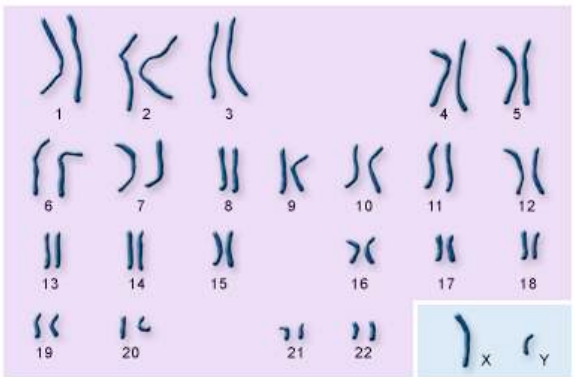
On Missing Data and Genotyping Errors in Association Studies

Ingo Ruczinski

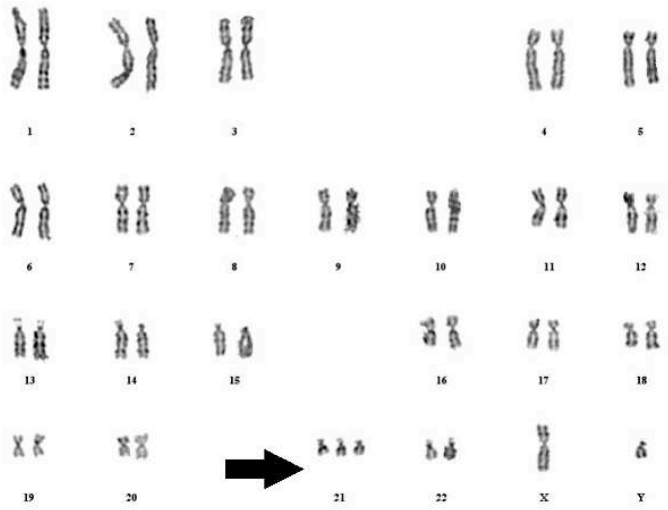
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

December 18, 2008

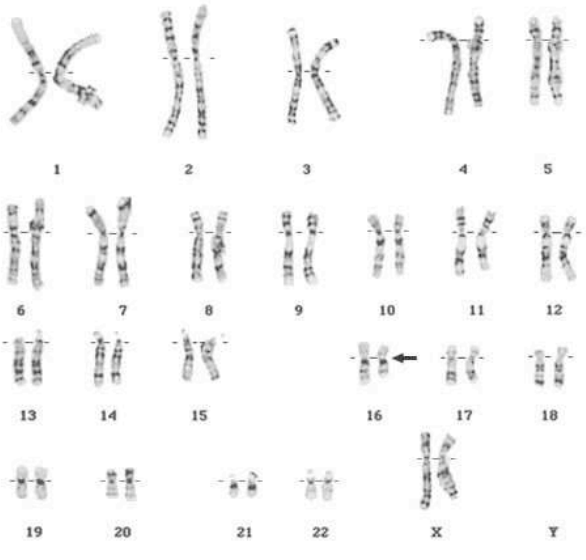
Karyotypes

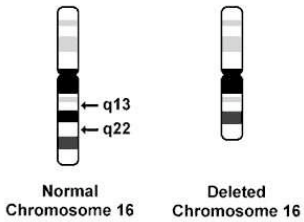


Trisomy



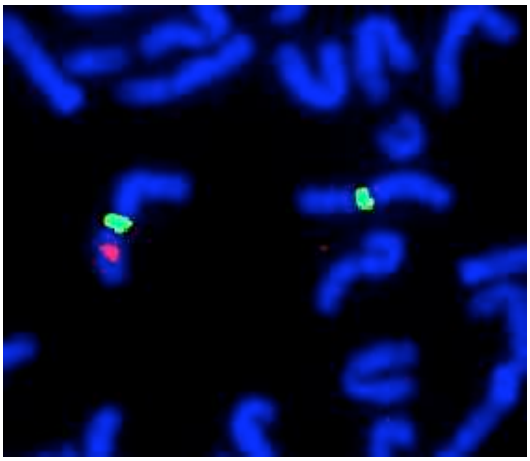
Karyotypes





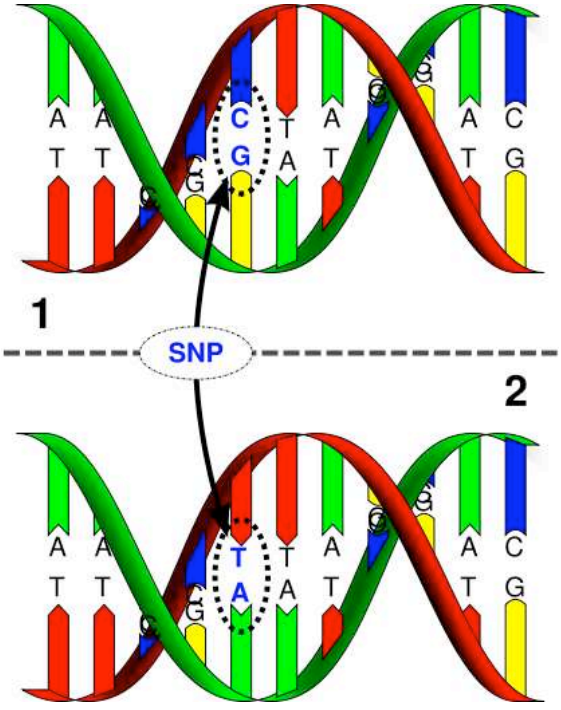
General Cytogenetics Information http://members.aol.com/chrominfo/

FISH



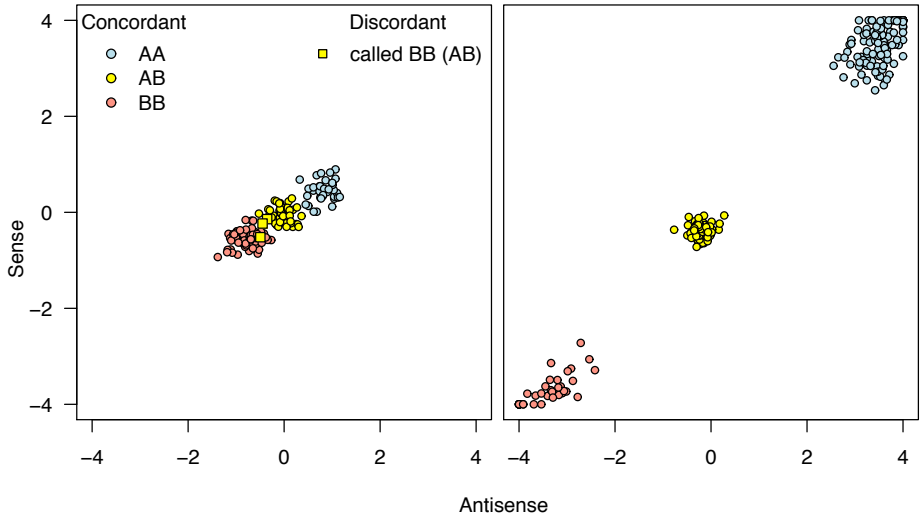
Courtesy of the Pevsner Laboratory

Single nucleotide polymorphisms



Genotype uncertainty

→ The confidence in genotype calls can differ substantially between SNPs!

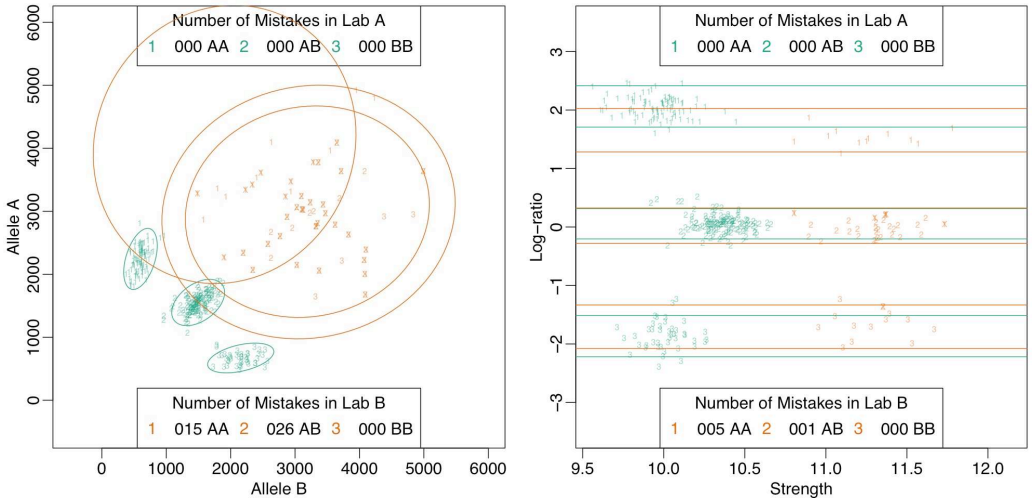


Missing at random?

Method	Confidence Threshold	Overall Call Rate	Hom Call Rate	Het Call Rate
DM	0.26	**94.16%**	97.24%	86.32%
DM	0.33	**95.96%**	98.24%	90.16%
BRLMM	0.3	**97.40%**	97.40%	97.75%
BRLMM	0.4	**98.27%**	98.30%	98.48%
BRLMM	0.5	**98.79%**	98.82%	98.93%
BRLMM	0.6	**99.15%**	99.18%	99.25%

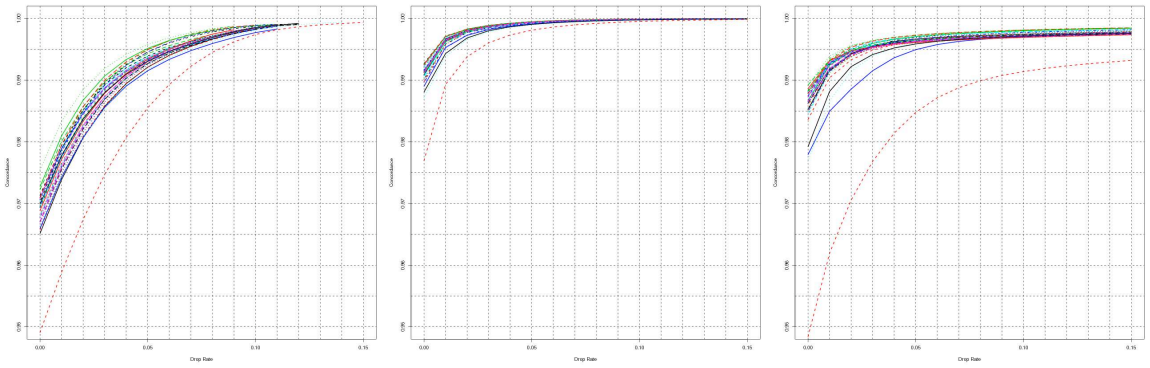
→ From the “white paper” available at
http://www.affymetrix.com/support/technical/productupdates/brlmm_algorithm.affx

Genotype calling

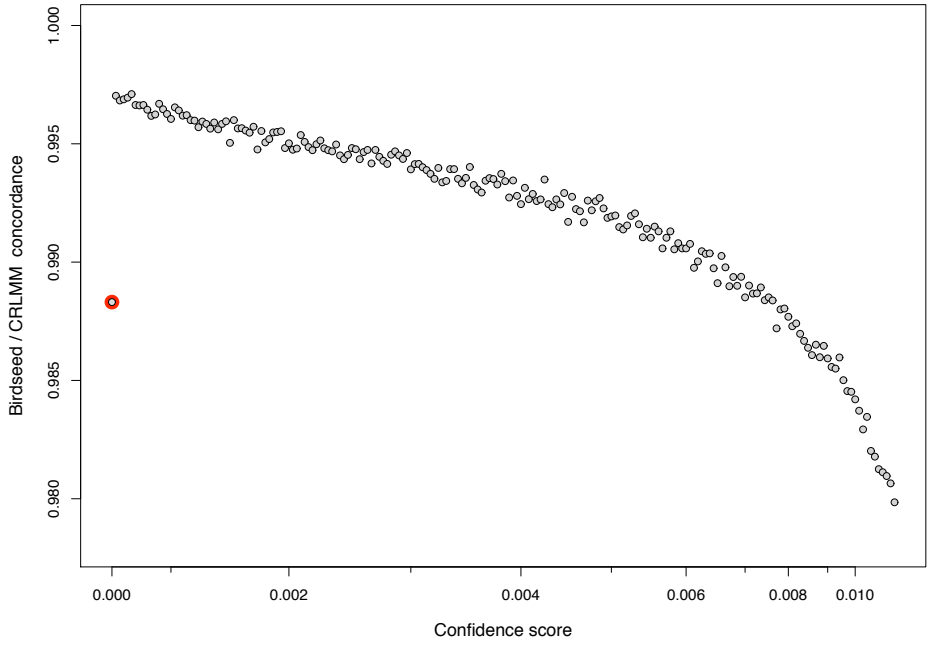


Genotype call comparison

BRLMM-P CRLMM BirdSeed



Concordance



Incorporating genotype uncertainty

- Let **Y** be the binary response, **d** be the score assigned to a genotype (reflecting the genetic model), **w** is the *working* genotype probability, and **t** is the *true* genotype probability.
- The score test statistic is

$$Z(\mathbf{w}, d) = \frac{\sum_i (w_{i1} d_1 + w_{i2} d_2)(Y_i - \hat{\pi})}{\sqrt{n \times \mathrm{Var}(\mathbf{w_1} d_1 + \mathbf{w_2} d_2) \times \hat{\pi}(1 - \hat{\pi})}}$$

- Here

$$\hat{\pi} = \bar{\mathbf{Y}} = \frac{1}{n} \sum Y_i$$

and

$$\mathrm{Var}(\mathbf{w_1} d_1 + \mathbf{w_2} d_2) =$$

$$d_1^2 \times \mathrm{Var}(\mathbf{w_1}) + d_2^2 \times \mathrm{Var}(\mathbf{w_2}) + 2 \times d_1 d_2 \times \mathrm{Cov}(\mathbf{w_1}, \mathbf{w_2})$$

Incorporating genotype uncertainty

- The variance term measures statistical information

$$n \times \mathrm{Var}(\mathbf{w_1} d_1 + \mathbf{w_2} d_2) =$$

$$\sum_{j=1}^{2} d_j^2 \bar{w}_j (1 - \bar{w}_j) - n \times 2 \times d_1 d_2 \bar{w}_1 \bar{w}_2 -$$

$$\sum_{j=1}^{2} d_j^2 \sum_i w_{ij}(1 - w_{ij}) + 2 \times d_1 d_2 \sum_i w_{i1} w_{i2}$$

- **Blue** is the variance when $w_{ij} = 0$ or 1.
- **Red** is the loss of information.
- Note that $w_{ij} = 0$ or 1 does not imply validity!

Incorporating genotype uncertainty

- For $\theta = \frac{\theta^*}{\sqrt{n}}$ (a local alternative), $Z \sim N(m(\cdot), 1)$
- For a logistic, for example, we have

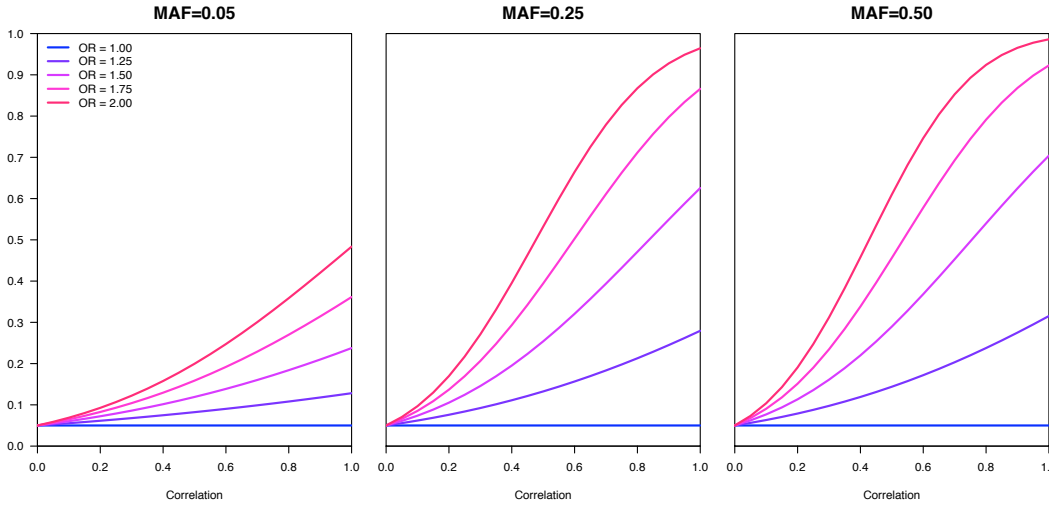
$$E(Z(\mathbf{w}, \mathbf{d}) \mid \mu, \theta^*, \mathbf{d}, \mathbf{w}, \mathbf{t}) \approx$$

$$\theta^* \{n \times \hat{\pi}(1 - \hat{\pi})\}^{\frac{1}{2}} \times \sqrt{\mathrm{Var}(\mathbf{t_1} d_1 + \mathbf{t_2} d_2)} \times$$

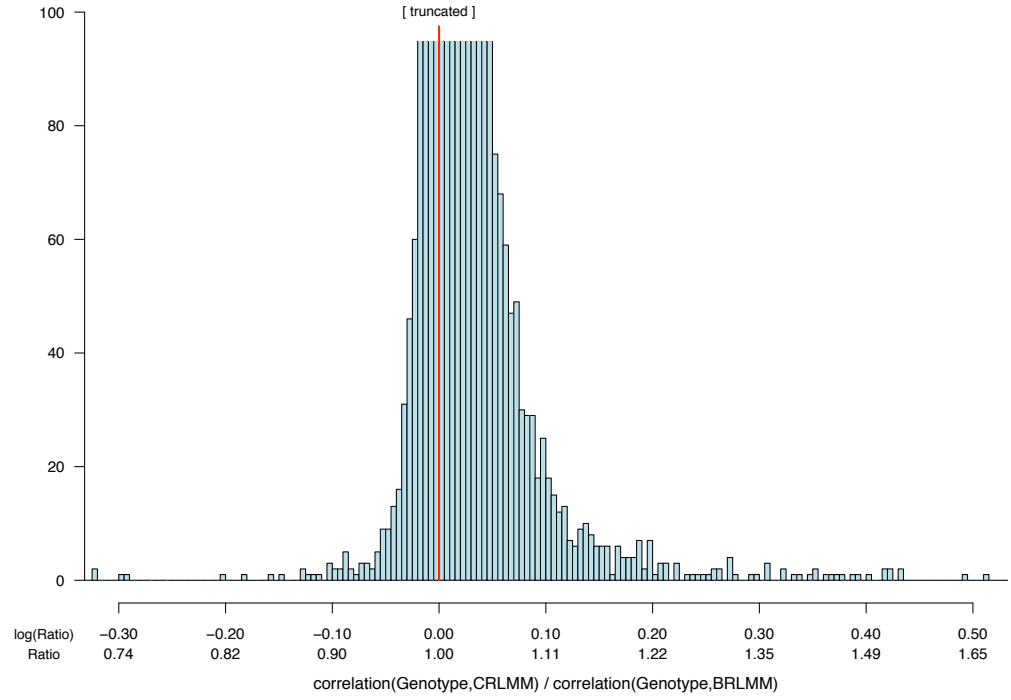
$$\mathrm{Corr}(\{\mathbf{w_1} d_1 + \mathbf{w_2} d_2\}, \{\mathbf{t_1} d_1 + \mathbf{t_2} d_2\})$$

- The correlation is the relative efficiency of using **w** rather than **t** (**t** can be **g**, the true genotype).
- For a fixed sample size and effect size, the power of the score test is a function of the correlation of the fuzzy genotype call with the true genotype.

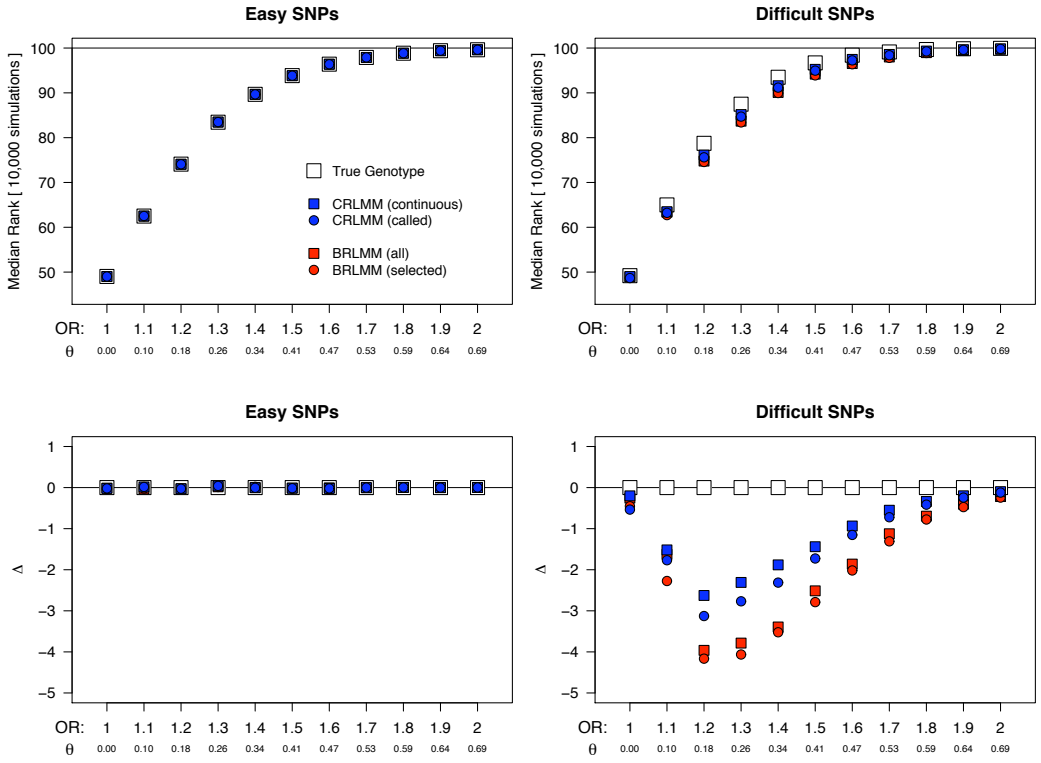
Incorporating genotype uncertainty



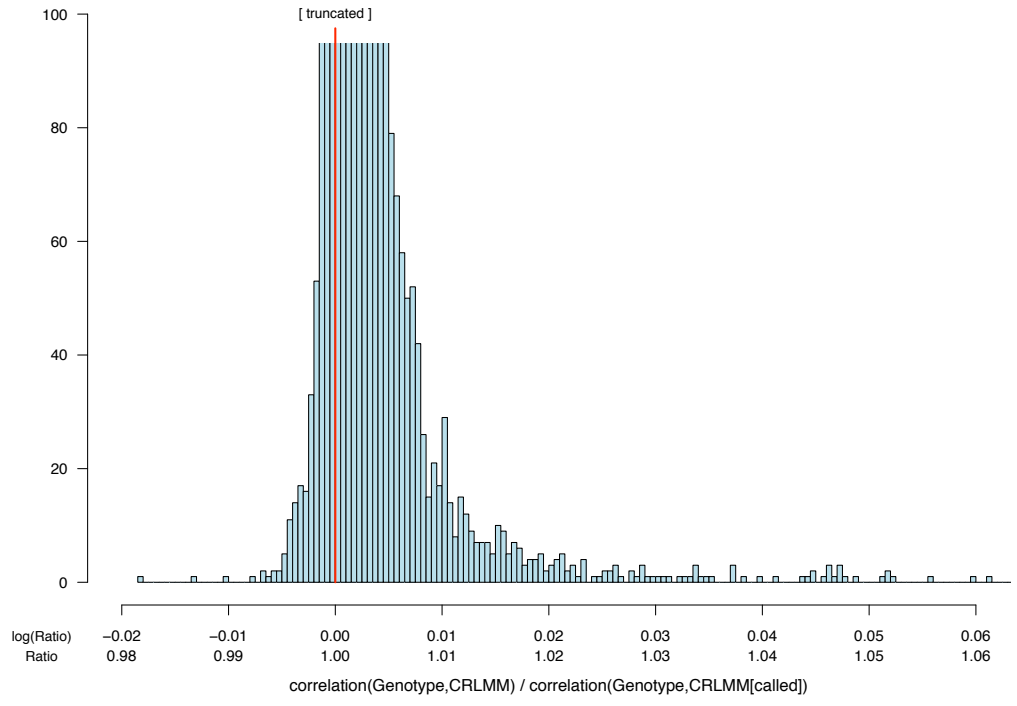
Incorporating genotype uncertainty



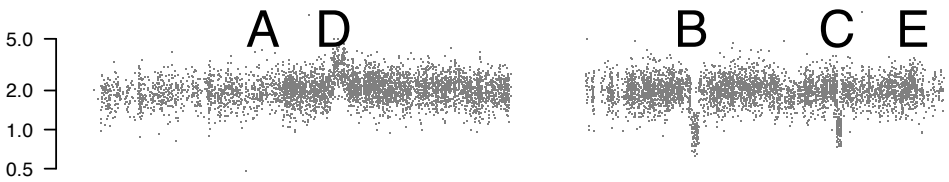
Incorporating genotype uncertainty



Incorporating genotype uncertainty

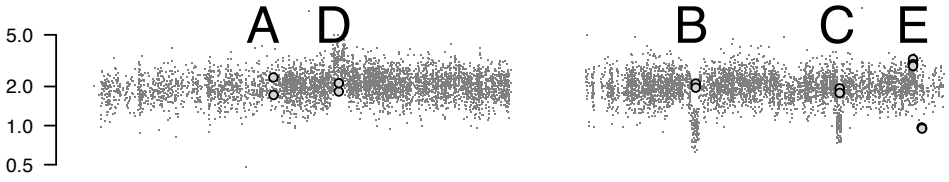


Copy number simulation study



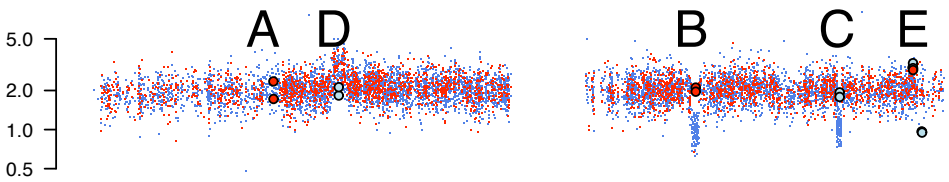
⟶ Scharpf et al (2008). *Hidden Markov Models for the Assessment of Chromosomal Alterations using High-throughput SNP Arrays.* The Annals of Applied Statistics, 2(2): 687-713.

Copy number simulation study



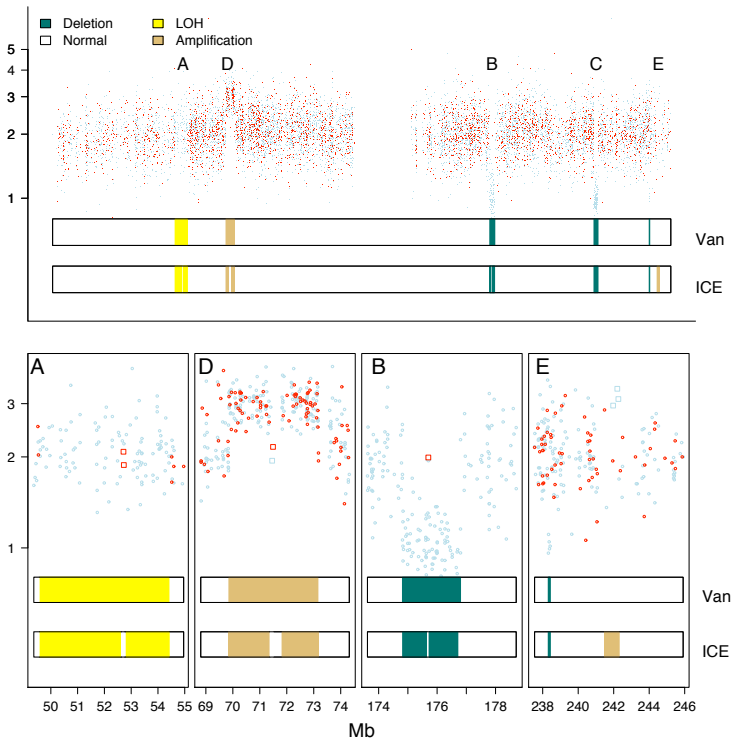
⟶ Scharpf et al (2008). *Hidden Markov Models for the Assessment of Chromosomal Alterations using High-throughput SNP Arrays.* The Annals of Applied Statistics, 2(2): 687-713.

Copy number simulation study

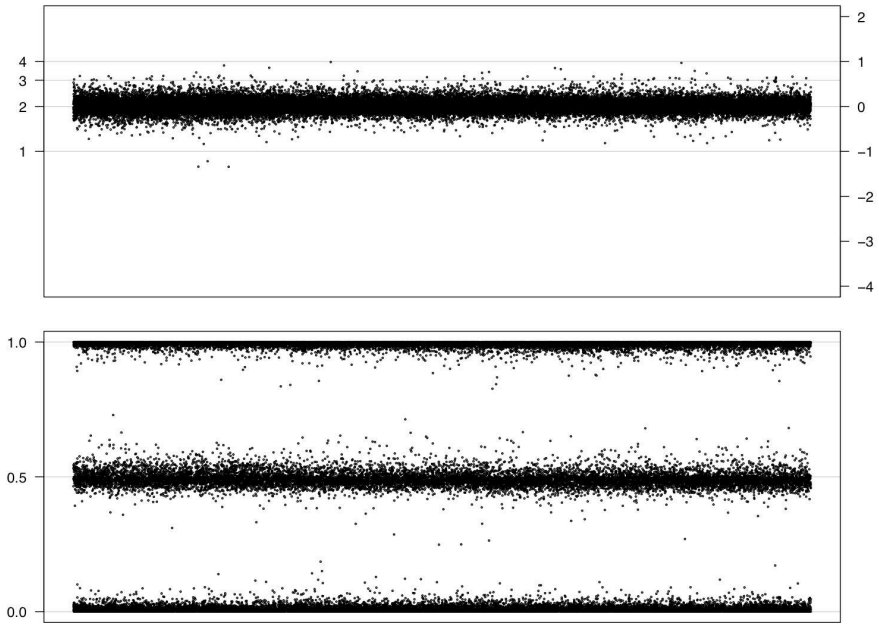


→ Scharpf et al (2008). *Hidden Markov Models for the Assessment of Chromosomal Alterations using High-throughput SNP Arrays.* The Annals of Applied Statistics, 2(2): 687-713.

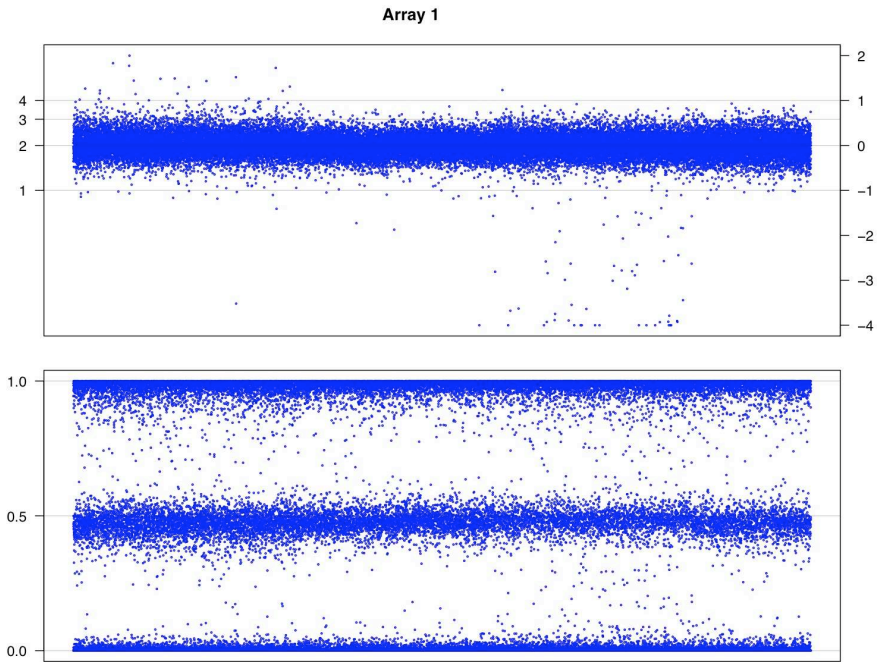
Copy number simulation study



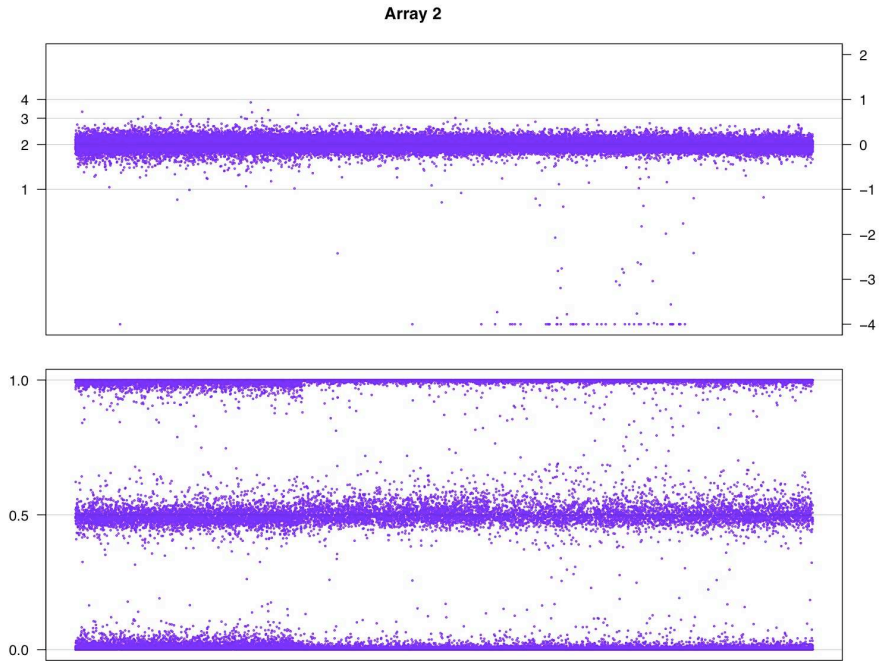
Illumina



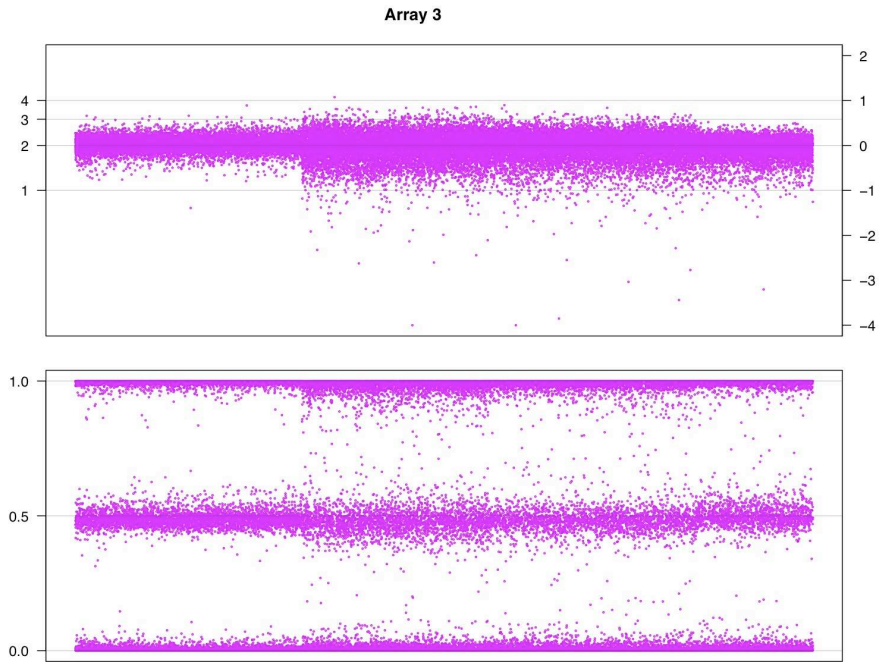
Spatial artifacts



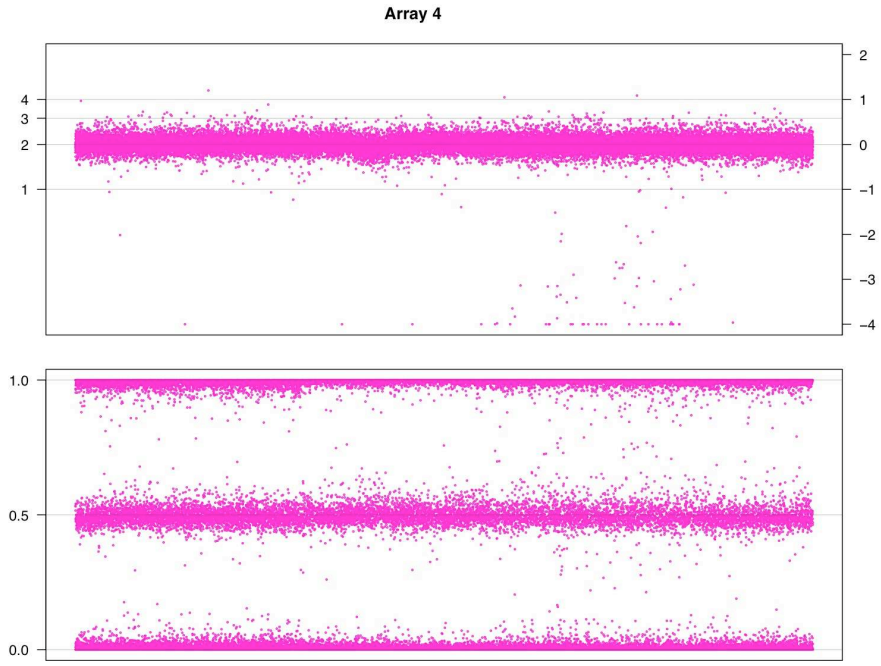
Spatial artifacts



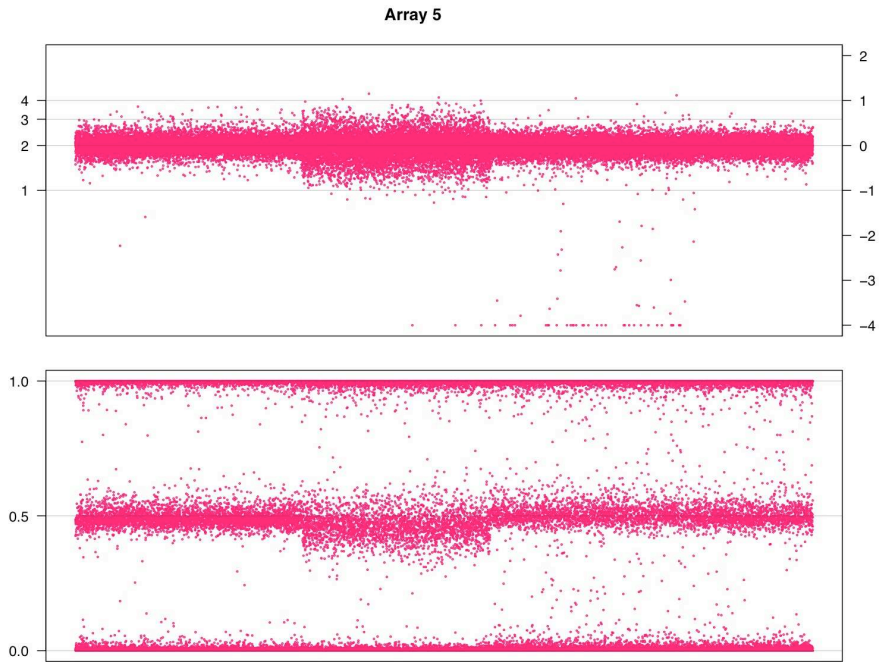
Spatial artifacts



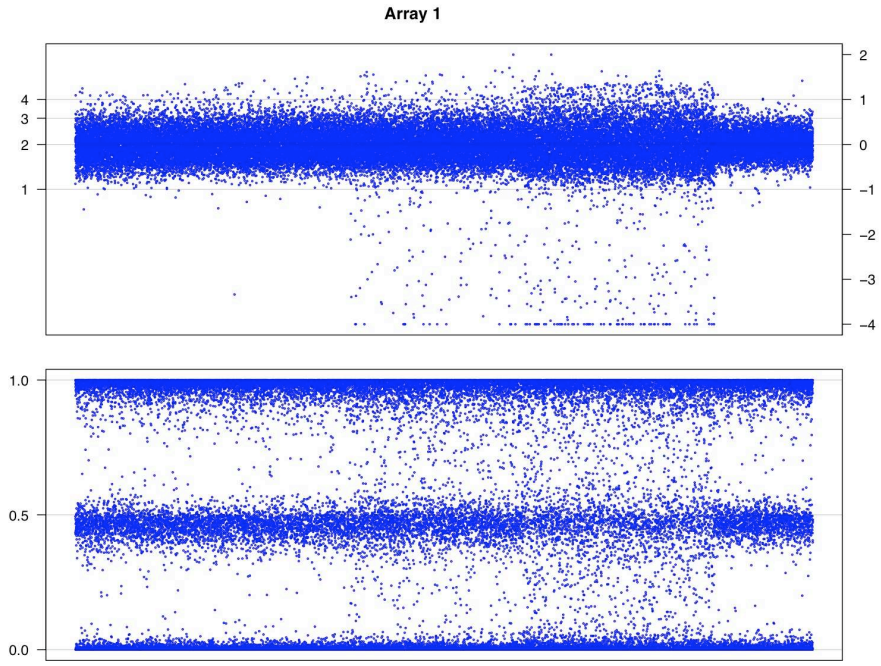
Spatial artifacts



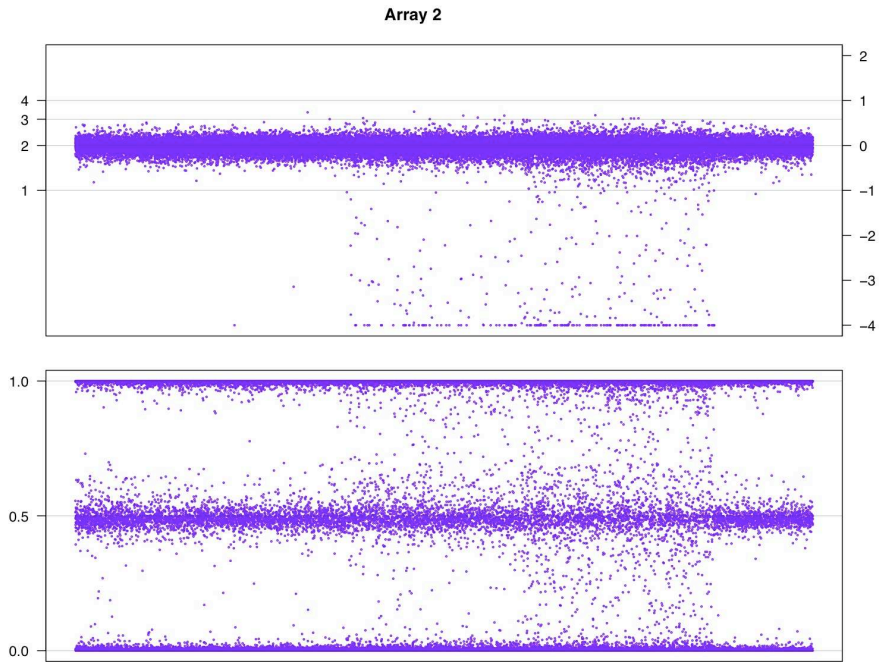
Spatial artifacts



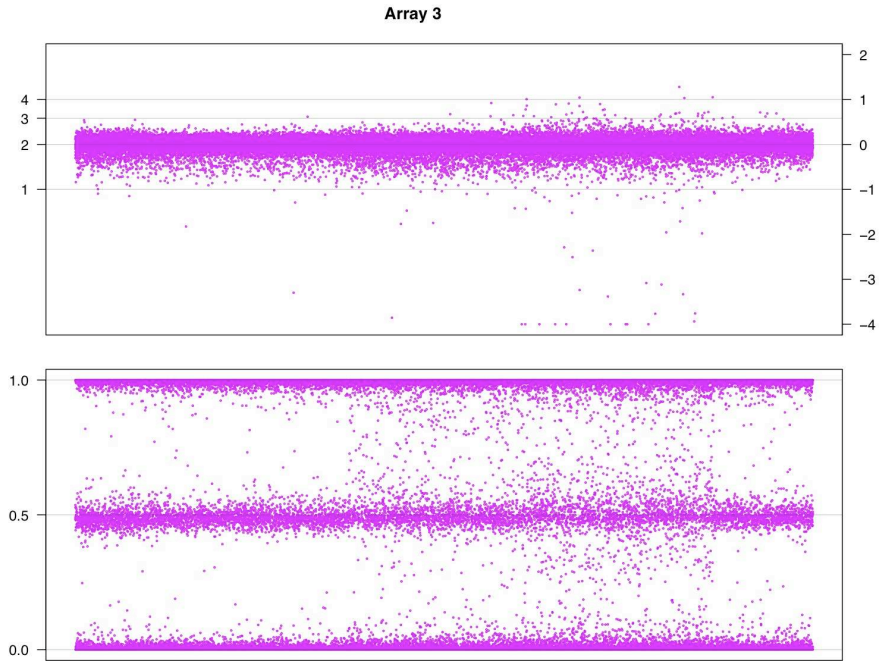
Spatial artifacts



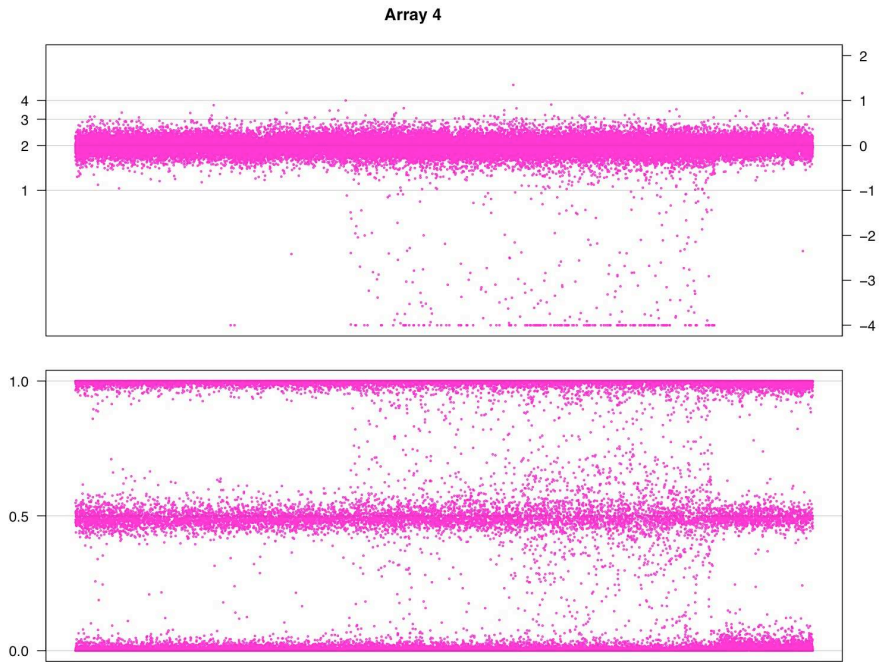
Spatial artifacts



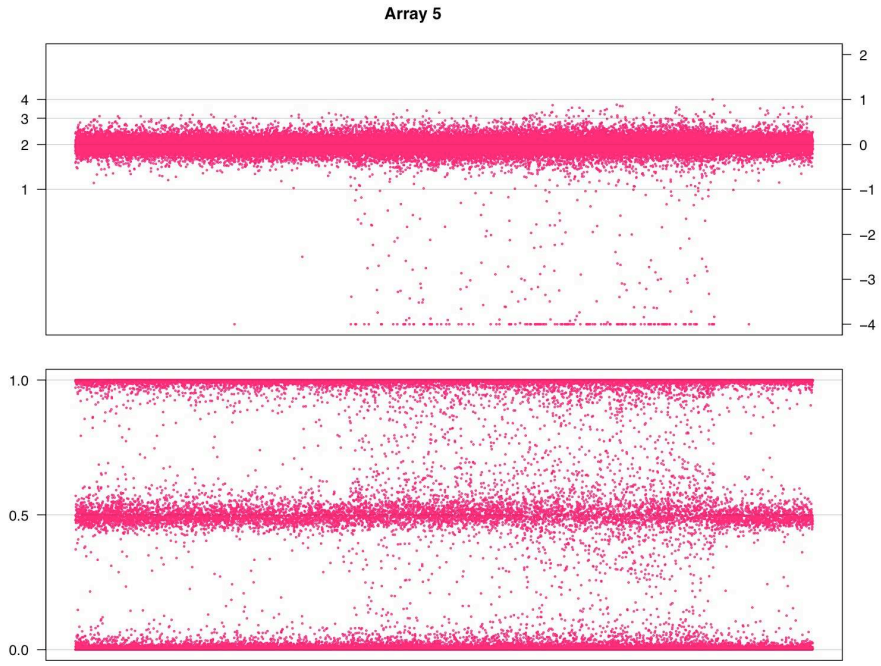
Spatial artifacts



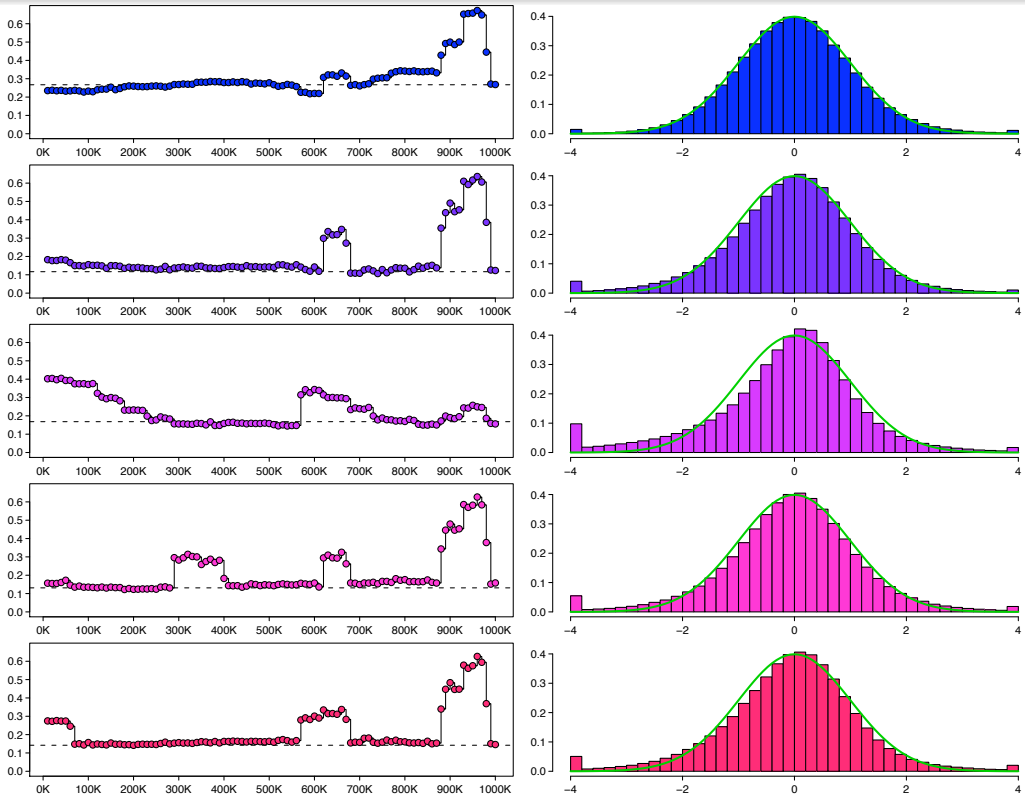
Spatial artifacts



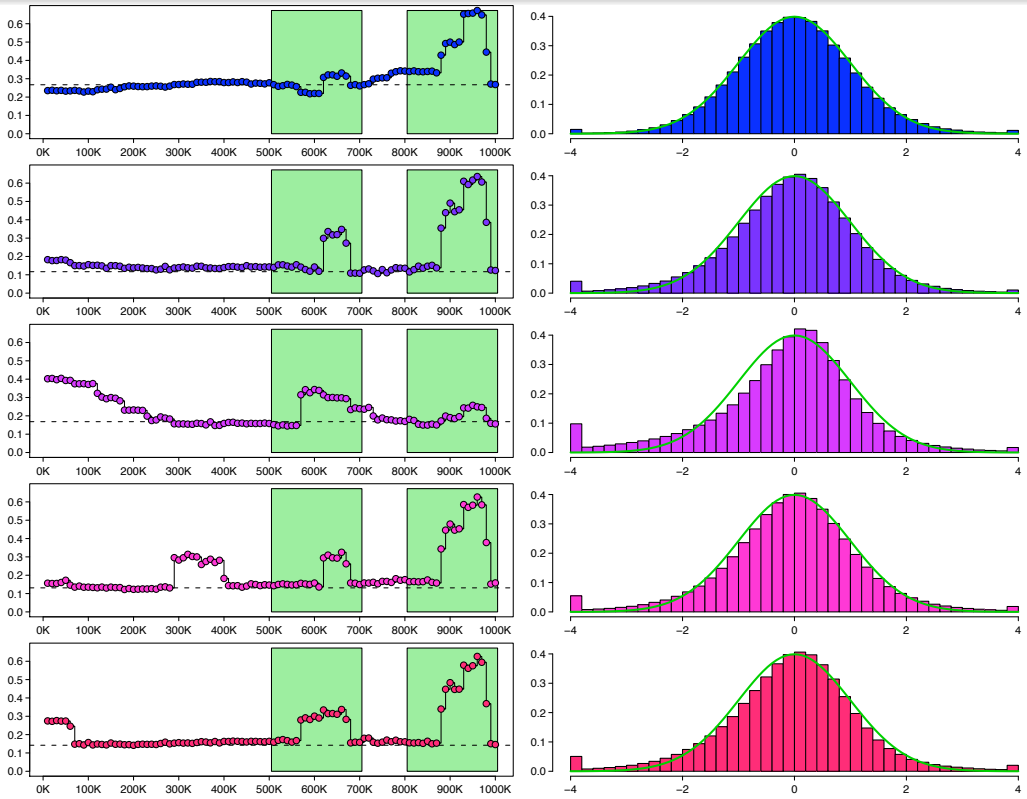
Spatial artifacts



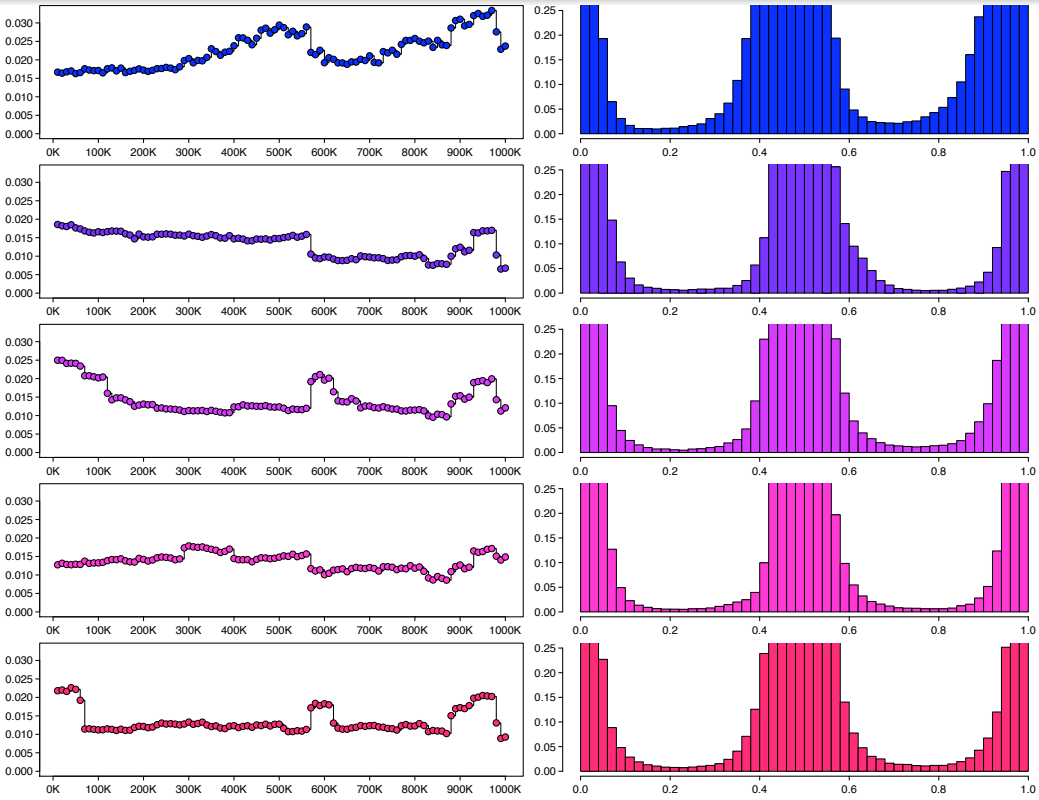
Spatial artifacts



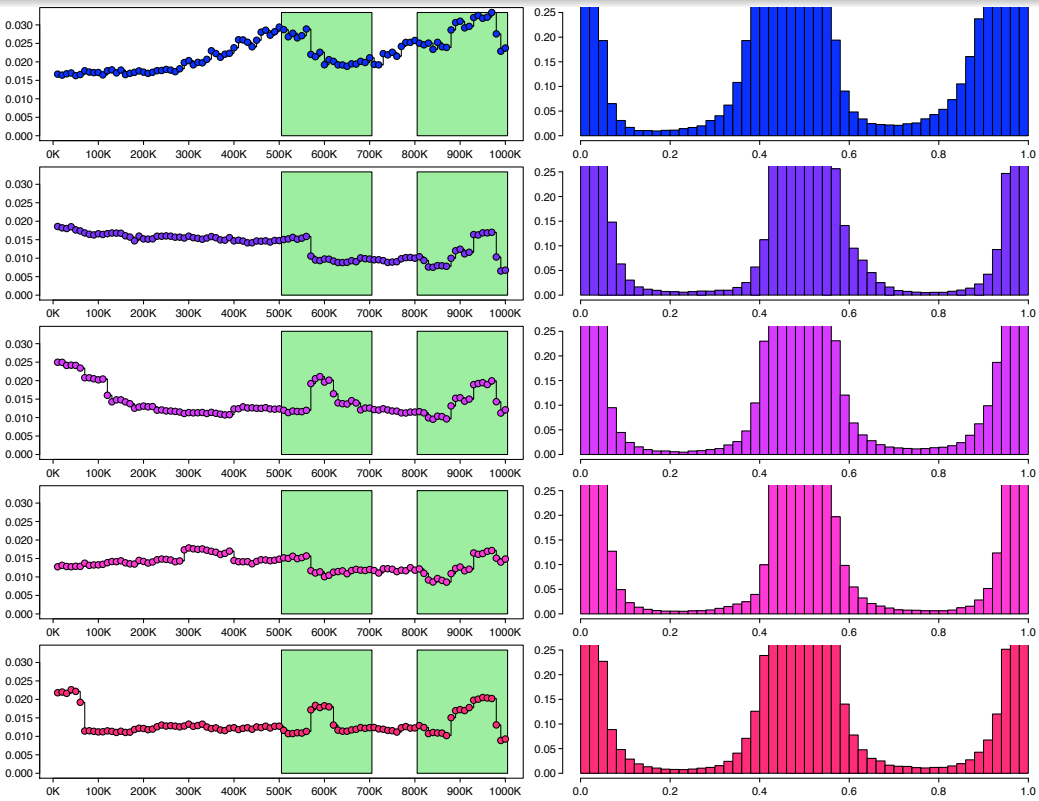
Spatial artifacts



Spatial artifacts



Spatial artifacts



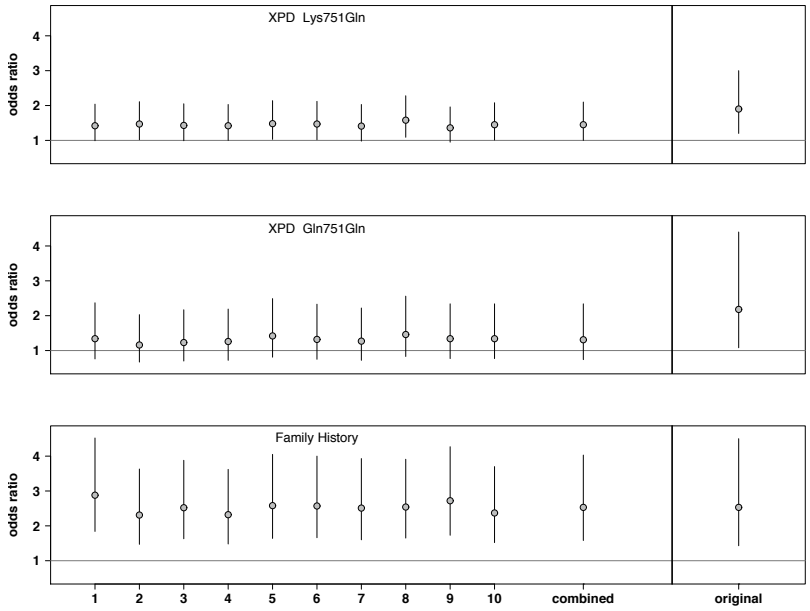
What's missing?

1. Phenotype delineation
2. Genotypes (unobserved, mis-typed)
3. Copy number variants
4. Genetic models (complex disease, biological interactions)
5. Epigenetics and such
6. Environmental data

Missing environmental data

	Number of Pairs	Odds Ratio	Confidence Interval
		XPD Lys751Gln	
original data set	202	1.90	(1.20 – 3.00)
		XPD Gln751Gln	
original data set	202	2.18	(1.08 – 4.40)
		Positive Family History	
original data set	202	2.53	(1.43 – 4.50)

Missing environmental data



The missing data were imputed using decision trees.

⟶ Dai J et al (2006). *Imputation Methods to Improve Inference in SNP Association Studies.* Genetic Epidemiology, 30(8): 690-702.

Missing environmental data

	Family History not complete				Family History complete			
	AA	AC	CC	na	AA	AC	CC	na
				raw numbers				
case	43	54	5	5	61	121	25	7
control	35	57	12	3	90	102	22	0
				percentages				
case	40.2	50.5	4.7	4.7	28.5	56.5	11.7	3.3
control	32.7	53.3	11.2	2.8	42.1	47.7	10.3	0.0

⟶ Brewster AM et al (2006). *Polymorphisms of the DNA Repair Genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): Relationship to Breast Cancer Risk and Familial Predisposition to Breast Cancer.* Breast Cancer Research and Treatment, 95(1): 73-80.

Missing environmental data

	Number of Pairs	Odds Ratio	Confidence Interval
		XPD Lys751Gln	
original data set	202	1.90	(1.20 – 3.00)
		XPD Gln751Gln	
original data set	202	2.18	(1.08 – 4.40)
		Positive Family History	
original data set	202	2.53	(1.43 – 4.50)

Missing environmental data

	Number of Pairs	Odds Ratio	Confidence Interval
		XPD Lys751Gln	
original data set	202	1.90	(1.20 – 3.00)
multiple imputations	321	1.45	(1.00 – 2.10)
		XPD Gln751Gln	
original data set	202	2.18	(1.08 – 4.40)
multiple imputations	321	1.31	(0.74 – 2.34)
		Positive Family History	
original data set	202	2.53	(1.43 – 4.50)
multiple imputations	321	2.53	(1.58 – 4.03)

Why become a Biostatistician?

Why become a Biostatistician?

⟶ Because people appreciate your help analyzing their data, and that means that people surely will like you.

Why become a Biostatistician?

⟶ Because people appreciate your help analyzing their data, and that means that people surely will like you.

From: [redacted]
Subject: **A curse on you and your progeny!!!**

To: Ingo Ruczinski <iruczins@jhsph.edu>

Ingo:

Curse you, Ingo! Yet another disappearing act!

The association between flame broiled food consumption and breast cancer disappears in the imputed dataset (see below). I'm beginning to hate this imputation stuff! I much prefer biased data. The findings are more interesting (and more publishable).

Acknowledgments

Copy number: Benilton Carvalho, Rafael Irizarry, Giovanni Parmigiani, Rob Scharpf.

Score test: Dani Fallin, Qing Li, Tom Louis.

Computing support: Marvin Newhouse, Jiong Yang.

Miscellaneous other: Abeena Brewster, Joe Coresh, Linda Kao, Anna Kottgen, Tim Jorgenson, Mandy Li, Yun Liu, Jon Pevsner, Forrest Spencer.

Funding: NIH R01 DK061662, GM083084, HL090577, and a CTSA grant to the Johns Hopkins Medical Institutions.

http://biostat.jhsph.edu/∼iruczins/