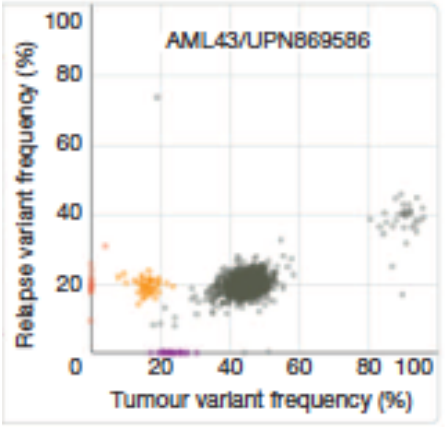
Two Aspects in Tumor Heterogeneity: Subclonal Mutations and Stromal Expression

Wenyi Wang, PhD

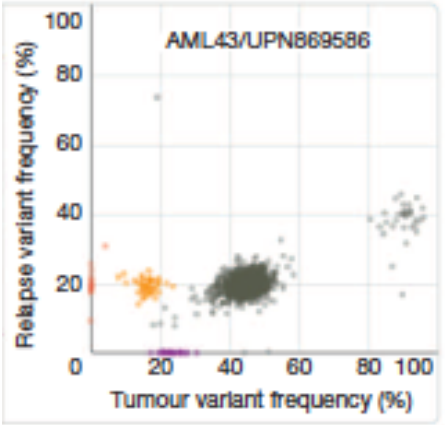
Department of Bioinformatics and Computational Biology
The University of Texas MD Anderson Cancer Center

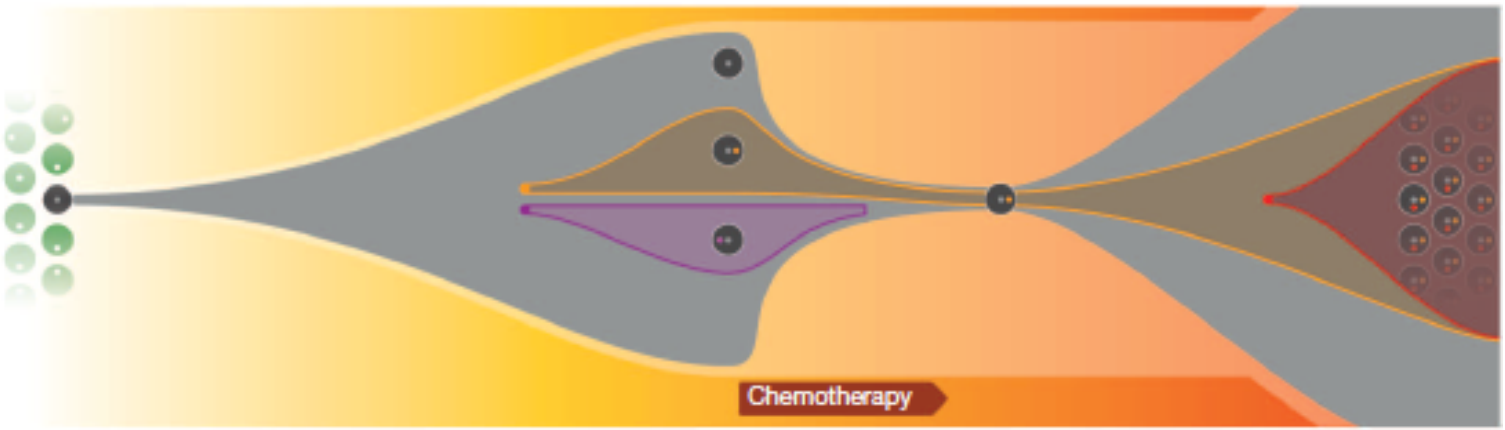
June 3, 2014

Evolution of subclonal mutations in Acute Myeloid Leukemia (Ding et al. Nature 2012)

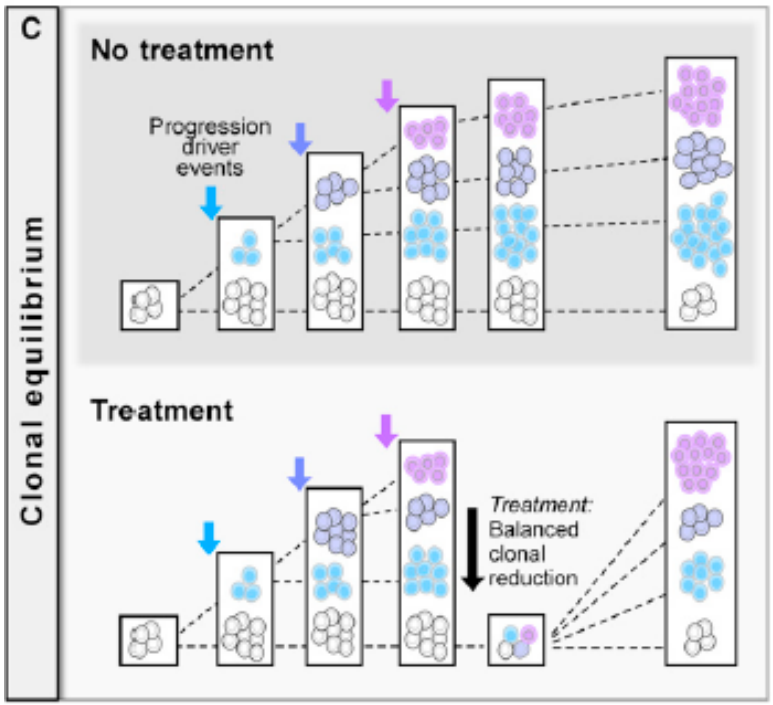


Evolution of subclonal mutations in Acute Myeloid Leukemia (Ding et al. Nature 2012)

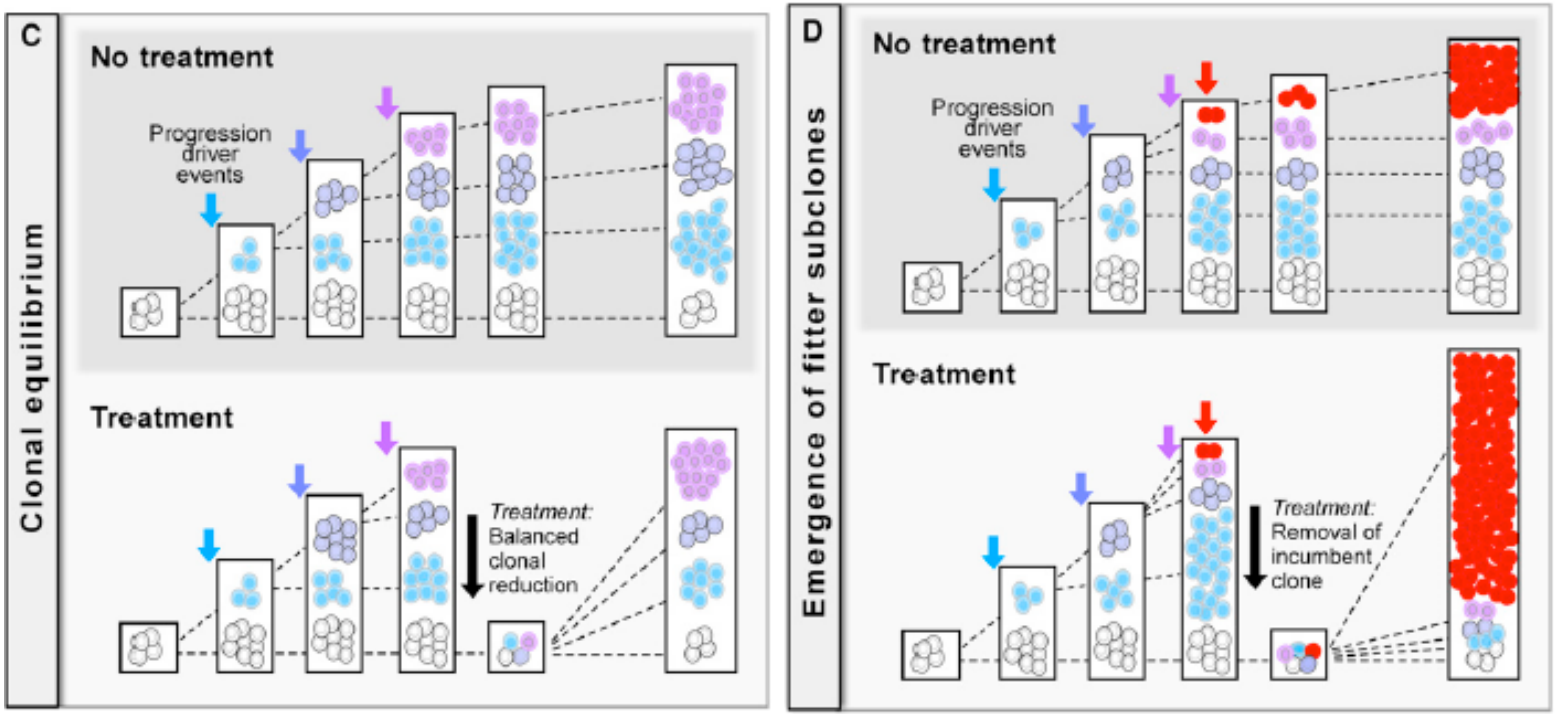




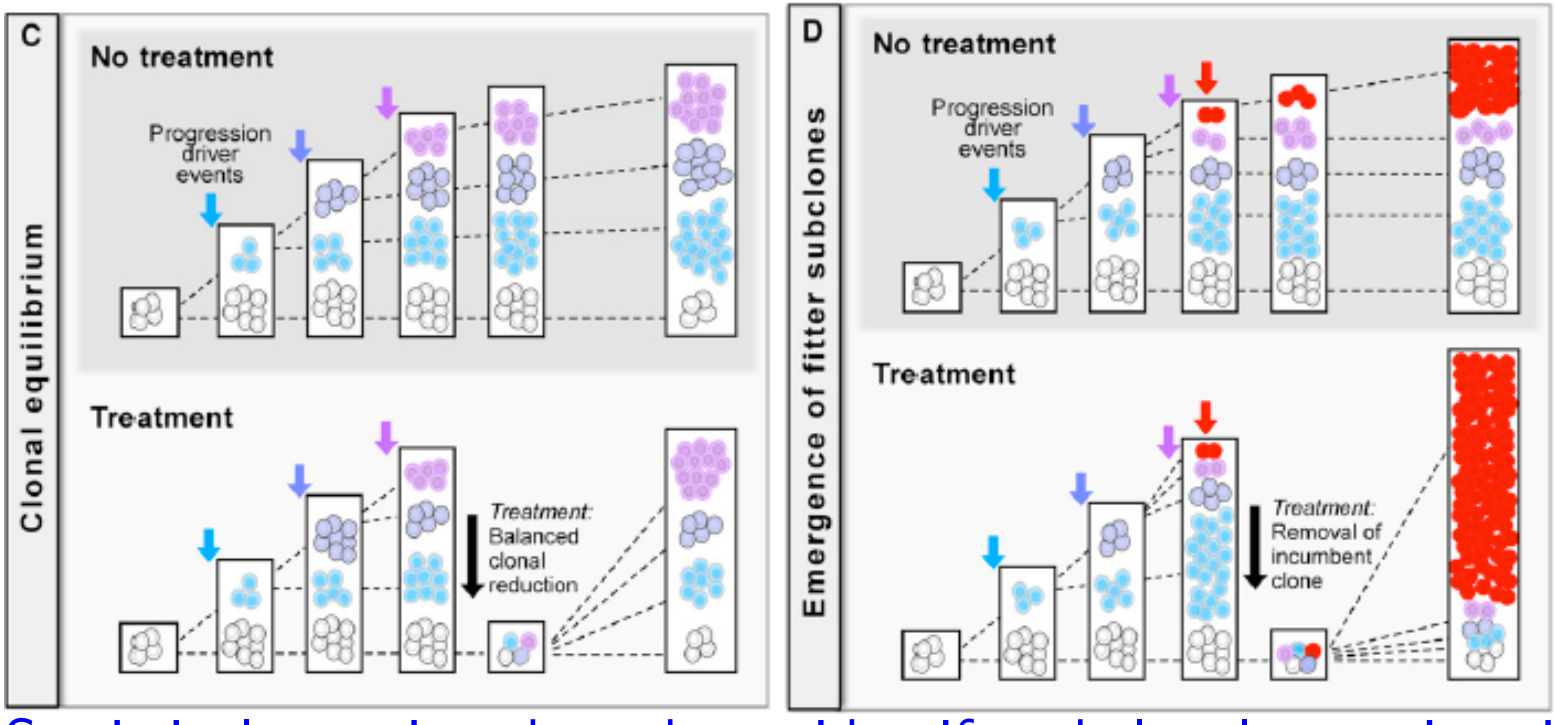
Evolution of subclonal mutations in Chronic Lymphocytic Leukemia (Landau et al. Cell 2013)



Evolution of subclonal mutations in Chronic Lymphocytic Leukemia (Landau et al. Cell 2013)



Evolution of subclonal mutations in Chronic Lymphocytic Leukemia (Landau et al. Cell 2013)

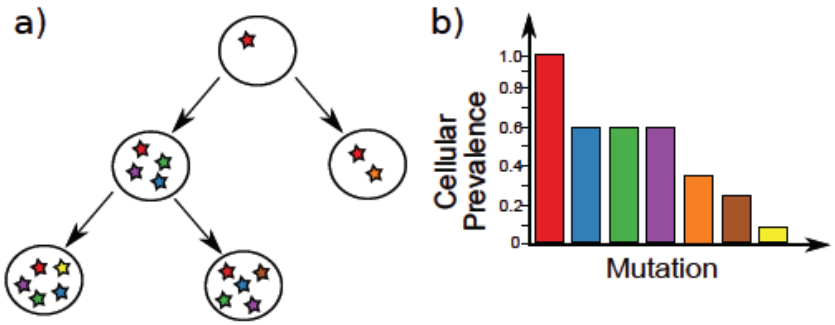


Statistical question: how do we identify subclonal mutations in the DNA sequencing data from tumor samples?

PyClone, Nature Methods 2014

The probability p of a read containing variant allele with mutation state $\psi = (g_N, g_R, g_V)$ and cellular prevalence ϕ is given by:

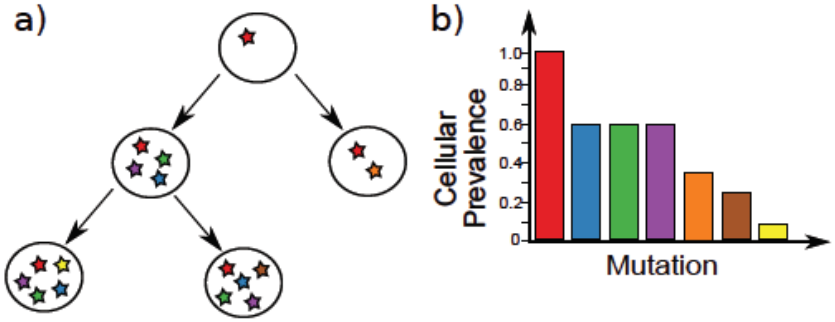
$$\begin{aligned} p(\psi, \phi, t) &= \frac{(1-t)c(g_N)}{Z}\mu(g_N) + \frac{t(1-\phi)c(g_R)}{Z}\mu(g_R) + \frac{t\phi c(g_V)}{Z}\mu(g_V) \\ Z &= (1-t)c(g_N) + t(1-\phi)c(g_R) + t\phi c(g_V) \end{aligned}$$



PyClone, Nature Methods 2014

The probability p of a read containing variant allele with mutation state $\psi = (g_N, g_R, g_V)$ and cellular prevalence ϕ is given by:

$$\begin{aligned} p(\psi, \phi, t) &= \frac{(1-t)c(g_N)}{Z}\mu(g_N) + \frac{t(1-\phi)c(g_R)}{Z}\mu(g_R) + \frac{t\phi c(g_V)}{Z}\mu(g_V) \\ Z &= (1-t)c(g_N) + t(1-\phi)c(g_R) + t\phi c(g_V) \end{aligned}$$



The cellular prevalence ϕ cannot be deconvoluted into subclonal fractions, unless under stringent assumptions of mutational evolution.

Lee et al. and Xu et al. 2014

Straightforward modeling of the fraction of cell clone c for sample t using w_{tc}:

$$p_{st} = w_{t0}p_0 + \sum_{1}^{C} w_{tc}z_{sc}$$

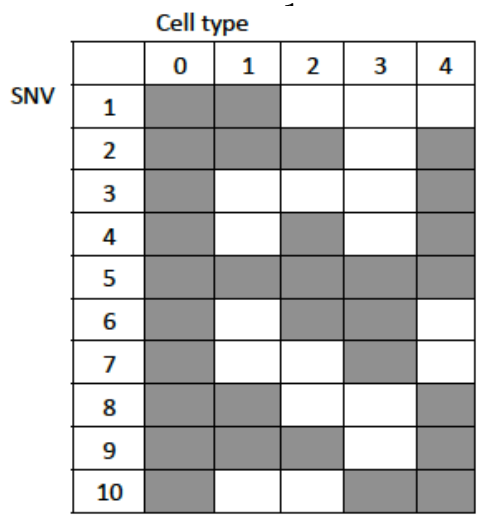
Cell type

SNV	0	1	2	3	4
1	■	■			
2	■	■	■		■
3	■				■
4	■		■		■
5	■	■	■	■	■
6	■		■	■	
7	■			■	
8	■	■			■
9	■	■	■		■
10	■			■	■

Lee et al. and Xu et al. 2014

Straightforward modeling of the fraction of cell clone c for sample t using w_{tc}:

$$p_{st} = w_{t0}p_0 + \sum_{1}^{C} w_{tc}z_{sc}$$

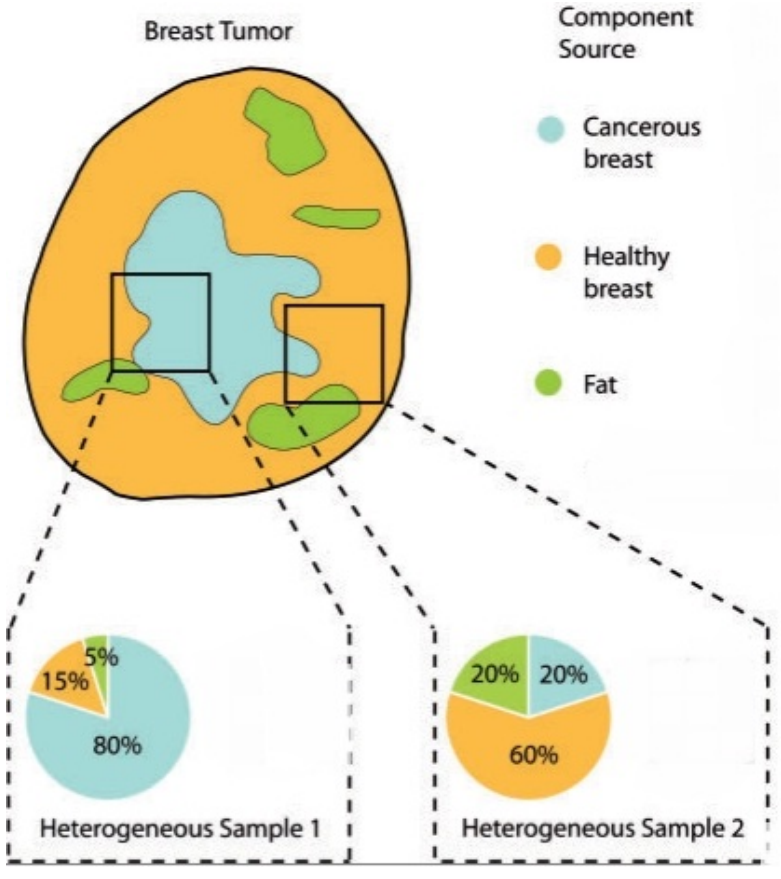


There is no genotype estimation. The interpretation of w_{tc} is therefore convoluted with one fixed type of mutation genotype, e.g. pairs of haplotypes, or extend z_{sc} to be catogorical: (0,1,2)

Summary

- Clonal evolution is a key feature of cancer progression and relapse.
- Identification of subclonal mutations in cancer studies consists of
  - Estimating total number of subclones and their mixing proportions
  - Finding mutations within each cellular subclone

Distinct compartments and changing proportions in tumor samples

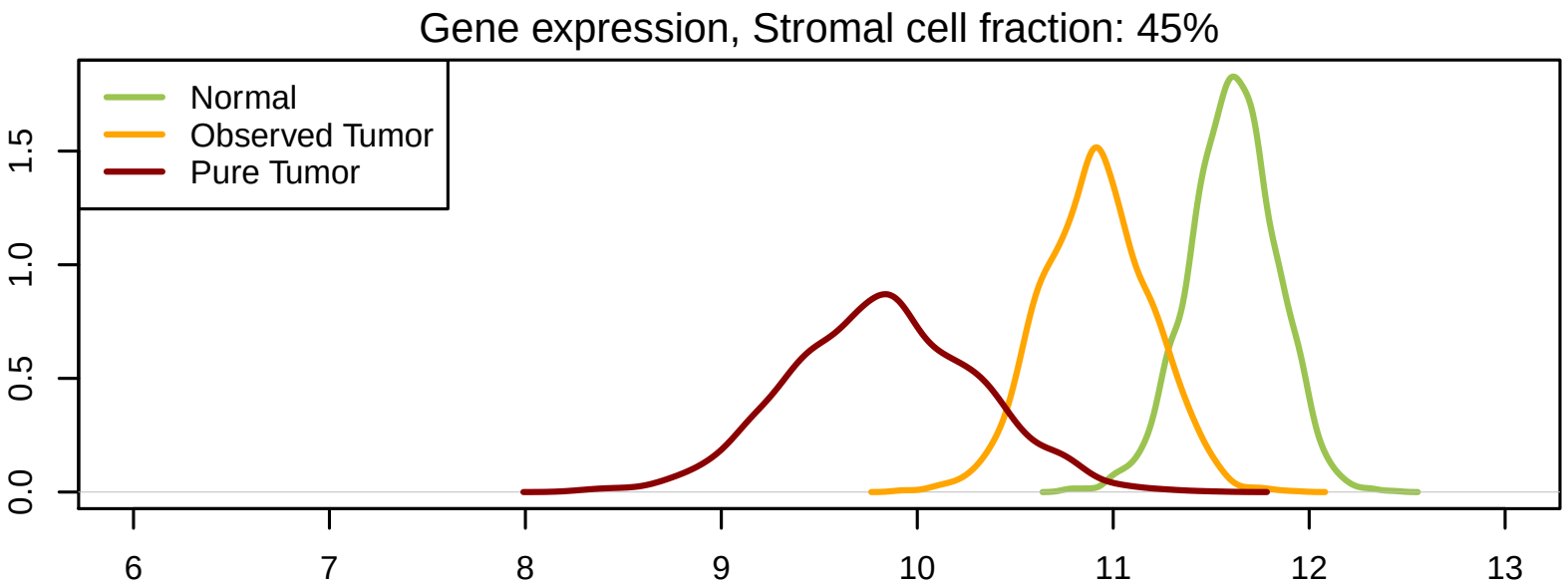


ref: Quon and Morris (2009)

Issue of tumor heterogeneity in gene expression

Traditional gene expression (GE) profiling

- True tumor cell gene expressions are masked by stromal cell gene expressions



Previous work

Gene expression deconvolution

- Experimental: Laser-capture microdissection (LCM, 1996)
- *In silico*: $AX = B$ for each mixed sample,

  where A is a matrix pure cell expressions, X is a vector of proportions, B is the observed heterogeneous expression data

Previous work

Gene expression deconvolution

- Experimental: Laser-capture microdissection (LCM, 1996)
- *In silico*: $AX=B$ for each mixed sample,

  where A is a matrix pure cell expressions, X is a vector of proportions, B is the observed heterogeneous expression data

Linear assumption

For gene g and sample i, π_i is an unknown cell fraction for sample i

$$Y_{ig} = \pi_i T_{ig} + (1 - \pi_i) N_{ig}.$$

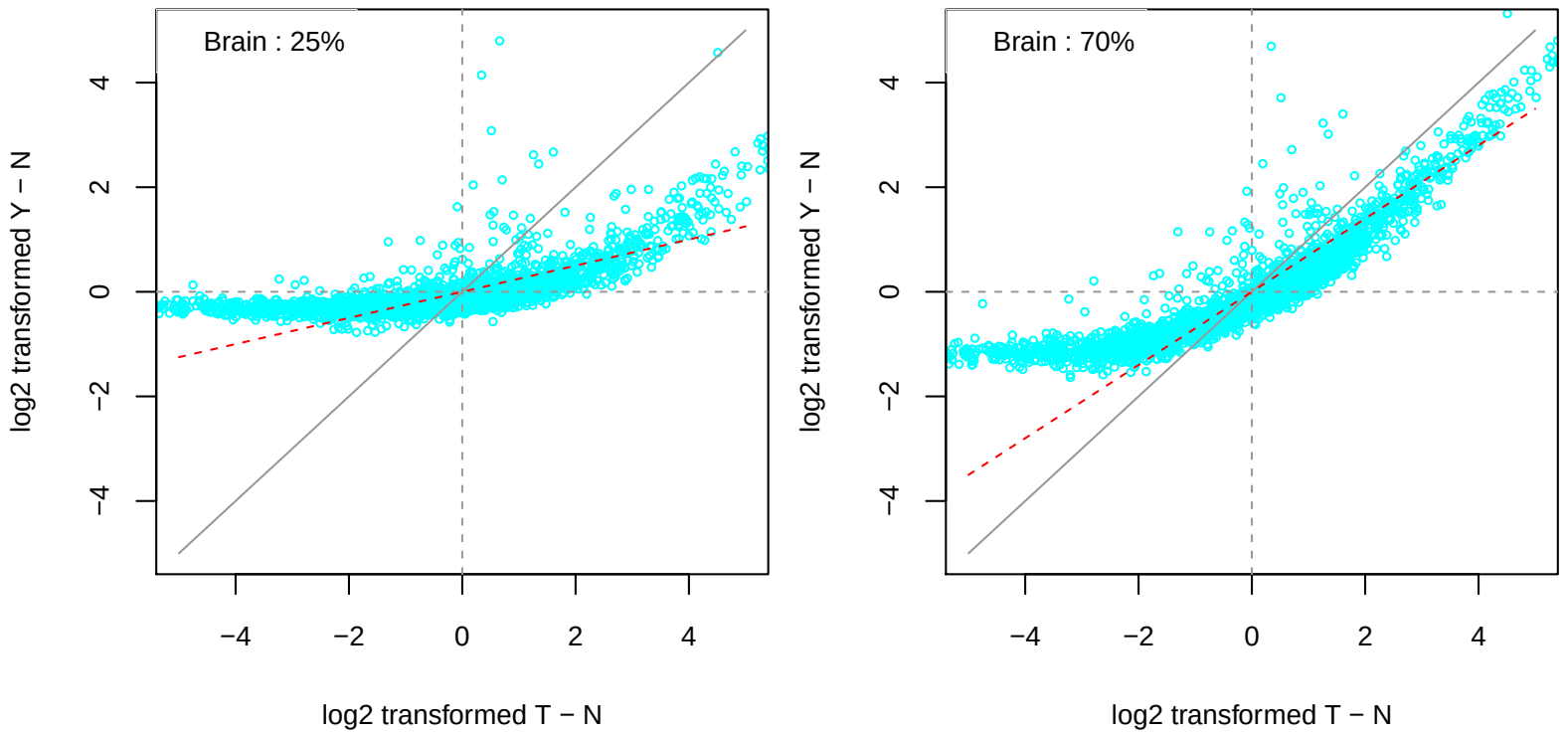
where T_{ig} represents *tumoral expression* and N_{ig} represents *stromal expression*.

Challenges with gene expression deconvolution

- Linearity assumption holds better with raw measured data (Liu and zhong, Nature Methods 2011)
- There is need for more practical way to jointly estimate the mean of A across samples and X.
- Estimating individual expression A is needed for clinical profiling.

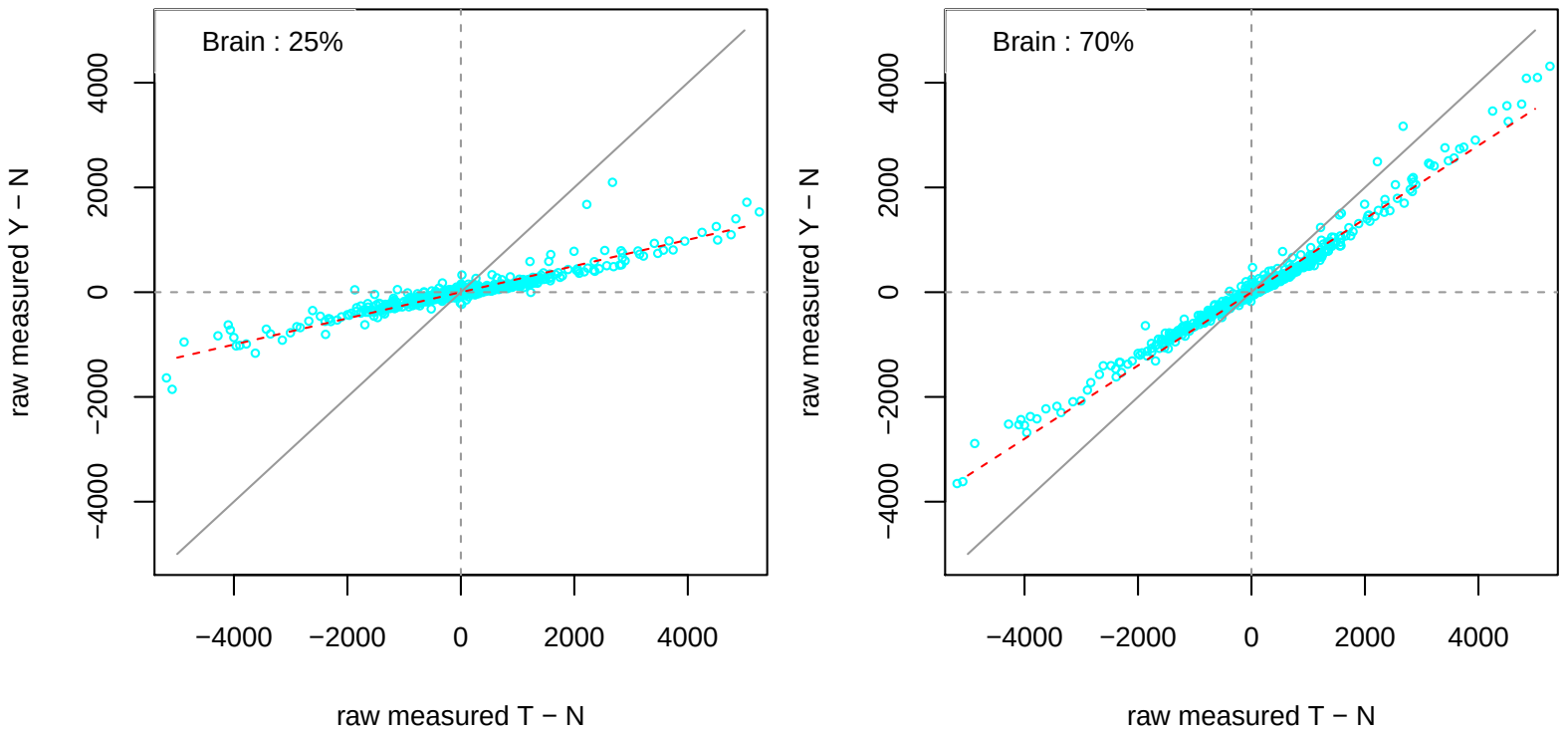
GSE19830 data with known proportions

Using log-transformed data, Linearity does not hold

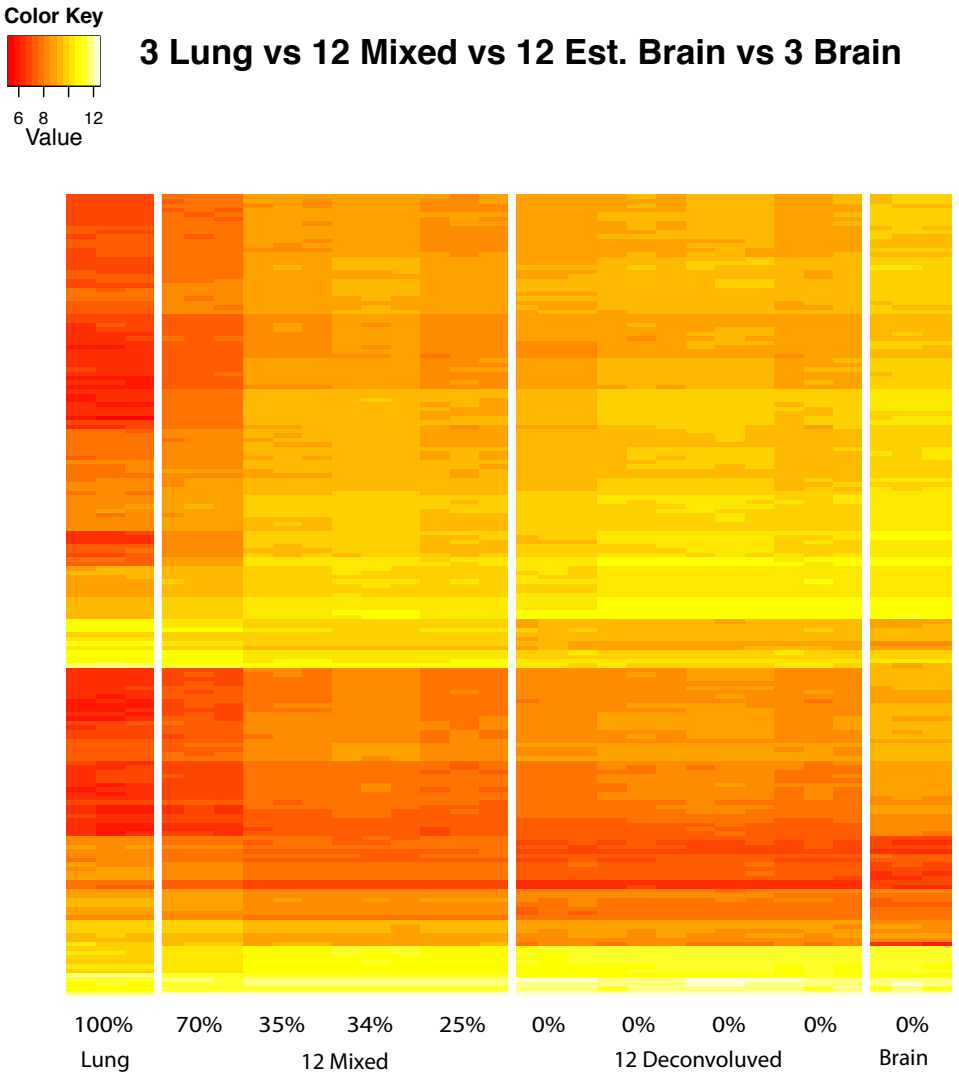


GSE19830 data with known proportions

Using raw-measured data, linearity holds



Clinical impact – deconvolved individual gene expressions



Related concepts

Matched versus unmatched samples

- $Y_{ig} = (1-\pi_i)N_{ig} + \pi_i T_{ig} \quad => \quad \pi_i = (Y_{ig} - N_{ig})/(T_{ig} - N_{ig})$,
  $\rightarrow$Applies to sample-specific and gene-specific Y's and N's (matched design).
- In practice, we often need to deconvolve for unmatched samples

Related concepts

Matched versus unmatched samples

- $Y_{ig} = (1-\pi_i)N_{ig} + \pi_i T_{ig} \quad => \quad \pi_i = (Y_{ig} - N_{ig})/(T_{ig} - N_{ig})$,
  $\rightarrow$Applies to sample-specific and gene-specific Y's and N's (matched design).
- In practice, we often need to deconvolve for unmatched samples

Reference genes

- Genes with known expression profiles for both tumor and normal samples.
- Available methods require knowledge of reference genes for deconvolution. What can we do when no reference genes are available?

Goals

Using raw-measured data, we develop a general framework that

- Estimates unobserved cell-type proportions in heterogeneous samples with/without knowledge of reference genes
- Reconstitute pure normal/tumor gene expressions for matched/unmatched individual samples

Ahn et al. Bioinformatics 2013

Assumption

For gene g and sample i, π_i is an unknown cell fraction for sample i

$$Y_{ig} = \pi_i T_{ig} + (1 - \pi_i) N_{ig}.$$

We assume $N_{i'g} \sim LN(\mu_{Ng}, \sigma^2_{Ng})$ and $T_{ig} \sim LN(\mu_{Tg}, \sigma^2_{Tg})$ where LN represents a $\log_2$ Normal distribution.

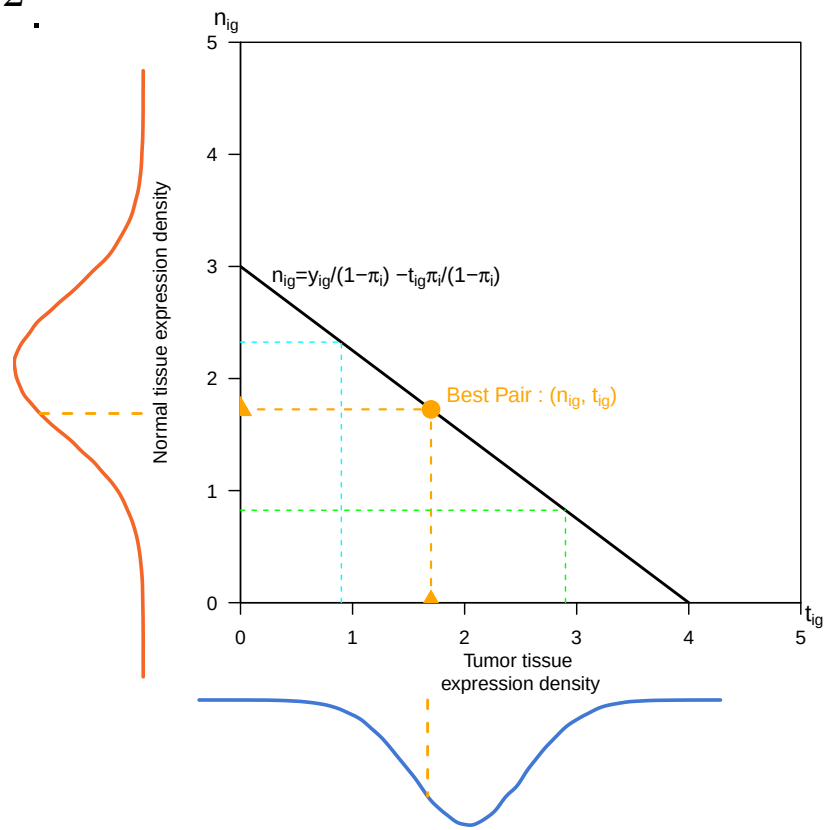
Step i. Given the Y's and the distribution of the N's, we search for a set of $\{\boldsymbol{\pi}\}$ that maximize the likelihood of observing Y, using the Nelder-Mead procedure.

Step ii. Given the $\hat{\pi}$'s and the distributions of the T's and N's, we estimate an individual pair of (t,n) for each sample and each gene.

Geometric interpretation of the individual deconvolution.

$$\text{argmax}_{t_{ig}} \phi(t_{ig}|\hat{\mu}_{Tg}, \hat{\sigma}^2_{Tg}) \phi\left(\frac{y_{ig} - \hat{\pi}_i t_{ig}}{1 - \hat{\pi}_i} \middle| \hat{\mu}_{Ng}, \hat{\sigma}^2_{Ng}\right)$$

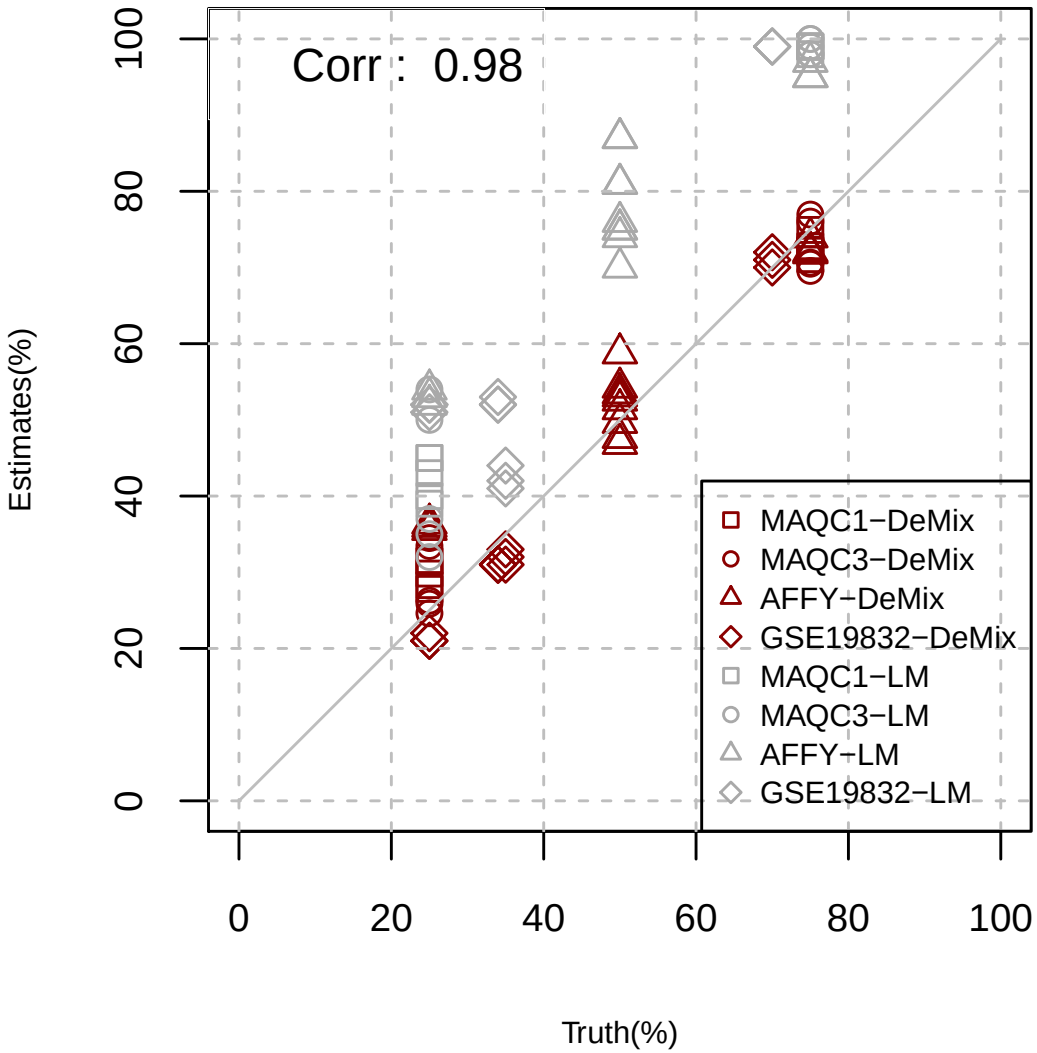
where $\phi(\cdot|\mu, \sigma^2)$ is a $\log_2$ Normal density with corresponding mean μ and variance σ^2.



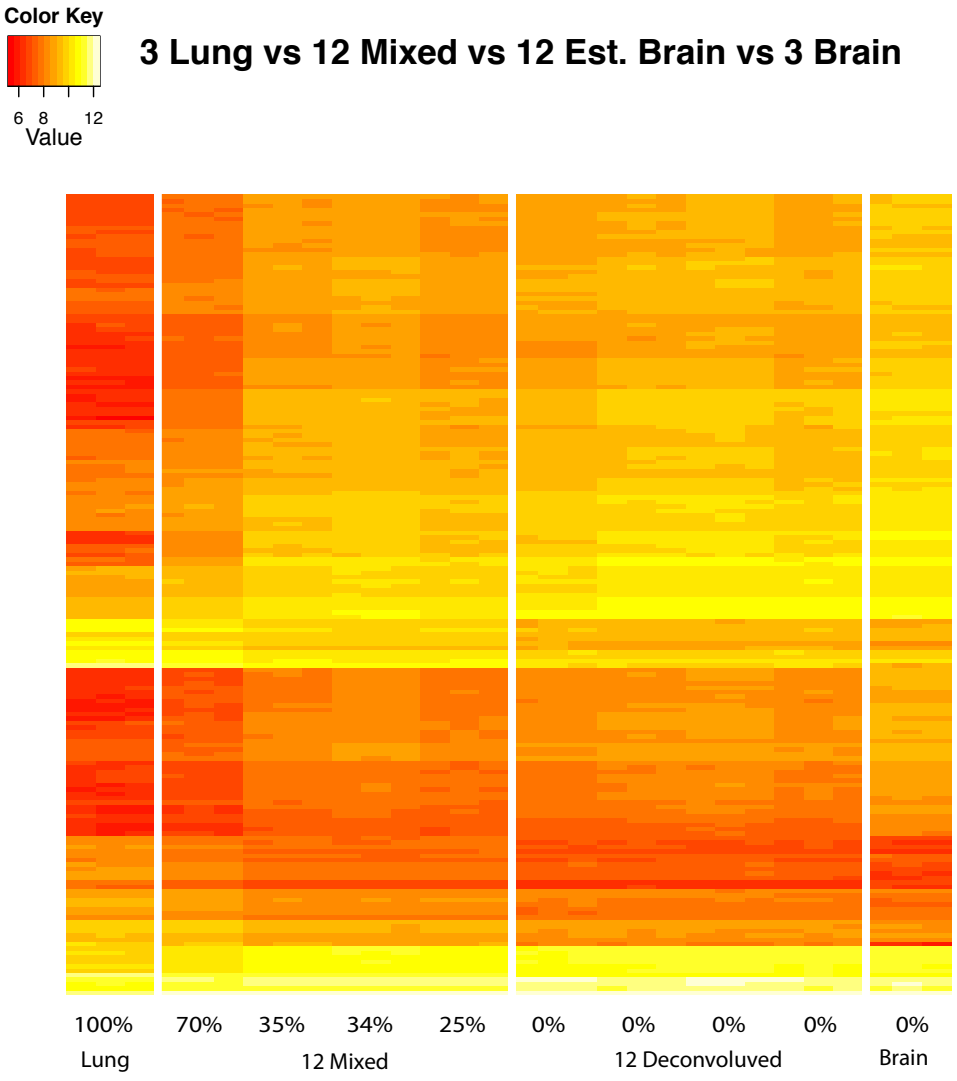
Real data with known proportions for validation

1. GSE19830 (Shen-Orr et al., 2010). Twelve liver-brain mixed samples.
2. GSE5350 from the MicroArray Quality Control (MAQC) project (MAQC Consortium, 2006). Ten mixed samples from Affymetrix and ten mixed samples from Illumina arrays.
3. Affymetrix Twelve brain-heart mixed samples.

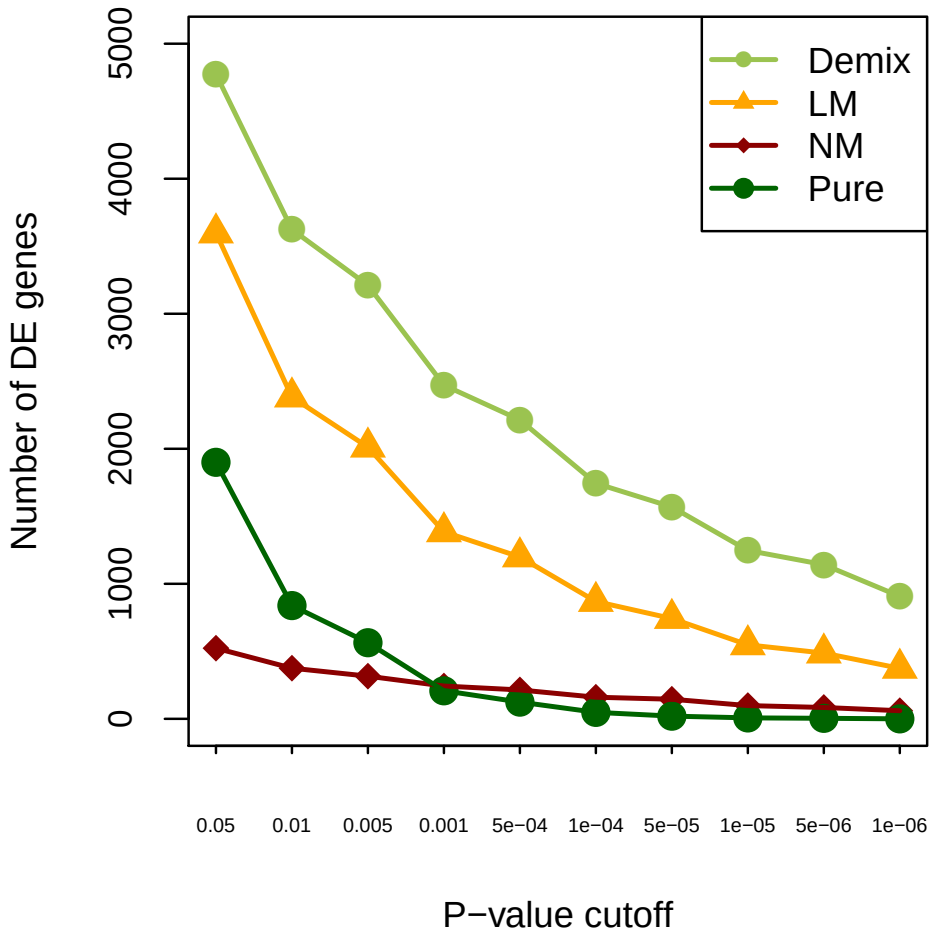
Proportion estimations



Deconvolved individual gene expressions



Expected changes



Our general framework for RNAseq data

Assumption

For gene g and sample i, π_i is an unknown cell fraction for sample i

$$Y_{ig} = \pi_i T_{ig} + (1 - \pi_i) N_{ig}.$$

We assume $N_{i'g}$ and T_{ig} follow 1) Negative binomial distribution, with overdispersion parameters η_{Tg} and η_{Ng}. or 2) Poisson distribution.

Prior

Noninformative or informative priors

$$\mu_{Ng}, \mu_{Tg} \overset{iid}{\sim} Normal(0, 10^5),$$
$$\eta_{Ng} \overset{iid}{\sim} IG(0.1, 0.1),$$
$$\eta_{Tg} \overset{iid}{\sim} IG(0.1, 0.1),$$
$$\pi_i|\cdot \overset{indep}{\sim} \begin{cases} Beta(a_\pi, b_\pi), \text{no prior knowledge} \\ Beta(a_{\pi_i}, b_{\pi_i}), \text{with prior knowledge} \end{cases}$$

We use the Metropolis algorithm with the random walk proposal distribution.

ISOpure, *Genome Medicine* 2013

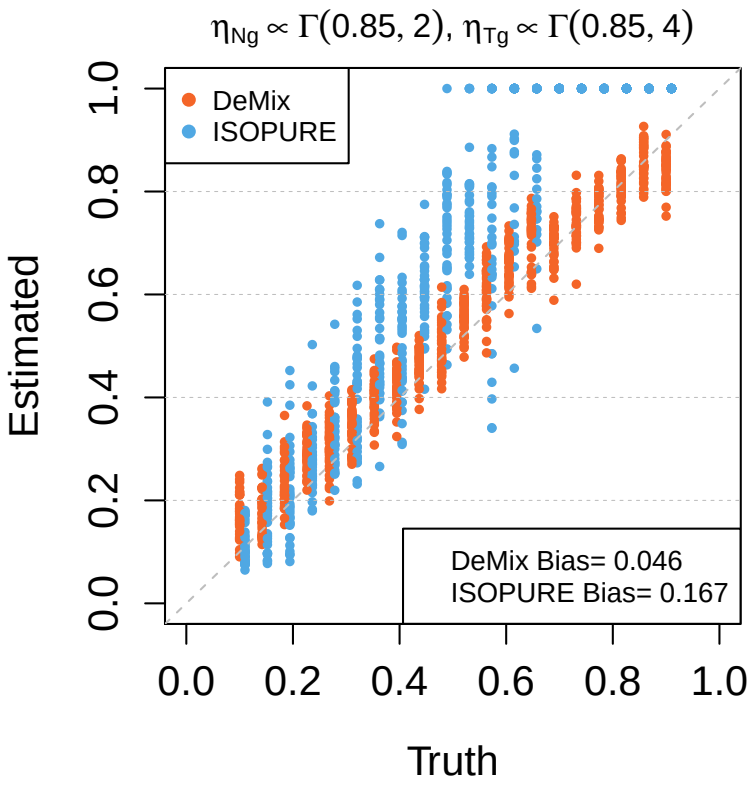
Model

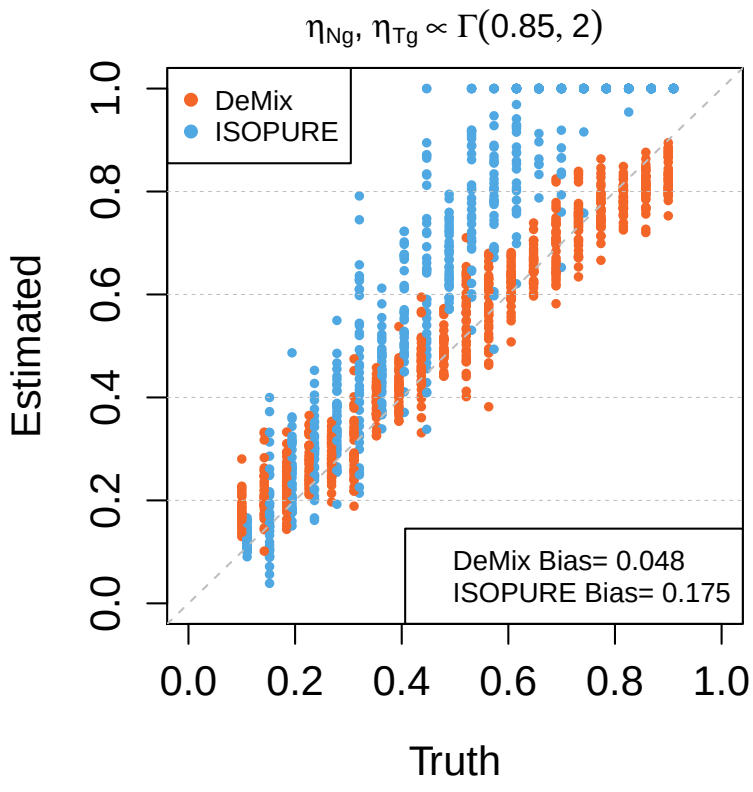
$$t_n = \alpha_n c_n + \sum_{r=1}^{R} \theta_{n,r} b_r + \epsilon_n,$$

where c_n represents the individual tumor expression level, b_r represents a tissue profile, α_n represents tumor proportion and $\theta_{n,r}$ represents proportion of tissue represented by profile b_r.

Simulations - RNA-seq

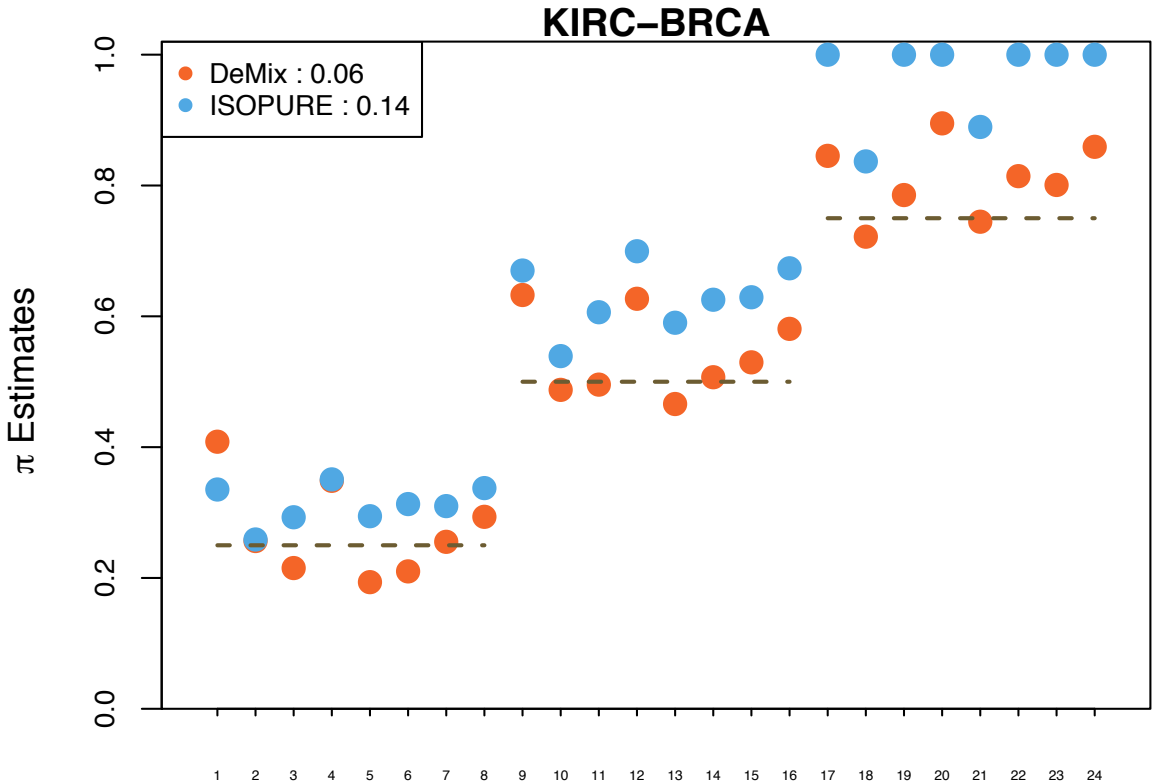
π estimation





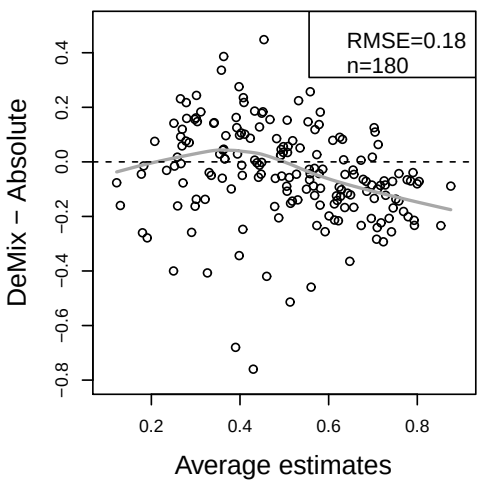
Data example with known truth: π estimation

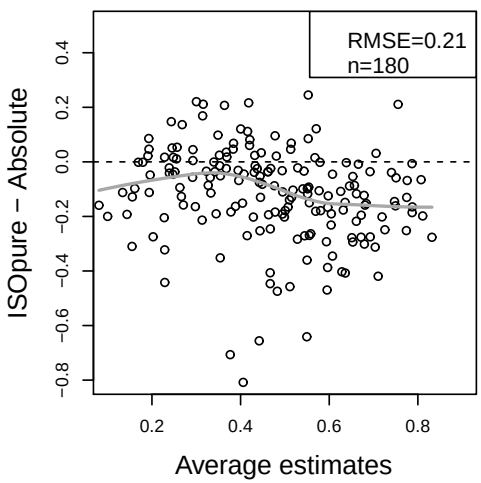
- TCGA RNAseq bam files: 8 samples with normal breast tissues, 8 samples with normal kidney tissues.

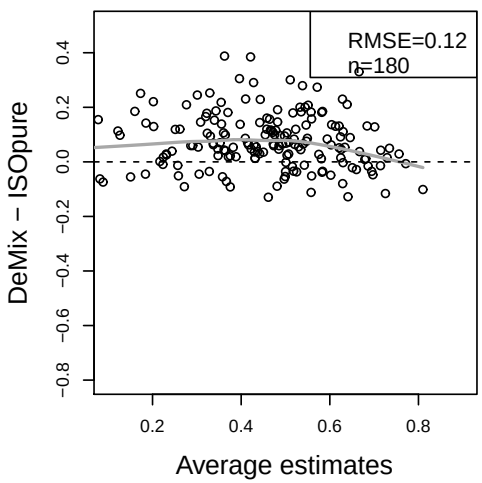


Data example with unknown truth: π estimation

- TCGA on-going study of prostate cancer. Matching DNA samples are available. ABSOLUTE estimates tumor proportions using SNP arrays.

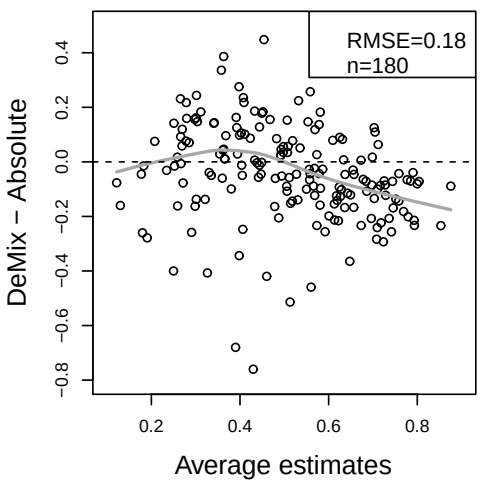


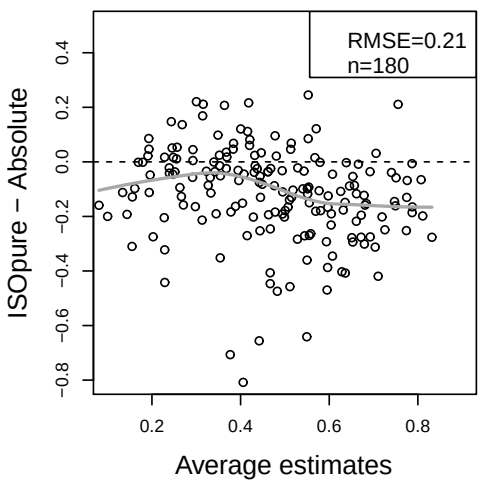


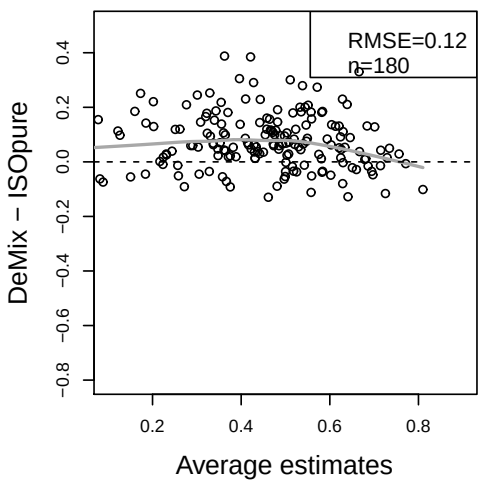


Data example with unknown truth: π estimation

- TCGA on-going study of prostate cancer. Matching DNA samples are available. ABSOLUTE estimates tumor proportions using SNP arrays.

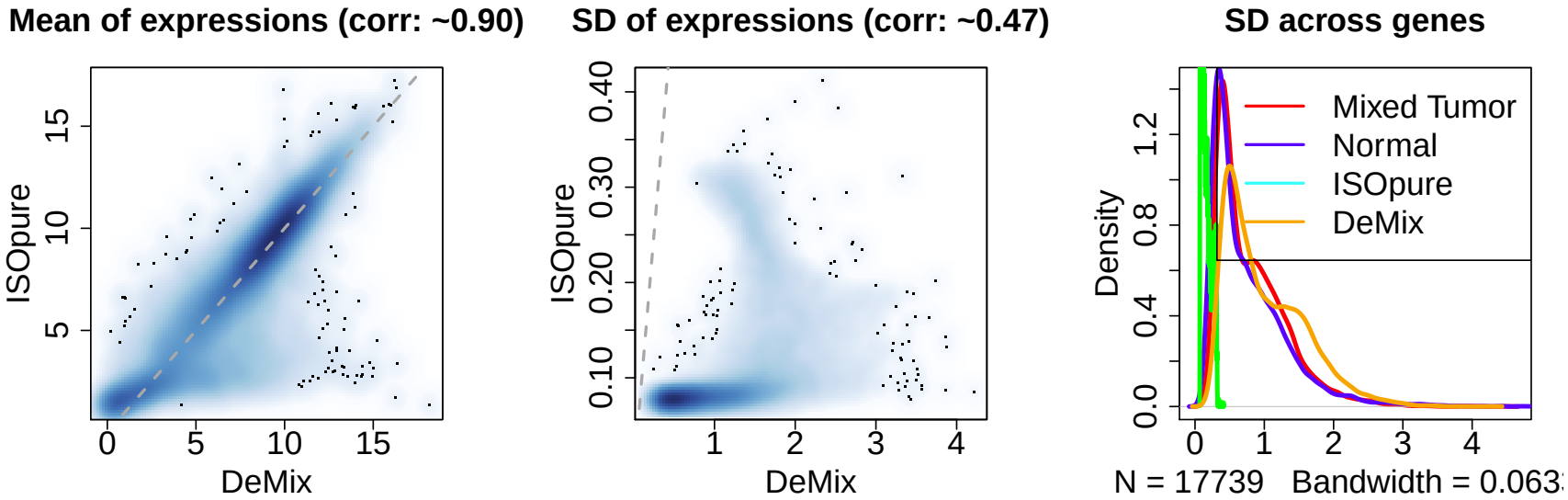






Little correlation between these proportions and the pathologists' estimates.

Data example with unknown truth: deconvolved expressions



Summary

Demix : statistical framework for gene expression deconvolution

- Only one mixture component is needed. Training sets, reference genes, and pathologists' guess are not required.
- Applicable to matched/unmatched sample designs. Individualized deconvolution is available.
- DeMix-Bayes is more flexible to include prior knowledge and provides uncertainty measure.

Summary

Demix : statistical framework for gene expression deconvolution

- Only one mixture component is needed. Training sets, reference genes, and pathologists' guess are not required.
- Applicable to matched/unmatched sample designs. Individualized deconvolution is available.
- DeMix-Bayes is more flexible to include prior knowledge and provides uncertainty measure.

- Noise in low abundance regions, normalization matters.
- Linearity assumption : empirically true. Beware of extreme values.
- Sensitivity of our model to the "normal" samples.

Acknowledgements

- MD Anderson Cancer Center
  - Ying Yuan
  - Yu Fan
  - Zeya Wang
  - Ignacio Wistuba
  - Milind Suraokar
  - John Heymach
- Georgetown University
  - Jaeil Ahn
- Texas A&M
  - Sara Algeri
- Dana Farber Cancer Institute
  - Giovanni Parmigiani
  - Svitlana Tyekucheva
- The Cancer Genome Atlas Project