Risk-prediction modelling in cancer with multiple genomic data sets: a Bayesian variable selection approach

Manuela Zucknick

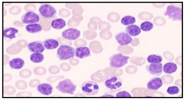
Division of Biostatistics, German Cancer Research Center

Biometry Workshop, Freising, 07 November 2013

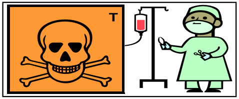
1. Introduction
2. Integration of gene expression and copy number variations in a logistic Bayesian variable selection (BVS) model
3. Proof-of-principle example: Glioblastoma
4. Simulation study: Influence of prior specifications
5. Application: medulloblastoma
6. Summary

From targeted therapy to personalised medicine

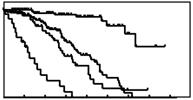
one diagnosis



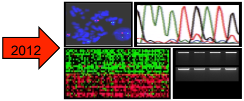
uniform therapy



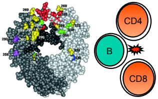
variable results



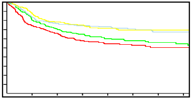
molecular subtypes



targeted therapy

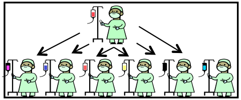


better results

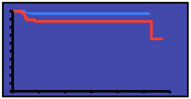


individual profile

personal therapy



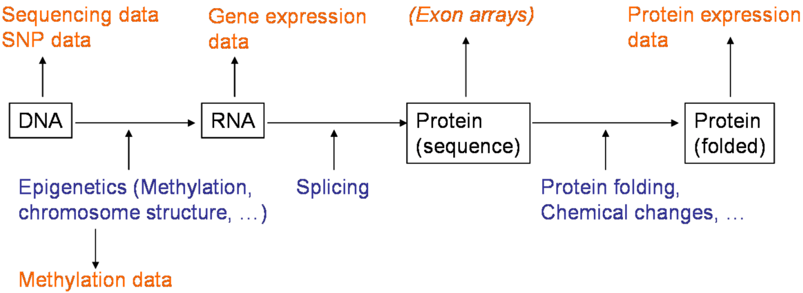
optimal results



Slide by S. Pfister (University Hospital Heidelberg, Germany)

Molecular data sources

Genomics → **Transcriptomics** → **Proteomics**



Epigenomics

Goal

Develop risk prediction models based on 'omics' data

- Prediction of clinical endpoints (therapy response, survival)
- with simultaneous selection of biomarkers
- combining several high-dimensional input 'omics' data sets.

Patient-based genome-wide data from several sources

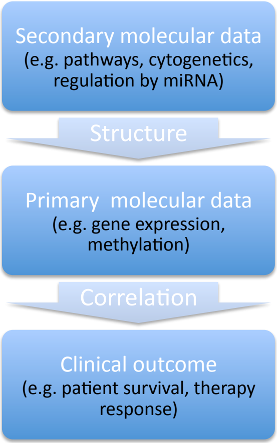
- **Transcriptomics:** gene expression
- **Epigenomics:** CpG methylation
- **Genomics:** copy number variation, SNPs/point mutations

Analyse data in a (model-based) integrative manner for

- a more comprehensive picture of the disease biology,
- improved performance of risk prediction models.

Possible approaches to data integration

Hierarchical approach



Secondary molecular data
(e.g. pathways, cytogenetics, regulation by miRNA)

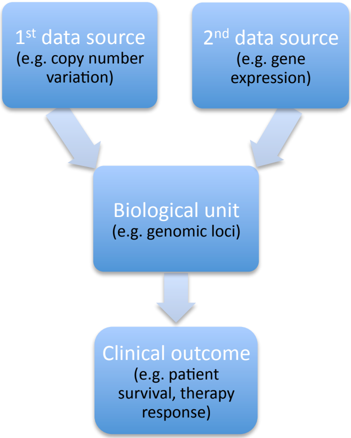
Structure

Primary molecular data
(e.g. gene expression, methylation)

Correlation

Clinical outcome
(e.g. patient survival, therapy response)

Same-level approach



1st data source
(e.g. copy number variation)

2nd data source
(e.g. gene expression)

Biological unit
(e.g. genomic loci)

Clinical outcome
(e.g. patient survival, therapy response)

Bayesian hierarchical model for variable selection (BVS)

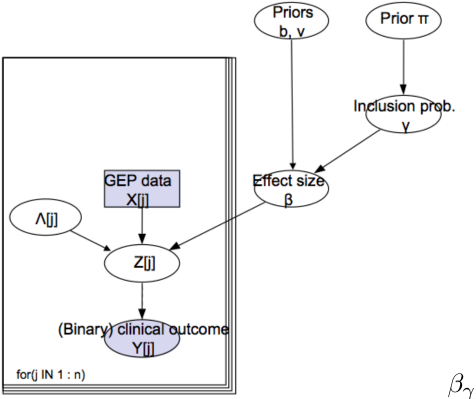
BVS model with indicator variable $\gamma_i = \begin{cases} 1 & \text{, i is included} \\ 0 & \text{, i is excluded} \end{cases}$

- The model space becomes huge, of size 2^p (when no interactions included) and full exploration is unfeasible
- For high-dimensional data ($p >> n$) many alternative models having similar explanatory power
- $\rightarrow$ Use of MCMC methods as stochastic search algorithms
- We favour sparse solutions via prior distribution for model size.

Frequentist alternatives:

- Penalised regression (lasso etc.), boosting,...

(1) Logistic BVS model based on gene expression alone



$$
\begin{aligned}
Y_j &= \begin{cases} 1 & \text{if } Z_j > 0 \\ 0 & \text{otherwise} \end{cases} \\
Z_j &= X_{\gamma j}\beta_\gamma + \epsilon_j \\
\epsilon_j &\sim N(0, \lambda_j) \\
\lambda_j &= (2\phi_j)^2 \\
\phi_j &\sim \text{Kolmogorov-Smirnov, i.i.d.} \\
\gamma &\sim p(\gamma) = \prod_{i=1}^{p} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i} \\
\beta_{\gamma=1} &\sim N(b_\gamma = 0, v_\gamma = I_{p_\gamma})
\end{aligned}
$$

Auxiliary variable representation for normal scale mixture distribution resulting in exact logistic regression model.

Holmes & Held (2006), Zucknick (2009)

(2) Incorporate copy number variations (CNV)

Modify the prior on the model space $p(\gamma)$

$$p(\gamma) = \prod_{i=1}^{p} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}$$

→ Assign prior individual variable inclusion probabilities π_i using information on association between CNV distribution and model endpoint Y.

Assumptions

- Genes in deleted regions will not be expressed. Equivalently, genes in amplified regions might have higher expression. →
- Genes in regions with differential copy numbers get larger inclusion probability π.

Prior specifications for π

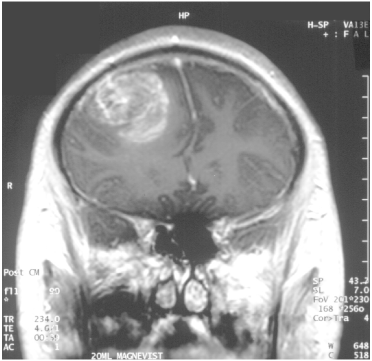
$$\pi_i \propto \min(1, \pi_0(1 + C \times f_{\text{dist}}(\text{CNV}, Y)))$$

- Base prior variable inclusion probability π_0
- Factor C
- Distance metrics $f_{\text{dist}}(\text{CNV}, Y)$ for ordinal distributions, e.g. *loss* $<$ *normal* $<$ *gain*:
    - "Modal states distance" (MOD): For a sample from CNV distribution, compute the modes for both classes (mode_0 and mode_1), then: $f_{\text{dist}}(\text{CNV}, Y) = 0.5 * |\text{mode}_1 - \text{mode}_0|$
    - Earth mover's distance (EMD) (Rubner et al., IJCV 2000): minimal cost that must be paid to transform one distribution into the other (moving within order 'loss' ↔ 'normal' ↔ 'gain')

Glioblastoma

Today, the number of children dying from brain tumours is similar to child lymphoma deaths, even though incidence of brain tumours is only half as high. → Better (targeted) treatment strategies needed!

- **Known prognostic factors** in glioblastoma:
  - Loss of chromosome arm 10q
  - Mutation in the IDH1 gene or in the H3.3 histone
- **Which genes** are associated with loss of chromosome 10q or with H3.3 mutations?



Wikipedia.org

Glioblastoma data

Data:

- 40 tumour samples with
  - gene expression (GE) array data (Agilent) and
  - copy number variation estimated from Illumina 450K arrays.
- $p = 2000$ top variable GE probes and corresponding CNV data
- Endpoint: loss of chromosome arm 10q versus no loss

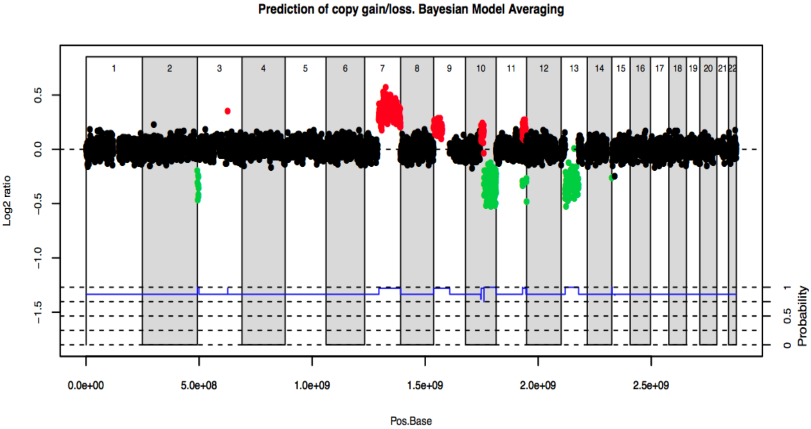
Prior specifications:

- C=100, f_{dist} = modal states distance and $\pi_0 = 5/p$

MCMC setting:

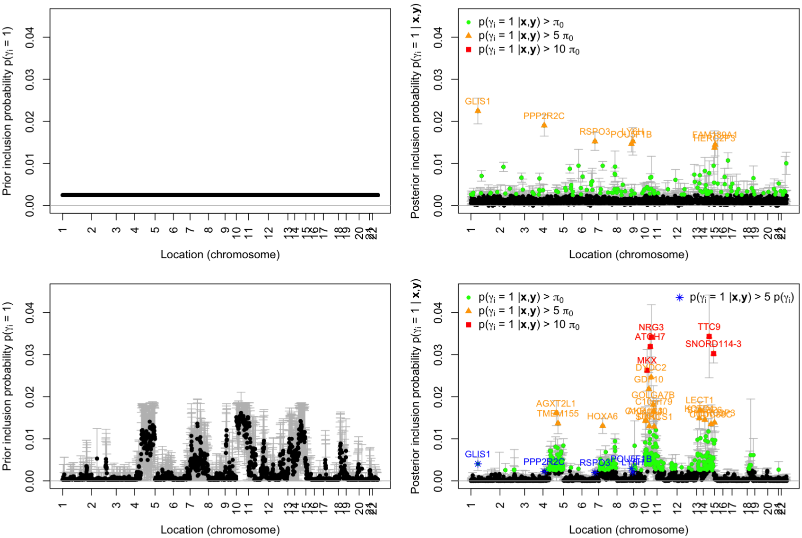
- $B = 10$ Markov chains from different starting points with
- $K = 100,000$ iterations ($10,000$ burn-in iterations discarded)

CNV data in typical sample with known loss of 10q



R package RJaCGH (Rueda and Diaz-Uriarte, 2007)

Glioblastoma: prior and posterior inclusion probabilities



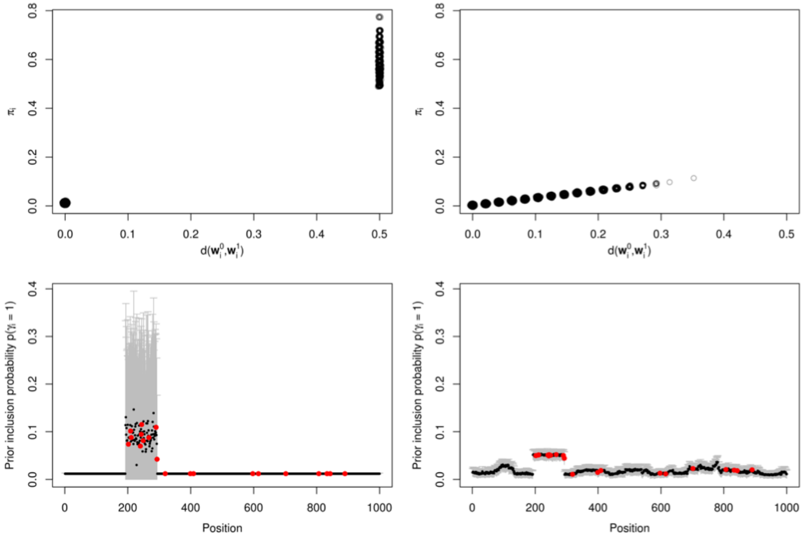
Simulate training and test data sets with:

- $n = 50$ samples with binary response (tumour versus normal)
- $p = 1,000$ variables (genes/genomic regions)
- Correlation $0.5^{|i_1 - i_2|}$ between two variables with IDs i_1 and i_2

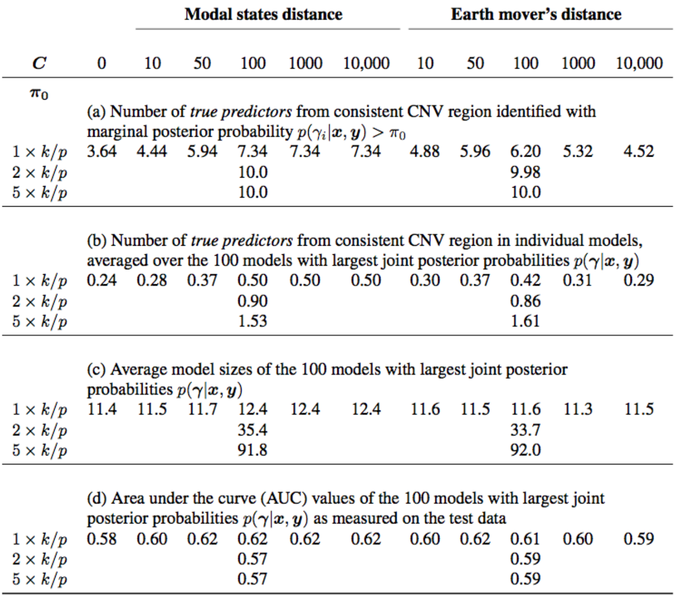
- Generate one random CNV region per sample (loss or gain)
- Generate consistent CNV region in 50% of all tumour samples (all gain)
- Add (gain) or subtract (loss) $\log_2(2)$ to $\log_2$ gene expression, if gene is expressed
- $p^* = 20$ variables are related to response y (true model) via logistic link with effect sizes β ($\rightarrow$ Prior $\pi_0 \propto \frac{20}{p}$)
- Add measurement noise: $\tilde{w}_{ij} = w_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, 0.25)$

- 10 genes from consistent CNV region with $\exp(\beta^*) = (1.5, 1.5, 2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.5, 3.5)$
- 10 genes from outside with $\exp(\beta^*)$

Simulation results: $f_{\text{dist}}(\text{CNV}, Y)$ versus π

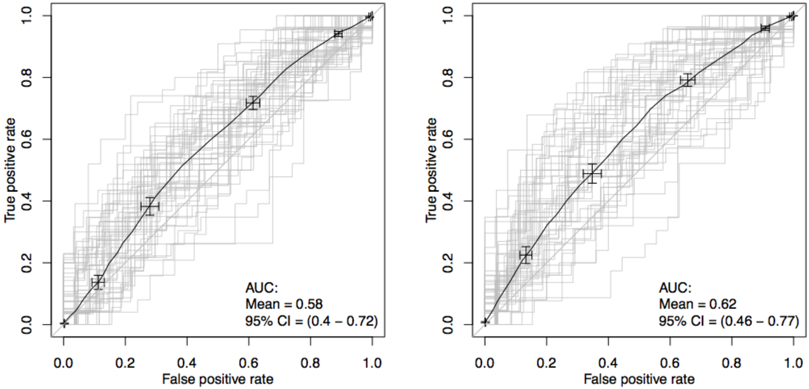


Simulation results: averages across 50 simulation runs



	Modal states distance						Earth mover's distance				
C	0	10	50	100	1000	10,000	10	50	100	1000	10,000
π_0											
	(a) Number of *true predictors* from consistent CNV region identified with marginal posterior probability $p(\gamma_i \mid \boldsymbol{x}, \boldsymbol{y}) > \pi_0$										
$1 \times k/p$	3.64	4.44	5.94	7.34	7.34	7.34	4.88	5.96	6.20	5.32	4.52
$2 \times k/p$				10.0					9.98		
$5 \times k/p$				10.0					10.0		
	(b) Number of *true predictors* from consistent CNV region in individual models, averaged over the 100 models with largest joint posterior probabilities $p(\boldsymbol{\gamma} \mid \boldsymbol{x}, \boldsymbol{y})$										
$1 \times k/p$	0.24	0.28	0.37	0.50	0.50	0.50	0.30	0.37	0.42	0.31	0.29
$2 \times k/p$				0.90					0.86		
$5 \times k/p$				1.53					1.61		
	(c) Average model sizes of the 100 models with largest joint posterior probabilities $p(\boldsymbol{\gamma} \mid \boldsymbol{x}, \boldsymbol{y})$										
$1 \times k/p$	11.4	11.5	11.7	12.4	12.4	12.4	11.6	11.5	11.6	11.3	11.5
$2 \times k/p$				35.4					33.7		
$5 \times k/p$				91.8					92.0		
	(d) Area under the curve (AUC) values of the 100 models with largest joint posterior probabilities $p(\boldsymbol{\gamma} \mid \boldsymbol{x}, \boldsymbol{y})$ as measured on the test data										
$1 \times k/p$	0.58	0.60	0.62	0.62	0.62	0.62	0.60	0.62	0.61	0.60	0.59
$2 \times k/p$				0.57					0.59		
$5 \times k/p$				0.57					0.59		

Simulation results: test data ROC curves of BMA* results



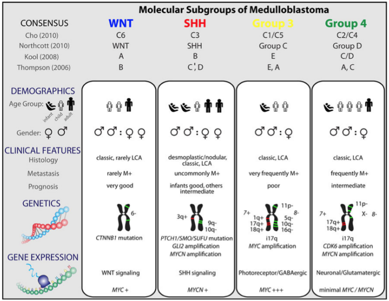
*Bayesian model averaging (BMA) of the 100 models with largest joint posterior probabilities

Simulation conclusions

- True predictors have high marginal probability $(p(\gamma_i = 1|D) > \pi_0)$.
- But no highest probability model contains all true predictors.
- → Bayesian model averaging (BMA) is important.
- Also important in order to “catch” all important variables, including correlated ones.
- Modest improvement in prediction accuracy.

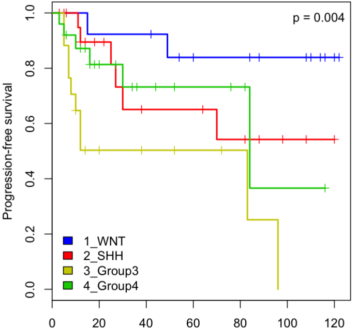
Medulloblastoma (Northcott et al., Nature 2012)

- Most common malignant brain tumours in children
- Current treatment: nonspecific cytotoxic therapy and surgery
- Four known molecular subgroups **WNT**, **SHH**, **Group 3**, **Group 4**, but currently no subgroup-specific targets for targeted therapy
- Known survival differences between subgroups, but not many known individual prognostic factors (ex: CTNNB1 mut in **WNT** group)



Molecular Subgroups of Medulloblastoma

CONSENSUS	WNT	SHH	Group 3	Group 4
Cho (2010)	C6	C3	C1/C5	C2/C4
Northcott (2010)	WNT	SHH	Group C	Group D
Kool (2008)	A	B	E	C/D
Thompson (2006)	B	C, D	E, A	A, C
DEMOGRAPHICS				
Age Group: infant child adult				
Gender: ♀ ♂	♂♂:♀♀	♂♂:♀♀	♂♂:♀	♂♂:♀
CLINICAL FEATURES				
Histology	classic, rarely LCA	desmoplastic/nodular, classic, LCA	classic, LCA	classic, LCA
Metastasis	rarely M+	uncommonly M+	very frequently M+	frequently M+
Prognosis	very good	infants good, others intermediate	poor	intermediate
GENETICS	6- *CTNNB1* mutation	3q+ 9q- 10q- *PTCH1/SMO/SUFU* mutation *GLI2* amplification *MYCN* amplification	7+ 1q+ 17q+ 18q+ 11p- 8- 5q- 10q- 16q- i17q *MYC* amplification	7+ 17q+ 18q+ 11p- X- 8- i17q *CDK6* amplification *MYCN* amplification
GENE EXPRESSION	WNT signaling *MYC* +	SHH signaling *MYCN* +	Photoreceptor/GABAergic *MYC* +++	Neuronal/Glutamatergic minimal *MYC* / *MYCN*



Medulloblastoma data set

Data:

- 55 training samples and 44 test samples (without CNV data)
  - gene expression (GE) array data (Affymetrix U133plus2) and
  - copy number variation estimated from Illumina 450K arrays.
- $p = 5000$ top variable GE probes (including 44 putative driver genes, Northcott et al., 2012) and corresponding CNV data
- Endpoint: disease progression (recurrence or death) three years after diagnosis

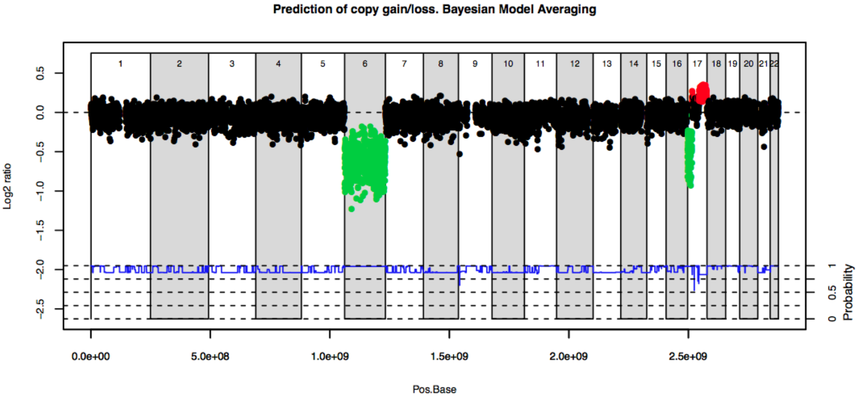
Prior specifications:

- C=100, f_{dist} = modal states distance and $\pi_0 = 10/p$

MCMC setting:

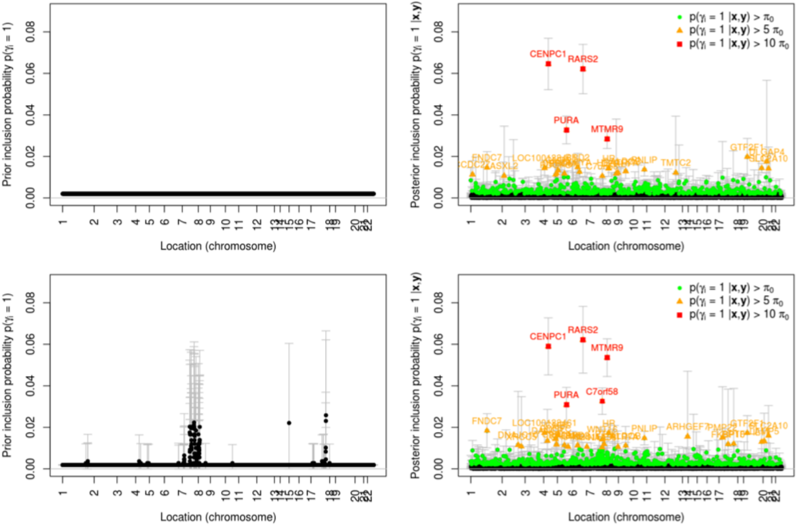
- $B = 5$ Markov chains with
- $K = 100,000$ iterations each

CNV data in typical **WNT** subgroup sample

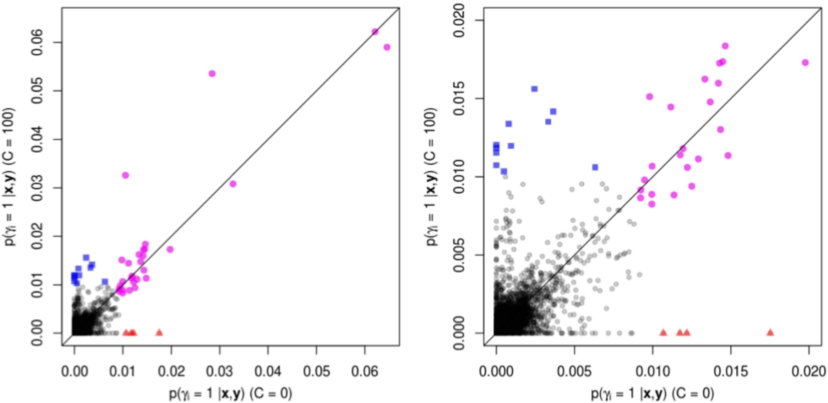


R package RJaCGH (Rueda and Diaz-Uriarte, 2007)

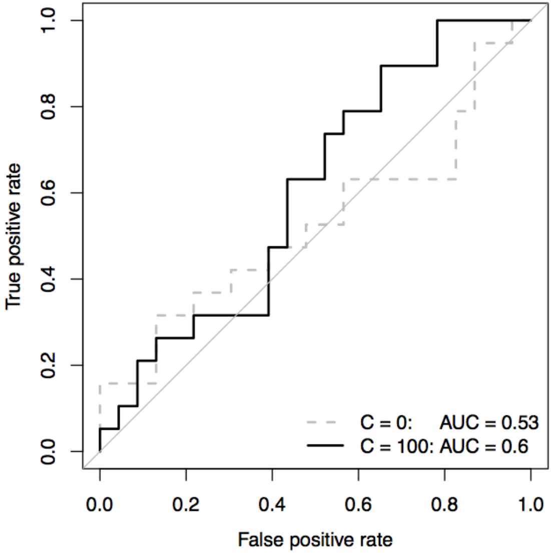
Medulloblastoma: prior and posterior inclusion probabilities



Medulloblastoma: prior and posterior inclusion probabilities



Medulloblastoma: test data ROC curves of BMA results



Summary

Advantages of Bayesian variable selection setup:

- Straightforward inclusion of external information through prior
- Fully probabilistic models
  - full posterior output: marginal and joint distributions
  - can be subjected to sensitivity analyses

Disadvantages:

- Larger computational burden than frequentist methods (computing time and memory usage)
- Users need to be more involved in model checking and interpretation

Software:

- R code with computationally intensive parts implemented in C
- R package BVSflex will be available on R-forge soon.

Acknowledgements

- Axel Benner (DKFZ Heidelberg)
- Stefan Pfister (DKFZ and University Hospital Heidelberg)
- Sylvia Richardson (MRC Biostatistics Cambridge)

Funding

- National Center for Tumor Diseases (NCT) Heidelberg
  Interdisciplinary Research Program,
  Project "Pediatric Brain Tumor Preclinical Testing"

Thank you!

References

- Holmes C, Held L (2006). *Bayesian Auxiliary Variable Models for Binary and Multinomial Regression*. Bayesian Analysis 1:145–168

- Northcott PA et al. (2012). *Subgroup-specific structural variation across 1,000 medulloblastoma genomes*. Nature 488:49–56

- Rueda OM, Diaz-Uriarte R (2007). *Flexible and Accurate Detection of Genomic Copy-Number Changes from aCGH*. PLoS Comput Biol. 3(6):e122

- Zucknick M (2009). *Multivariate analysis of tumour gene expression profiles applying regularisation and Bayesian variable selection techniques*. PhD dissertation. `http://www.ukpubmedcentral.ac.uk/theses/ETH/506406`

- Zucknick M (2013). *Integrating copy number variation information in gene expression models for two-class prediction and biomarker selection: a Bayesian variable selection approach*, submitted