Spatial clustering of array CGH features in combination with hierarchical multiple testing

Mark A. van de Wiel*[1,2]

[1]Dep. of Epidemiology and Biostatistics
VU University medical center, Amsterdam

[2]Dep. of Mathematics
VU University, Amsterdam

Cancer Bioinformatics Workshop, September 2010

* Joint work with Kyung In Kim (NCI/NIH) and Etienne Roquain (Univ. Paris 06), to be published in SAGMB (2010).

Setting

Data

- Discretized (called) high-resolution DNA copy number data: -1 (loss), 0 (normal), 1 (gain).
- Relevant clinical response, such as group label

Aim

Detect DNA regions with significant association between copy number and clinical response

Main problems

- DNA regions relevant for copy number not a priori defined
- Increase of resolution challenges multiple testing corrections

Setting

Data

- Discretized (called) high-resolution DNA copy number data: -1 (loss), 0 (normal), 1 (gain).
- Relevant clinical response, such as group label

Aim

Detect DNA regions with significant association between copy number and clinical response

Main problems

- DNA regions relevant for copy number not a priori defined
- Increase of resolution challenges multiple testing corrections

Setting

Data

- Discretized (called) high-resolution DNA copy number data: -1 (loss), 0 (normal), 1 (gain).
- Relevant clinical response, such as group label

Aim

Detect DNA regions with significant association between copy number and clinical response

Main problems

- DNA regions relevant for copy number not a priori defined
- Increase of resolution challenges multiple testing corrections

Increasing resolution

$\rightarrow$ Samples

probe1	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
probe2	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
probe3	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1

$\Downarrow$

probe1	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
newpr1	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
probe2	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
newpr2	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
probe3	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1

Collapsing

Collapse highly repetitious probes to one row

Loss of information can be controlled (Van de Wiel & Van Wieringen (2007), *Cancer Informatics*)

Handles locally ‘too high’ technical resolution

$\rightarrow$ Samples

probe1	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
probe2	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1
probe3	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1

$\Downarrow$

region1	0	1	1	0	0	1	1	-1	0	0	1	1	0	0	-1

Why clustering?

Collapsing: number of features reduces from several 100.000s to several 100s.

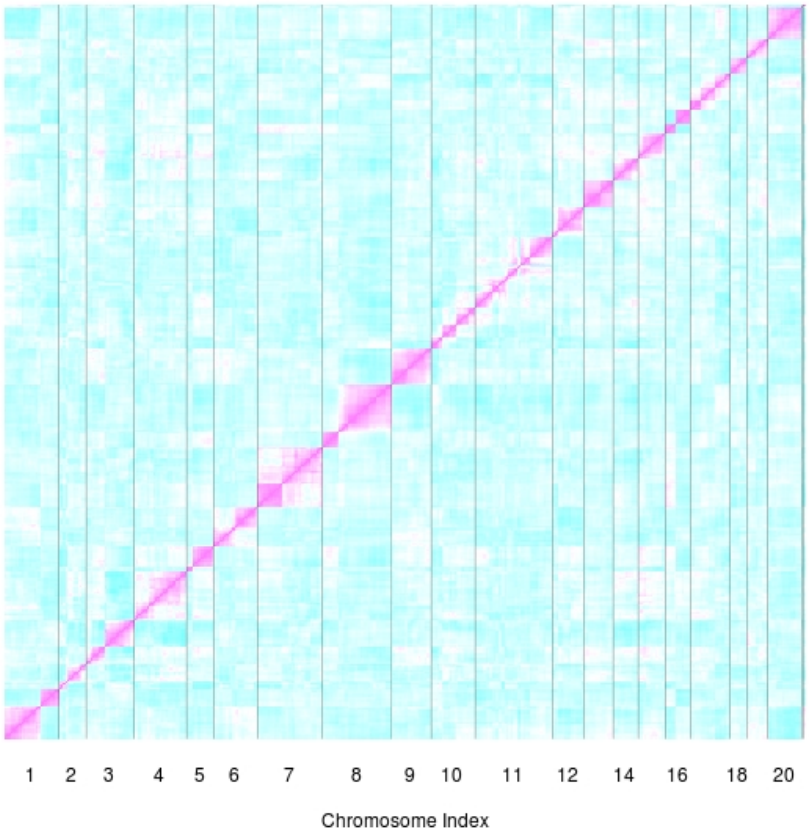
Resulting regions still possess a large degree of correlation.

Chin breast cancer data

- Chin et al. (2006), *Cancer Cell*
- Collapsing at 0.5% information loss: 383 regions
- 96 ER+, 49 ER- samples

Correlation between DNA regions

Correlation (Kendall's τ) heatmap for the Chin data set



Regions in order of chromosomal position. Colors represent correlations from -1 (cyan) to 1 (pink).

Cluster model

Aim

Find optimal partition of regions $\{1, \ldots, p\}$.

A: cluster of contiguous regions. x_A: possible realization of region data for cluster A. E.g. $A = \{1, 2, 3, 4\}$ and $x_A = (1, 1, 0, 1)^T$

Quadratic exponential model (Cox and Wermuth, 1994, *Biometrika*)

Model for cluster A (dropping sample index):

$$\log p_A(x_A; \alpha_A, \vec{\beta}_A, \gamma_A) = \alpha_A + \sum_{j \in A} \beta_{A,j} x_j + \gamma_A \sum_{j<k, j,k \in A} d_{jk} f(x_j, x_k),$$

where $f(x_j, x_k) = -1$, if $x_j \neq x_k$ and $f(x_j, x_k) = x_j x_k$, otherwise and d_{jk} is a distance function.

Cluster model

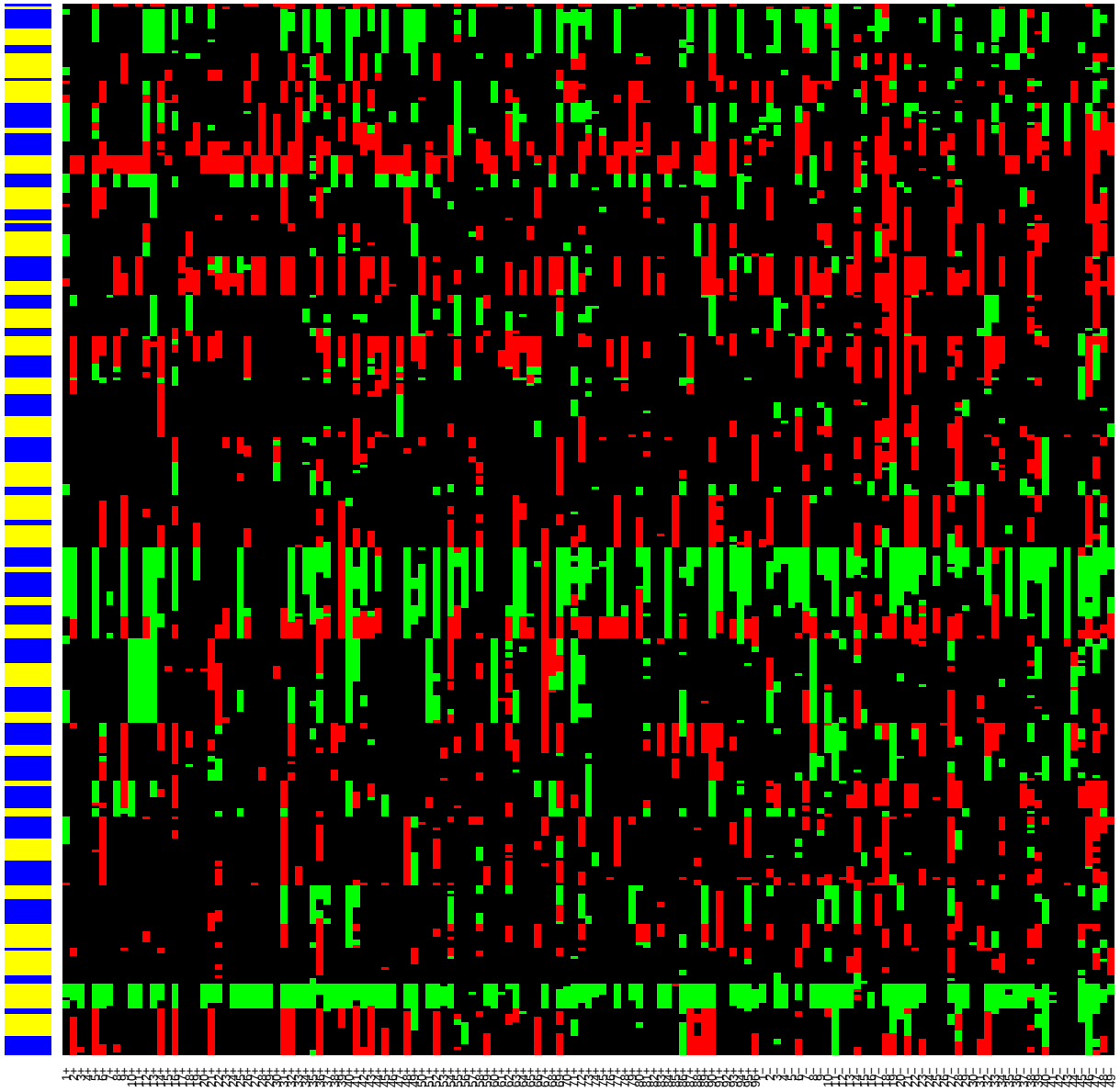
$$\log p_A(x_A; \alpha_A, \vec{\beta}_A, \gamma_A) = \alpha_A + \sum_{j \in A} \beta_{A,j} x_j + \gamma_A \sum_{j<k, j,k \in A} d_{jk} f(x_j, x_k),$$

Full log-likelihood model: sum over samples ànd over clusters (implying independence).

Given partition $\mathcal{A} = \{A_1, \ldots, A_C\}$, $\alpha_A, \vec{\beta}_A, \gamma_A$ are easy to estimate by ML. Difficult parameter is the clustering, $\mathcal{A}$, itself.

Full model contains intrinsic trade-off for dividing cluster into two sub-clusters.

Cluster results Chin data set



Cluster results, 'validation'

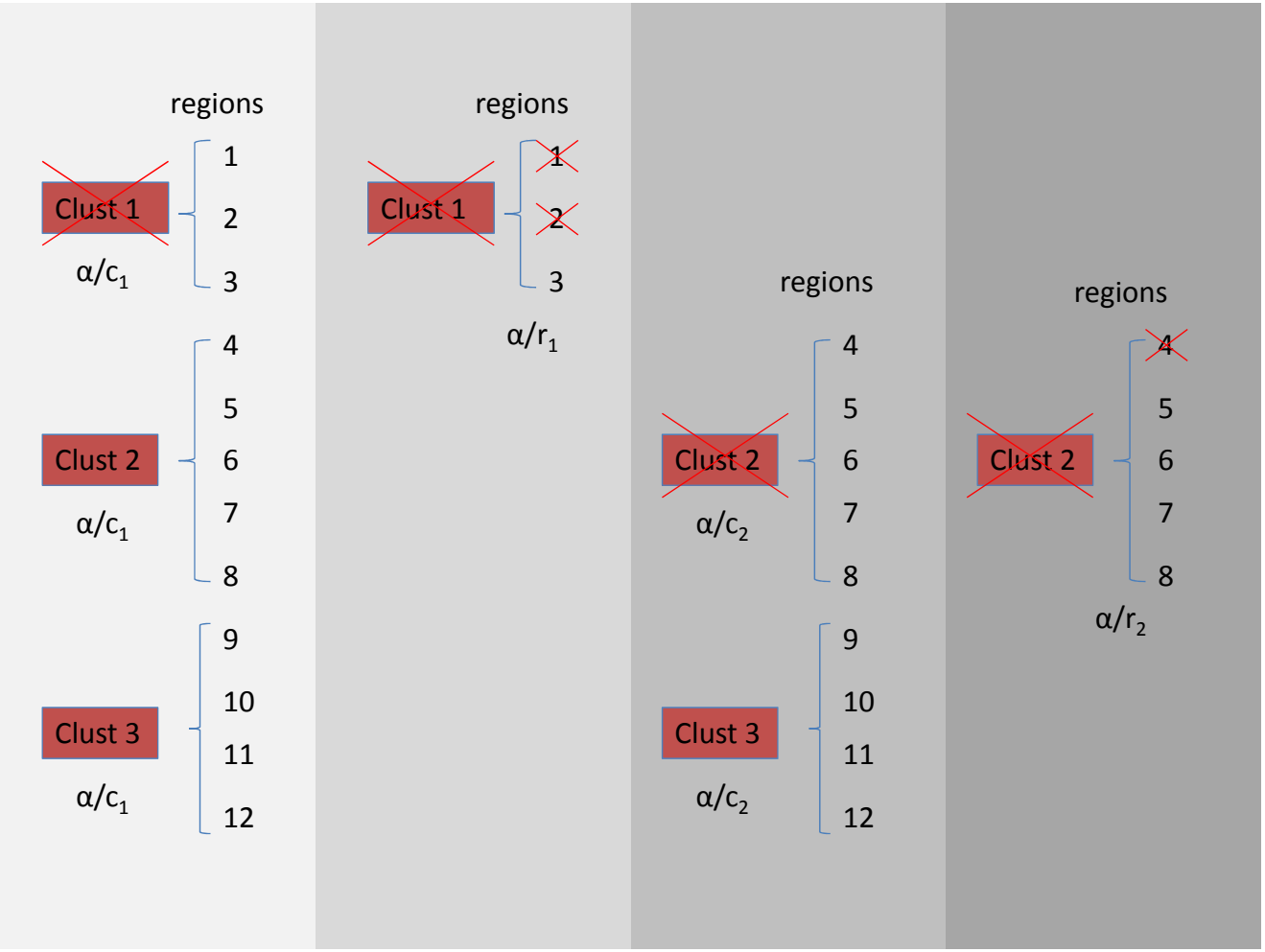
Results are robust

Adjusted average Rand index very high: ≈ 0.96 (10-fold CV)

Coincidental clustering is rare

- Clustering of two *independent* consecutive regions due to similar aberration pattern.
- Shuffle regions, consecutive regions from different chromosomes
- Then, consecutive regions are necessarily independent units
- Cluster algorithm should not cluster any regions
- Result on 383 regions: on average 382.2 clusters are formed.

Hierarchical testing



FWER $\leq \alpha$ (Meinshausen (2008). *Biometrika*)

Hierarchical testing: our setting

- Test statistic regions: χ^2; clusters: min p (max χ^2)
- Clustering permutation-invariant: testing using **same** samples

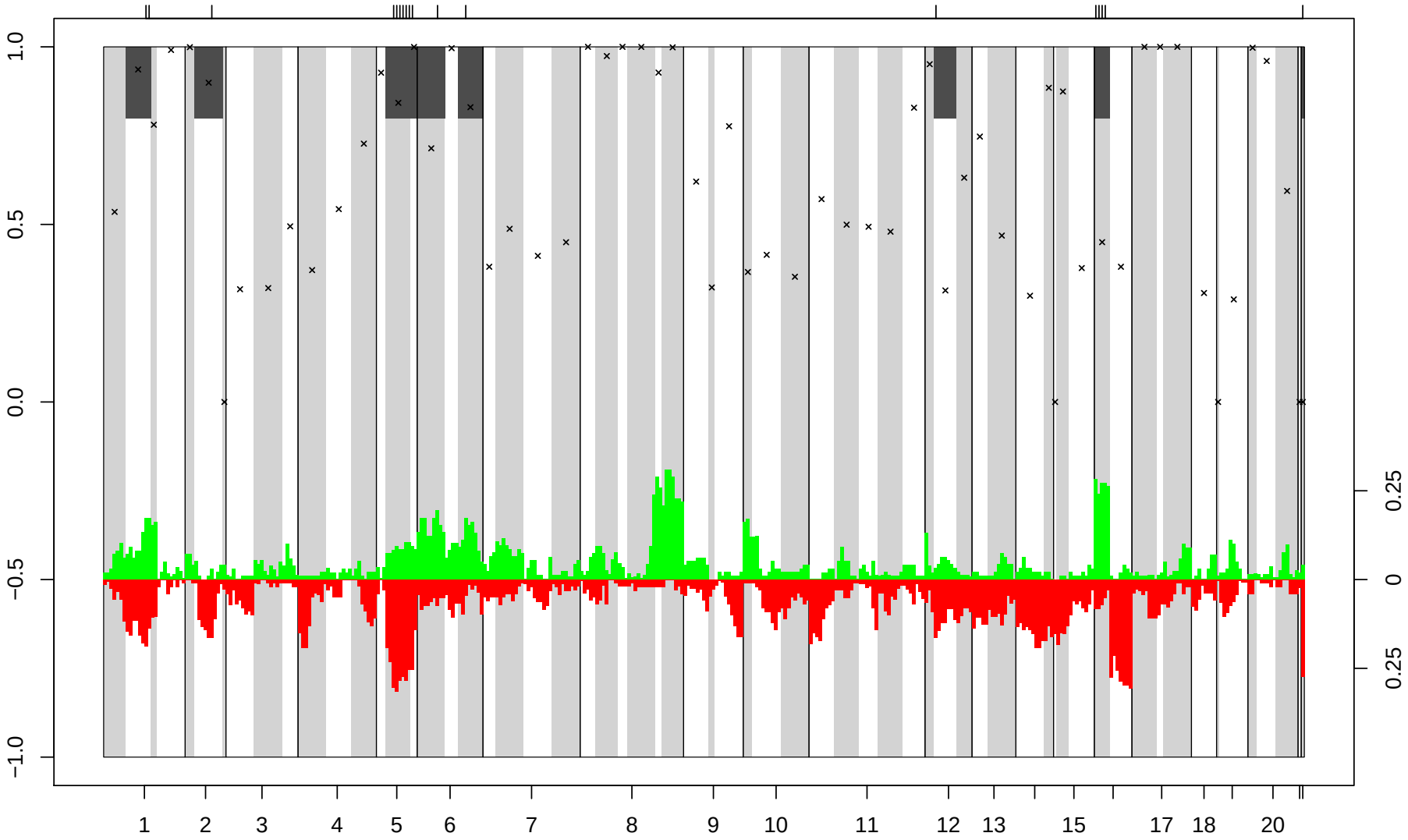
Reg	Clust	Chr.	BP
1	1	1	$3.8*10^5$
2	1	1	$6.1*10^5$
3	1	1	$2.4*10^6$
...	...	...	...
319	55	18	$4.3*10^5$
320	55	18	$7.3*10^5$
321	55	18	$3.9*10^6$
322	55	18	$4.8*10^6$
...	...	...	...
383	63	23	$2.0*10^5$

S1	S2	...	S96	S97	S98	...	S145
0	0	...	0	0	0	...	0
0	0	...	0	1	0	...	0
0	0	...	-1	1	1	...	1
...	...	...	...	...	...	...	...
1	0	...	1	-1	0	...	0
1	0	...	1	-1	0	...	0
1	1	...	1	0	0	...	0
1	1	...	1	0	0	...	0
...	...	...	...	...	...	...	...
0	0	...	1	0	1	...	1

Permutations

		...				...	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		...				...	

Results



Results: power

- Eight clusters were detected at $\alpha = 0.05$
- Two of those do not contain any significant regions according to ordinary FWER correction (Holm)
- On the region level, the procedure is as powerful as an ordinary FWER correction (Holm)
- Benefit is even larger for data containing less samples

Discussion

Use of clustering for other purposes

- probe design for low-dimensional platforms (MLPA)
- clustering of samples
- prediction/classification

- Inclusion of amplification state in the cluster model
- Conditional vs. unconditional testing
- Software: www.few.vu.nl/∼markvdw

Discussion

Use of clustering for other purposes

- probe design for low-dimensional platforms (MLPA)
- clustering of samples
- prediction/classification

- Inclusion of amplification state in the cluster model
- Conditional vs. unconditional testing
- Software: www.few.vu.nl/∼markvdw

References

- Chin, K. et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.
- Cox, D. R., and N. Wermuth (1994). A note on the quadratic exponential binary distribution. *Biometrika*, **81**, 403–408.
- Kim, K.I., E. Roquain and M.A. van de Wiel (2010). Spatial clustering of array CGH features in combination with hierarchical multiple testing. *Statist. Appl. Genet. Mol. Biol.*
- Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, **95**, 265–278.
- Van de Wiel, M.A., and W.N. van Wieringen (2007).CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Informatics*, **2**, 55–63.