Bayesian analysis of iTRAQ data with nonrandom missing: Identification of differentially expressed proteins

Ruiyan Luo, Hongyu Zhao

Dec 3, 2008

Outline

iTRAQ

- iTRAQ: isobaric tag for relative and absolute quantitation.
- Compare multiple samples: 4 (or 8) isobaric tags are used to lable peptides.
- Proteins from samples of interests are digested independently prior to labeling.
- The labeled peptides from each sample are then mixed, separated, and studied by MS and MS/MS.

8-plex workflow

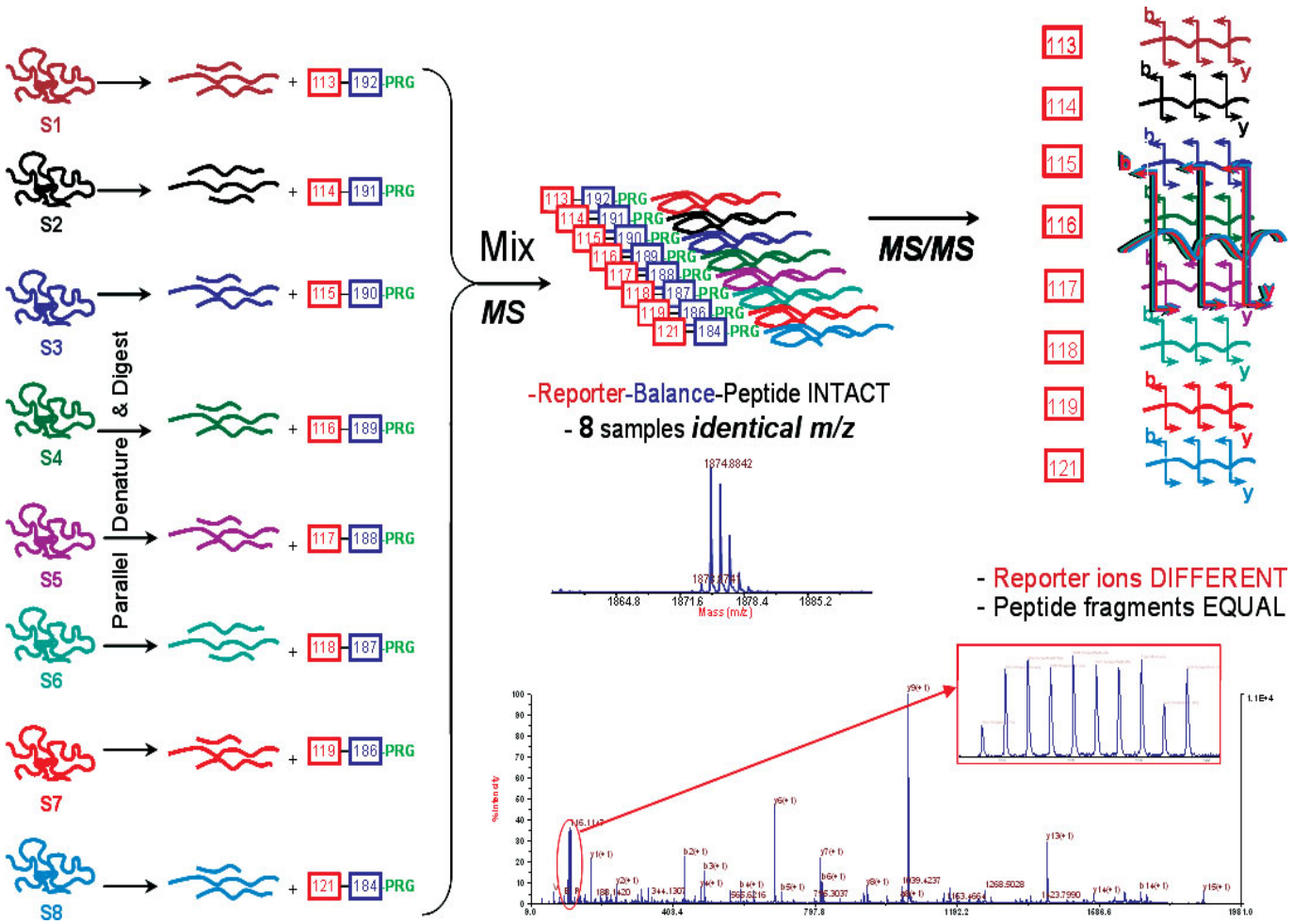


Figure 1. A schematic depiction of the workflow for using 8-plex isobaric reagents in shotgun proteomics studies. The basic workflow is the same as used for 4-plex isobaric reagents.

Choe etal (2007)

Protein identification

- Spectra are scanned against comprehensive protein sequence database.
- Calculate the significance of the match between the observed spectrum and the sequences contained in a database.

Accessions	Peptide Sequence	Area114	Area115	Area116	Area117
IPI00798592.1	ADVVESWIGEK	22.03	29.88	29.08	36.89
IPI00798592.1	ADVVESWIGEK	6.32	6.91	6.8	8.13
IPI00798592.1	ADVVESWIGEK	5.3	3.84	3.66	10.26
IPI00798592.1	DLAALEDKVK	22.2	40.98	49.51	71.88
IPI00798592.1	DLAALEDKVK	33.96	42.87	35.32	44.43
IPI00798592.1	DLTSWVTEMK	16.5	25.53	42.21	36.56
IPI00798592.1	DVDEIEAWISEK	0	7.11	13.6	0
IPI00798592.1	DVDETIGWIK	15.33	32.09	75.23	33.78
IPI00798592.1	DVTGAEALLER				
IPI00798592.1	EAFLNTEDKGDSLDSVEALIK	19.54	28.86	65.76	59.58
IPI00798592.1	EAIVTSEELGQDLEHVEVLQK	9.86	12.72	19.43	6.71
IPI00798592.1	EKEPIVGSTDYGKDEDSAEALLK	39.79	79.8	145.38	145.74
...					

Features of iTRAQ data

1. Hierarchical structure.
2. Missing data.
   - Liu et al (2004): controlled study with 9 technical replicate global proteomic experiments. Only 35.4% of total 1751 proteins were found in every experiment, and 24% found in 1 experiment.
   - Wang et al (2006): the total number of features identified in an experiment decreased over time by 49–73%.
   - The probability of missing data for a protein is not random. It is related to abundance.

Missing data patterns

- Compare three Caveolin-1 knock-out mice with three wild type mice.
- In all three experiments,
  - iTRAQ 114 and 115 label wild type samples.
  - iTRAQ 116 and 117 label knock-out samples.

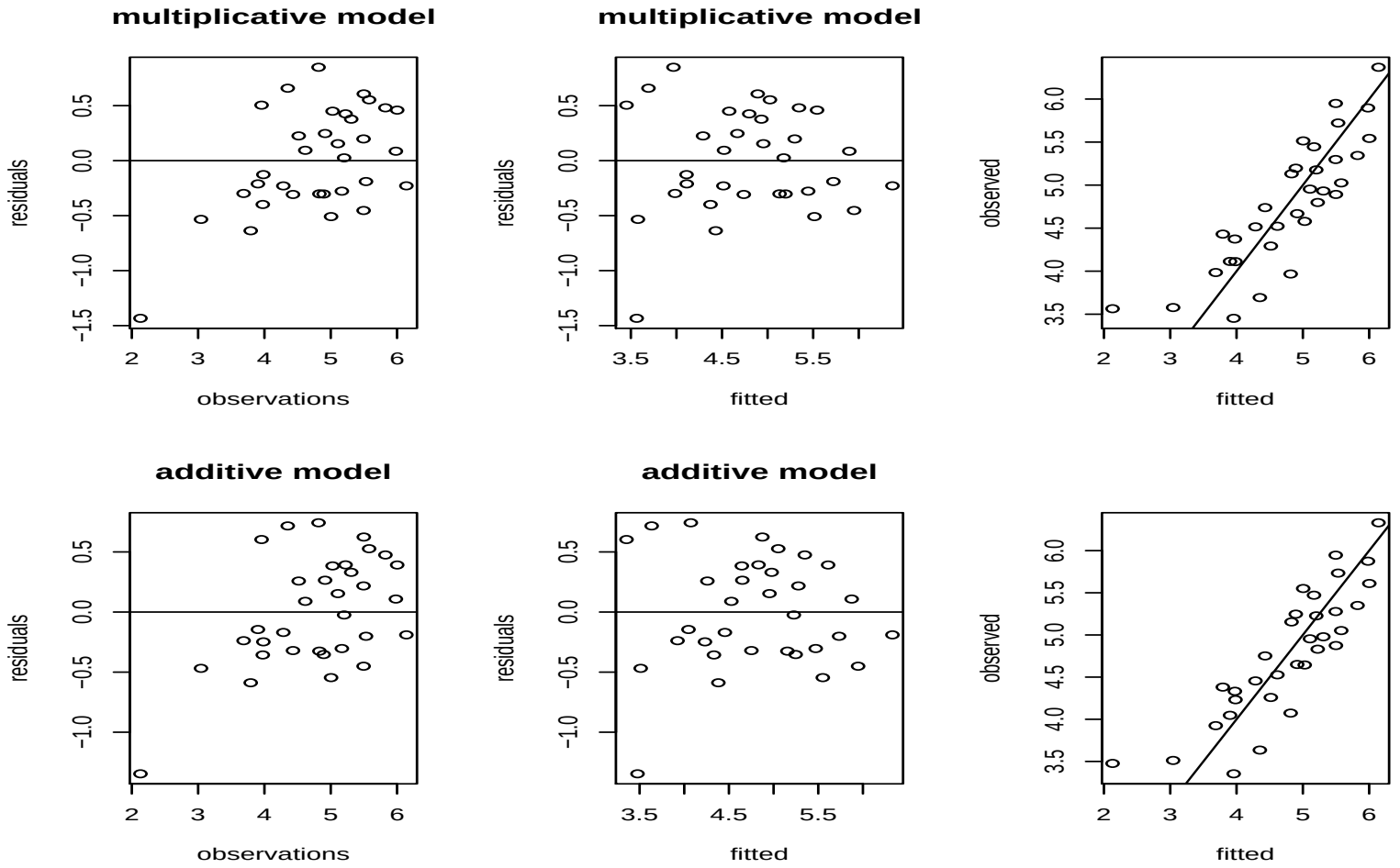
		number of experiments protein/peptide is present		
	counts	1	2	3
proteins	424	192 (45.3%)	94 (22.2%)	138 (32.5%)
peptides	8045	4765 (59.2%)	1156 (14.4%)	2124 (26.4%)

Outline

Preliminary study: additive vs multiplicative

- Protein IPI109044.8, 11 peptides observed in 3 experiments.
- m: marker (sample); j: peptide.
- x_m: protein concentration; z_j: peptide effect; y_{mj}: peptide observation.
- Additive model: $y_{mj} = x_m + z_j + \epsilon_{mj}$.
- Multiplicative model: $y_{mj} = x_m \times z_j + \epsilon_{mj}$.

Preliminary study: multiplicative vs additive



R^2: 0.69 (multiplicative) vs 0.73 (additive)

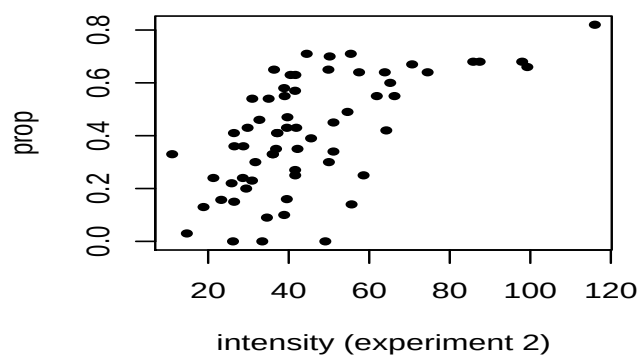
Model: notation

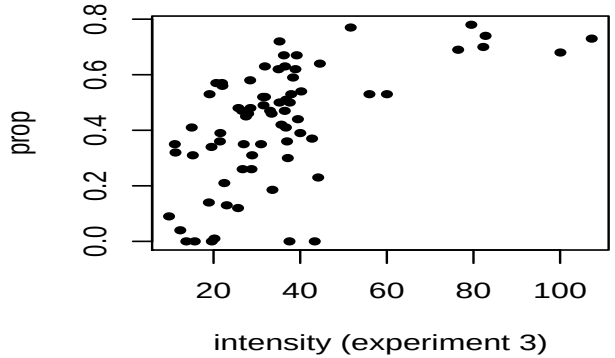
- K: the number of experiments.
- I: the number of proteins.
- J_i: the number of peptides for the ith protein.
- y_{kmijn}: the log-transformed value of the nth measured intensity for the jth peptide of the ith protein in the kth experiment and the mth marker.
- x_{kmi}: the log-transformed expression value of the ith protein in the kth experiment and the mth marker (sample).
- z_{kij}: the peptide effect for the jth peptide of the ith protein in the kth experiment.

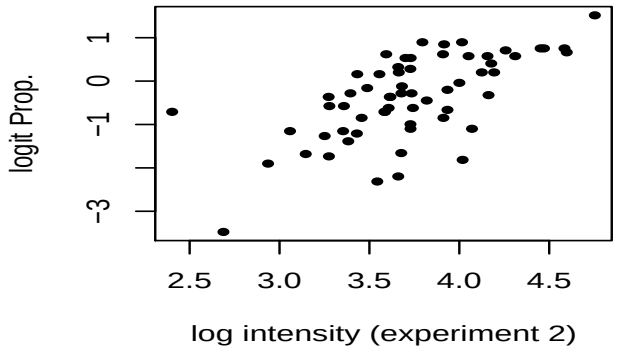
Additive model:

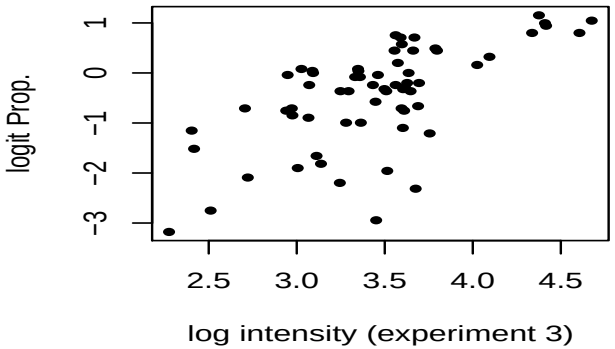
$$y_{kmijn} = x_{kmi} + z_{kij} + \delta_{kmijn}$$

Preliminary study: missing









For peptides observed in one experiment, what proportions of them are observed in other experiments?

Model

- Additive model:

$$y_{kmijn} = x_{kmi} + z_{kij} + \delta_{kmijn} \tag{1}$$

- Missing mechanism:

$$logit(P(I_{kmijn} = 1 | y_{kmijn}, a, b)) = a + b \times y_{kmijn}, \tag{2}$$

where $I_{kmijn} = 1$ indicates that the jth peptide of the ith protein is measured in the kth experiment, the mth sample and nth MS run.

Model: priors

First level of priors:

$$x_{kmi} \sim N(x_{mi}, \sigma_x) \text{ for } m > 1,$$
$$z_{kij} \sim N(z_{ij}, \sigma_z),$$

which leads to an equivalent form of (11):

$$y_{kmijn} = x_{mi} + z_{ij} + e^x_{kmi} + e^z_{kij} + \delta_{kmijn},$$

where e^x_{kmi} and e^z_{kij} denote the random effects across experiments. Restrict $x_{k1i} = x_{1i} = 0$. x_{mi}: the expression of protein i at the mth marker relative to the first marker.

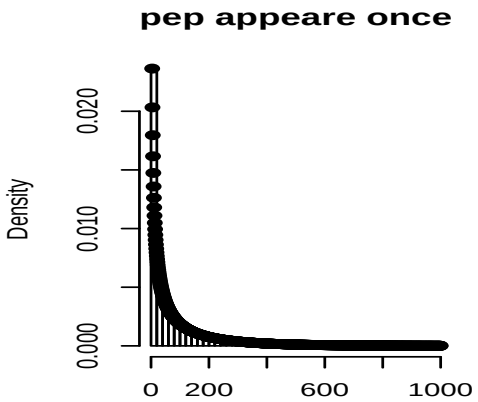
Model: priors

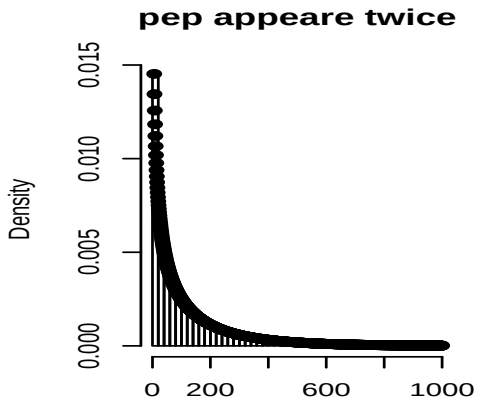
The second level of priors:

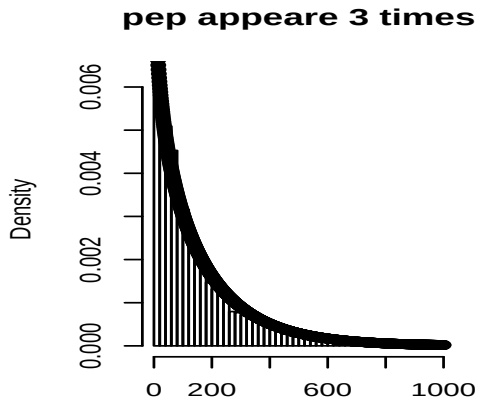
$$x_{mi} \sim N(0, \tau) \text{ for } m > 1,$$
$$z_{ij}|r \sim logGamma(l_r, sh_r, sc_r).$$

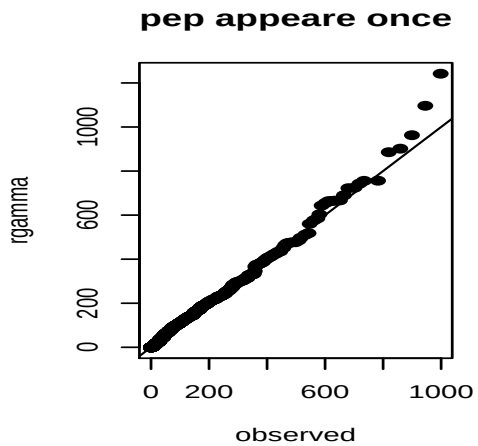
- r: the frequency that the peptide is measured in K experiments
- $\phi = (l_r, sh_r, sc_r)$ denote the location, shape and scale parameters of a logGamma distribution.

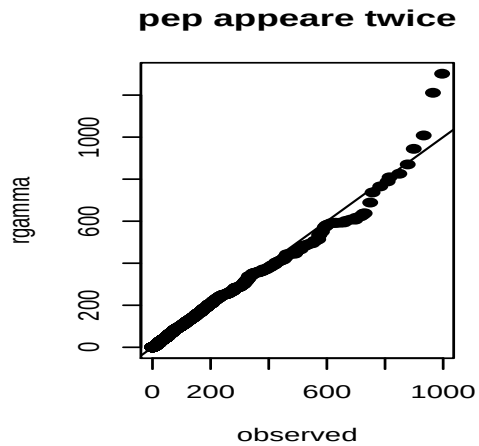
Model: priors

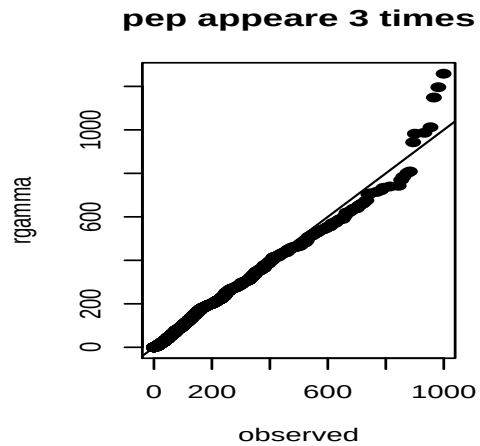












Model

- Priors on a, b, σ_{kmij} and δ_{kmijn}.
- MCMC to simulate posterior of x_{mi} for $m > 1$.

Outline

Simulation Study

We simulate data from a 4-plex version of iTRAQ. Consider one protein with 10 peptides. Let $K = 3$.

1. Specify model parameters a, b, l_r, sh_r, sc_r, σ_x, σ_z, σ_δ and N_{kmij}.
2. For each peptide and each experiment, simulate the presence of the peptide. Then calculate the frequency of presence (r) of the peptide in K experiments.
3. For each peptide, simulate $z_{ij}|r \sim \text{logGamma}(l_r, sh_r, sc_r)$.
4. Simulate $x_{kmi} \sim N(x_{mi}, \sigma_x)$, $z_{kij} \sim N(z_{ij}, \sigma_z)$.
5. Calculate $P(I_{kmijn} = 1)$ and simulate I_{kmijn}. Simulate $y_{kmijn} \sim N(x_{kmi} + z_{kij}, \sigma_\delta)$ when $I_{kmijn} = 1$.

Simulation Study

$x = (0, -0.04, -0.48, -0.66)$, $\sigma_x = 0.01$.

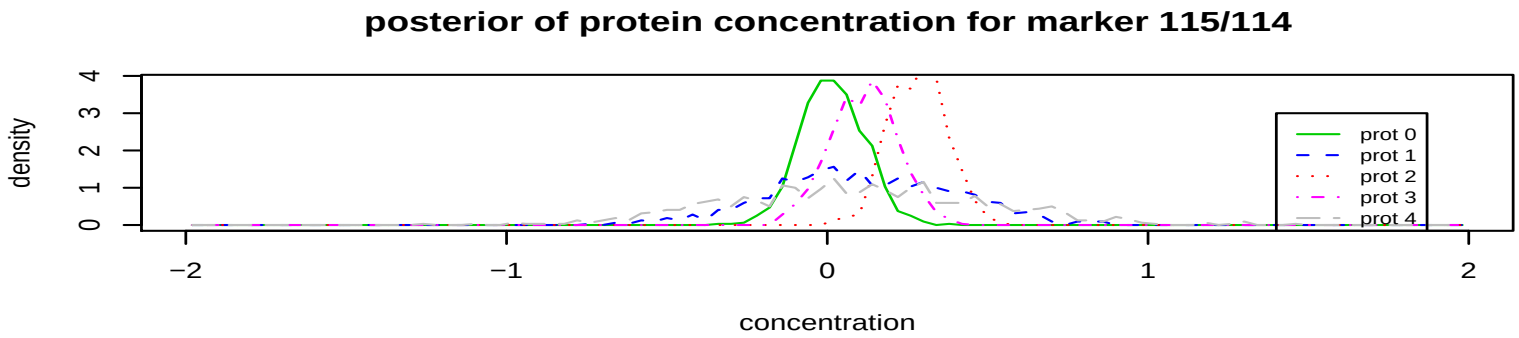
σ_z	σ_δ	method	$log(\frac{S2}{S1})$	$log(\frac{S3}{S1})$	$log(\frac{S4}{S1})$
0.01	0.01	posterior	-0.033 (-0.05,0.01)	-0.465 (-0.48,-0.45)	-0.659 (-0.68,-0.64)
		fold change	-0.033 (-0.06,0.00)	-0.465 (-0.49,-0.44)	-0.660 (-0.70,-0.62)
0.1	0.1	posterior	-0.040 (-0.13,0.04)	-0.467 (-0.55,-0.38)	-0.661 (-0.75,-0.58)
		fold change	-0.039 (-0.30,0.21)	-0.458 (-0.72,-0.16)	-0.663 (-0.99, -0.30)
1	1	posterior	0.058 (-0.19,0.30)	-0.379 (-0.65,-0.13)	-0.562 (-0.84,-0.29)
		fold change	0.079 (-3.03,2.39)	-0.348 (-3.26,2.70)	-0.576 (-3.49,1.83)

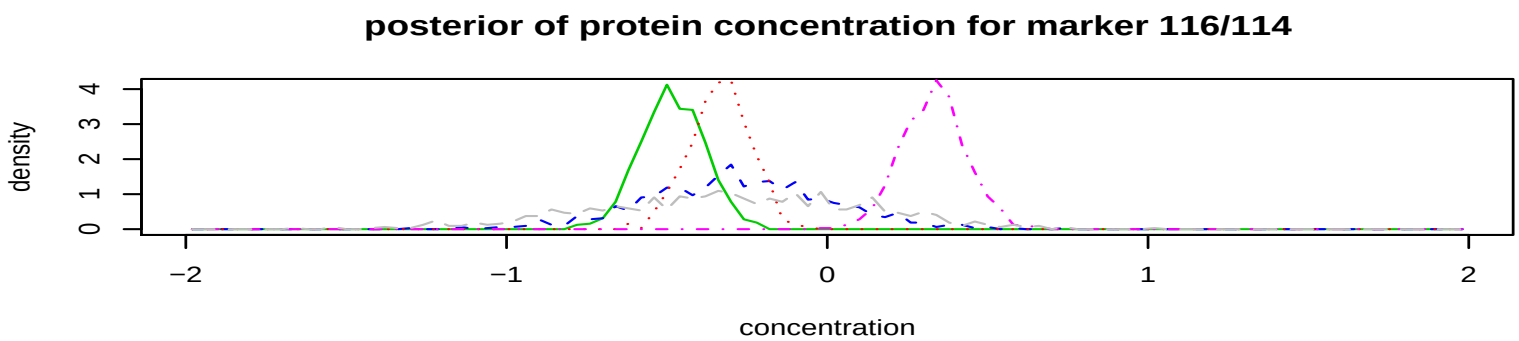
Simulation Study

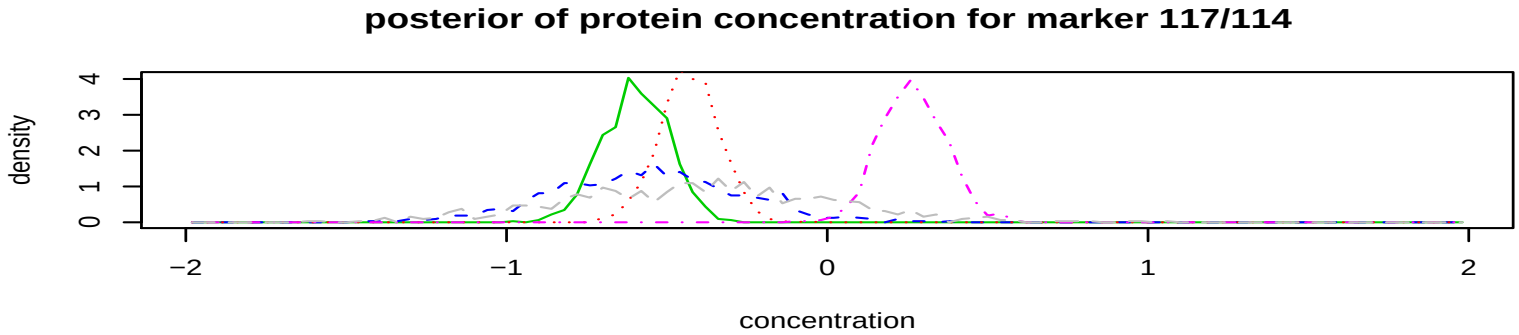
$x = (0, -0.04, -0.48, -0.66)$, $\sigma_x = 0.1$.

σ_z	σ_δ	method	$log(\frac{S2}{S1})$	$log(\frac{S3}{S1})$	$log(\frac{S4}{S1})$
0.1	0.1	posterior	-0.082 (-0.22,0.05)	-0.476 (-0.61,0.33)	-0.689 (-0.83,0.54)
		fold change	-0.082 (-0.44,0.23)	-0.477 (-0.80,-0.16)	-0.698 (-1.01,-0.38)
1	1	posterior	-0.095 (-0.29,0.11)	-0.513 (-0.73,-0.29)	-0.527 (-0.76,-0.33)
		fold change	-0.127 (-3.00,2.71)	-0.561 (-3.04,1.90)	-0.455 (-3.23,2.60)
4	1	posterior	0.104 (-0.18,0.38)	-0.461 (-0.75,-0.19)	-0.640 (-0.91,-0.38)
		fold change	0.117 (-2.78,2.70)	-0.459 (-3.52,2.21)	-0.586 (-3.17,2.00)

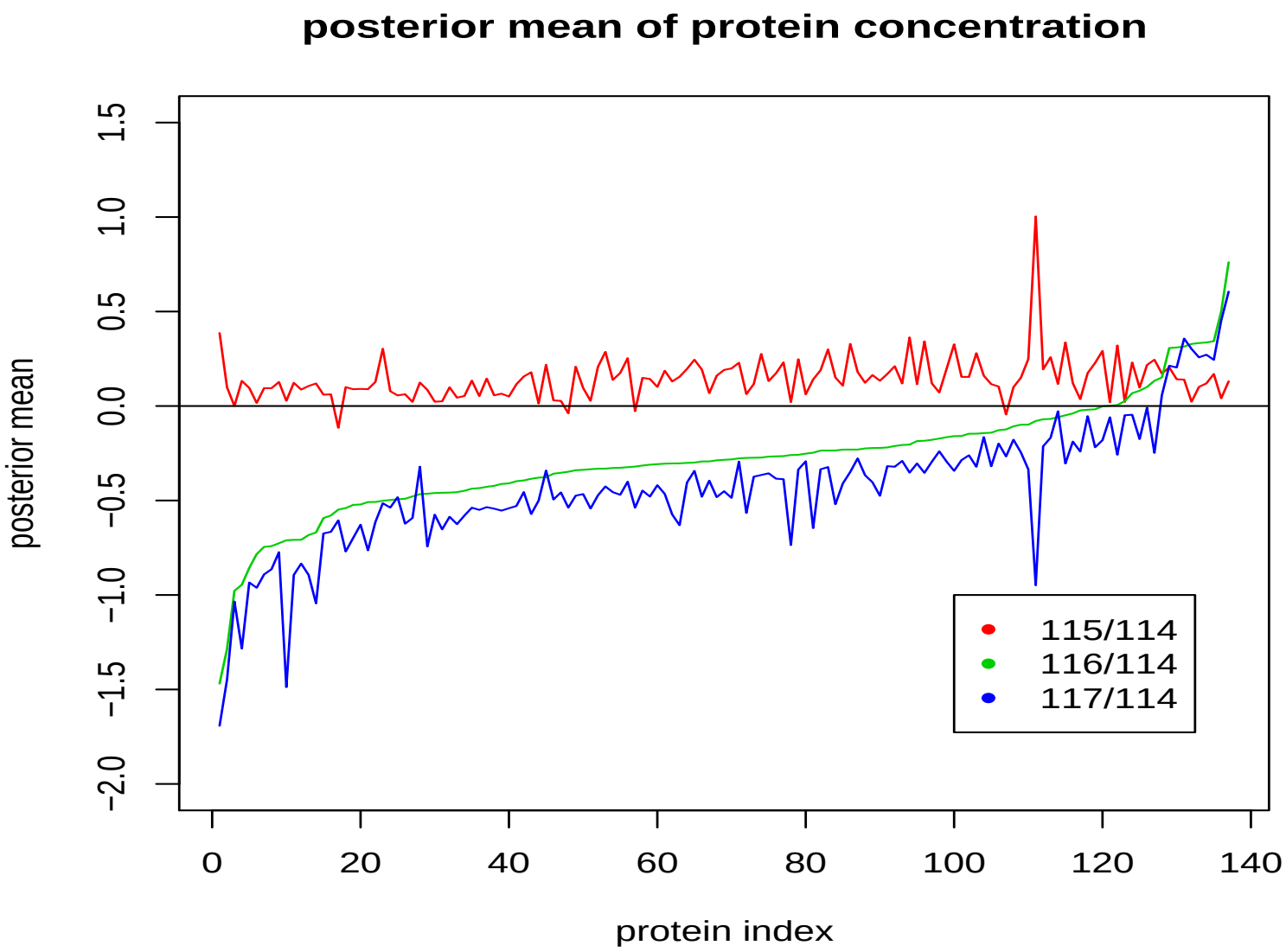
Results I


posterior of protein concentration for marker 115/114
density
0 1 2 3 4
−2 −1 0 1 2
concentration
prot 0
prot 1
prot 2
prot 3
prot 4


posterior of protein concentration for marker 116/114
density
0 1 2 3 4
−2 −1 0 1 2
concentration


posterior of protein concentration for marker 117/114
density
0 1 2 3 4
−2 −1 0 1 2
concentration

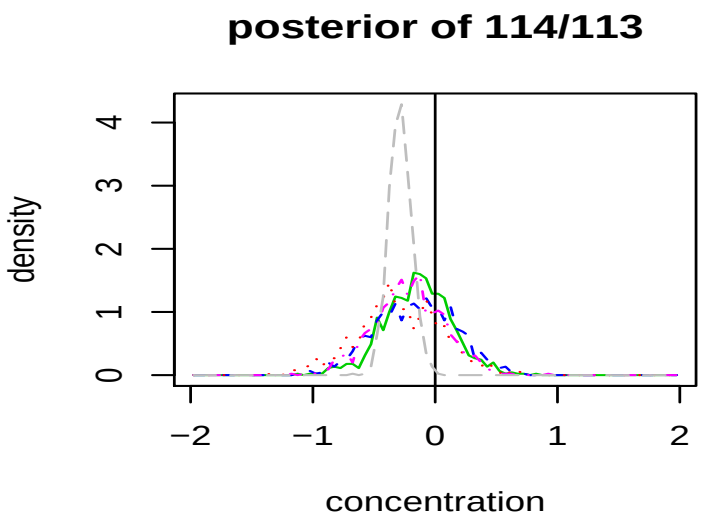
Results I

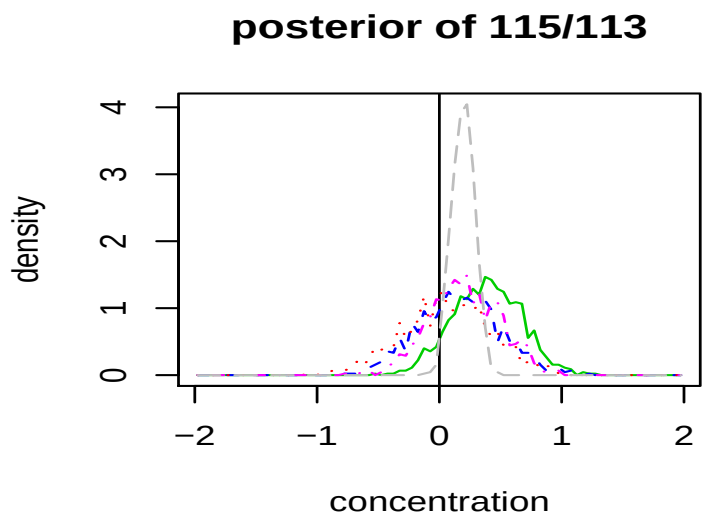

posterior mean of protein concentration
posterior mean
protein index
115/114
116/114
117/114
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
20
40
60
80
100
120
140

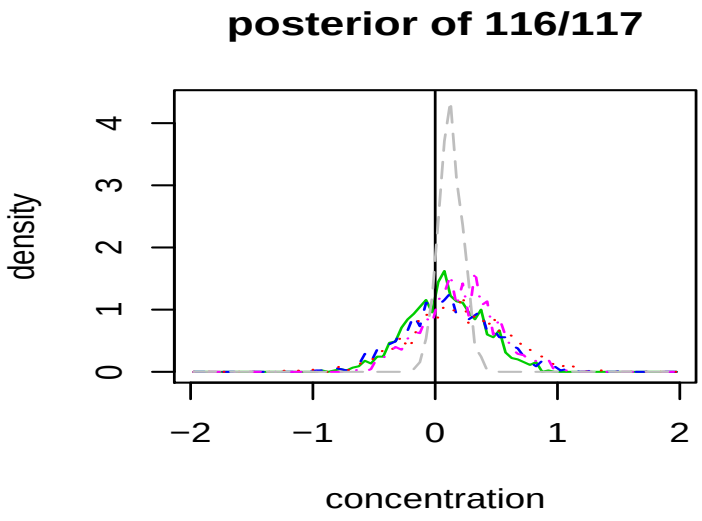
Results II

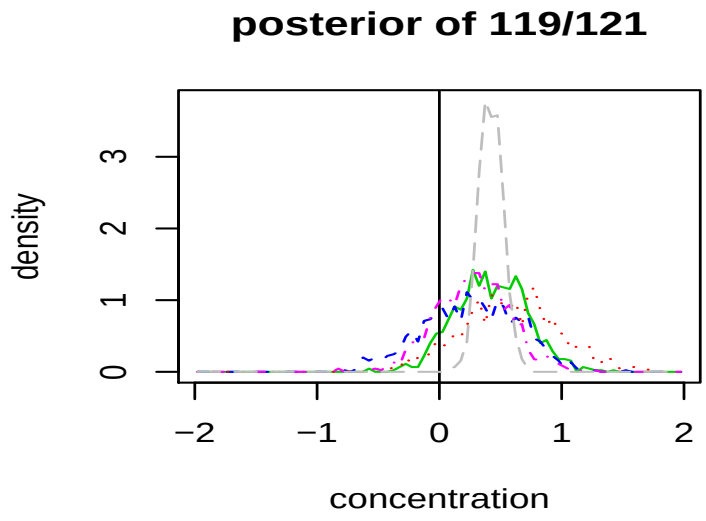
1 experiment (Angus Nairn):
iTRAQ 113, 114, 115 : Cortex
iTRAQ 116, 117, 118: Striatum
iTRAQ 119, 121: Hippocampus

Results II

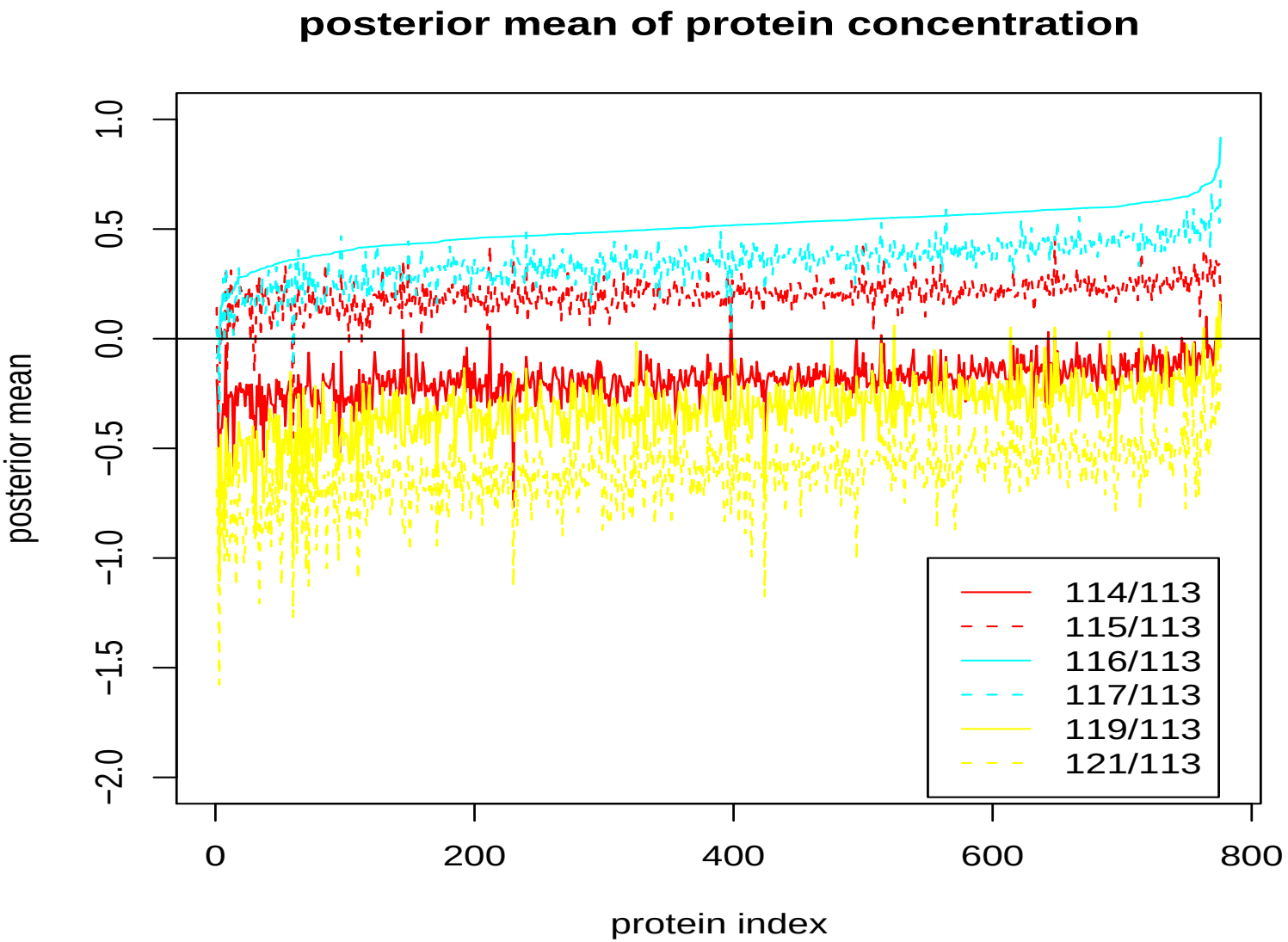

posterior of 114/113
density
0
1
2
3
4
−2
−1
0
1
2
concentration


posterior of 115/113
density
0
1
2
3
4
−2
−1
0
1
2
concentration


posterior of 116/117
density
0
1
2
3
4
−2
−1
0
1
2
concentration


posterior of 119/121
density
0
1
2
3
−2
−1
0
1
2
concentration

Results II


posterior mean of protein concentration
posterior mean
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
200
400
600
800
protein index
114/113
115/113
116/113
117/113
119/113
121/113