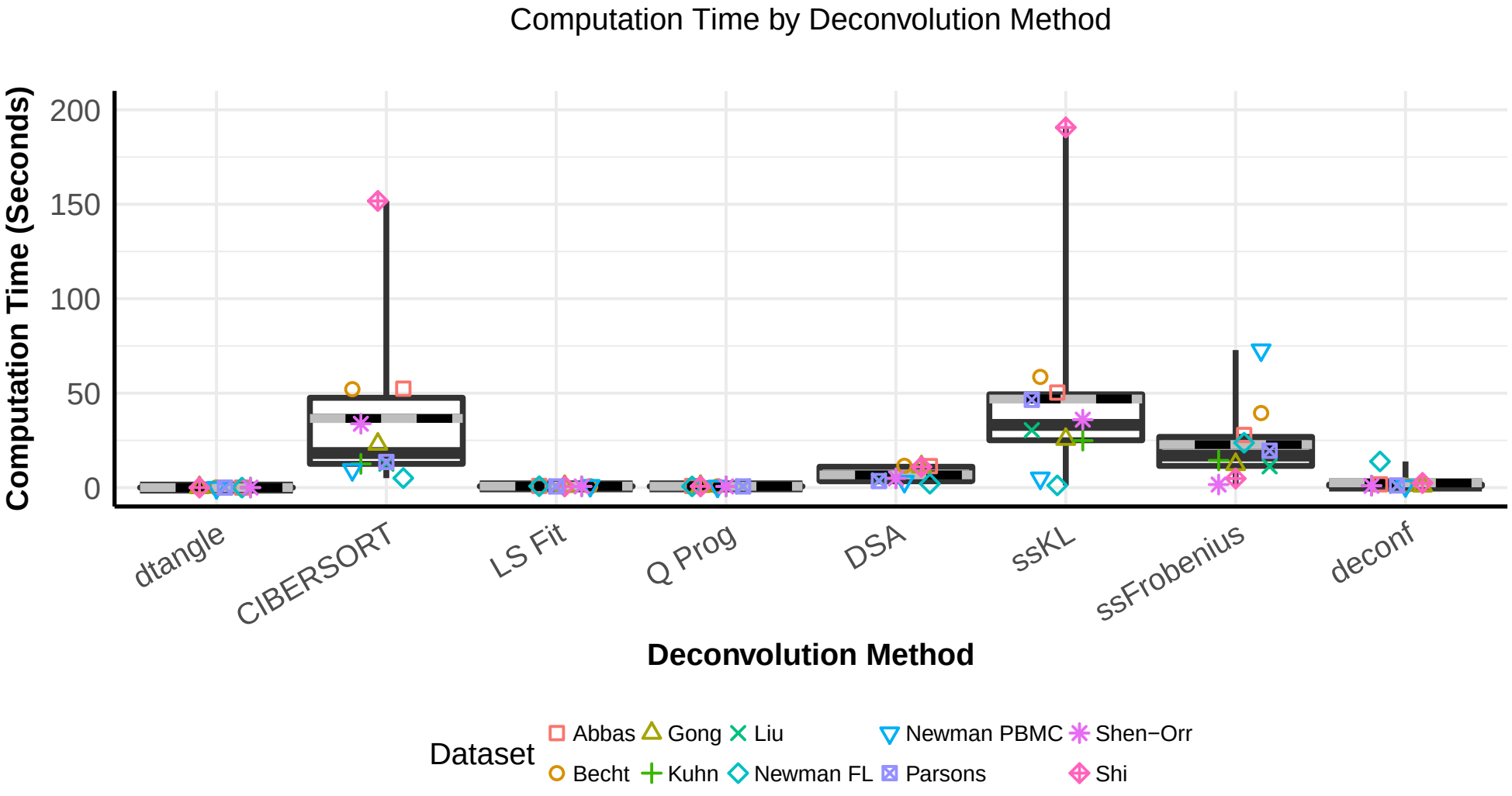
$$\widehat{p_k} = \xi_k \, p_k$$

for some error term ξ_k.

dtangle is consistently accurate



dtangle is fast

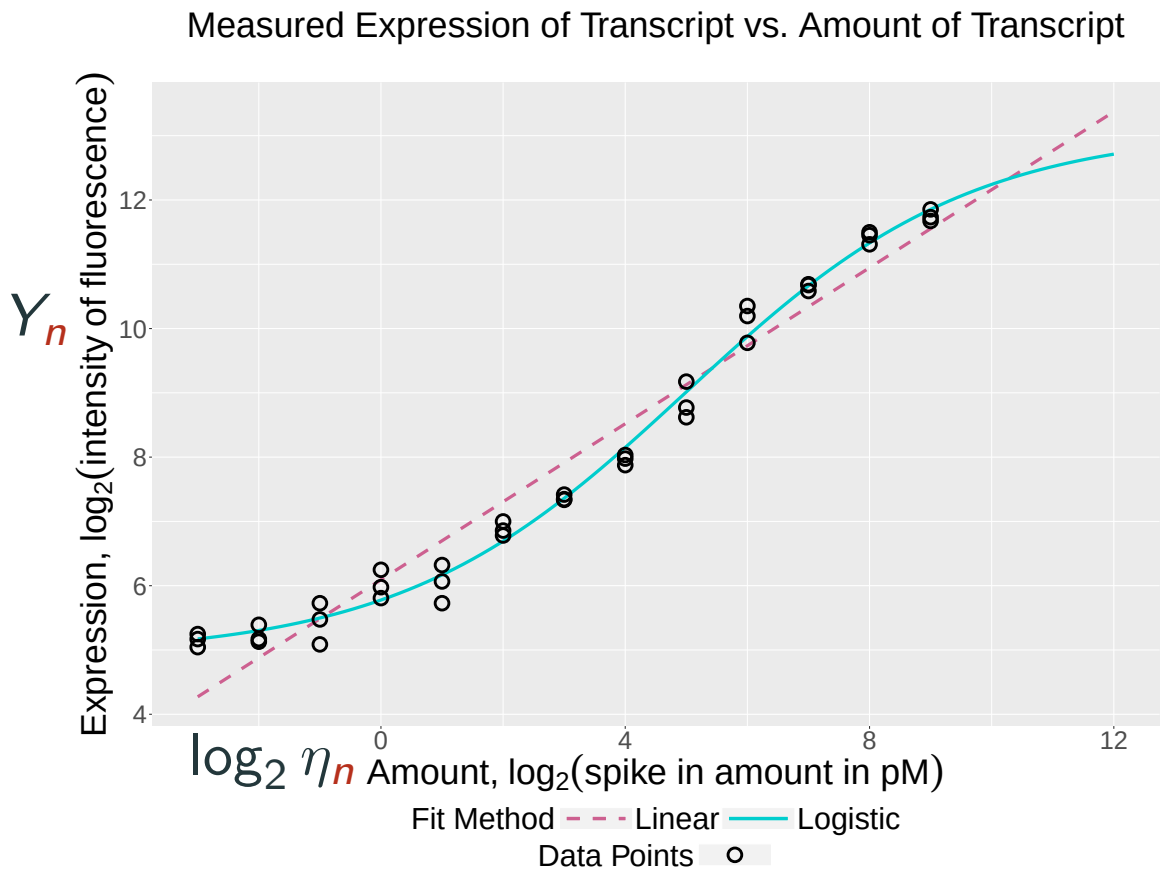


Thanks!

dtangle is on CRAN

Estimate γ from Benchmark Data Sets

1. For a given technology estimate γ globally using reference data sets where we know the amount and the expressions.
2. We simply take median of regression slopes for a large set of spiked-in genes.



Choose Markers From Reference Pure Samples

Goal: Determine which genes are highly expressed in some cell types but not others.

Method: Some type of differential expression analysis among pure sample expressions.

dtangle isn't very sensitive to precisely how the differential expression analysis is done.

dtangle is robust to changes in it's tuning parameters