On Missing Data and Genotyping Errors in Association Studies

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

August 4, 2008

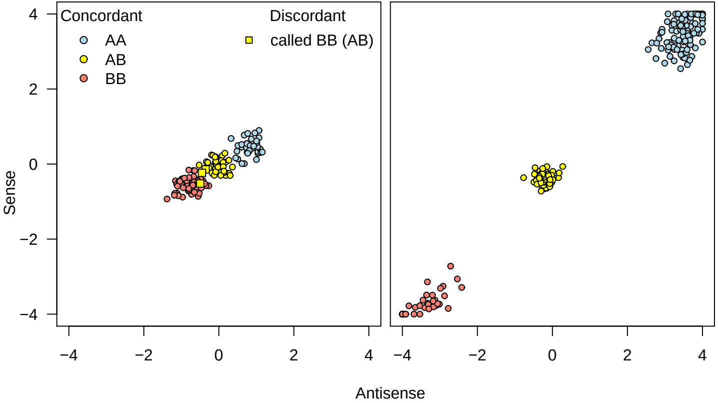
Missing Data

There are mainly three types of missing / unobserved data in genetic association studies:

1. Missing observations in some environmental variables.
2. Missing data at SNPs selected for genotyping.
3. Genotypes of SNPs not selected.

Genotype Uncertainty

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



VanillaICE HMM for CNVs

Important HMM features:

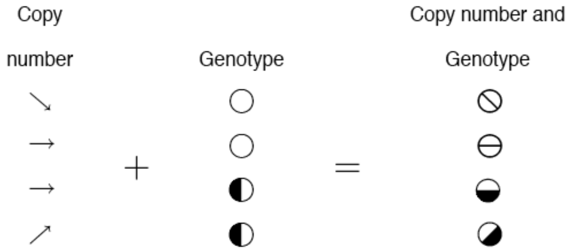
1. Model the observation sequence of genotype calls and copy number jointly (Vanilla)
2. Integrate confidence estimates of the genotype calls and copy number estimates (ICE)

QuantiSNP and *PennCNV* also model genotype and copy number jointly!

Colella et al (2007) *QuantiSNP: an objective Bayes hidden-Markov model...* Nucleic Acids Res 35(6): 2013-25.

Wang et al (2008) *PennCNV: An integrated hidden-Markov model designed for...* Genome Research 17: 1665-74.

Hidden States



Copy number

Genotype

Copy number and Genotype

Emission Probabilities

We assume conditional independence between copy number estimates and the genotype calls.

For example:

$$f(\widehat{\mathrm{CN}},\widehat{\mathrm{GT}}|\oslash) \quad = \quad f(\widehat{\mathrm{CN}}|\oslash)\times f(\widehat{\mathrm{GT}}|\oslash) \quad = \quad f\left\{\ \widehat{\mathrm{CN}}\ |\ \searrow\ \right\}\times f\left\{\ \widehat{\mathrm{GT}}\ |\ \bigcirc\ \right\}$$

Integrating Confidence Estimates for Genotype Calls

Let $S_{\widehat{\mathrm{GT}}}$ be the confidence score for the genotype estimate.

We can estimate from Hapmap the following densities:

$$f\left\{ \mathrm{S}_{\widehat{\mathrm{HOM}}} \mid \widehat{\mathrm{HOM}}, \mathrm{HOM} \right\}, f\left\{ \mathrm{S}_{\widehat{\mathrm{HOM}}} \mid \widehat{\mathrm{HOM}}, \mathrm{HET} \right\}, f\left\{ \mathrm{S}_{\widehat{\mathrm{HET}}} \mid \widehat{\mathrm{HET}}, \mathrm{HOM} \right\}, f\left\{ \mathrm{S}_{\widehat{\mathrm{HET}}} \mid \widehat{\mathrm{HET}}, \mathrm{HET} \right\}.$$

$\longrightarrow$ Note:

$$\begin{aligned} f\left\{ \mathrm{S}_{\widehat{\mathrm{HOM}}} \mid \widehat{\mathrm{HOM}}, \bigcirc \right\} &\approx f\left\{ \mathrm{S}_{\widehat{\mathrm{HOM}}} \mid \widehat{\mathrm{HOM}}, \mathrm{HOM} \right\} \\ f\left\{ \mathrm{S}_{\widehat{\mathrm{HET}}} \mid \widehat{\mathrm{HET}}, \bigcirc \right\} &\approx f\left\{ \mathrm{S}_{\widehat{\mathrm{HET}}} \mid \widehat{\mathrm{HET}}, \mathrm{HOM} \right\}. \end{aligned}$$

Emission Probabilities - Loss

Recall that

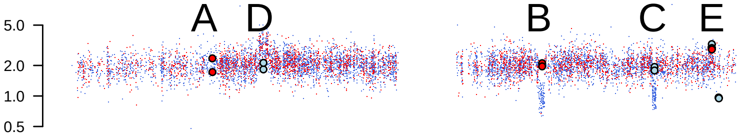
$$f(\widehat{\mathrm{CN}}, \widehat{\mathrm{GT}} | \obslash) \quad = \quad f(\widehat{\mathrm{CN}} | \obslash) \times f(\widehat{\mathrm{GT}} | \obslash) \quad = \quad f\left\{ \widehat{\mathrm{CN}} \mid \searrow \right\} \times f\left\{ \widehat{\mathrm{GT}} \mid \bigcirc \right\}$$

If the state for a particular SNP is *Loss*, we have

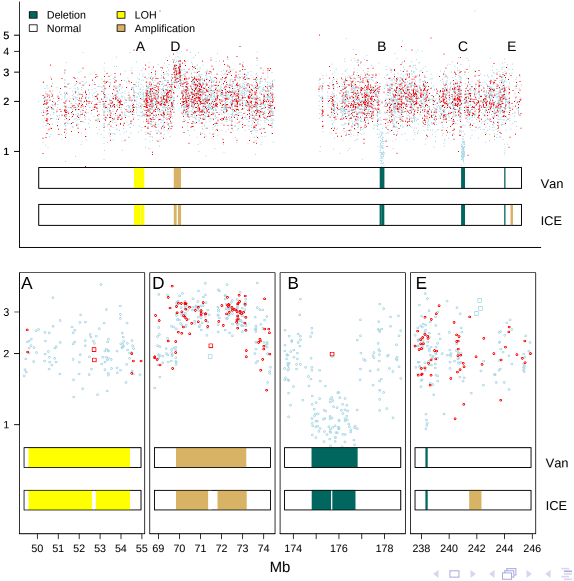
$$f\left\{ \widehat{\mathrm{GT}}, \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \bigcirc \right\} \quad = \quad f\left\{ \widehat{\mathrm{GT}} \mid \bigcirc \right\} \times f\left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \widehat{\mathrm{GT}}, \bigcirc \right\}.$$

Incorporating Genotype Uncertainty - CNVs

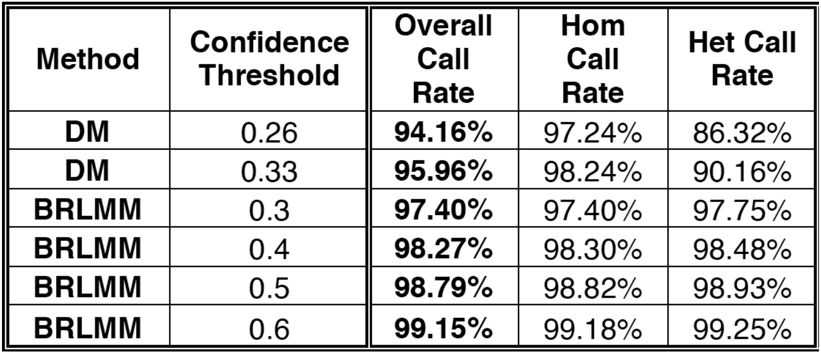
Scharpf et al (2008) *Hidden Markov models for the assessment...* The Annals of Applied Statistics, 2(2): 687-713.



Incorporating Genotype Uncertainty - CNVs



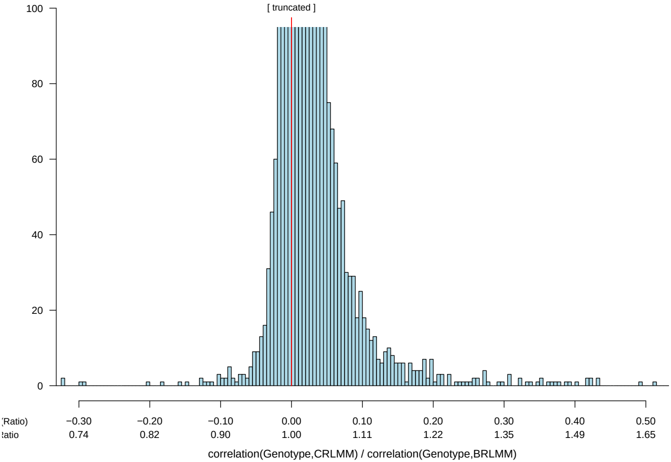
Missingness at Random?



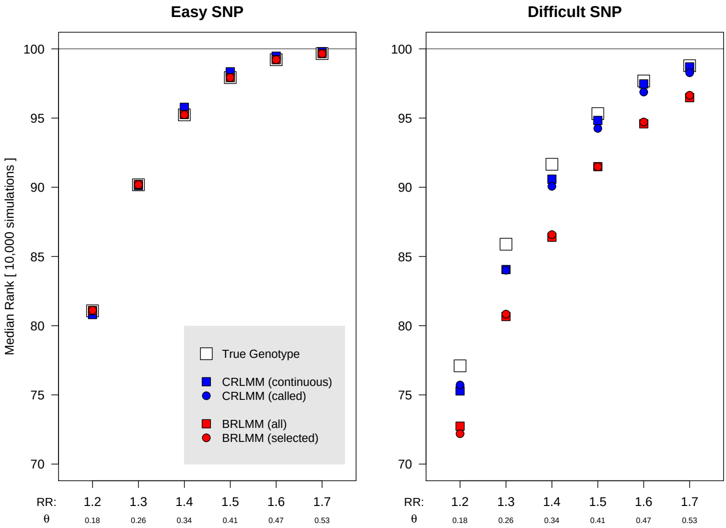
Method	Confidence Threshold	Overall Call Rate	Hom Call Rate	Het Call Rate
DM	0.26	**94.16%**	97.24%	86.32%
DM	0.33	**95.96%**	98.24%	90.16%
BRLMM	0.3	**97.40%**	97.40%	97.75%
BRLMM	0.4	**98.27%**	98.30%	98.48%
BRLMM	0.5	**98.79%**	98.82%	98.93%
BRLMM	0.6	**99.15%**	99.18%	99.25%

From the “white paper”,
`http://www.affymetrix.com/support/technical/product_updates/brlmm_algorithm.affx`

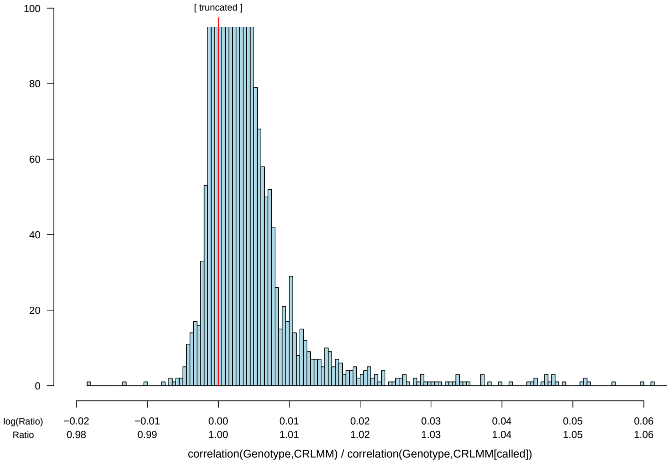
Incorporating Genotype Uncertainty



Incorporating Genotype Uncertainty



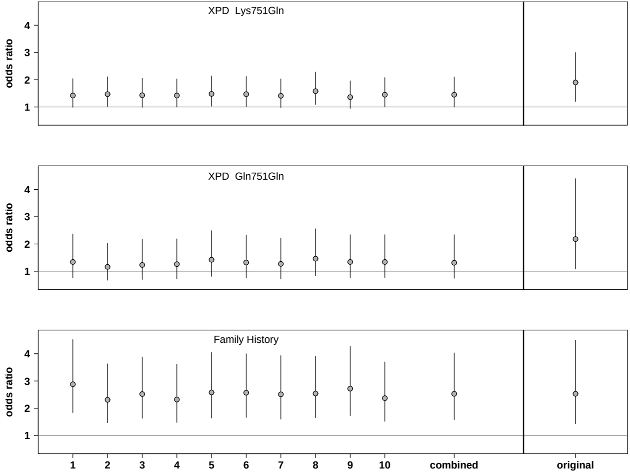
Incorporating Genotype Uncertainty



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 et al (2006) *Imputation methods to improve inference...* Genetic Epidemiology, 30(8): 690-702.

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)

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 et al (2006) *Polymorphisms of the DNA repair genes XPD and...* Breast Cancer Res Treat, 95(1): 73-80.

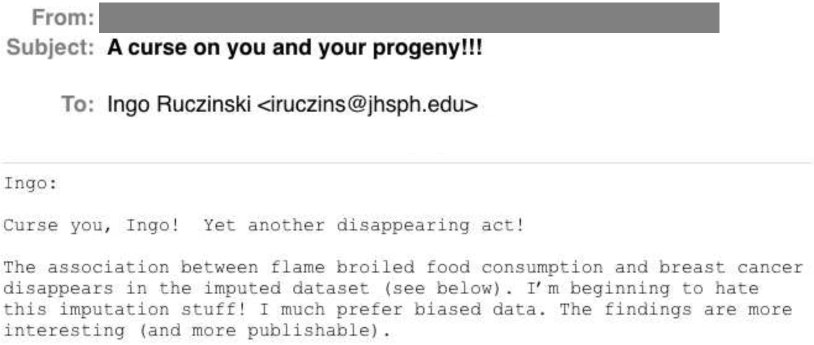
Why Become a Biostatistician?

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

- Qing Li, Tom Louis, Dani Fallin
- Rob Scharpf, Giovanni Parmigiani
- Rafael Irizarry, Benilton Carvalho
- Marvin Newhouse, Jiong Yang

http://biostat.jhsph.edu/∼iruczins/