Estimating Cell-type Proportions Using Gene Expressions

Greg Hunt[1], Saskia Freytag[2,3], Melanie Bahlo[2,3] and Johann Gagnon-Bartsch[1]

December 12, 2017

[1]Statistics at the University of Michigan

[2]Population Health and Immunity Division at the Walter and Eliza Hall Institute of Medical Research

[3]Department of Medical Biology at the University of Melbourne

Estimating Cell-type Proportions Using Gene Expressions

Saskia Freytag[2,3]

Melanie Bahlo[2,3]

Johann Gagnon-Bartsch[1]

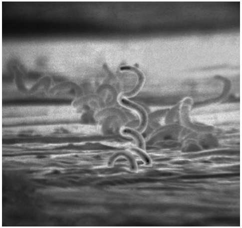


[1]Statistics at the University of Michigan
[2]Population Health and Immunity Division at the Walter and Eliza Hall Institute of Medical Research

[3]Department of Medical Biology at the University of Melbourne

Understanding The Immune Response to Lyme

Adult deer tick.
Scott Bauer [1].



A typical spirochete.
CDC/Dr. David Cox [2].

Lyme disease: bacterial infection spread by ticks.

1. treatable with antibiotics
2. patients report fatigue, arthritis, muscle soreness and memory problems
3. can lead to worse conditions like Lyme encephalopathy, insomnia, or depression

Bouquet et al: try to understand the immune progression of Lyme.

Study WBCs to Understand Immune Response to Lyme

Bouquet et al: collect gene expression measurements or "profiles" (GEPs) of white blood cells (WBCs) of

1. 28 Lyme patients
2. and 13 healthy controls.

The analysis compares GEPs across groups.

Study WBCs to Understand Immune Response to Lyme

Bouquet et al: collect gene expression measurements or "profiles" (GEPs) of white blood cells (WBCs) of

1. 28 Lyme patients
2. and 13 healthy controls.

The analysis compares GEPs across groups.

WBCs encompass many types: B, T, NK, monocytes, ...

Understanding these sub-types would be helpful:

1. tracking subtype composition changes over disease course
2. adjusting expression comparisons across groups

Study WBCs to Understand Immune Response to Lyme

Bouquet et al: collect gene expression measurements or "profiles" (GEPs) of white blood cells (WBCs) of

1. 28 Lyme patients
2. and 13 healthy controls.

The analysis compares GEPs across groups.

WBCs encompass many types: B, T, NK, monocytes, ...

Understanding these sub-types would be helpful:

1. tracking subtype composition changes over disease course
2. adjusting expression comparisons across groups

Problem: estimate the cell-type proportions of the samples using the gene expression data.

Gene Expression Data

"Gene expression measurements" $=$ What genes the cells are using

Measure expression using mRNA:

Gene Expressed $\rightarrow$ mRNA transcript created

Gene 1 $\longrightarrow$ mRNA 1 1

Gene 2 $\longrightarrow$ mRNA 2 2

Gene Expression Data

"Gene expression measurements" = What genes the cells are using

Measure expression using mRNA:

Gene Expressed → mRNA transcript created

Gene 1 ⟶ mRNA 1 1

Gene 2 ⟶ mRNA 2 2

Measuring gene expression = Quantifying mRNA transcripts

"Gene expression measurements" = What genes the cells are using

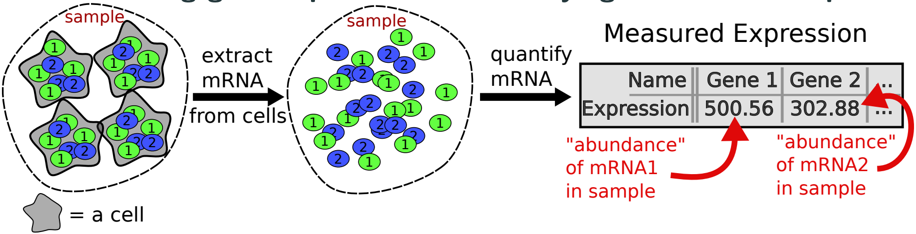
Measure expression using mRNA:

Gene Expressed → mRNA transcript created

Gene 1 → mRNA 1 (1)

Gene 2 → mRNA 2 (2)

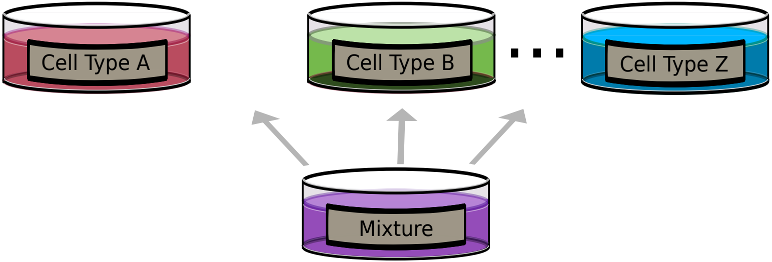
Measuring gene expression = Quantifying mRNA transcripts



Estimating Cell-type Proportions

Given: Gene Expression Profiles (GEPs) of:

1. sample that is mixture of cell types A,B,C,...Z
2. reference samples of types A,B,C,...,Z

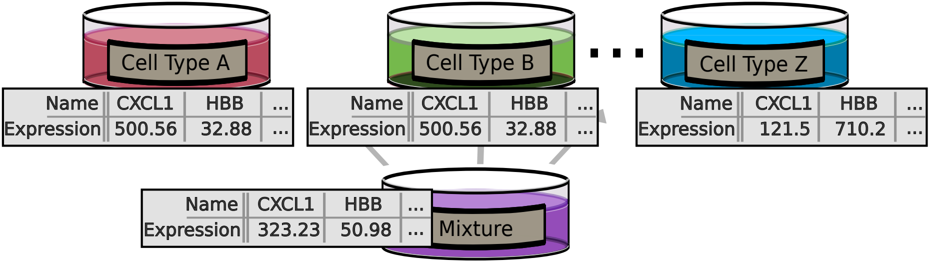


Goal: estimate cell-type proportions

	Type A	Type B	⋯	Type Y	Type Z
Mixture	5%	20%	⋯	30%	0%

Given: Gene Expression Profiles (GEPs) of:

1. sample that is mixture of cell types A,B,C,...Z
2. reference samples of types A,B,C,...,Z



Goal: estimate cell-type proportions

	Type A	Type B	⋯	Type Y	Type Z
Mixture	5%	20%	⋯	30%	0%

Previous Work

A Linear Model is Commonly Used

General model: $M \approx PR$, predict P with known M and R

$$\underbrace{\overset{\textstyle M}{\begin{array}{c} \\ \textbf{sample}_1 \\ \textbf{sample}_2 \\ \textbf{sample}_3 \\ \vdots \\ \textbf{sample}_S \end{array} \begin{array}{c} \begin{array}{cccc} \textit{gene}_1 & \textit{gene}_2 & \cdots & \textit{gene}_G \end{array} \\ \begin{bmatrix} 6.6 & 8.9 & \cdots & 3.5 \\ 3.2 & 5.4 & \cdots & 4.8 \\ 7.3 & 7.7 & & \\ \vdots & & \ddots & \\ 4.1 & & & \end{bmatrix} \end{array}}}_{\text{mixture expressions}} \approx \underbrace{\overset{\textstyle P}{\begin{array}{c} \\ \textbf{sample}_1 \\ \textbf{sample}_2 \\ \textbf{sample}_3 \\ \vdots \\ \textbf{sample}_S \end{array} \begin{array}{c} \begin{array}{cccc} \textit{type}_1 & \textit{type}_2 & \cdots & \textit{type}_K \end{array} \\ \begin{bmatrix} .5 & .2 & \cdots & .1 \\ 0 & .01 & \cdots & .95 \\ .35 & .45 & & 0 \\ \vdots & & \ddots & \\ .1 & & & \end{bmatrix} \end{array}}}_{\text{mixing proportions}} \underbrace{\overset{\textstyle R}{\begin{array}{c} \\ \textit{type}_1 \\ \textit{type}_2 \\ \textit{type}_3 \\ \vdots \\ \textit{type}_K \end{array} \begin{array}{c} \begin{array}{cccc} \textit{gene}_1 & \textit{gene}_2 & \cdots & \textit{gene}_G \end{array} \\ \begin{bmatrix} 9.3 & 4.1 & \cdots & 3.6 \\ 3.7 & 5.4 & \cdots & 9.3 \\ 2.9 & 3.6 & & \\ \vdots & & \ddots & \\ 8.6 & & & \end{bmatrix} \end{array}}}_{\text{reference expressions}}$$

A Linear Model is Commonly Used

General model: $M \approx PR$, predict P with known M and R

$$\underbrace{\overset{\Large\text{M}}{\begin{array}{cc} & \begin{array}{cccc} \textit{gene}_1 & \textit{gene}_2 & \cdots & \textit{gene}_G \end{array} \\ \begin{array}{c} \textbf{sample}_1 \\ \textbf{sample}_2 \\ \textbf{sample}_3 \\ \vdots \\ \textbf{sample}_S \end{array} & \begin{bmatrix} 6.6 & 8.9 & \cdots & 3.5 \\ 3.2 & 5.4 & \cdots & 4.8 \\ 7.3 & 7.7 & & \\ \vdots & & \ddots & \\ 4.1 & & & \end{bmatrix} \end{array}}}_{\text{mixture expressions}} \approx \underbrace{\overset{\Large\text{P}}{\begin{array}{cc} & \begin{array}{cccc} \textit{type}_1 & \textit{type}_2 & \cdots & \textit{type}_K \end{array} \\ \begin{array}{c} \textbf{sample}_1 \\ \textbf{sample}_2 \\ \textbf{sample}_3 \\ \vdots \\ \textbf{sample}_S \end{array} & \begin{bmatrix} .5 & .2 & \cdots & .1 \\ 0 & .01 & \cdots & .95 \\ .35 & .45 & & 0 \\ \vdots & & \ddots & \\ .1 & & & \end{bmatrix} \end{array}}}_{\text{mixing proportions}} \underbrace{\overset{\Large\text{R}}{\begin{array}{cc} & \begin{array}{cccc} \textit{gene}_1 & \textit{gene}_2 & \cdots & \textit{gene}_G \end{array} \\ \begin{array}{c} \textit{type}_1 \\ \textit{type}_2 \\ \textit{type}_3 \\ \vdots \\ \textit{type}_K \end{array} & \begin{bmatrix} 9.3 & 4.1 & \cdots & 3.6 \\ 3.7 & 5.4 & \cdots & 9.3 \\ 2.9 & 3.6 & & \\ \vdots & & \ddots & \\ 8.6 & & & \end{bmatrix} \end{array}}}_{\text{reference expressions}}$$

Solutions:

1. **Regression: regress M on R.**
   (Abbas *et al.*; Gong *et al.*; Lu *et al.*; Wang *et al.*; Qiao *et al.*; Altboum *et al.*; Newman *et al.*)
2. **Bayesian: Similar to LDA. Estimate as MAP.**
   (Quon and Morris; Qiao *et al.*; Quon *et al.*)

Marker Genes are Genes Expressed in Only One Cell Type

A marker gene is one which is predominantly expressed in one cell type and not the others.

Main Idea: Find marker genes for each cell type. Incorporate them in the model.

1. Can be as simple as fitting using sub-matrices.
2. Many different ways to select markers. Usually chosen by looking at reference samples.

Empirically models have better fit if restricted to marker genes.

dtangle

a new cell-type proportion estimator

dtangle in a Simple Setting

1. Two cell types: A and B

dtangle in a Simple Setting

1. Two cell types: A and B
2. mixture sample M with unknown mixing proportions p_A, p_B
   $p_A + p_B = 1$

dtangle in a Simple Setting

1. Two cell types: A and B
2. mixture sample M with unknown mixing proportions p_A, p_B
   $p_A + p_B = 1$
3. a and b are marker genes of cell types A and B.

dtangle in a Simple Setting

1. Two cell types: A and B
2. mixture sample M with unknown mixing proportions p_A, p_B
   $p_A + p_B = 1$
3. a and b are marker genes of cell types A and B.
4. $M_a, M_b = \log_2(\text{expr. of markers } a, b \text{ in mixture})$

dtangle in a Simple Setting

1. Two cell types: A and B
2. mixture sample M with unknown mixing proportions p_A, p_B
   $p_A + p_B = 1$
3. a and b are marker genes of cell types A and B.
4. $M_a, M_b = \log_2$ (expr. of markers a, b in mixture)
5. $R_{Aa}, R_{Bb} = \log_2$ (expr. of markers a, b in refs. A, B, resp.)

dtangle in a Simple Setting

1. Two cell types: A and B
2. mixture sample M with unknown mixing proportions p_A, p_B
   $p_A + p_B = 1$
3. a and b are marker genes of cell types A and B.
4. $M_a, M_b = \log_2(\text{expr. of markers } a, b \text{ in mixture})$
5. $R_{Aa}, R_{Bb} = \log_2(\text{expr. of markers } a, b \text{ in refs. A, B, resp.})$

The **dtangled** estimator of p_A is

$$\widehat{p_A} = \text{logistic}_2\left(\frac{(M_a - R_{Aa})}{\widehat{\gamma}} - \frac{(M_b - R_{Bb})}{\widehat{\gamma}}\right)$$

and similarly for p_B where $\text{logistic}_2(x) = 1/(1 + 2^{-x})$, and $\widehat{\gamma}$ is a sensitivity parameter.

dtangle in a Simple Setting

1. Two cell types: A and B
2. mixture sample M with unknown mixing proportions p_A, p_B
   $p_A + p_B = 1$
3. a and b are marker genes of cell types A and B.
4. $M_a, M_b = \log_2$ (expr. of markers a, b in mixture)
5. $R_{Aa}, R_{Bb} = \log_2$ (expr. of markers a, b in refs. A, B, resp.)

The **dtangled** estimator of p_A is

$$\widehat{p_A} = \text{logistic}_2 \overbrace{\left(\frac{(M_a - R_{Aa})}{\widehat{\gamma}} - \frac{(M_b - R_{Bb})}{\widehat{\gamma}} \right)}^{\text{norm. expr. a - norm. expr. b}}$$

and similarly for p_B where $\text{logistic}_2(x) = 1/(1 + 2^{-x})$, and $\widehat{\gamma}$ is a sensitivity parameter.

dtangle in a Simple Setting

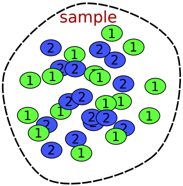
1. Two cell types: A and B
2. mixture sample M with unknown mixing proportions p_A, p_B
   $p_A + p_B = 1$
3. a and b are marker genes of cell types A and B.
4. $M_a, M_b = \log_2$ (expr. of markers a, b in mixture)
5. $R_{Aa}, R_{Bb} = \log_2$ (expr. of markers a, b in refs. A, B, resp.)

The **dtangled** estimator of p_A is

$$\widehat{p_A} = \text{logistic}_2 \left(\overbrace{\frac{(M_a - R_{Aa})}{\widehat{\gamma}} - \frac{(M_b - R_{Bb})}{\widehat{\gamma}}}^{\text{how much more type } A \text{ than } B} \right)$$

and similarly for p_B where $\text{logistic}_2(x) = 1/(1 + 2^{-x})$, and $\widehat{\gamma}$ is a sensitivity parameter.

dtangle is a New Cell-type Estimation Method



Actual Expression
(Unobserved)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	550	290	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	9.10	8.18	...

GEP technology

Measured Expression
(Observed)

Linear Scale

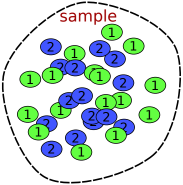
Name	Gene 1	Gene 2	...
Expression	500.56	302.88	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	8.96	8.23	...

Log Transform

dtangle is a New Cell-type Estimation Method



Actual Expression (Unobserved)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	550	290	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	9.10	8.18	...

GEP technology →

Measured Expression (Observed)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	500.56	302.88	...

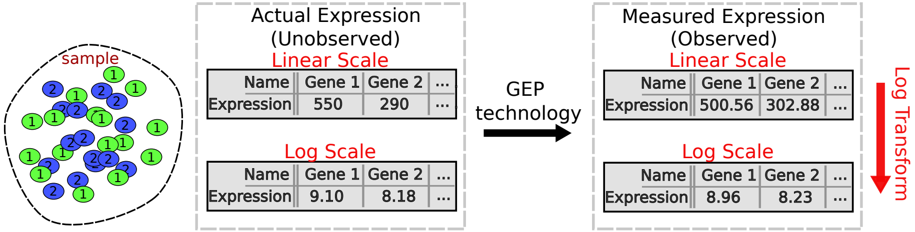
Log Scale

Name	Gene 1	Gene 2	...
Expression	8.96	8.23	...

Log Transform

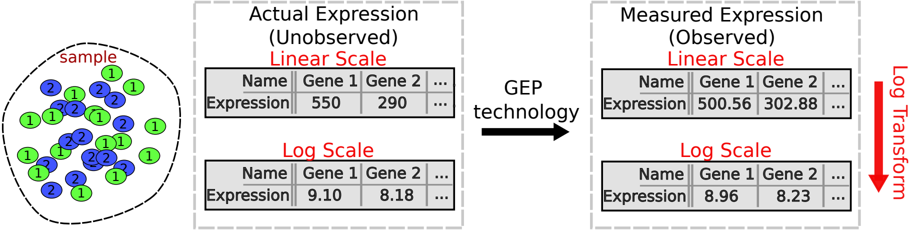
1. **Existing approach:** model and fit measured exprs. as $M \approx PR$ on **linear** (biologically plausible, not robust) or **log** scale (robust, not biologically plausible)

dtangle is a New Cell-type Estimation Method



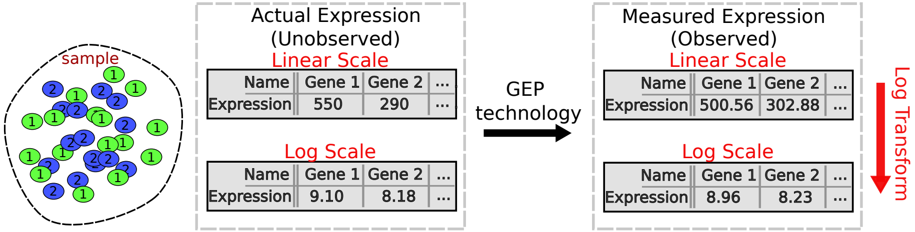
1. **Existing approach:** model and fit measured exprs. as $M \approx PR$ on **linear** (biologically plausible, not robust) or **log** scale (robust, not biologically plausible)
2. **dtangle's approach:**
   (1) model actual exprs. as $M \approx PR$ on **linear** scale

dtangle is a New Cell-type Estimation Method



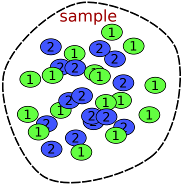
1. **Existing approach:** model and fit measured exprs. as $M \approx PR$ on **linear** (biologically plausible, not robust) or **log** scale (robust, not biologically plausible)
2. **dtangle's approach:**
   (1) model actual exprs. as $M \approx PR$ on **linear** scale (biologically plausible)

dtangle is a New Cell-type Estimation Method



1. **Existing approach:** model and fit measured exprs. as $M \approx PR$ on **linear** (biologically plausible, not robust) or **log** scale (robust, not biologically plausible)
2. **dtangle's approach:**
   (1) model actual exprs. as $M \approx PR$ on **linear** scale (biologically plausible)
   (2) model GEP tech. as linear on **log** scale

dtangle is a New Cell-type Estimation Method



Actual Expression (Unobserved)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	550	290	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	9.10	8.18	...

GEP technology →

Measured Expression (Observed)

Linear Scale

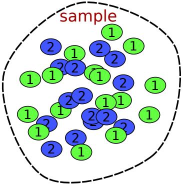
Name	Gene 1	Gene 2	...
Expression	500.56	302.88	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	8.96	8.23	...

Log Transform

1. **Existing approach:** model and fit measured exprs. as $M \approx PR$ on **linear** (biologically plausible, not robust) or **log** scale (robust, not biologically plausible)
2. **dtangle's approach:**
   (1) model actual exprs. as $M \approx PR$ on **linear** scale (biologically plausible)
   (2) model GEP tech. as linear on **log** scale (robust)



Actual Expression (Unobserved)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	550	290	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	9.10	8.18	...

GEP technology →

Measured Expression (Observed)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	500.56	302.88	...

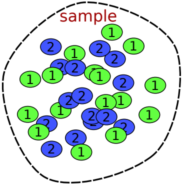
Log Transform ↓

Log Scale

Name	Gene 1	Gene 2	...
Expression	8.96	8.23	...

1. **Existing approach:** model and fit measured exprs. as $M \approx PR$ on **linear** (biologically plausible, not robust) or **log** scale (robust, not biologically plausible)
2. **dtangle's approach:**
   (1) model actual exprs. as $M \approx PR$ on **linear** scale (biologically plausible)
   (2) model GEP tech. as linear on **log** scale (robust)
   (3) combine and simplify (1) and (2) with **marker genes**, fit on log scale

dtangle is a New Cell-type Estimation Method



Actual Expression (Unobserved)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	550	290	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	9.10	8.18	...

GEP technology →

Measured Expression (Observed)

Linear Scale

Name	Gene 1	Gene 2	...
Expression	500.56	302.88	...

Log Scale

Name	Gene 1	Gene 2	...
Expression	8.96	8.23	...

Log Transform ↓

1. **Existing approach:** model and fit measured exprs. as $M \approx PR$ on **linear** (biologically plausible, not robust) or **log** scale (robust, not biologically plausible)
2. **dtangle's approach:**
   (1) model actual exprs. as $M \approx PR$ on **linear** scale (biologically plausible)
   (2) model GEP tech. as linear on **log** scale (robust)
   (3) combine and simplify (1) and (2) with **marker genes**, fit on log scale (plausible, robust, closed form, fast)

(Step 1) dtangle Models Actual Expression Mixing

Existing approach: model mixing of measured expressions:

$$M_g = p_A R_{Ag} + p_B R_{Bg}$$

on either the log or linear scale.

(Step 1) dtangle Models Actual Expression Mixing

Existing approach: model mixing of measured expressions:

$$M_g = p_A R_{Ag} + p_B R_{Bg}$$

on either the log or linear scale.

dtangle: model mixing of actual expressions on the linear scale:

$$\widetilde{M}_g = \text{actual expression of gene } g \text{ in mixture sample}$$

(and similarly for $\widetilde{R}_{Ag}$ and $\widetilde{R}_{Bg}$),

$$\widetilde{M}_g = p_A \widetilde{R}_{Ag} + p_B \widetilde{R}_{Bg}.$$

(Step 1) dtangle Models Actual Expression Mixing

Existing approach: model mixing of measured expressions:

$$M_g = p_A R_{Ag} + p_B R_{Bg}$$

on either the log or linear scale.

dtangle: model mixing of actual expressions on the linear scale:

$$\widetilde{M}_g = \text{actual expression of gene } g \text{ in mixture sample}$$

(and similarly for $\widetilde{R}_{Ag}$ and $\widetilde{R}_{Bg}$),

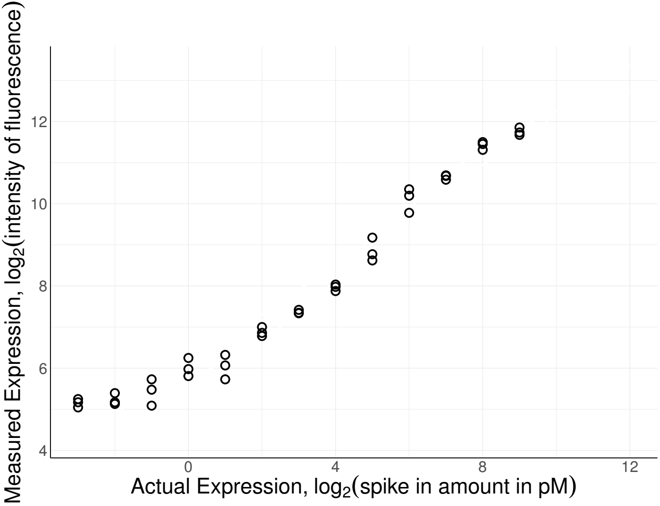
$$\widetilde{M}_g = p_A \widetilde{R}_{Ag} + p_B \widetilde{R}_{Bg}.$$

Compare:

$M_g = \log_2(\text{measured expression of gene } g \text{ in mixture sample})$

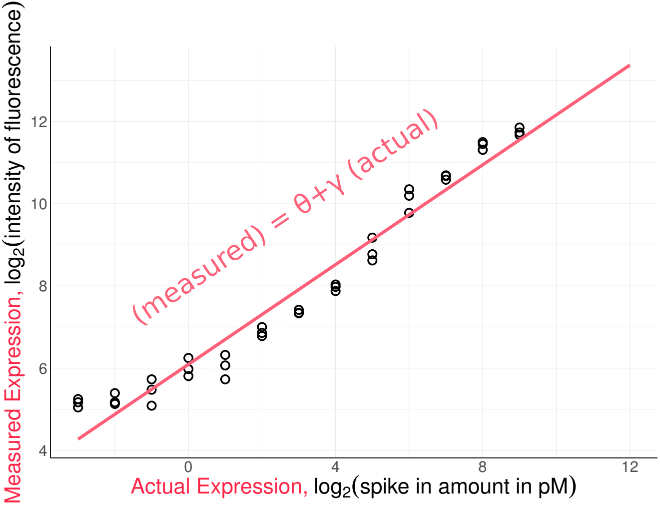
(Step 2) dtangle's Models GEP Technology on the log Scale

We model measured expression on actual expression as linear:



We model measured expression on actual expression as linear:

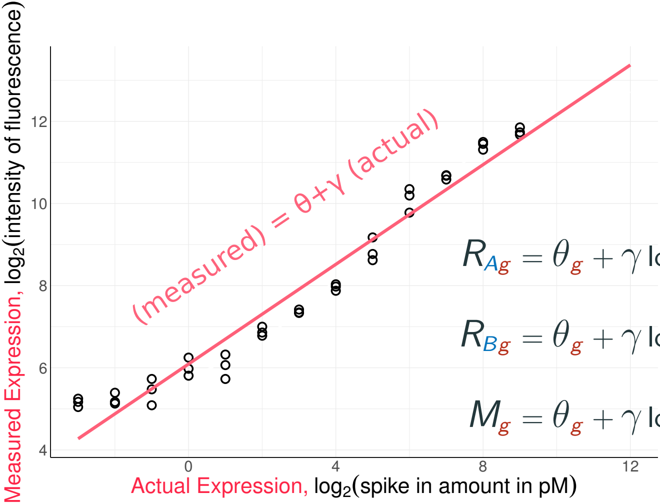
M_g



$\log_2 \widetilde{M}_g$

We model measured expression on actual expression as linear:

M_g



$$R_{Ag} = \theta_g + \gamma \log_2\left(\widetilde{R}_{Ag}\right) + \epsilon_{Ag}$$

$$R_{Bg} = \theta_g + \gamma \log_2\left(\widetilde{R}_{Bg}\right) + \epsilon_{Bg}$$

$$M_g = \theta_g + \gamma \log_2\left(\widetilde{M}_g\right) + \epsilon_g$$

$\log_2 \widetilde{M}_g$

(Step 3) dtangle Precisely Defines Marker Genes

(Defn) Marker gene: actually expressed in only one type.

$$\widetilde{R}_{Ab} = 0 \text{ and } \widetilde{R}_{Ba} = 0.$$

i.e. the actual expression of a in ref B is zero
and the actual expression of b in ref A is zero

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_a = p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}$.

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_a = p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}$.
(Step 2) model of GEP technology:

$$M_a = \theta_a + \gamma \log_2\left(\widetilde{M}_a\right) + \epsilon_a$$

$$R_{Aa} = \theta_a + \gamma \log_2\left(\widetilde{R}_{Aa}\right) + \epsilon_{Aa}$$

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_{a} = p_{A}\widetilde{R}_{Aa} + p_{B}\widetilde{R}_{Ba}$.
(Step 2) model of GEP technology:

$$M_{a} = \theta_{a} + \gamma \log_2\left(\widetilde{M}_{a}\right) + \epsilon_{a}$$
$$R_{Aa} = \theta_{a} + \gamma \log_2\left(\widetilde{R}_{Aa}\right) + \epsilon_{Aa}$$

(Step 3) define marker genes: $\widetilde{R}_{Ab} = 0$ and $\widetilde{R}_{Ba} = 0$

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_a = p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}$.
(Step 2) model of GEP technology:

$$M_a = \theta_a + \gamma \log_2 \left(\widetilde{M}_a \right) + \epsilon_a$$

$$R_{Aa} = \theta_a + \gamma \log_2 \left(\widetilde{R}_{Aa} \right) + \epsilon_{Aa}$$

(Step 3) define marker genes: $\widetilde{R}_{Ab} = 0$ and $\widetilde{R}_{Ba} = 0$

$$M_a = \theta_a + \gamma \log_2 \left(\widetilde{M}_a \right) + \epsilon_a$$

$$R_{Aa} = \theta_a + \gamma \log_2 \left(\widetilde{R}_{Aa} \right) + \epsilon_{Aa}$$

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_a = p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}$.
(Step 2) model of GEP technology:

$$M_a = \theta_a + \gamma \log_2\left(\widetilde{M}_a\right) + \epsilon_a$$
$$R_{Aa} = \theta_a + \gamma \log_2\left(\widetilde{R}_{Aa}\right) + \epsilon_{Aa}$$

(Step 3) define marker genes: $\widetilde{R}_{Ab} = 0$ and $\widetilde{R}_{Ba} = 0$

$$M_a = \theta_a + \gamma \log_2\left(p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}\right) + \epsilon_a$$
$$R_{Aa} = \theta_a + \gamma \log_2\left(\widetilde{R}_{Aa}\right) + \epsilon_{Aa}$$

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_a = p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}$.
(Step 2) model of GEP technology:

$$M_a = \theta_a + \gamma \log_2\left(\widetilde{M}_a\right) + \epsilon_a$$

$$R_{Aa} = \theta_a + \gamma \log_2\left(\widetilde{R}_{Aa}\right) + \epsilon_{Aa}$$

(Step 3) define marker genes: $\widetilde{R}_{Ab} = 0$ and $\widetilde{R}_{Ba} = 0$

$$M_a = \theta_a + \gamma \log_2\left(p_A \widetilde{R}_{Aa}\right) + \epsilon_a$$

$$R_{Aa} = \theta_a + \gamma \log_2\left(\widetilde{R}_{Aa}\right) + \epsilon_{Aa}$$

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_a = p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}$.
(Step 2) model of GEP technology:

$$M_a = \theta_a + \gamma \log_2 \left(\widetilde{M}_a \right) + \epsilon_a$$
$$R_{Aa} = \theta_a + \gamma \log_2 \left(\widetilde{R}_{Aa} \right) + \epsilon_{Aa}$$

(Step 3) define marker genes: $\widetilde{R}_{Ab} = 0$ and $\widetilde{R}_{Ba} = 0$

$$\frac{M_a - R_{Aa}}{\gamma} = \log_2 (p_A) + \epsilon$$

Combining dtangle's Models

(Step 1) model mixing $\widetilde{M}_a = p_A \widetilde{R}_{Aa} + p_B \widetilde{R}_{Ba}$.
(Step 2) model of GEP technology:

$$M_a = \theta_a + \gamma \log_2\left(\widetilde{M}_a\right) + \epsilon_a$$

$$R_{Aa} = \theta_a + \gamma \log_2\left(\widetilde{R}_{Aa}\right) + \epsilon_{Aa}$$

(Step 3) define marker genes: $\widetilde{R}_{Ab} = 0$ and $\widetilde{R}_{Ba} = 0$

$$\exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) \approx p_A$$

dtangle: A Re-normalization of the Exponential Terms.

We can show that,

$$p_A \approx \exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) \text{ and } p_B \approx \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)$$

dtangle: A Re-normalization of the Exponential Terms.

We can show that,

$$p_A \approx \exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) \text{ and } p_B \approx \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)$$

they are not nice since

1. they are not bounded above by 1
2. they do not sum to 1.

dtangle: A Re-normalization of the Exponential Terms.

We can show that,

$$p_A \approx \exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) \text{ and } p_B \approx \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)$$

they are not nice since

1. they are not bounded above by 1
2. they do not sum to 1.

We can fix this by re-normalizing each by their sum:

dtangle: A Re-normalization of the Exponential Terms.

We can show that,

$$p_A \approx \exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) \text{ and } p_B \approx \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)$$

they are not nice since

1. they are not bounded above by 1
2. they do not sum to 1.

We can fix this by re-normalizing each by their sum:

$$\widehat{p_A} = \frac{\exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right)}{\exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) + \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)}$$

dtangle: A Re-normalization of the Exponential Terms.

We can show that,

$$p_A \approx \exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) \text{ and } p_B \approx \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)$$

they are not nice since

1. they are not bounded above by 1
2. they do not sum to 1.

We can fix this by re-normalizing each by their sum:

$$\widehat{p_A} = \frac{\exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right)}{\exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) + \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)}$$
$$= \text{logistic}_2\left(\frac{(M_a - R_{Aa})}{\gamma} - \frac{(M_b - R_{Bb})}{\gamma}\right)$$

dtangle: A Re-normalization of the Exponential Terms.

We can show that,

$$p_A \approx \exp_2\left(\frac{M_a - R_{Aa}}{\gamma}\right) \text{ and } p_B \approx \exp_2\left(\frac{M_b - R_{Bb}}{\gamma}\right)$$

they are not nice since

1. they are not bounded above by 1
2. they do not sum to 1.

We can fix this by re-normalizing each by their sum:

$$\widehat{p_A} = \frac{\exp_2\left(\frac{M_a - R_{Aa}}{\widehat{\gamma}}\right)}{\exp_2\left(\frac{M_a - R_{Aa}}{\widehat{\gamma}}\right) + \exp_2\left(\frac{M_b - R_{Bb}}{\widehat{\gamma}}\right)}$$
$$= \text{logistic}_2\left(\frac{(M_a - R_{Aa})}{\widehat{\gamma}} - \frac{(M_b - R_{Bb})}{\widehat{\gamma}}\right)$$

dtangle is Generalizable

The general setting: (1) K cell types, (2) ν_k reference samples of type k, (3) set of marker genes G_k for each cell type. Want to estimate mixing proportions $p_1, \ldots, p_K$. For the simple case we had

$$\widehat{p_A} = \text{logistic}_2\left(\frac{(M_a - R_{Aa})}{\widehat{\gamma}} - \frac{(M_b - R_{Bb})}{\widehat{\gamma}}\right)$$

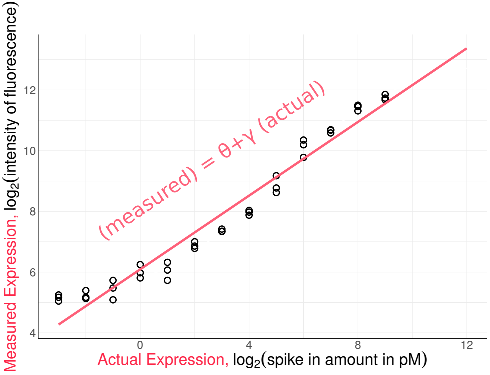
there is a direct generalization

$$\widehat{p_k} = L_k\left(\frac{(\overline{M_{G_k}} - \overline{R_{G_k}})}{\widehat{\gamma}} - \frac{(\overline{M_{G_1}} - \overline{R_{G_1}})}{\widehat{\gamma}}, \ldots, \frac{(\overline{M_{G_k}} - \overline{R_{G_k}})}{\widehat{\gamma}} - \frac{(\overline{M_{G_K}} - \overline{R_{G_K}})}{\widehat{\gamma}}\right)$$

1. $L_k(x) = 1/(1+\sum_{t \neq k} 2^{-x_t})$, a generalized logistic function
2. $\overline{M_{G_k}} = \frac{1}{|G_k|}\sum_{g \in G_k} M_g$, average marker genes in the mixture sample
3. $\overline{R_{G_k}} = \frac{1}{|G_k|\nu_k}\sum_{g \in G_k}\sum_{r=1}^{\nu_k} Z_{krg}$, average marker genes in references

Marker Genes and γ

1. Estimate γ from benchmark data sets:

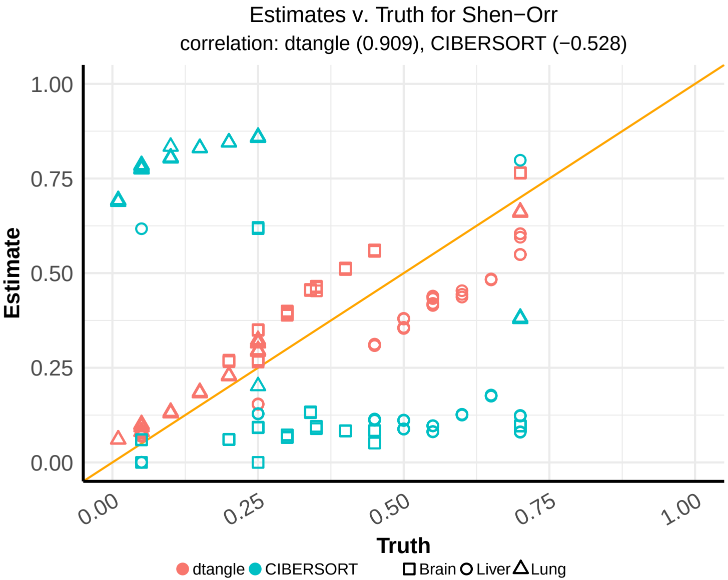


2. We find marker genes through differential expression analysis on the references.

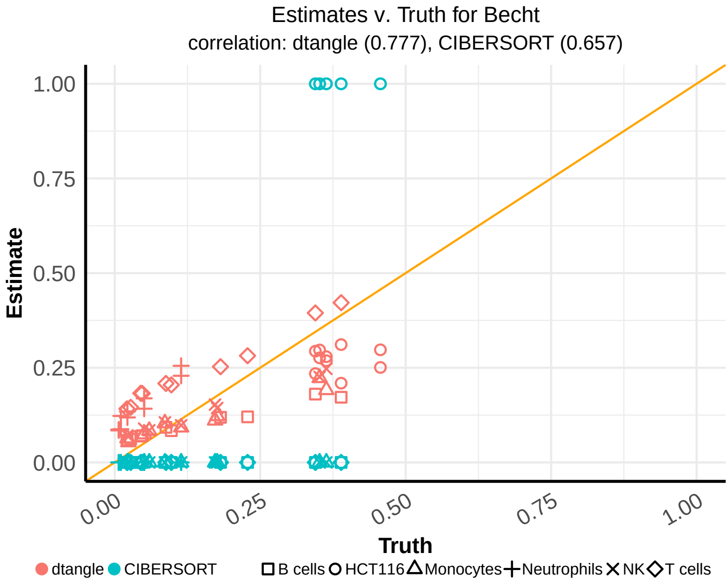
dtangle is robust to changes in γ and marker genes.

Benchmarking dtangle

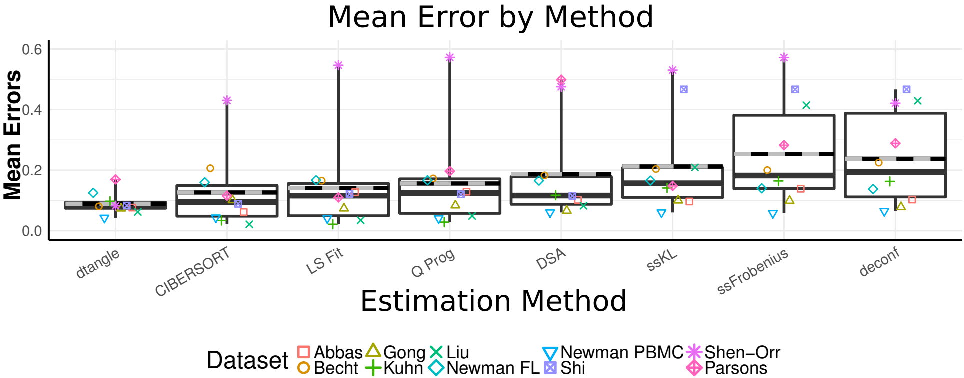
dtangle Works Well (Shen-Orr et al.)



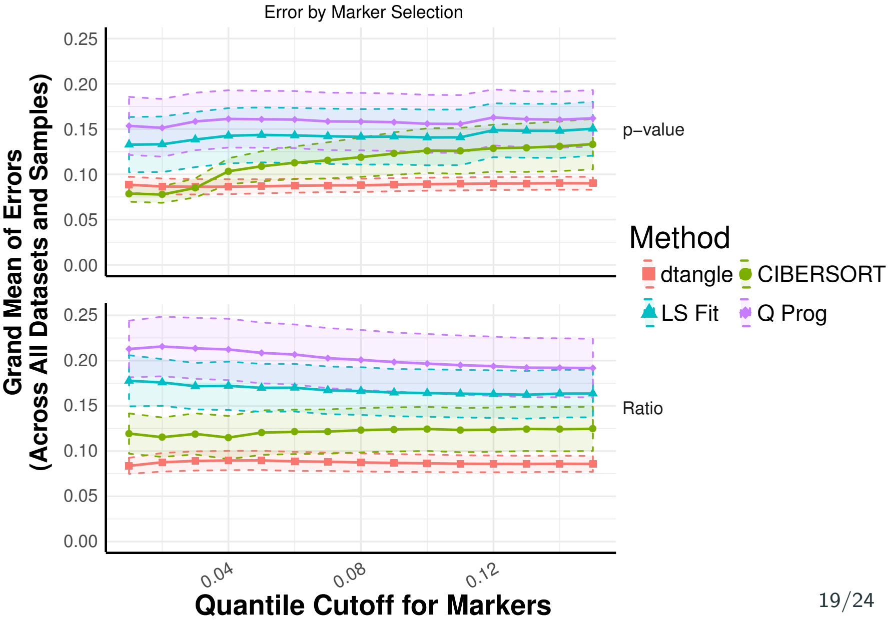
dtangle Works With Complicated Data (Becht et al.)



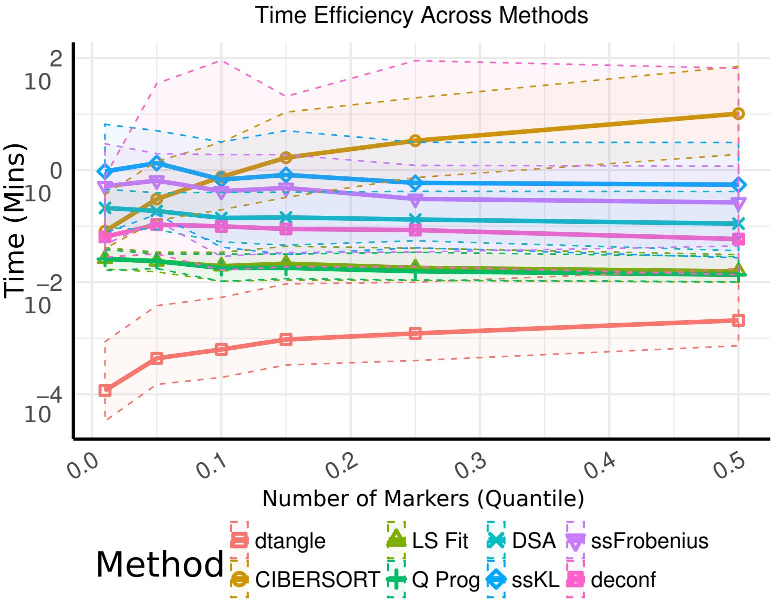
dtangle is Consistently Good



dtangle is Robust



dtangle is Fast

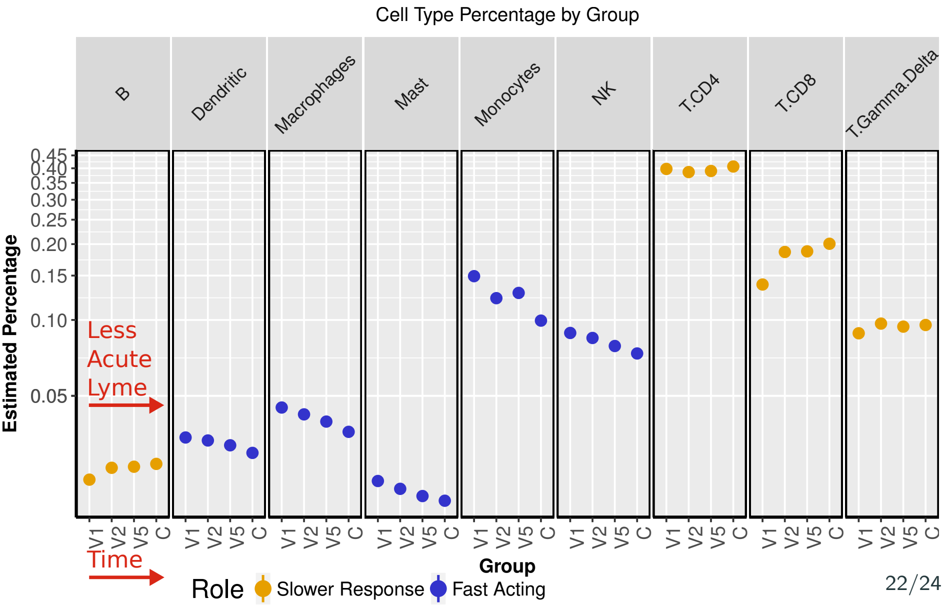


Revisiting the Lyme Example

Gene expression measurements of white blood cells from Bouquet et al.

1. Gene expression measurements of 28 patients at three points:
   - (V1) at diagnosis
   - (V2) after antibiotic treatment
   - (V5) 6 months post treatment
2. Gene expressions of 13 healthy controls (C)

dtangle on the Lyme Data



Future Work

Future research directions:

1. estimating proportion of unknown cell-types
2. removing unwanted latent factors as part of estimation
3. extension to high-throughput methylation data
4. variance estimate and goodness-of-fit

dtangle is Available!

An R package is available

on github

`dtangle.github.io`

or on CRAN

`cran.r-project.org/package=dtangle`

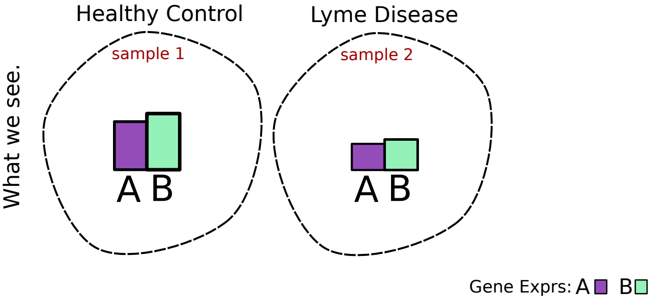
Hopefully rolling out to stemformatics soon!

`www.stemformatics.org`

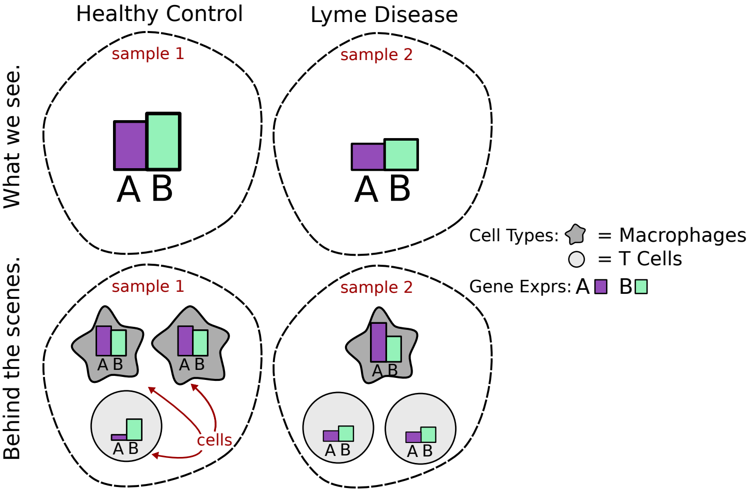
Thanks!

Extras

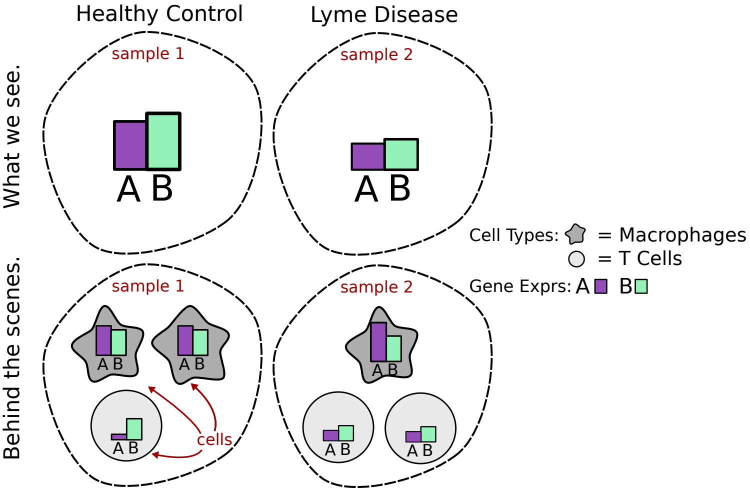
Cell-types Can Confound Differential Expression Analysis



Cell-types Can Confound Differential Expression Analysis



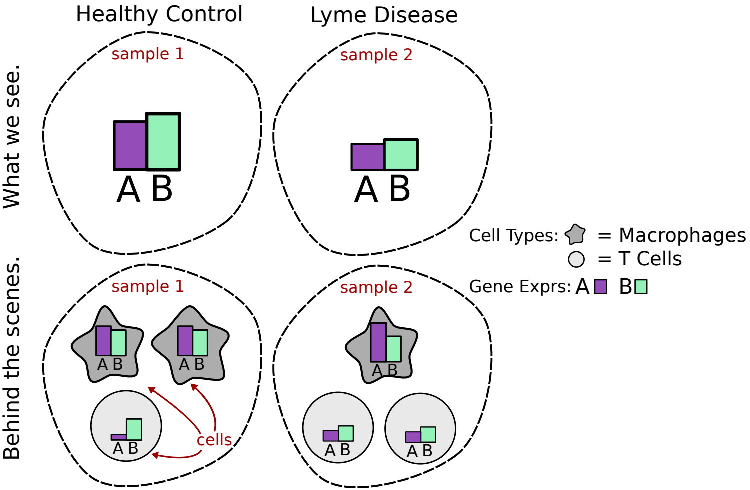
Cell-types Can Confound Differential Expression Analysis



Differences we see come from

1. differences across samples of GEPs for each cell type

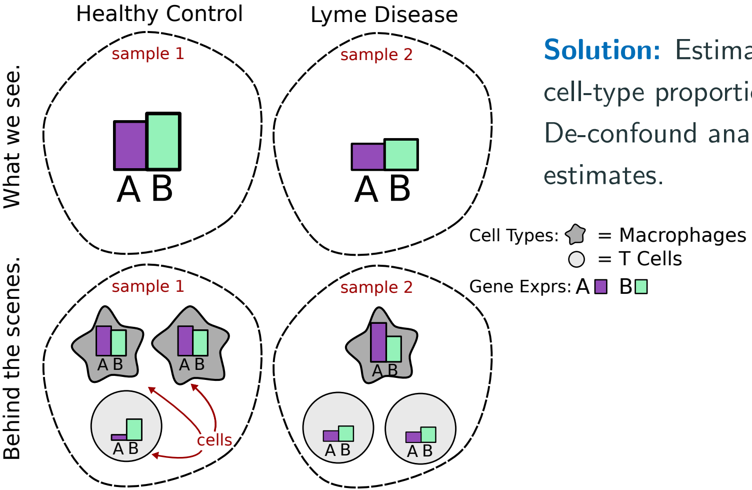
Cell-types Can Confound Differential Expression Analysis



Differences we see come from

1. differences across samples of GEPs for each cell type
2. differences across samples of cell-type composition

Cell-types Can Confound Differential Expression Analysis



Solution: Estimate the cell-type proportions. De-confound analysis with estimates.

Differences we see come from

1. differences across samples of GEPs for each cell type
2. differences across samples of cell-type composition

Accounting for Cell Types Drastically Changes Results

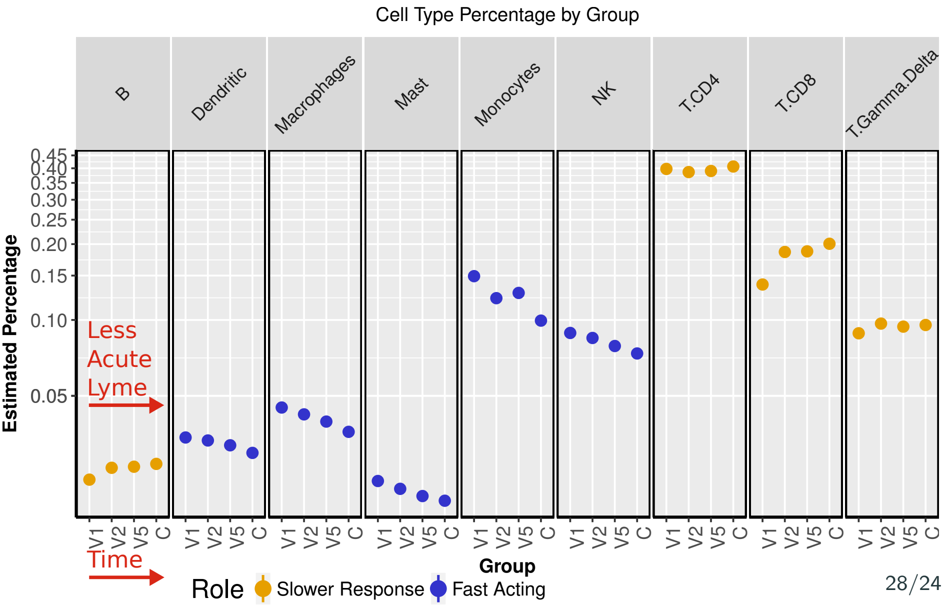
We compare the control group to Lyme patients:

1. **un-adjusted:** there are 399 differentially expressed genes
2. **cell-type adjusted:** there are 158 differentially expressed genes after adding in covariates to account for cell types

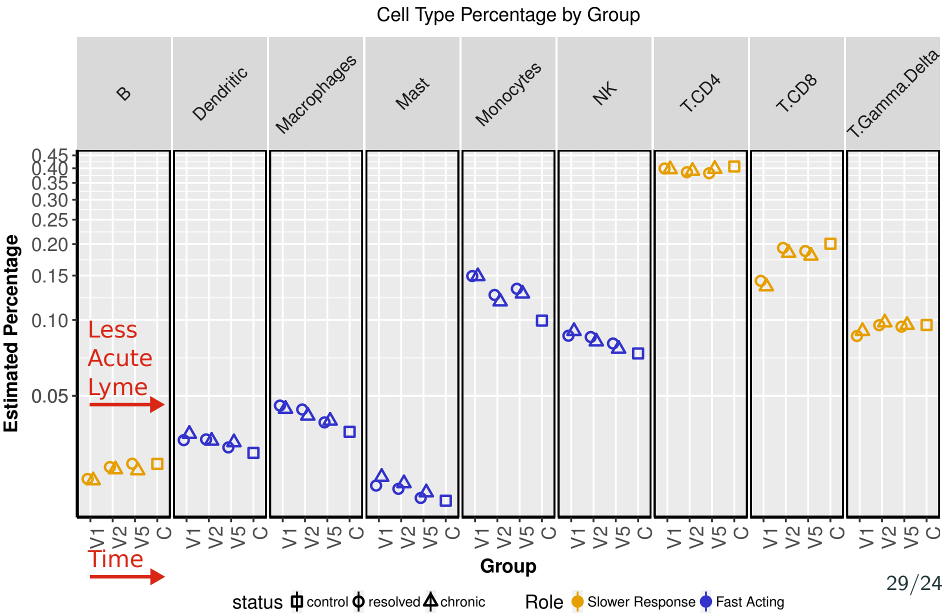
Number of diff. expressed genes changes by a factor of 2.5!

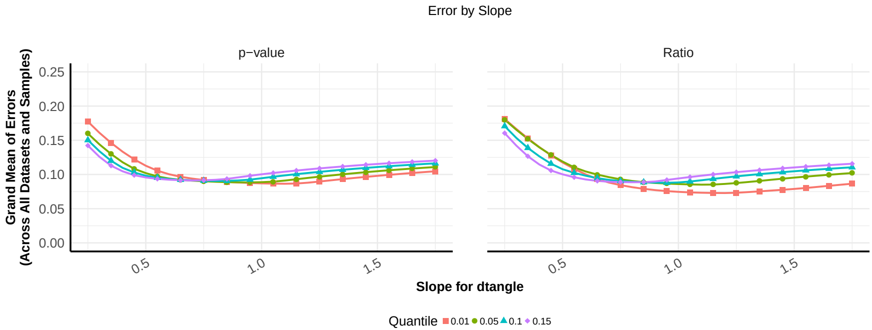
Some of the un-adjusted genes probably due to cell type.

dtangle on the Lyme data



dtangle on the Lyme data




Error by Slope
p−value
Ratio
Grand Mean of Errors
(Across All Datasets and Samples)
0.25
0.20
0.15
0.10
0.05
0.00
0.5
1.0
1.5
Slope for dtangle
Quantile
0.01
0.05
0.1
0.15

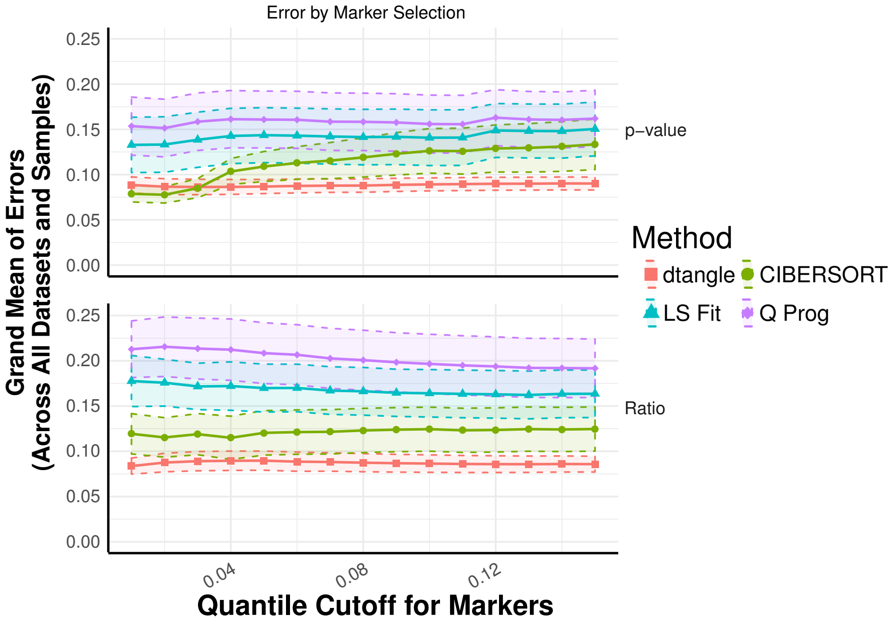

Error by Marker Selection
Grand Mean of Errors
(Across All Datasets and Samples)
0.25
0.20
0.15
0.10
0.05
0.00
p−value
Ratio
Method
dtangle
CIBERSORT
LS Fit
Q Prog
0.04
0.08
0.12
Quantile Cutoff for Markers

Image Attribution

1 Adult deer tick, "Ixodes scapularis".;Source: `http://www.ars.usda.gov/is/graphics/photos/mar98/k8002-3.htm`;Image Number: K8002-3 ;Credits: Photo by Scott Bauer. PD-USGov-USDA-ARS

2 Electron micrograph of "Treponema pallidum". From `http://phil.cdc.gov/phil/home.asp` ID 1977.

References

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE*, **4**(7).

Altboum, Z., Steuerman, Y., David, E., Barnett-Itzhaki, Z., Valadarsky, L., Keren-Shaul, H., Meningher, T., Mendelson, E., Mandelboim, M., Gat-Viks, I., and Amit, I. (2014). Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular Systems Biology*, **10**(2), 1–14.

Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., and Szustakowski, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE*, **6**(11).

Lu, P., Nakorchevskiy, A., and Marcotte, E. M. (2003). Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(18), 10370–5.

Newman, A. M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, W. F., Yue Xu, C. D. H., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.*, **12**(5), 193–201.

Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q., and Zandstra, P. W. (2012). PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Computational Biology*, **8**(12).

Quon, G. and Morris, Q. (2009). ISOLATE: A computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, **25**(21), 2882–2889.

Quon, G., Haider, S., Deshwar, A. G., Cui, A., Boutros, P. C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome medicine*, **5**(3), 29.

Wang, M., Master, S. R., and Chodosh, L. a. (2006). Computational expression deconvolution in a complex mammalian organ. *BMC bioinformatics*, **7**, 328.