PepBay: Implementation of Bayesian inference in the analysis of peptide arrays

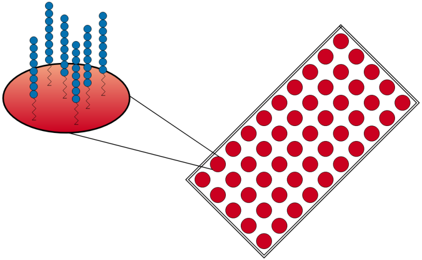
Katarzyna Sidorczuk

University of Wrocław, Faculty of Biotechnology, Department of Bioinformatics and Genomics

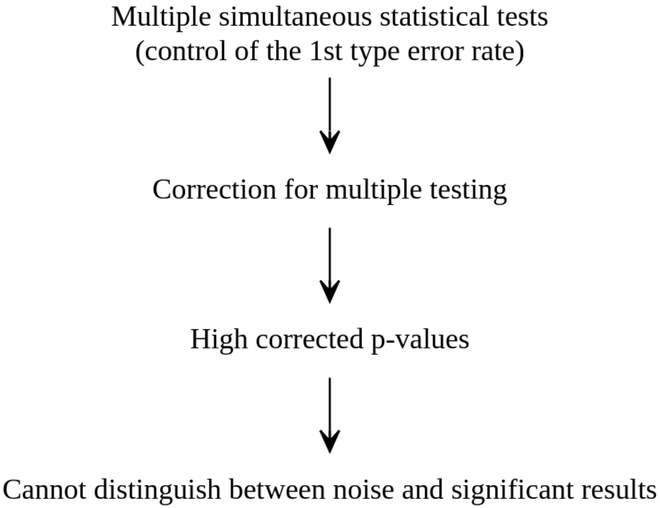
Why R? 2019, Warsaw

What are peptide arrays?

- Collections of short protein fragments;
- Efficient tool for search of new biomarkers;
- Peptide array data:
    - very small sample size (patients),
    - large number of variables (peptides),
    - correlated.



Why traditional methods fail when $p >> n$?



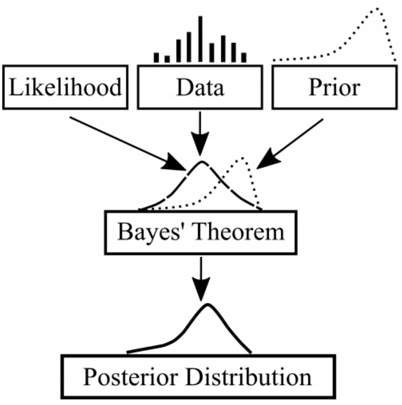
Multiple simultaneous statistical tests
(control of the 1st type error rate)

Correction for multiple testing

High corrected p-values

Cannot distinguish between noise and significant results

Solution: Bayesian inference

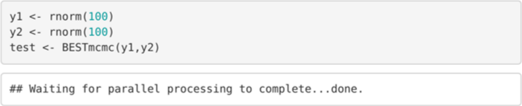


(Doll, J. C. and Jacquemin, S. J. 2018)

Bayes' theorem:

$$P(\theta|X) = \frac{P(X|\theta) \times P\theta}{P(X)}$$

Implementation using package BEST



```
y1 <- rnorm(100)
y2 <- rnorm(100)
test <- BESTmcmc(y1,y2)
```

```
## Waiting for parallel processing to complete...done.
```

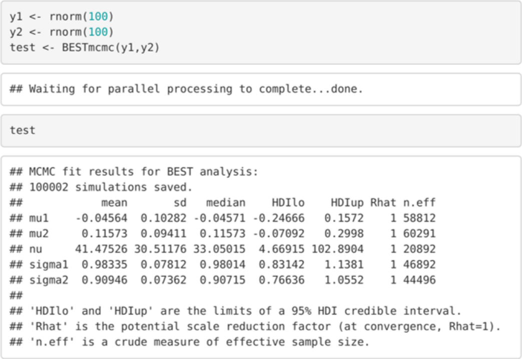
BEST package:

- Based on JAGS;
- Core function: BESTmcmc;
- Convenient wrapper: tidybayes;
- Alternative: rSTAN.

Implementation using package BEST

BEST package:

- Based on JAGS;
- Core function: BESTmcmc;
- Convenient wrapper: tidybayes;
- Alternative: rSTAN.



```
y1 <- rnorm(100)
y2 <- rnorm(100)
test <- BESTmcmc(y1,y2)
```

```
## Waiting for parallel processing to complete...done.
```

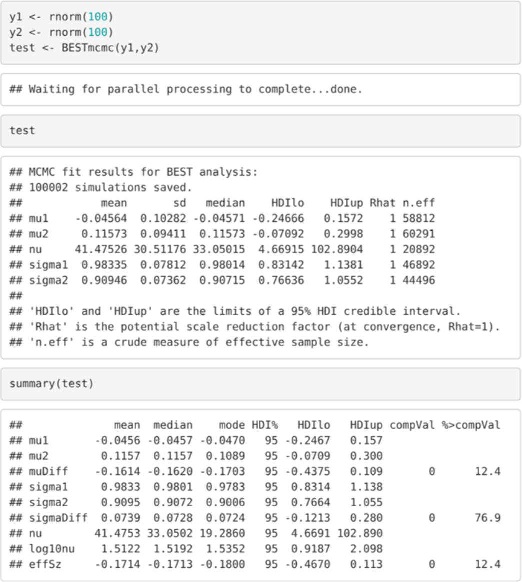
```
test
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##             mean       sd   median    HDIlo    HDIup Rhat n.eff
## mu1     -0.04564  0.10282 -0.04571 -0.24666   0.1572    1 58812
## mu2      0.11573  0.09411  0.11573 -0.07092   0.2998    1 60291
## nu      41.47526 30.51176 33.05015  4.66915 102.8904    1 20892
## sigma1   0.98335  0.07812  0.98014  0.83142   1.1381    1 46892
## sigma2   0.90946  0.07362  0.90715  0.76636   1.0552    1 44496
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## 'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
## 'n.eff' is a crude measure of effective sample size.
```

Implementation using package BEST

BEST package:

- Based on JAGS;
- Core function: BESTmcmc;
- Convenient wrapper: tidybayes;
- Alternative: rSTAN.



```
y1 <- rnorm(100)
y2 <- rnorm(100)
test <- BESTmcmc(y1,y2)
```

```
## Waiting for parallel processing to complete...done.
```

```
test
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##              mean       sd    median     HDIlo    HDIup Rhat n.eff
## mu1      -0.04564  0.10282 -0.04571  -0.24666   0.1572    1 58812
## mu2       0.11573  0.09411  0.11573  -0.07092   0.2998    1 60291
## nu       41.47526 30.51176 33.05015   4.66915 102.8904    1 20892
## sigma1    0.98335  0.07812  0.98014   0.83142   1.1381    1 46892
## sigma2    0.90946  0.07362  0.90715   0.76636   1.0552    1 44496
##
## 'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
## 'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
## 'n.eff' is a crude measure of effective sample size.
```

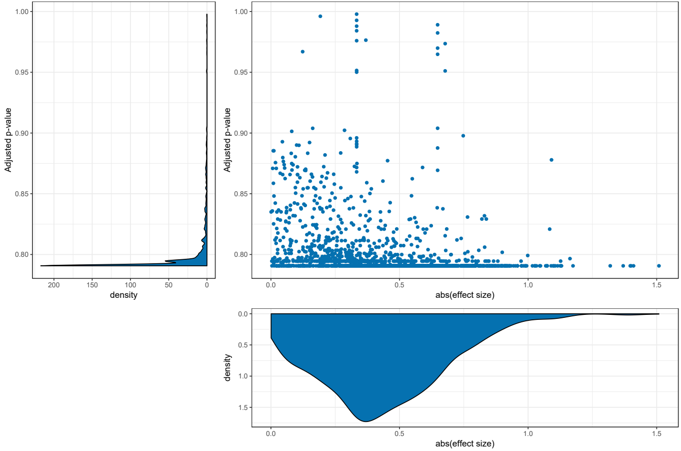
```
summary(test)
```

```
##              mean  median    mode HDI%   HDIlo   HDIup compVal %>compVal
## mu1       -0.0456 -0.0457 -0.0470   95 -0.2467   0.157
## mu2        0.1157  0.1157  0.1089   95 -0.0709   0.300
## muDiff    -0.1614 -0.1620 -0.1703   95 -0.4375   0.109       0      12.4
## sigma1     0.9833  0.9801  0.9783   95  0.8314   1.138
## sigma2     0.9095  0.9072  0.9006   95  0.7664   1.055
## sigmaDiff  0.0739  0.0728  0.0724   95 -0.1213   0.280       0      76.9
## nu        41.4753 33.0502 19.2860   95  4.6691 102.890
## log10nu    1.5122  1.5192  1.5352   95  0.9187   2.098
## effSz     -0.1714 -0.1713 -0.1800   95 -0.4670   0.113       0      12.4
```

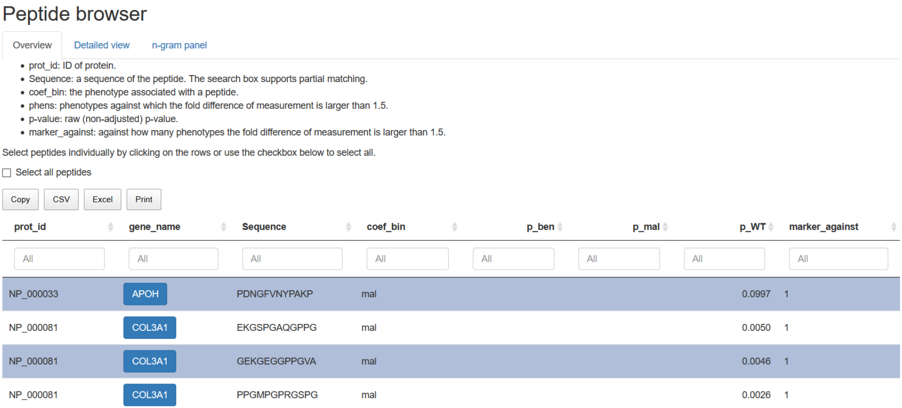
Bayesian inference vs. frequentist methods

Advantages of Bayesian inference:

- complete distributions of reliable values;
- effect size instead of p-value.



PepBay app screenshots



Peptide browser

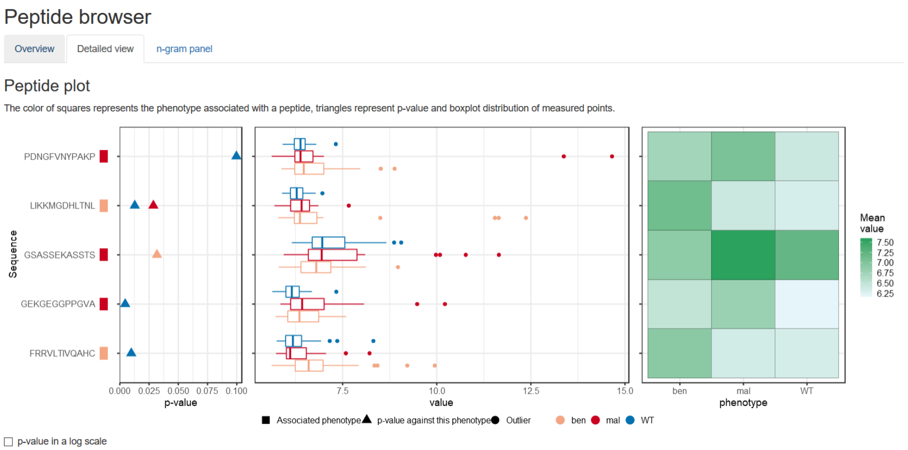
- prot_id: ID of protein.
- Sequence: a sequence of the peptide. The seearch box supports partial matching.
- coef_bin: the phenotype associated with a peptide.
- phens: phenotypes against which the fold difference of measurement is larger than 1.5.
- p-value: raw (non-adjusted) p-value.
- marker_against: against how many phenotypes the fold difference of measurement is larger than 1.5.

Select peptides individually by clicking on the rows or use the checkbox below to select all.

☐ Select all peptides

prot_id	gene_name	Sequence	coef_bin	p_ben	p_mal	p_WT	marker_against
All	All	All	All	All	All	All	All
NP_000033	APOH	PDNGFVNYPAKP	mal			0.0997	1
NP_000081	COL3A1	EKGSPGAQGPPG	mal			0.0050	1
NP_000081	COL3A1	GEKGEGGPPGVA	mal			0.0046	1
NP_000081	COL3A1	PPGMPGPRGSPG	mal			0.0026	1

PepBay app screenshots



Peptide browser

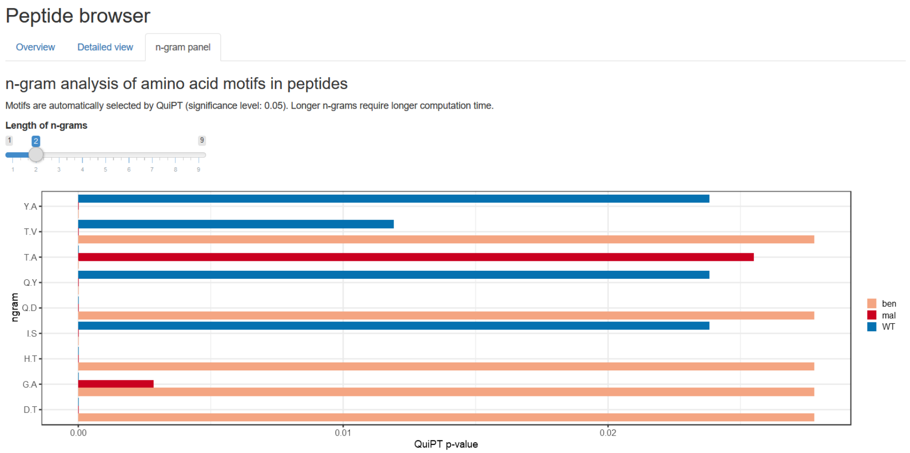
Overview | Detailed view | n-gram panel

Peptide plot

The color of squares represents the phenotype associated with a peptide, triangles represent p-value and boxplot distribution of measured points.

□ p-value in a log scale

PepBay app screenshots



Peptide browser

n-gram analysis of amino acid motifs in peptides

Motifs are automatically selected by QuiPT (significance level: 0.05). Longer n-grams require longer computation time.

Length of n-grams: 2

Acknowledgments

- Andreas Weinhäusel
  Austrian Instutute of Technology, Department of Molecular Diagnostics

- Michał Burdukiewicz
  Warsaw University of Technology, Faculty of Mathematics and Information Science