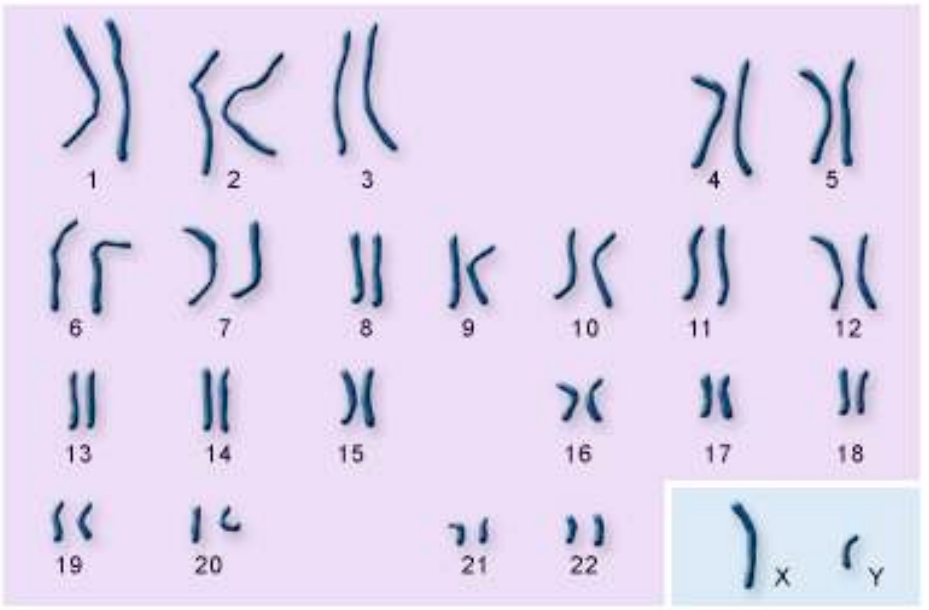
Inference for SNPchip Data in the Presence of Genotype and Copy Number Uncertainty

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

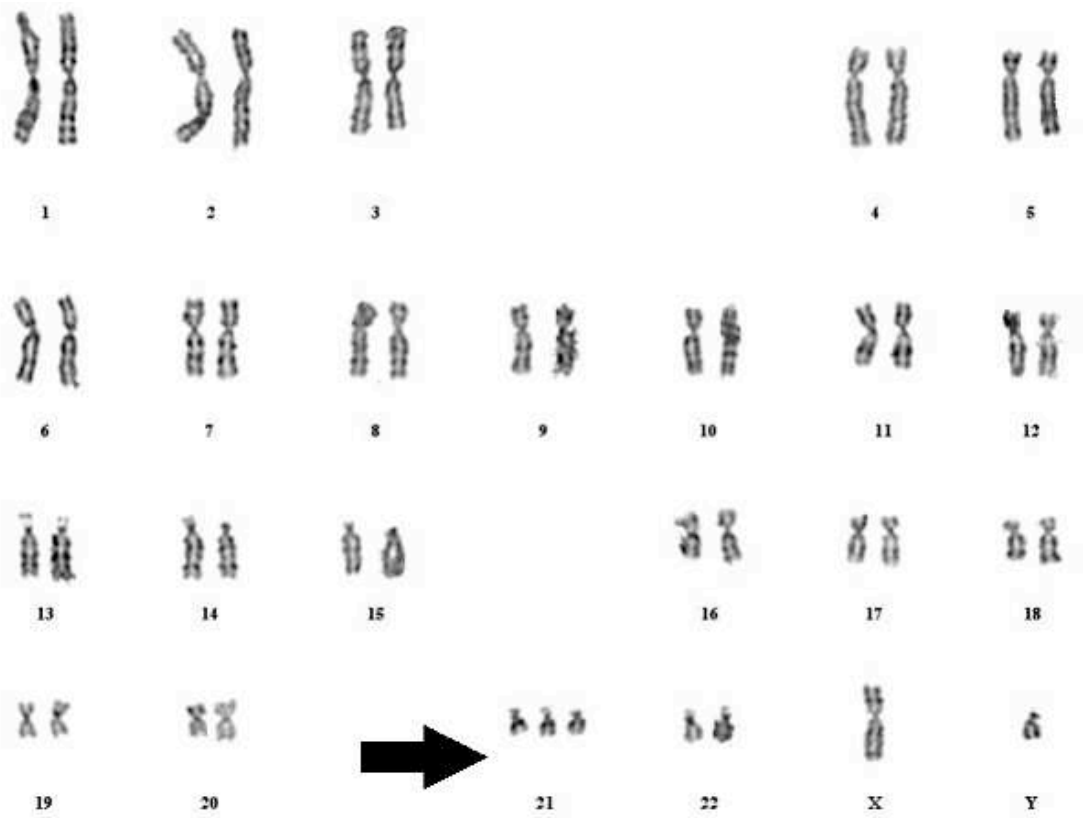
July 27, 2009

Karyotypes

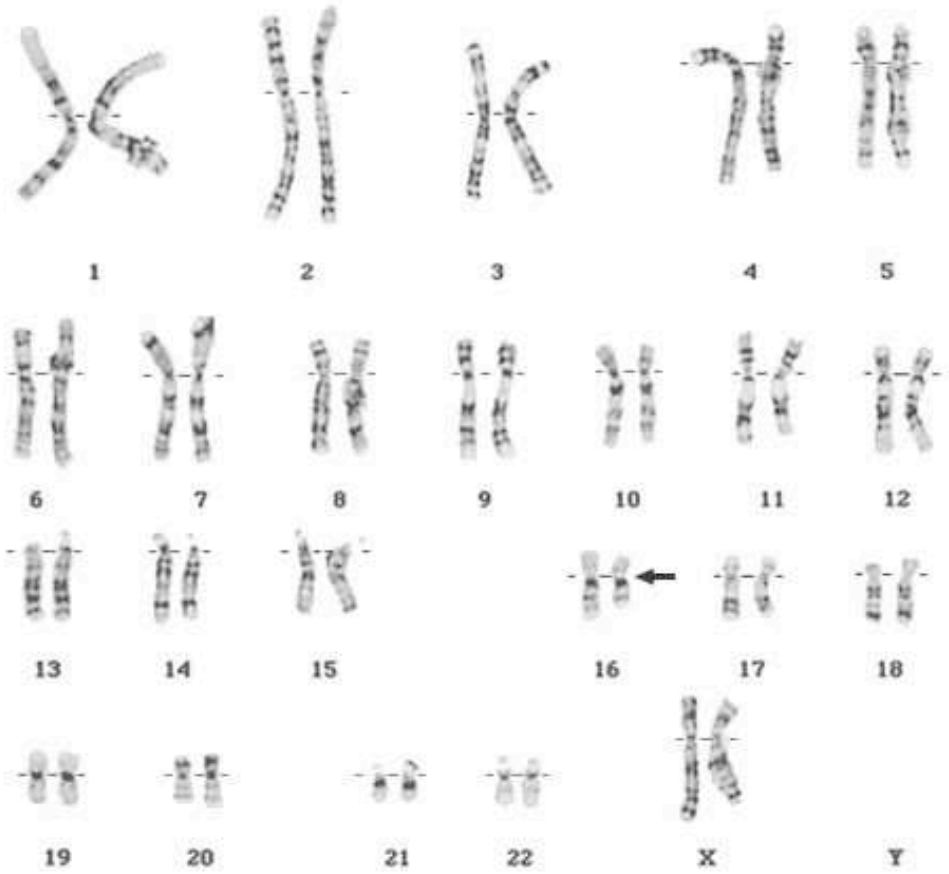


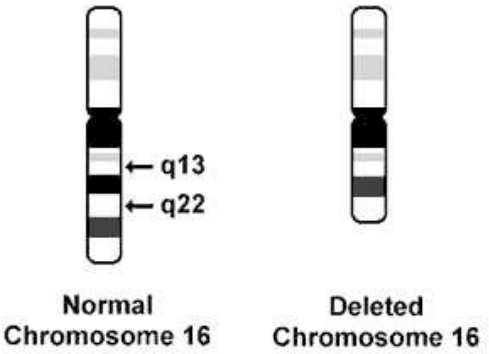
U.S. National Library of Medicine

Trisomy



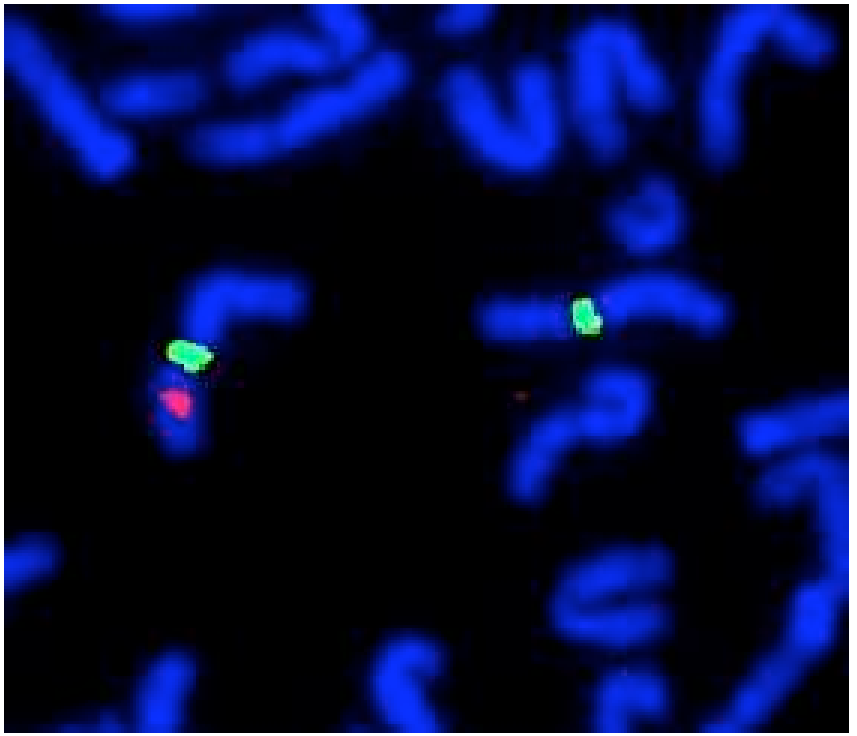
Karyotypes





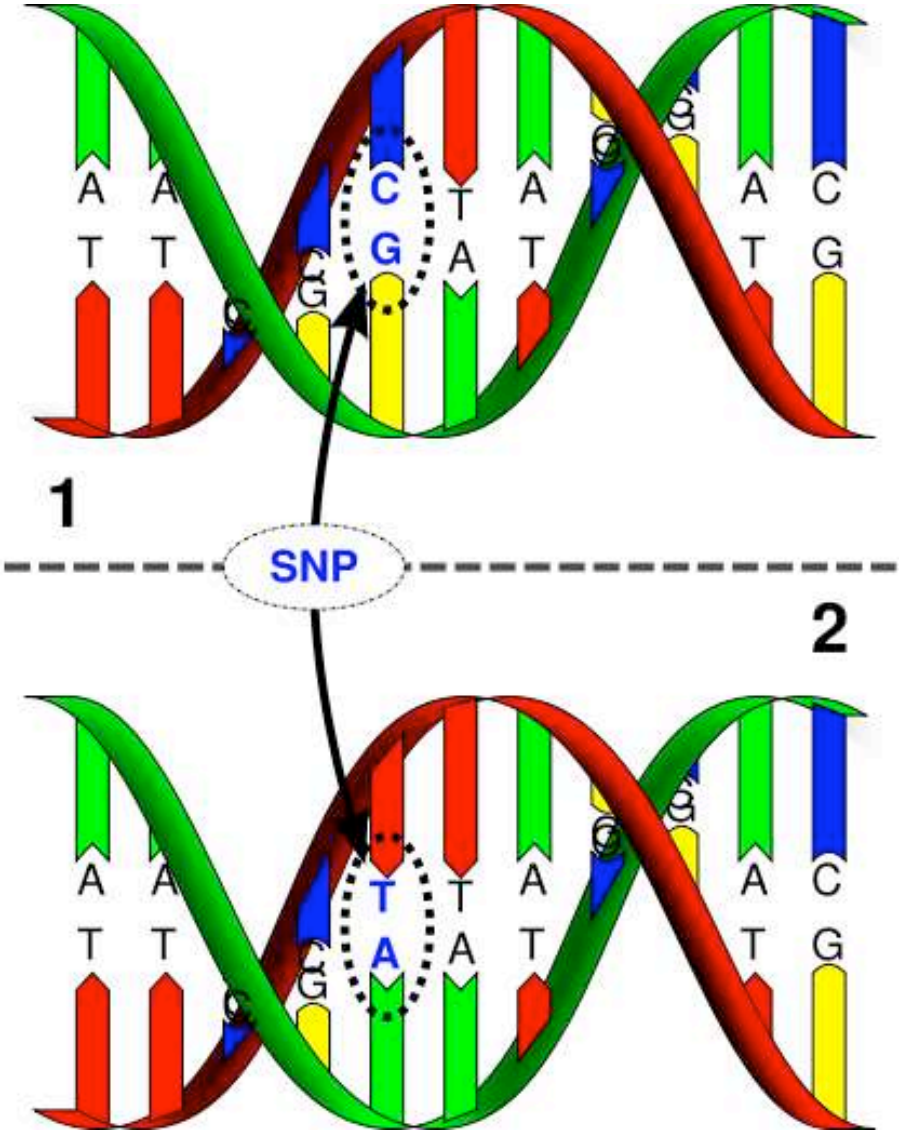
General Cytogenetics Information http://members.aol.com/chrominfo/

FISH

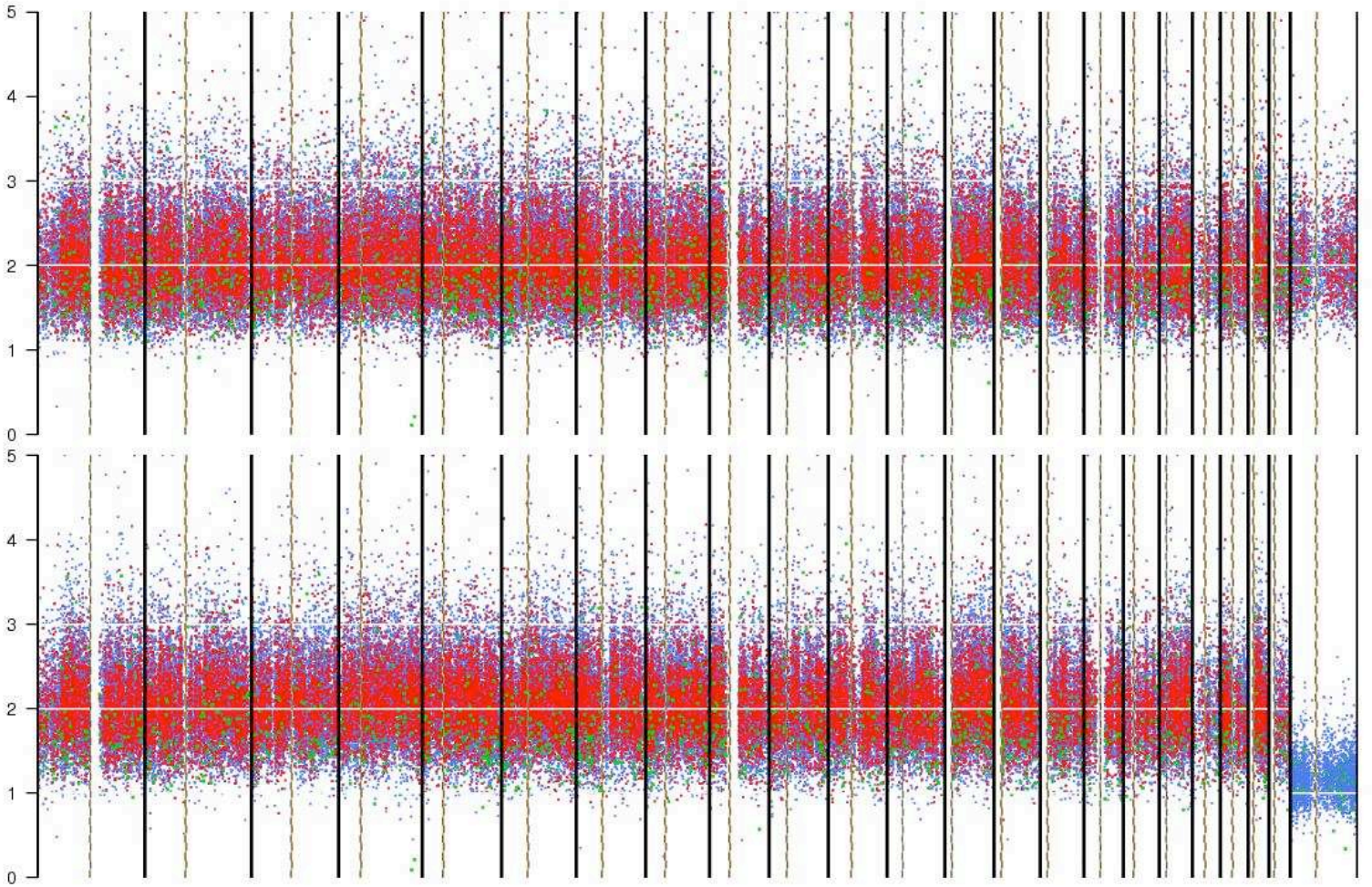


Courtesy of the Pevsner Laboratory

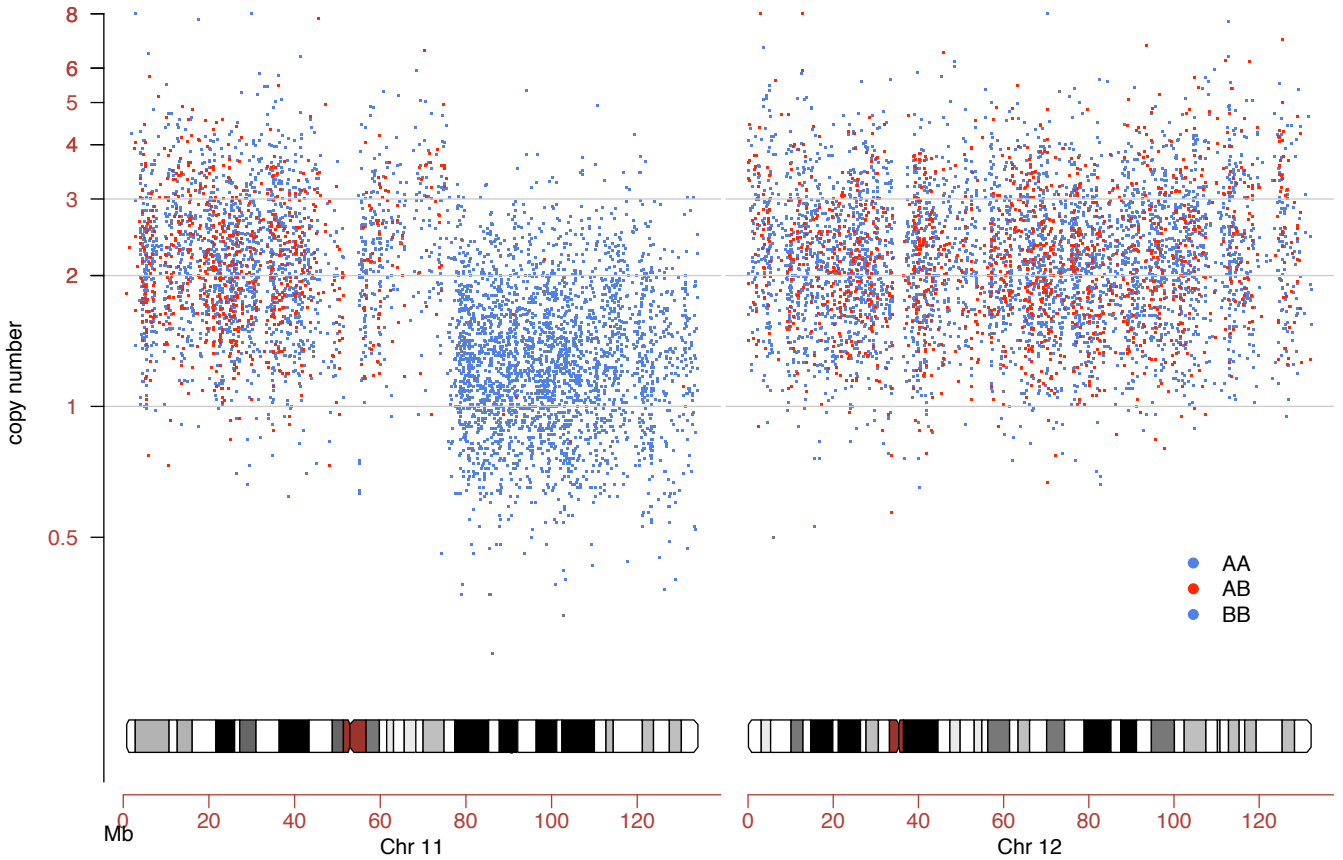
Single nucleotide polymorphisms



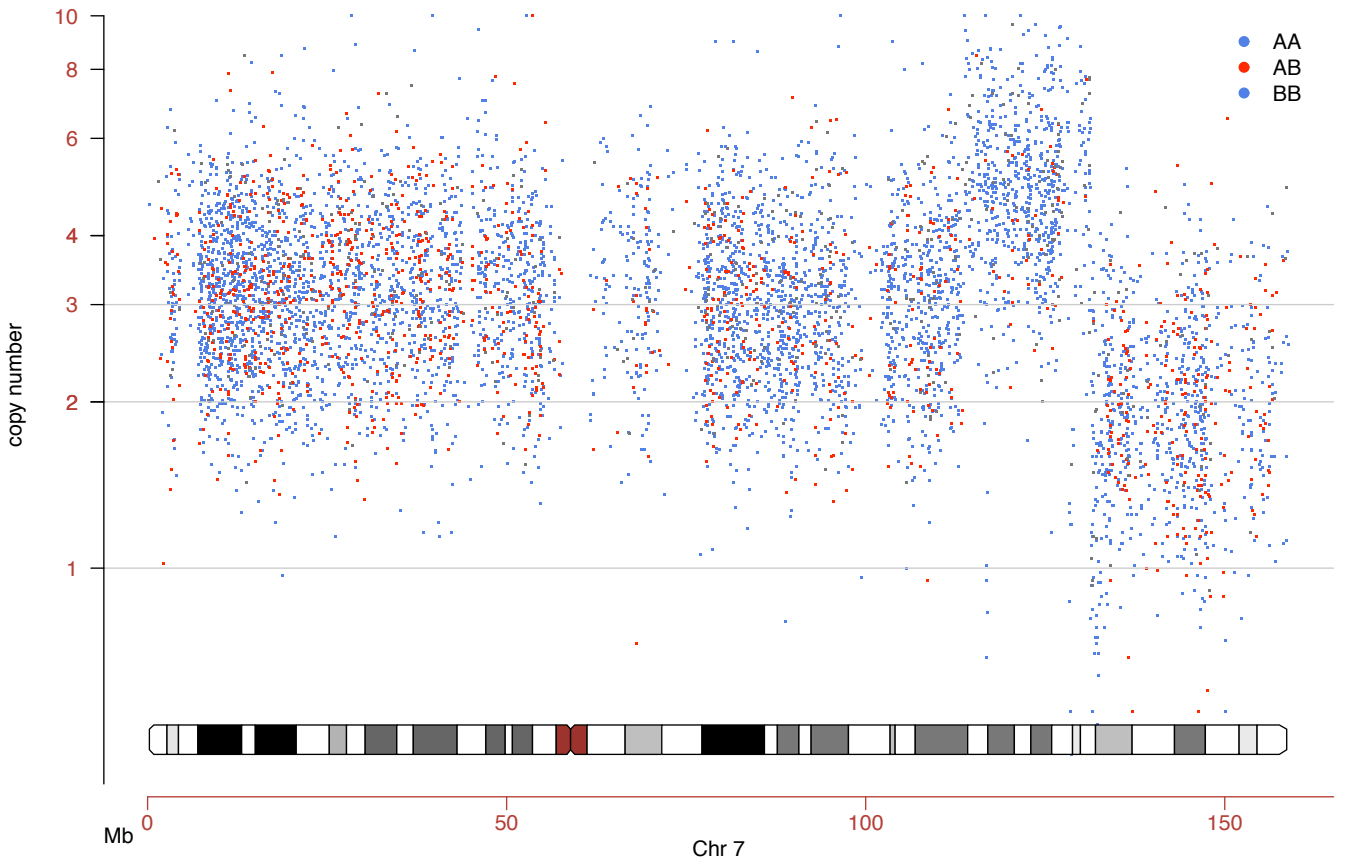
SNP chip data



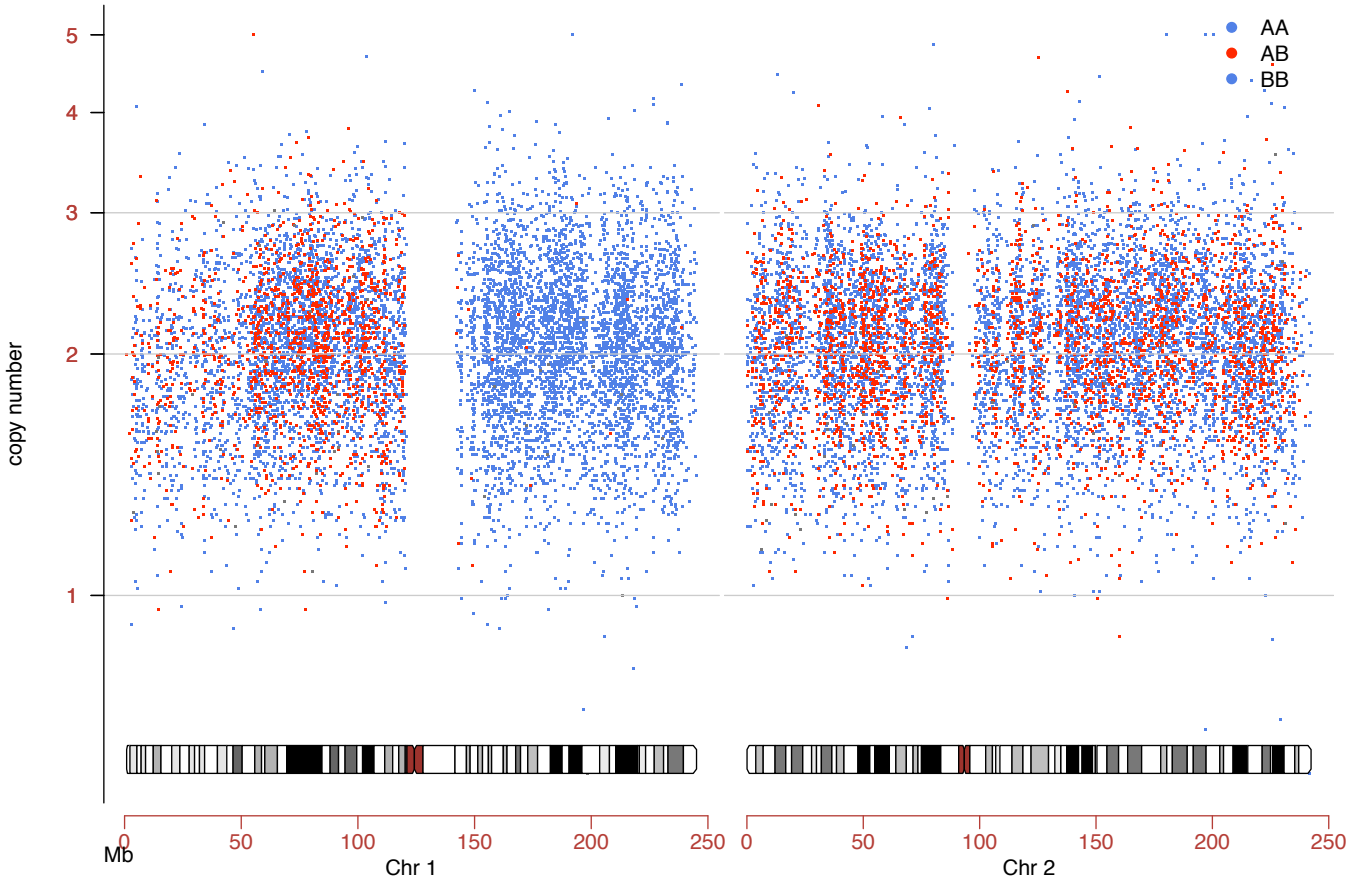
Deletion


8
6
5
4
3
2
1
0.5
copy number
AA
AB
BB
0
20
40
60
80
100
120
Mb
Chr 11
Chr 12

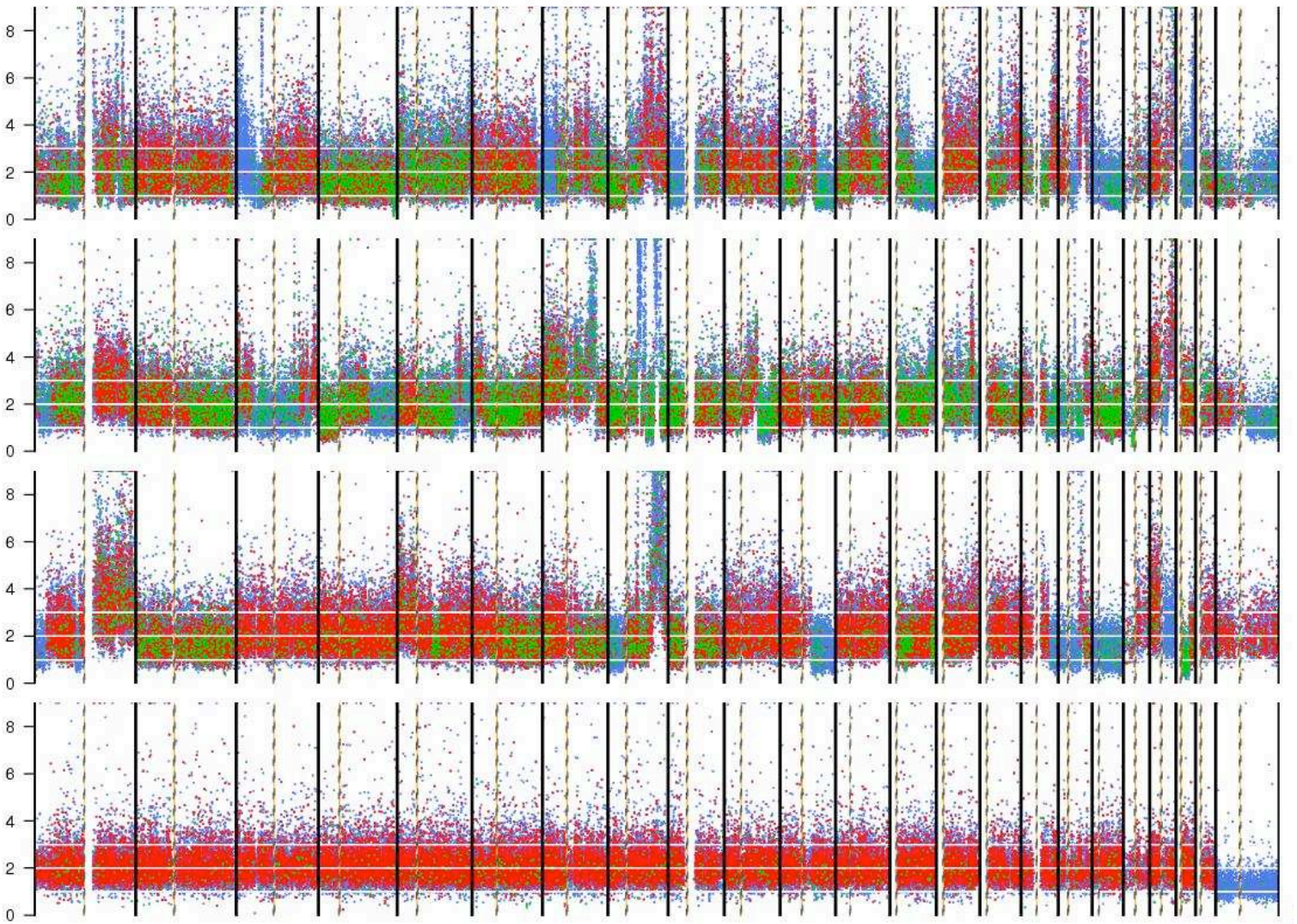
Amplification



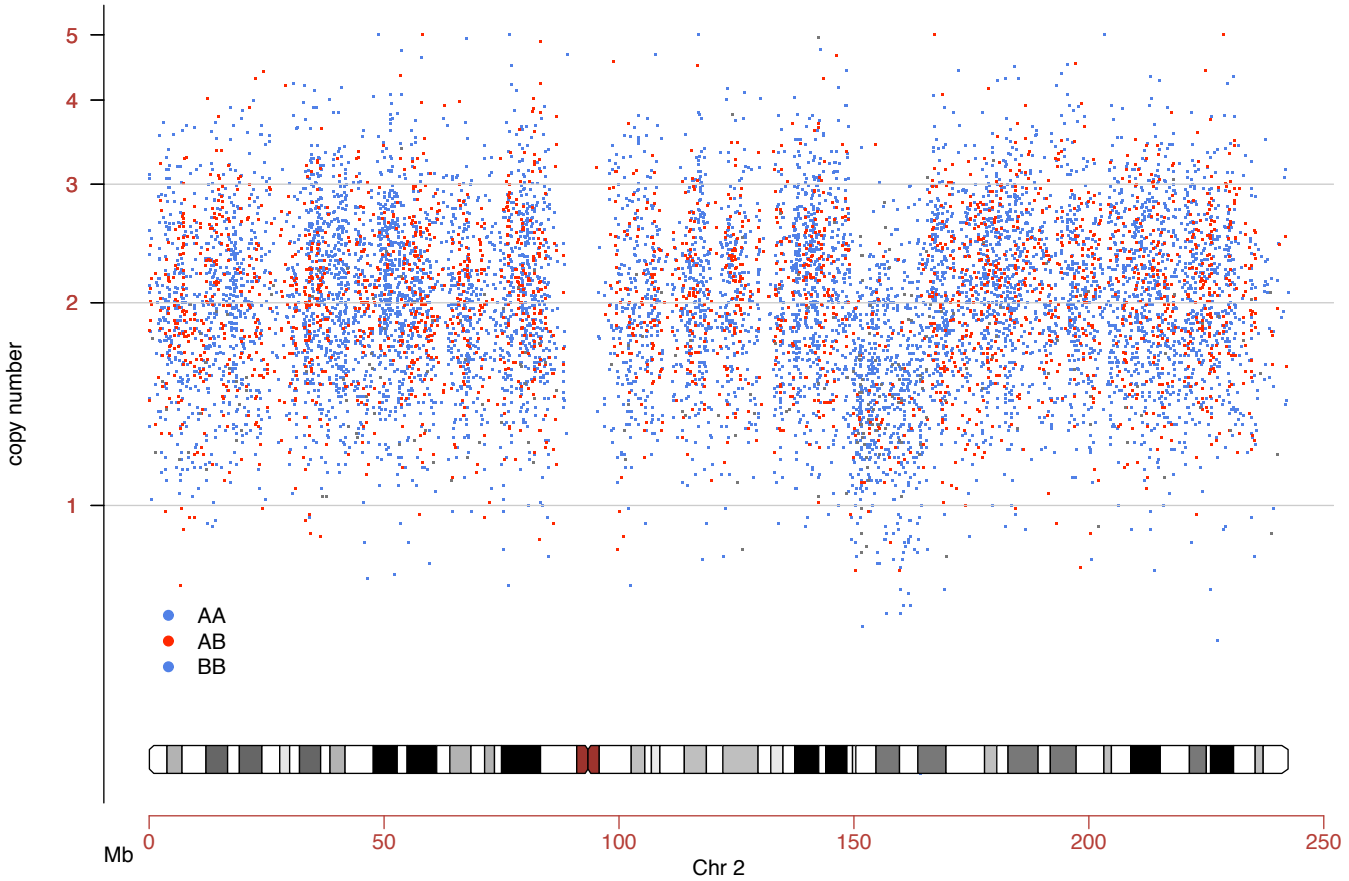
Uniparental Isodisomy



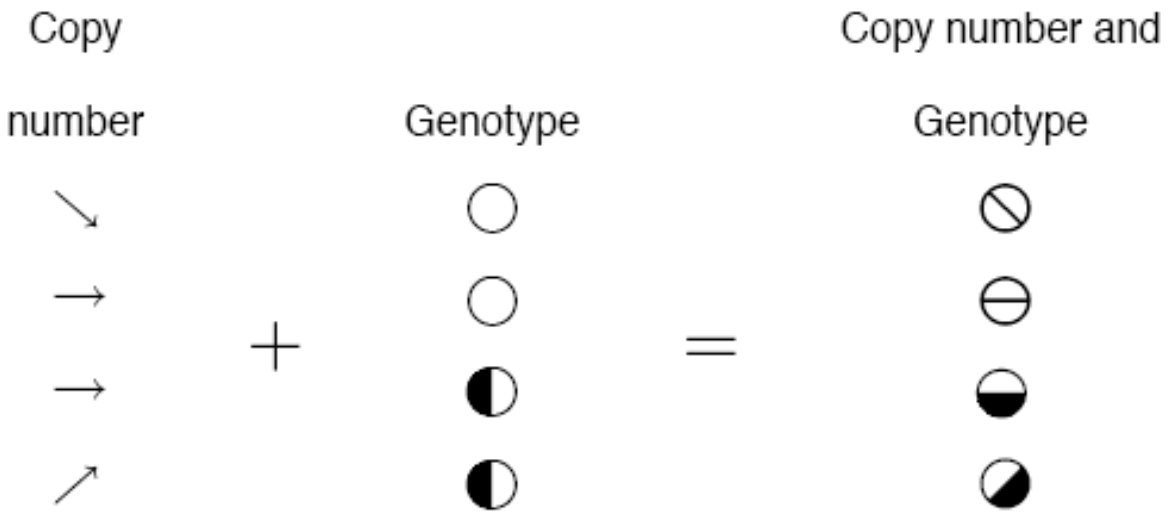
Cancer samples



Mosaicism

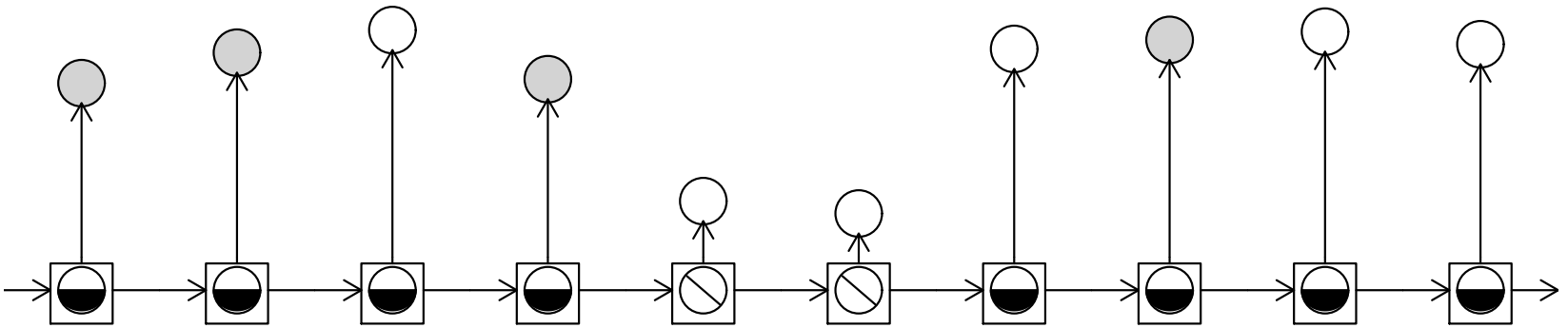


Hidden states



The structure of the data we observe

→ At each SNP, we observe a noisy measure of the true copy number and genotype (and possibly also measures of confidence in those estimates).



Hidden Markov models

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.

The Vanilla HMM components

- Observations $\widehat{\text{CN}}$ and $\widehat{\text{GT}}$
- Hidden states
- Initial state probability distribution
- Transition probabilities
- Emission probabilities

Transition probabilities

Following suggestions in the literature, we model the transition probabilities as a function of the distance d between SNPs.

Specifically, let $\theta(d) \equiv 1 - e^{-2d}$ denote the probability that SNP i is not informative (I^c) for SNP at $i+1$.

For example:

$$
\begin{aligned}
\tau_{\blacklozenge|\oslash}(d) &= P\{\blacklozenge_{i+1} \mid \oslash_i, d\} \\
&= P\{\blacklozenge_{i+1}, I \mid \oslash_i, d\} + P\{\blacklozenge_{i+1}, I^c \mid \oslash_i, d\} \\
&= P\{\blacklozenge_{i+1} \mid I, \oslash_i, d\} \times P\{I \mid \oslash_i, d\} + \\
&\quad P\{\blacklozenge_{i+1} \mid I^c, \oslash_i, d\} \times P\{I^c \mid \oslash_i, d\} \\
&= P\{\blacklozenge\} \times \theta(d).
\end{aligned}
$$

Emission probabilities

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

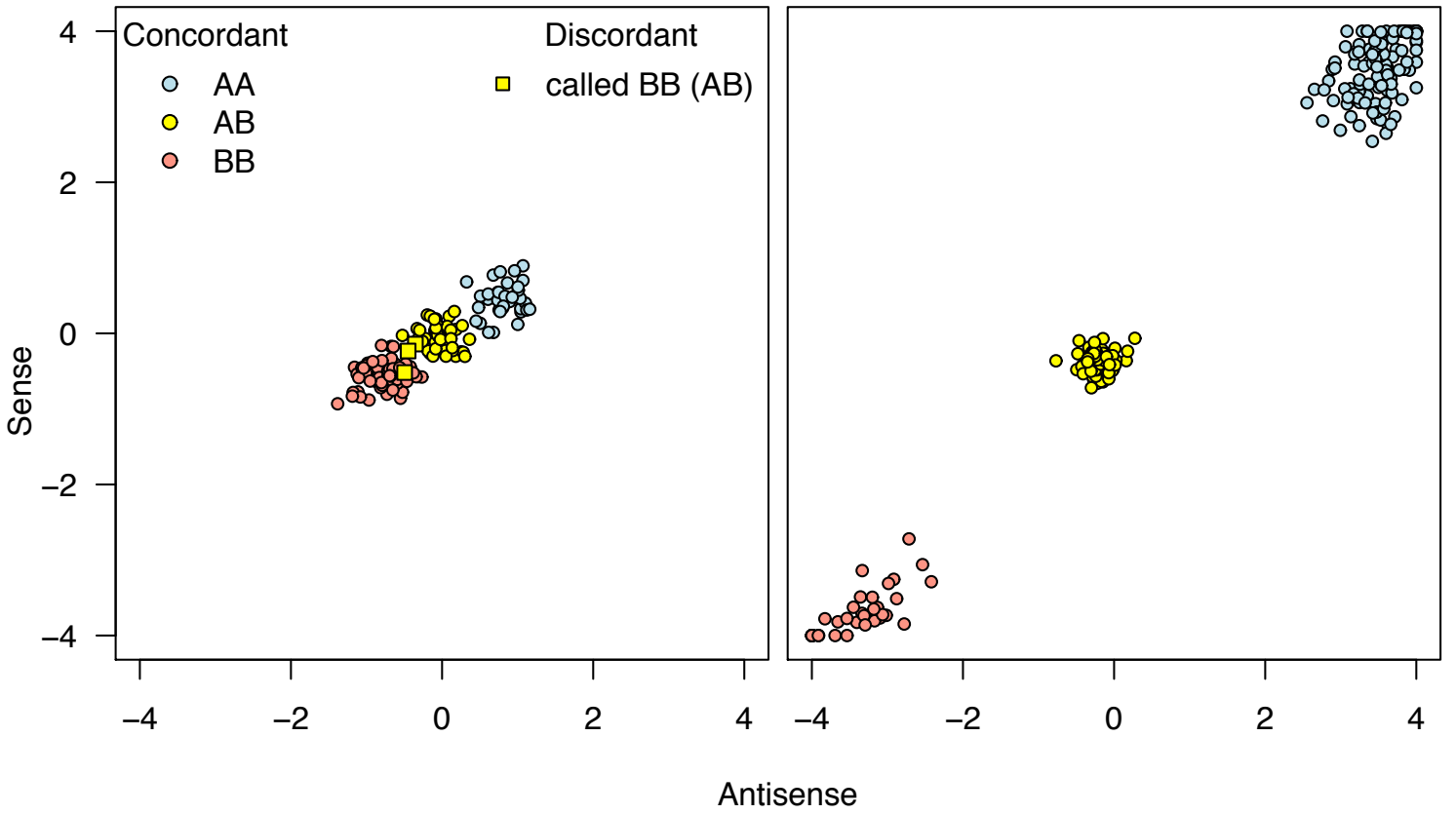
For example:

$$
\begin{aligned}
f(\widehat{\mathrm{CN}}, \widehat{\mathrm{GT}} | \oslash) &= f(\widehat{\mathrm{CN}} | \oslash) \times f(\widehat{\mathrm{GT}} | \oslash) \\
&= f\left\{ \widehat{\mathrm{CN}} \mid \searrow \right\} \times f\left\{ \widehat{\mathrm{GT}} \mid \bigcirc \right\} \\
&= \beta_{\searrow} \left\{ \widehat{\mathrm{CN}} \right\} \times \beta_{\bigcirc} \left\{ \widehat{\mathrm{GT}} \right\}.
\end{aligned}
$$

Some geezer down the hallway...

More information

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



Integrating confidence estimates for genotype calls

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

We can estimate from Hapmap the following densities:

$$f\left\{ S_{\widehat{\text{HOM}}} \mid \widehat{\text{HOM}}, \text{HOM} \right\}, f\left\{ S_{\widehat{\text{HOM}}} \mid \widehat{\text{HOM}}, \text{HET} \right\}, f\left\{ S_{\widehat{\text{HET}}} \mid \widehat{\text{HET}}, \text{HOM} \right\}, f\left\{ S_{\widehat{\text{HET}}} \mid \widehat{\text{HET}}, \text{HET} \right\}.$$

$\longrightarrow$ Note:

$$f\left\{ S_{\widehat{\text{HOM}}} \mid \widehat{\text{HOM}}, \bigcirc \right\} \approx f\left\{ S_{\widehat{\text{HOM}}} \mid \widehat{\text{HOM}}, \text{HOM} \right\}$$

$$f\left\{ S_{\widehat{\text{HET}}} \mid \widehat{\text{HET}}, \bigcirc \right\} \approx f\left\{ S_{\widehat{\text{HET}}} \mid \widehat{\text{HET}}, \text{HOM} \right\}.$$

Emission Probabilities - Loss

Recall that

$$\begin{aligned} f(\widehat{\mathrm{CN}}, \widehat{\mathrm{GT}} | \oslash) &= f(\widehat{\mathrm{CN}} | \oslash) \times f(\widehat{\mathrm{GT}} | \oslash) \\ &= f\left\{ \widehat{\mathrm{CN}} \mid \searrow \right\} \times f\left\{ \widehat{\mathrm{GT}} \mid \bigcirc \right\} \\ &= \beta_{\searrow}\left\{\widehat{\mathrm{CN}}\right\} \times \beta_{\bigcirc}\left\{\widehat{\mathrm{GT}}\right\}. \end{aligned}$$

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

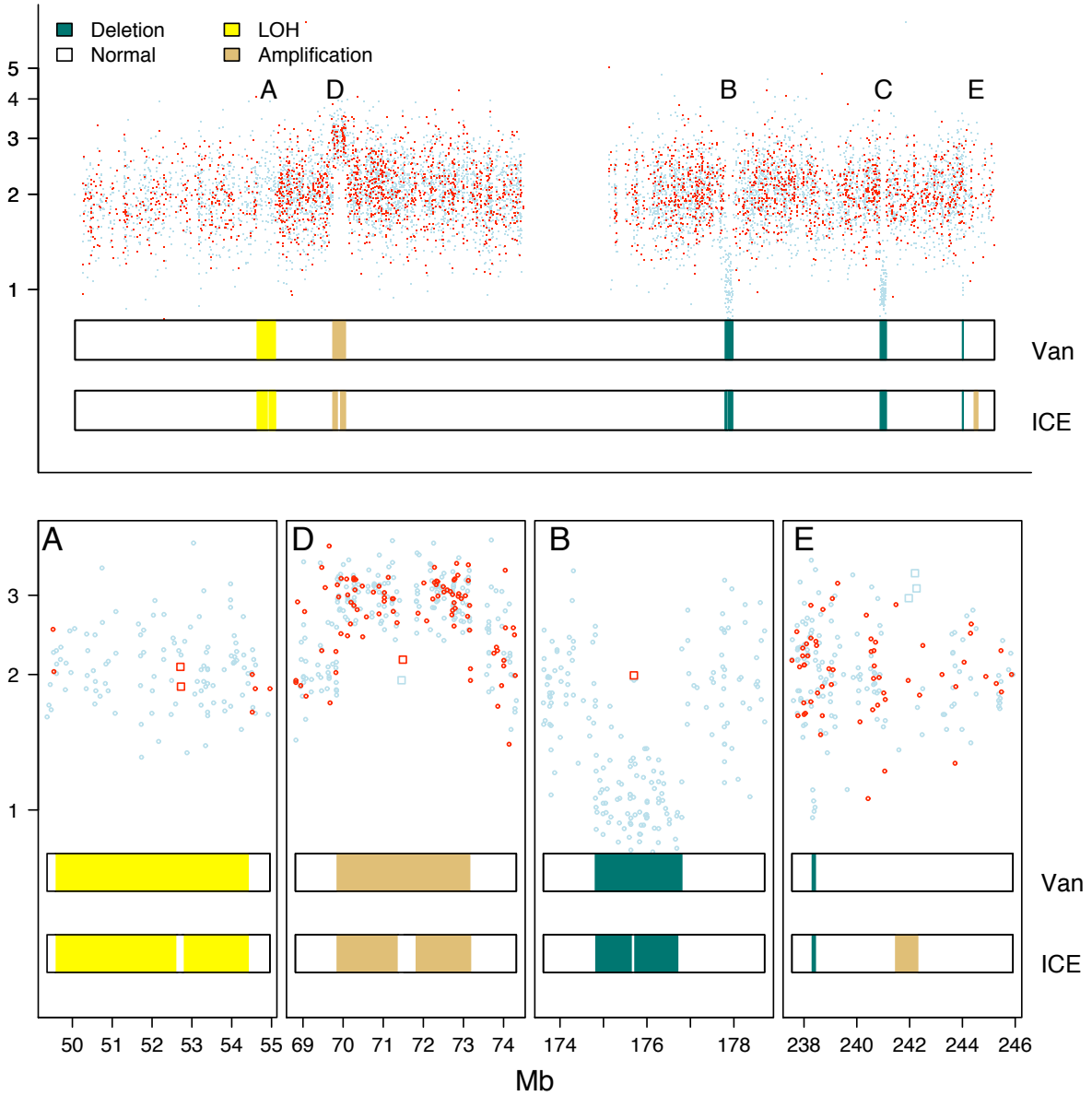
$$\beta_{\bigcirc}\left\{ \widehat{\mathrm{GT}}, \mathrm{S}_{\widehat{\mathrm{GT}}} \right\} = f\left\{ \widehat{\mathrm{GT}} \mid \bigcirc \right\} \times f\left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \widehat{\mathrm{GT}}, \bigcirc \right\}.$$

Emission Probabilities - Retention

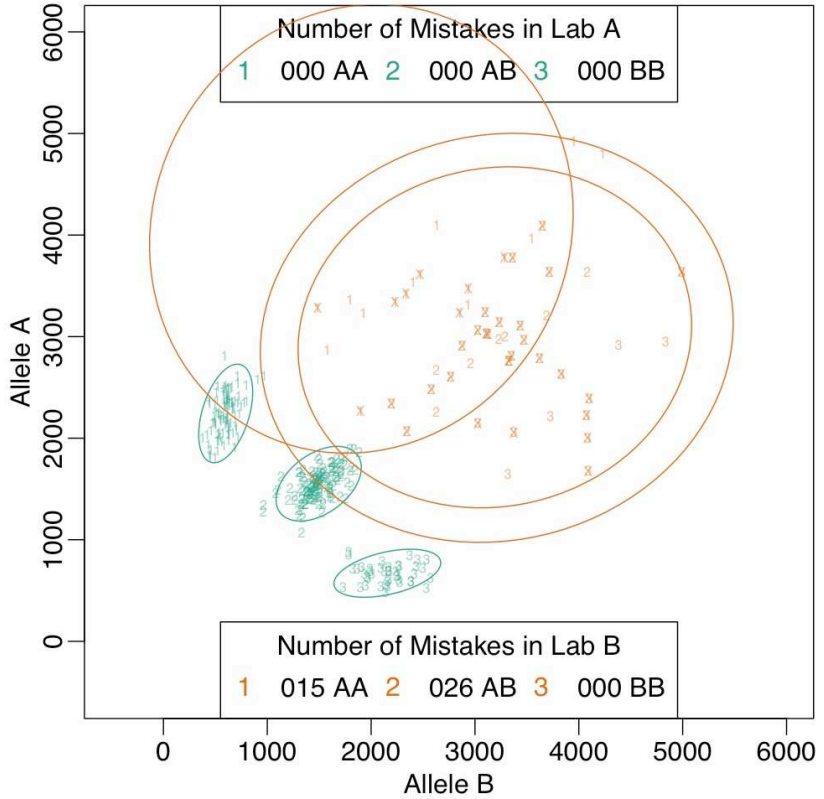
For retention, the true genotype can be HET or HOM:

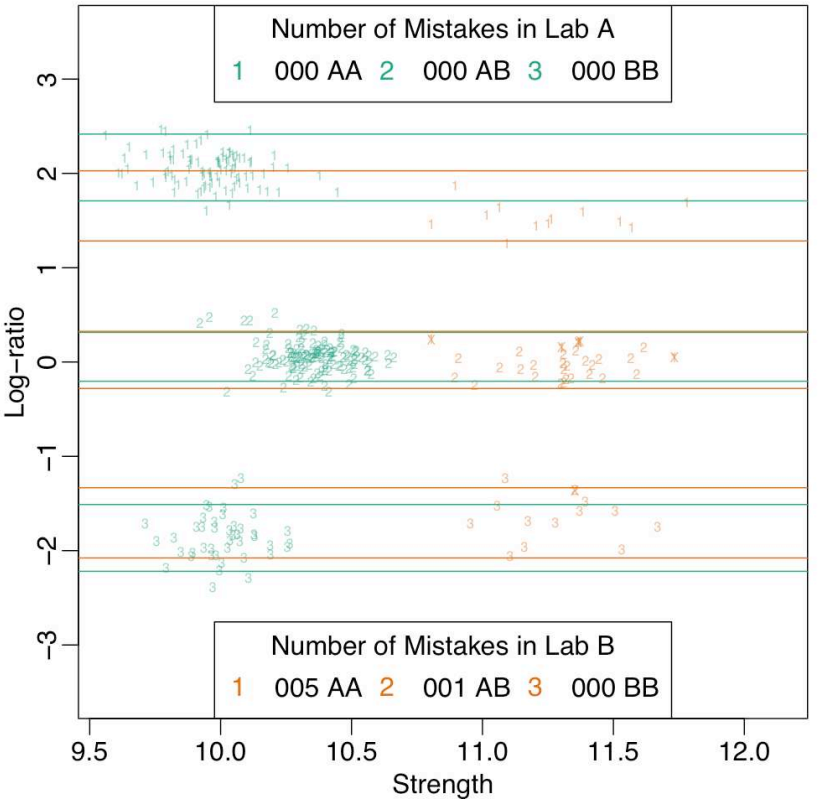
$$
\begin{aligned}
& \beta_{\CIRCLE} \left\{ \widehat{\mathrm{GT}}, \mathrm{S}_{\widehat{\mathrm{GT}}} \right\} \\
= \quad & f \left\{ \widehat{\mathrm{GT}} \mid \CIRCLE \right\} f \left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \widehat{\mathrm{GT}}, \CIRCLE \right\} \\
= \quad & f \left\{ \widehat{\mathrm{GT}} \mid \CIRCLE \right\} \left(f \left\{ \mathrm{S}_{\widehat{\mathrm{GT}}}, \mathrm{HOM} \mid \widehat{\mathrm{GT}}, \CIRCLE \right\} + f \left\{ \mathrm{S}_{\widehat{\mathrm{GT}}}, \mathrm{HET} \mid \widehat{\mathrm{GT}}, \CIRCLE \right\} \right) \\
= \quad & f \left\{ \widehat{\mathrm{GT}} \mid \CIRCLE \right\} \left(f \left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \mathrm{HOM}, \widehat{\mathrm{GT}}, \CIRCLE \right\} f \left\{ \mathrm{HOM} \mid \widehat{\mathrm{GT}}, \CIRCLE \right\} + f \left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \mathrm{HET}, \widehat{\mathrm{GT}}, \CIRCLE \right\} f \left\{ \mathrm{HET} \mid \widehat{\mathrm{GT}}, \CIRCLE \right\} \right) \\
= \quad & f \left\{ \widehat{\mathrm{GT}} \mid \CIRCLE \right\} \left(f \left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \mathrm{HOM}, \widehat{\mathrm{GT}} \right\} f \left\{ \mathrm{HOM} \mid \widehat{\mathrm{GT}}, \CIRCLE \right\} + f \left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \mathrm{HET}, \widehat{\mathrm{GT}} \right\} f \left\{ \mathrm{HET} \mid \widehat{\mathrm{GT}}, \CIRCLE \right\} \right)
\end{aligned}
$$

Vanilla and ICE HMMs for genotype and copy number



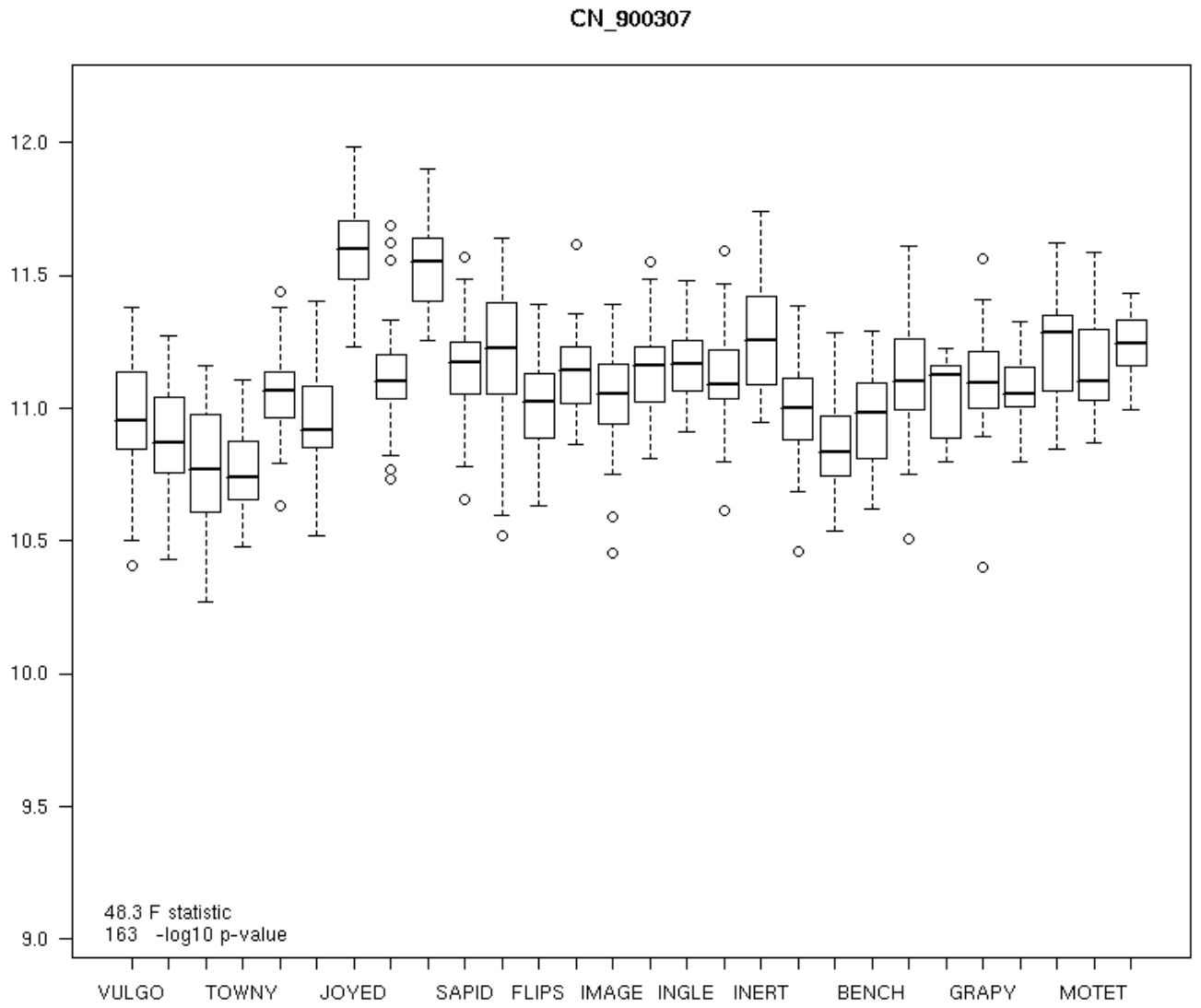
Genotypes and copy numbers





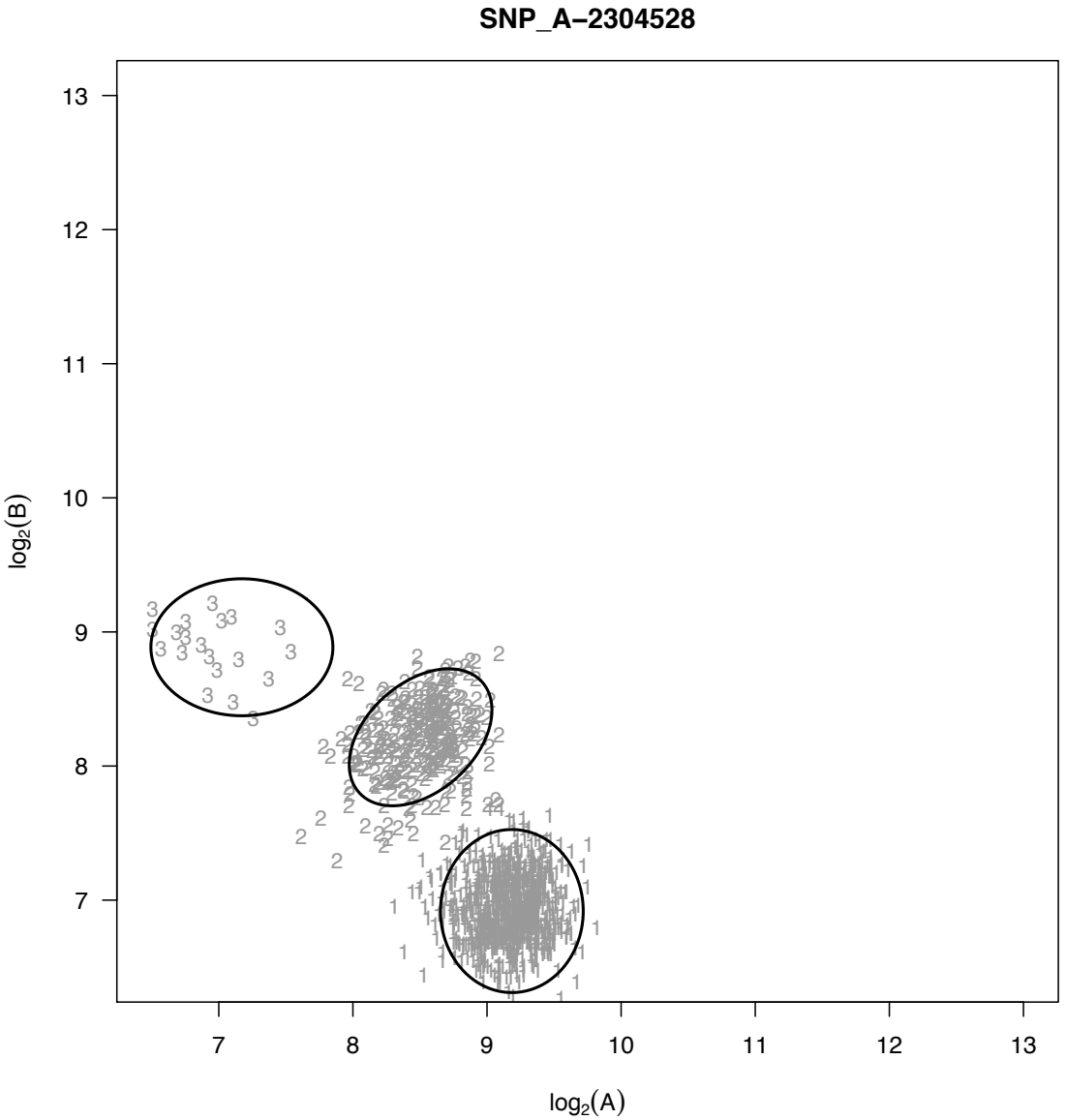
From Benilton Carvalho and Rafa Irizarry

Plate effects

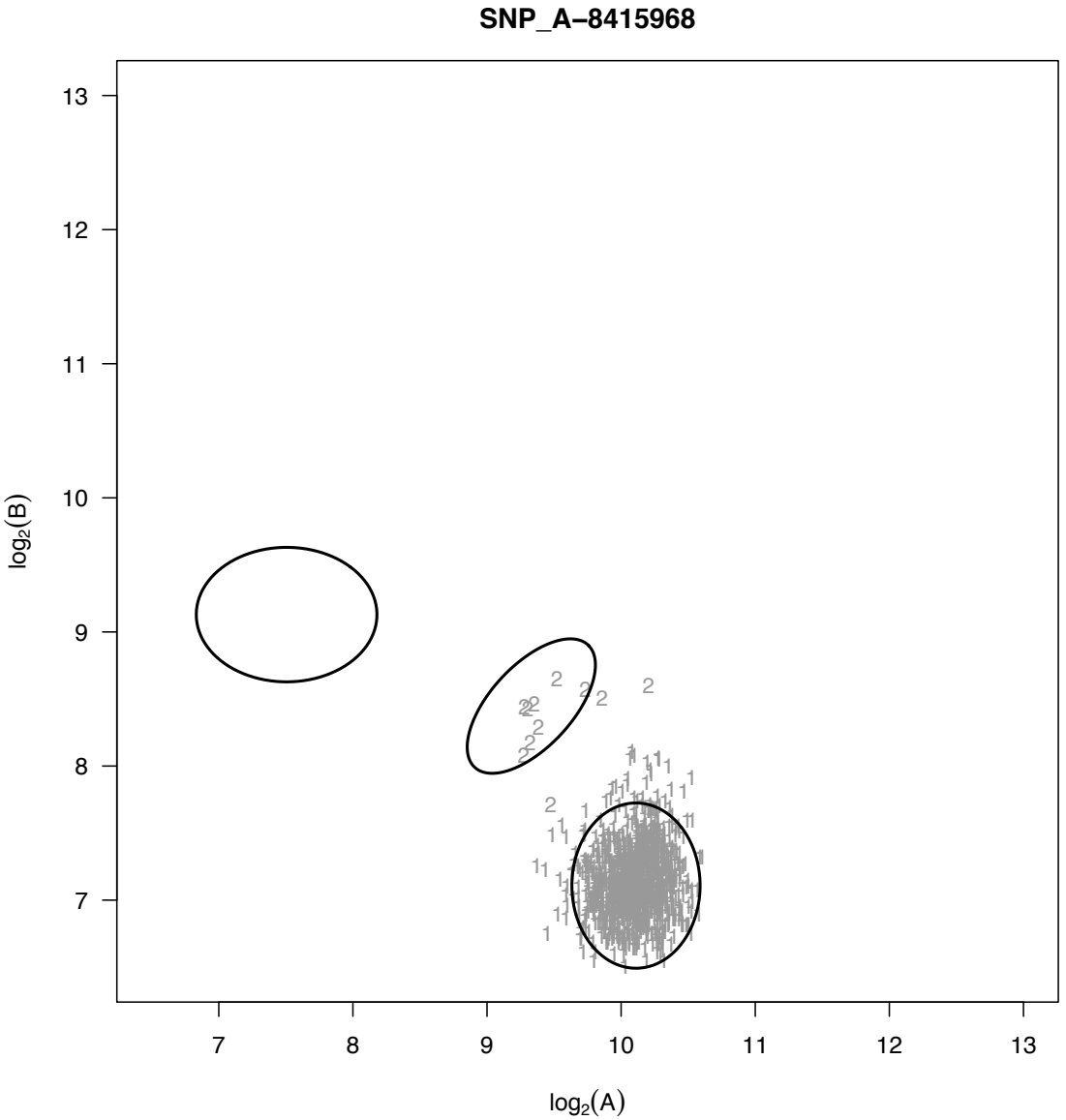


Bipolar GWAS (EA controls) from dbGap

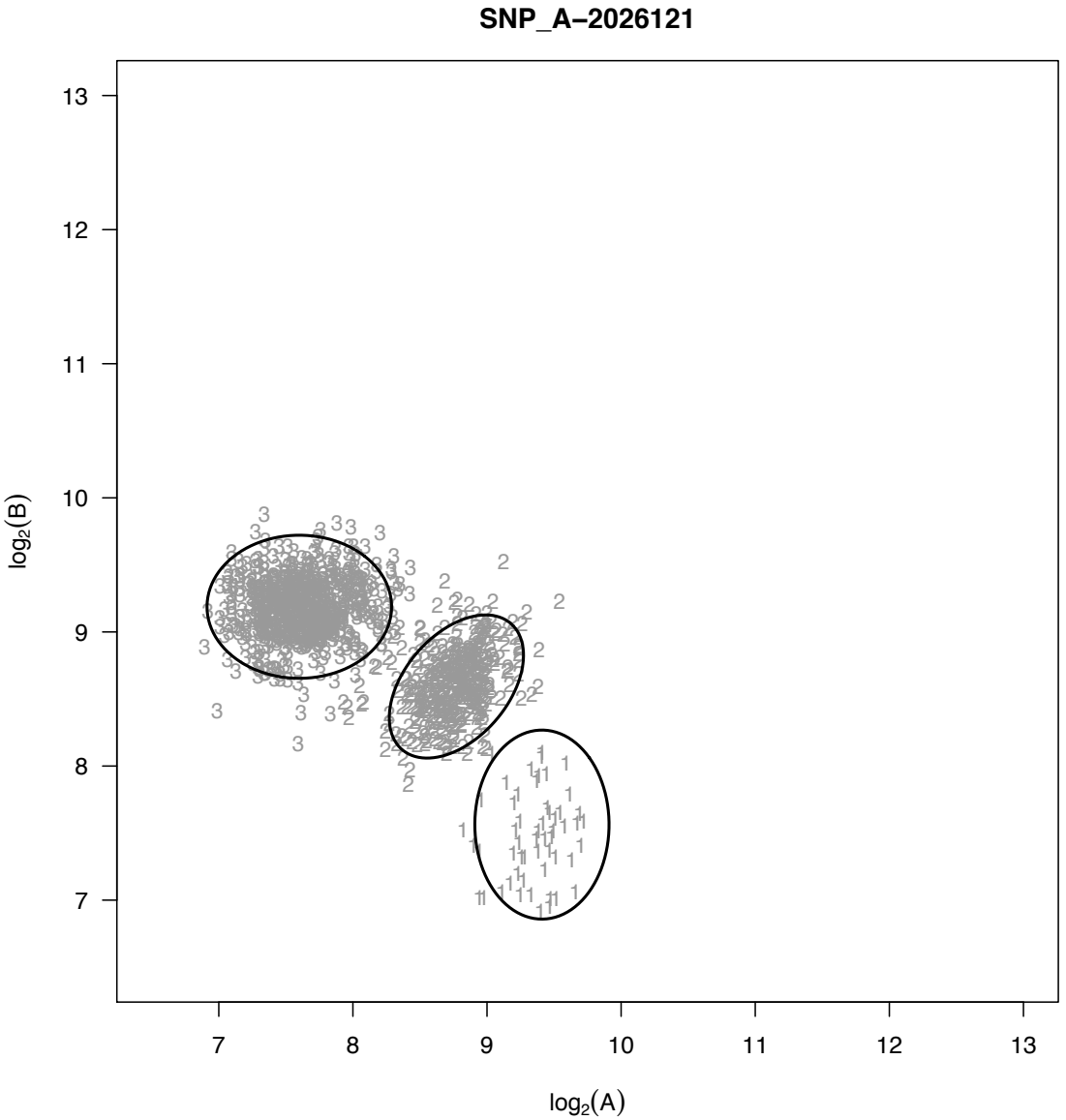
A versus B plots



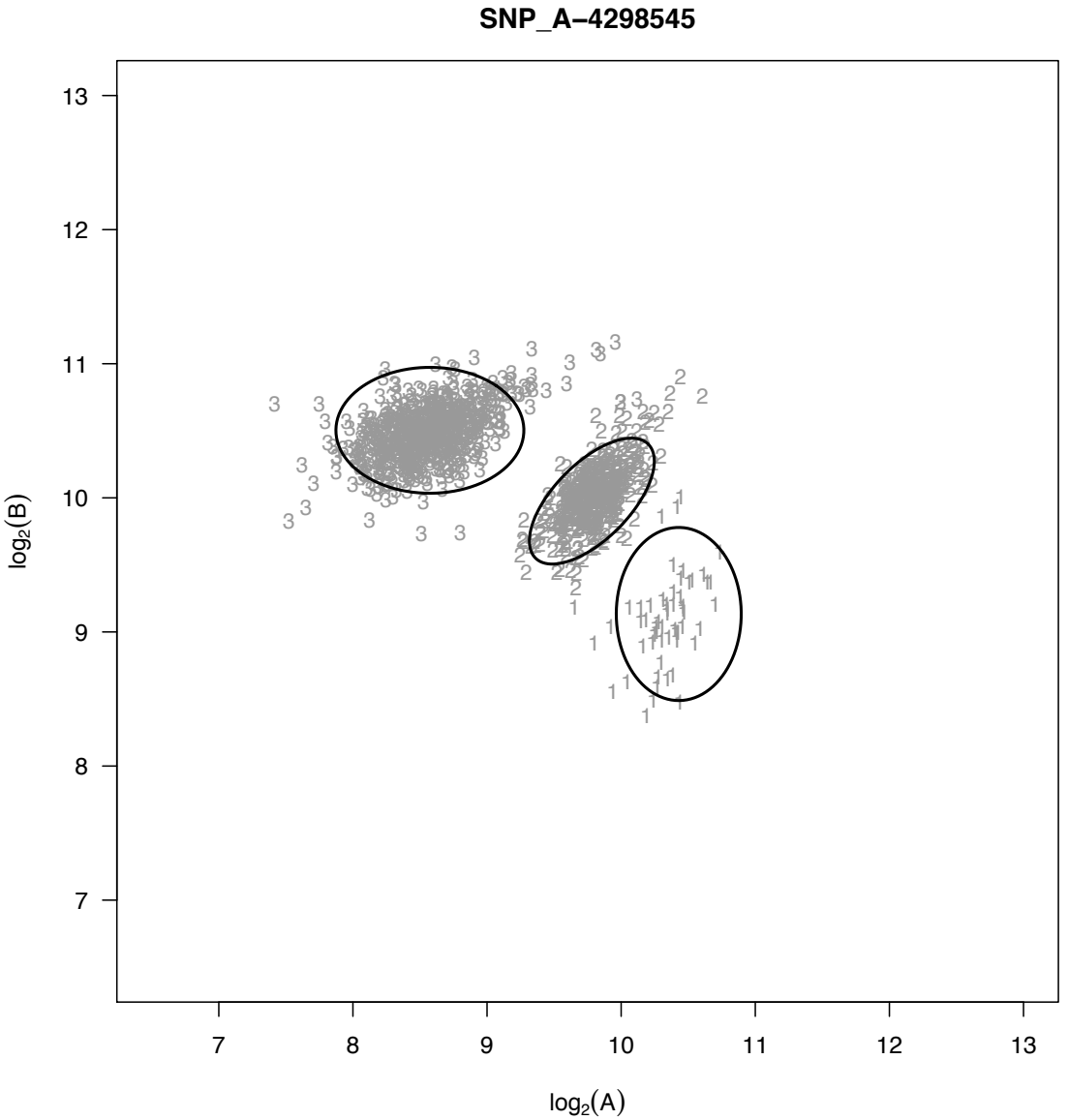
A versus B plots



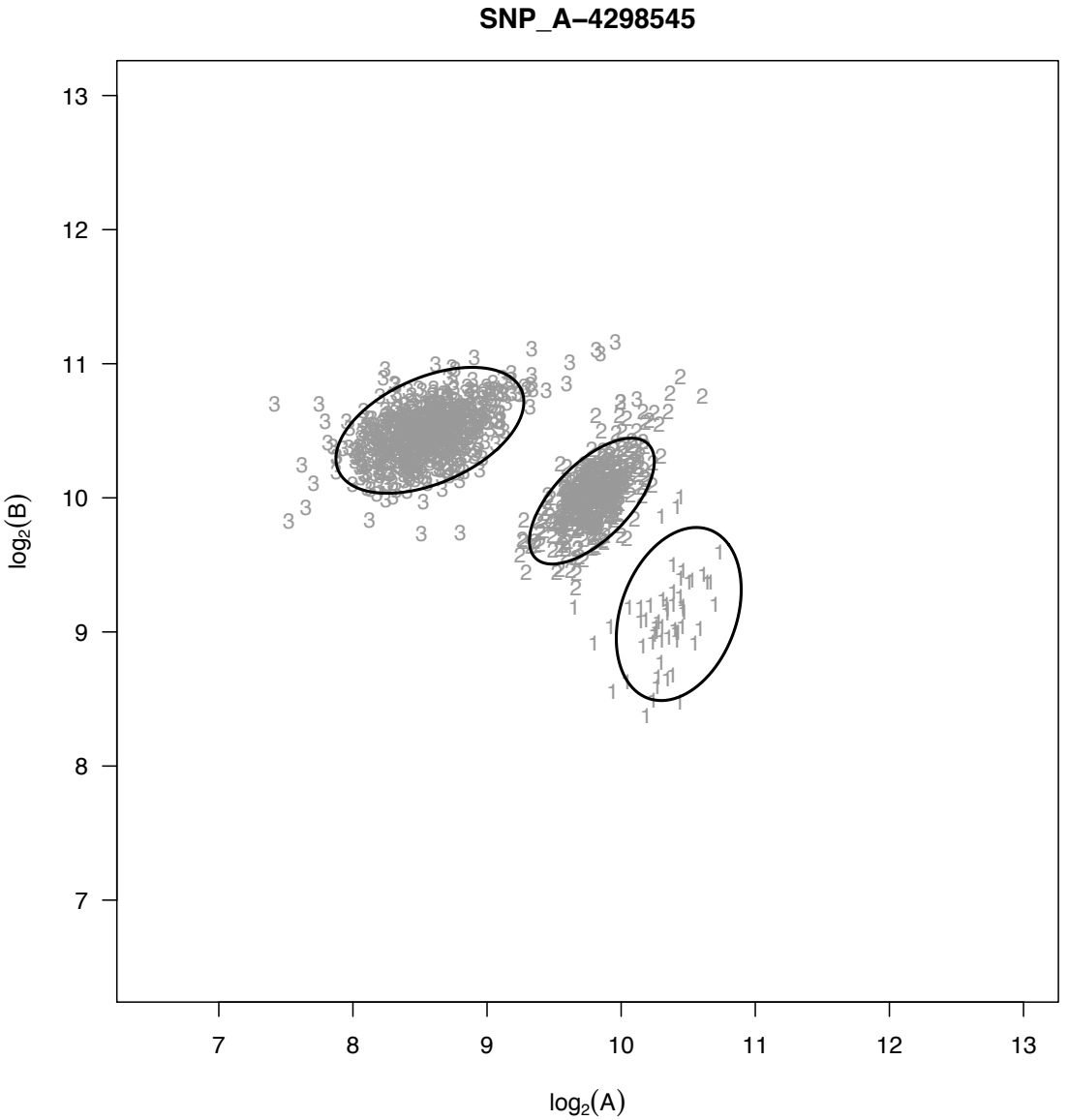
A versus B plots



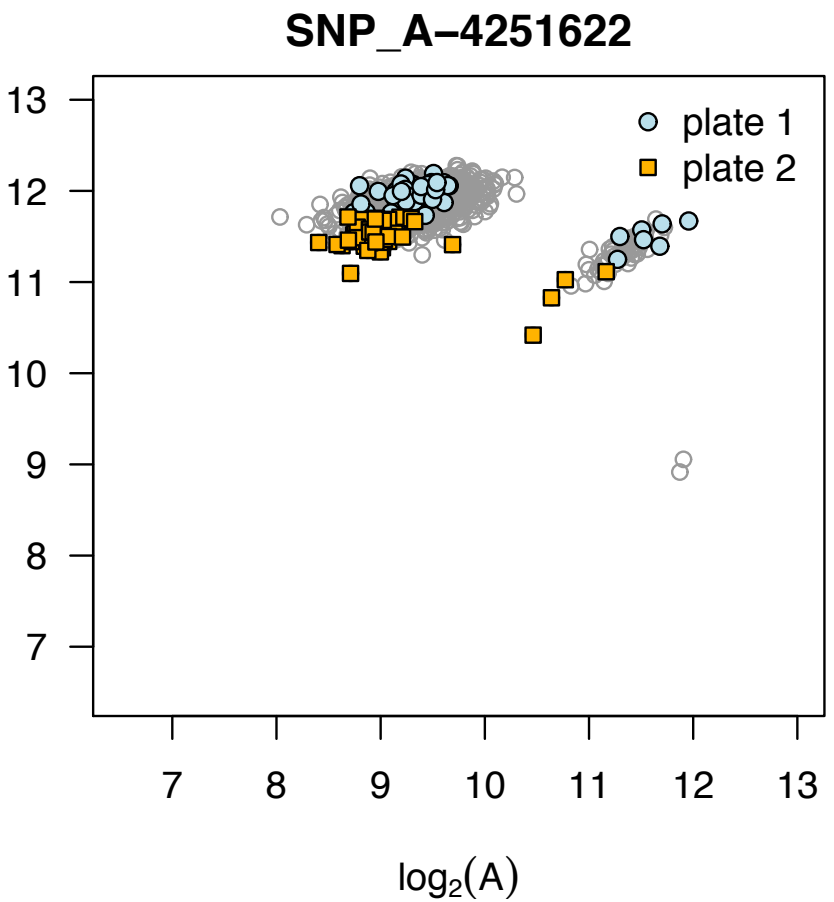
A versus B plots

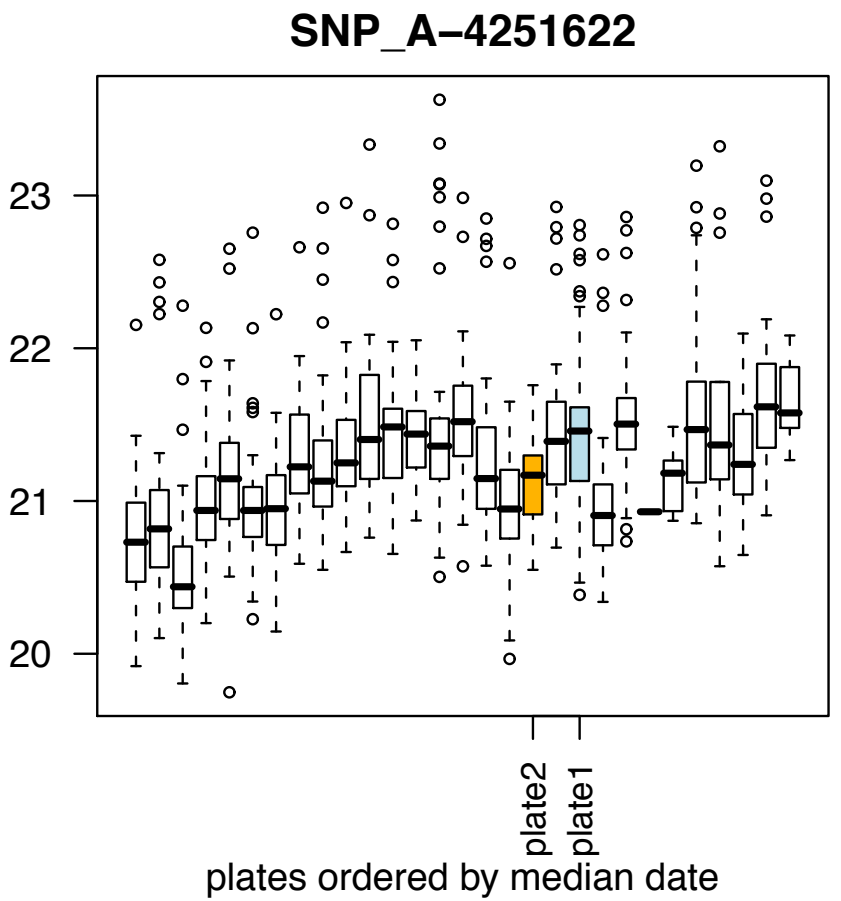


A versus B plots

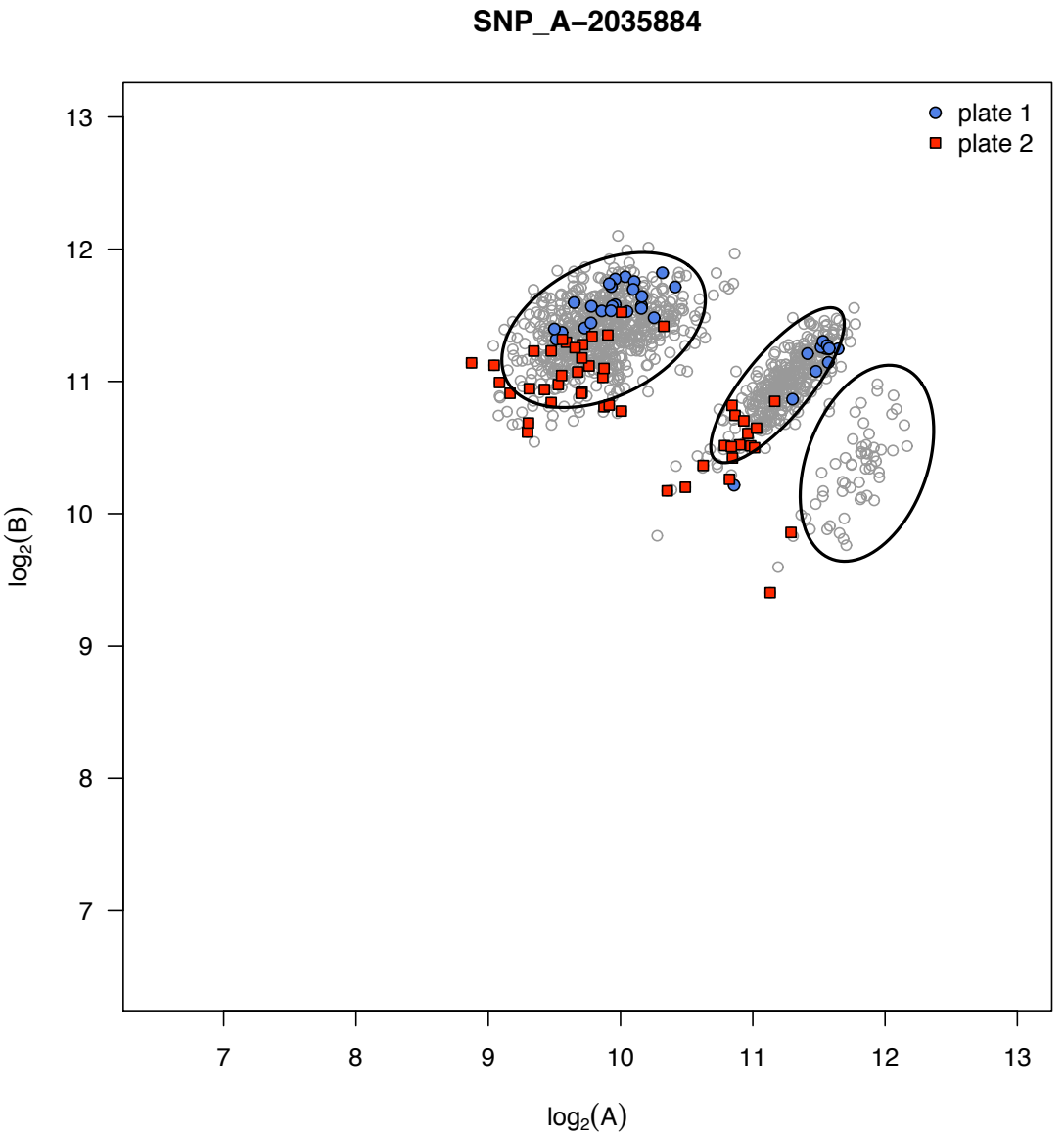


A versus B plots

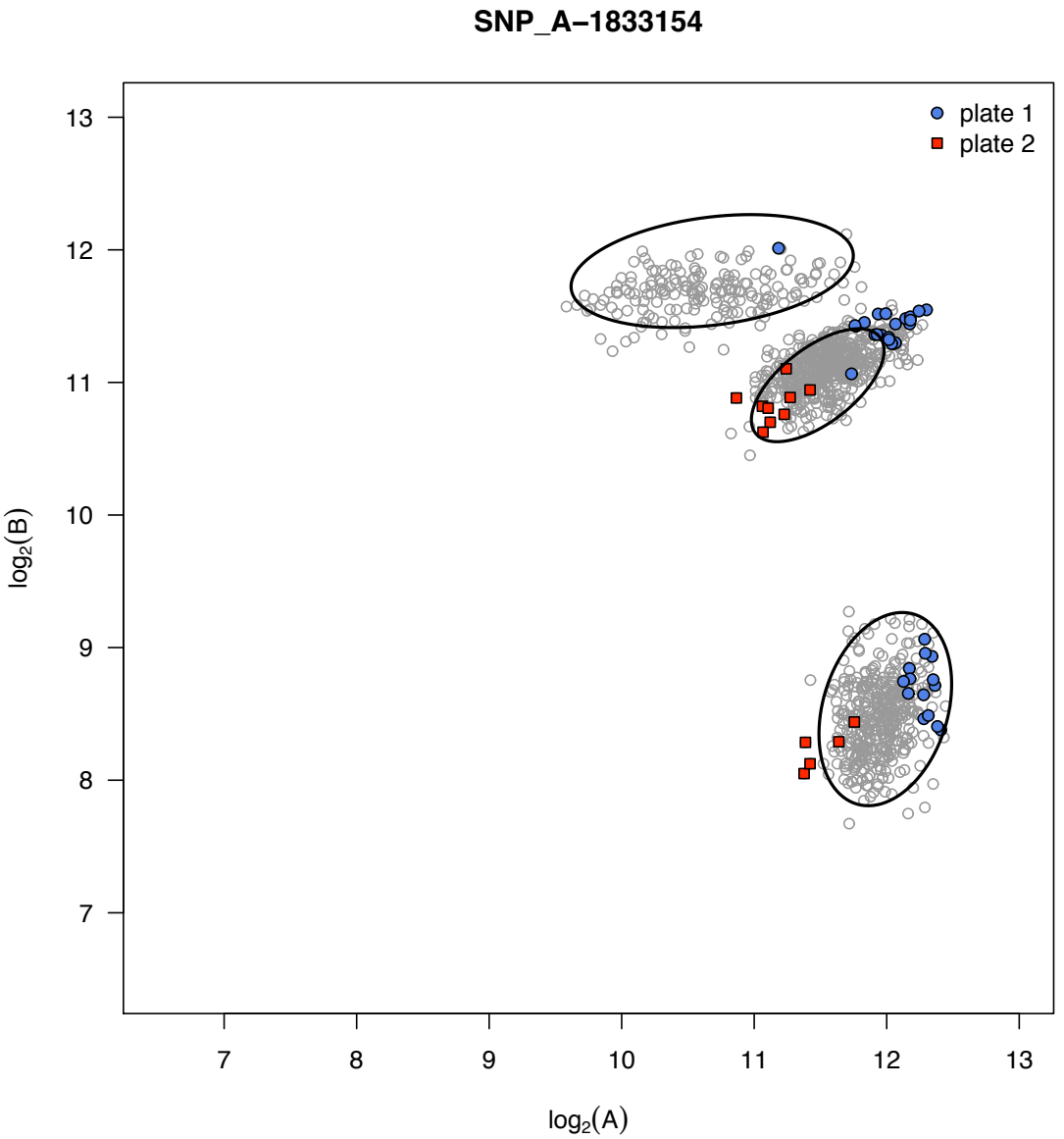




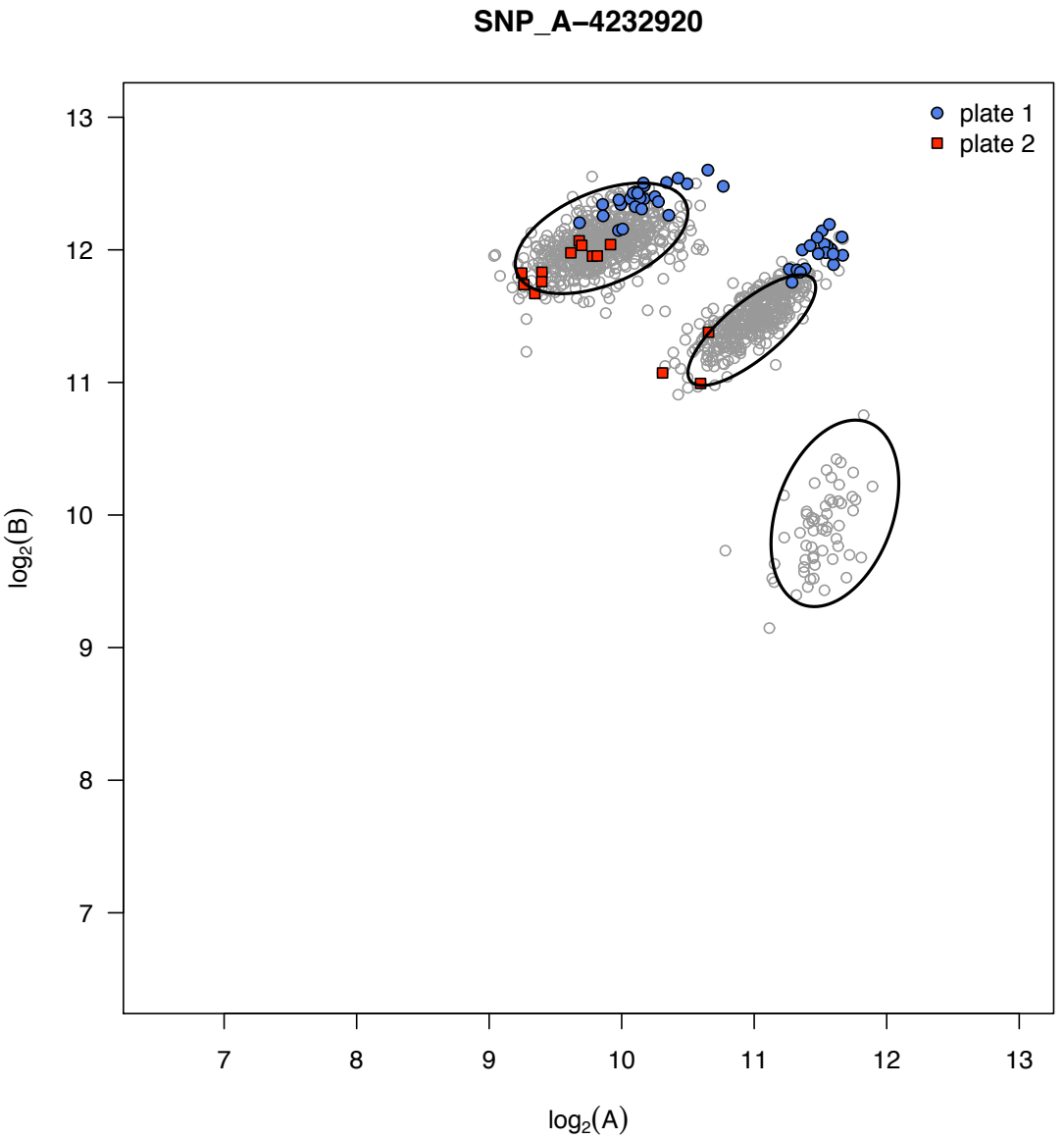
A versus B plots



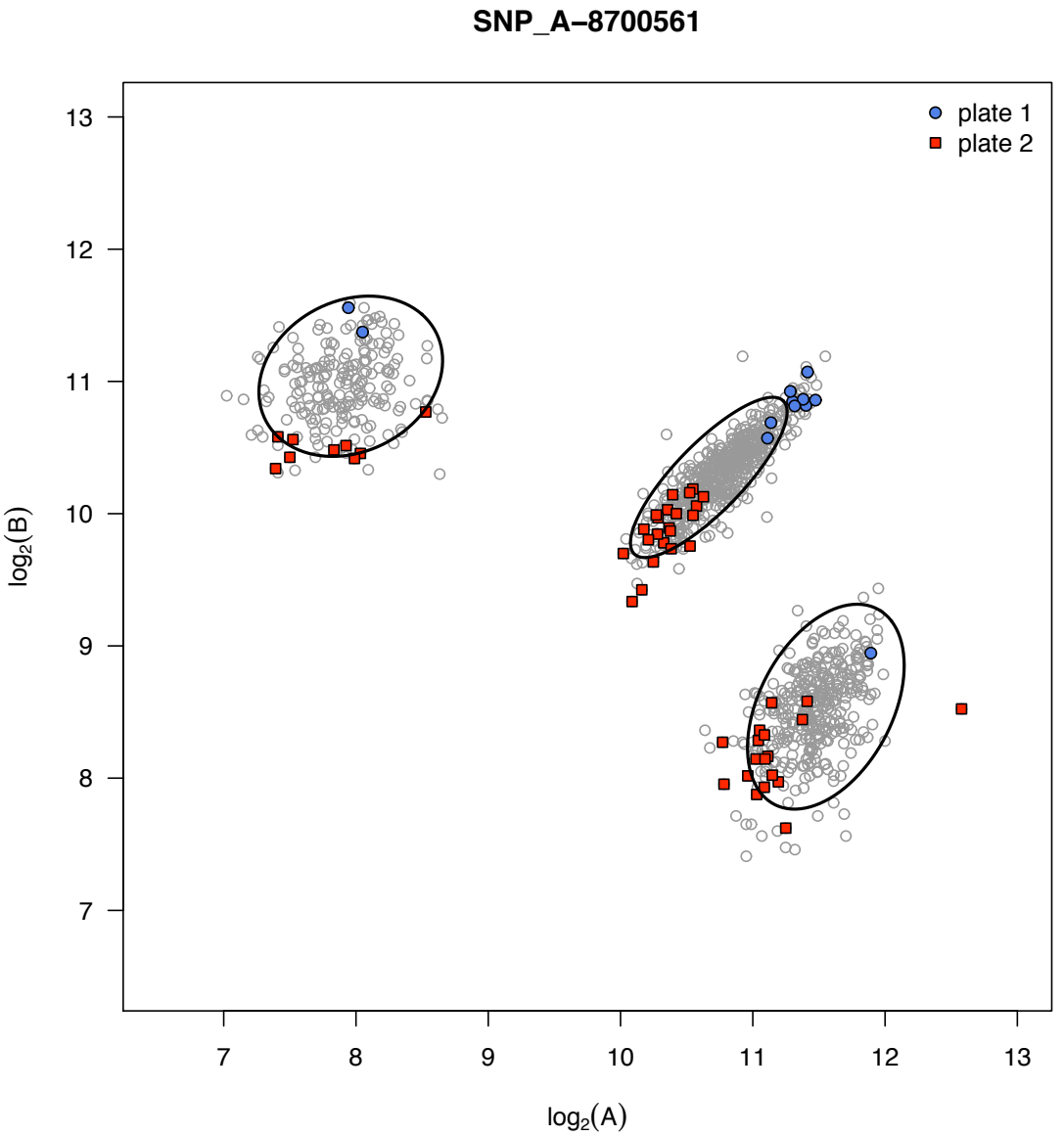
A versus B plots



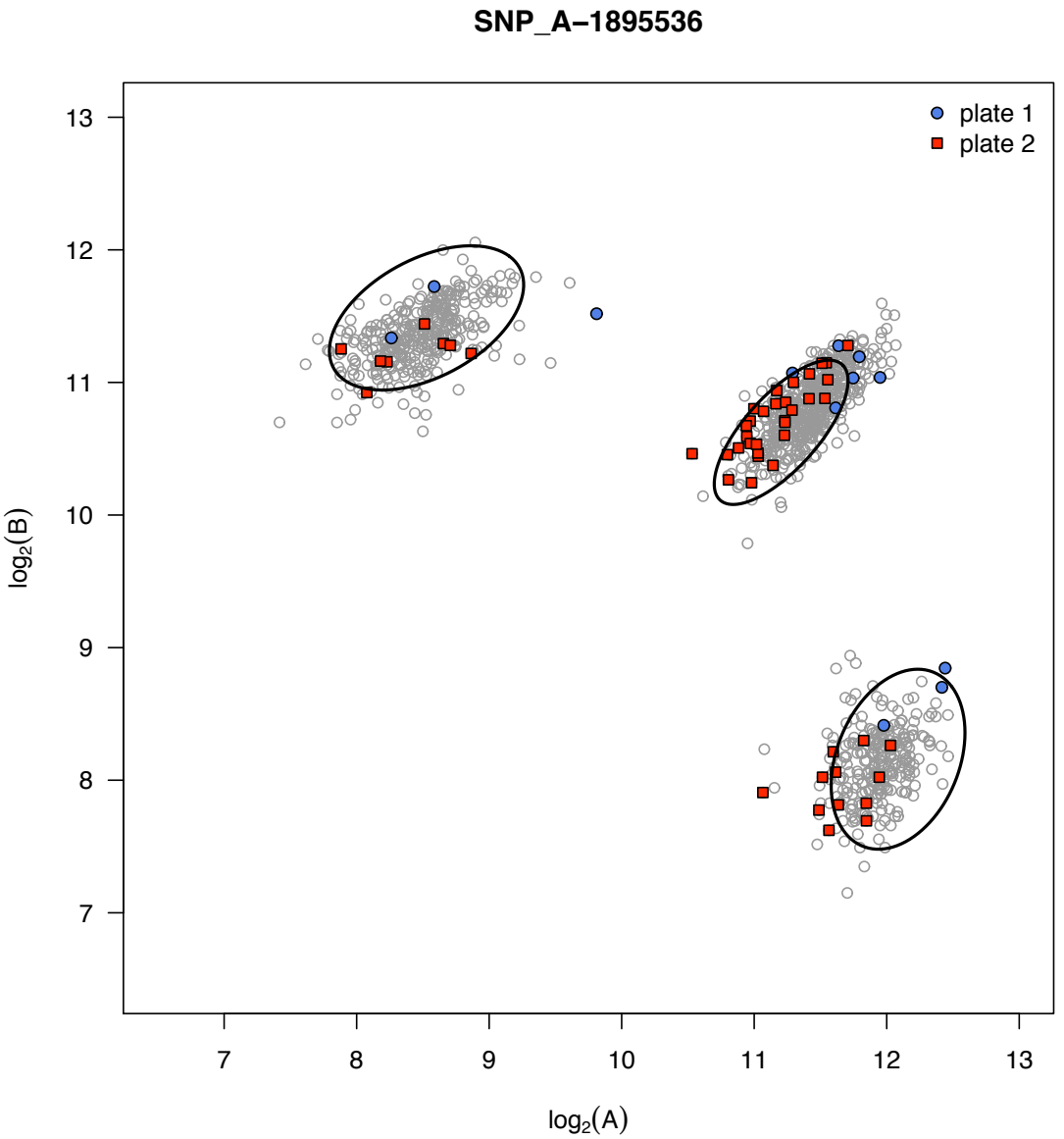
A versus B plots



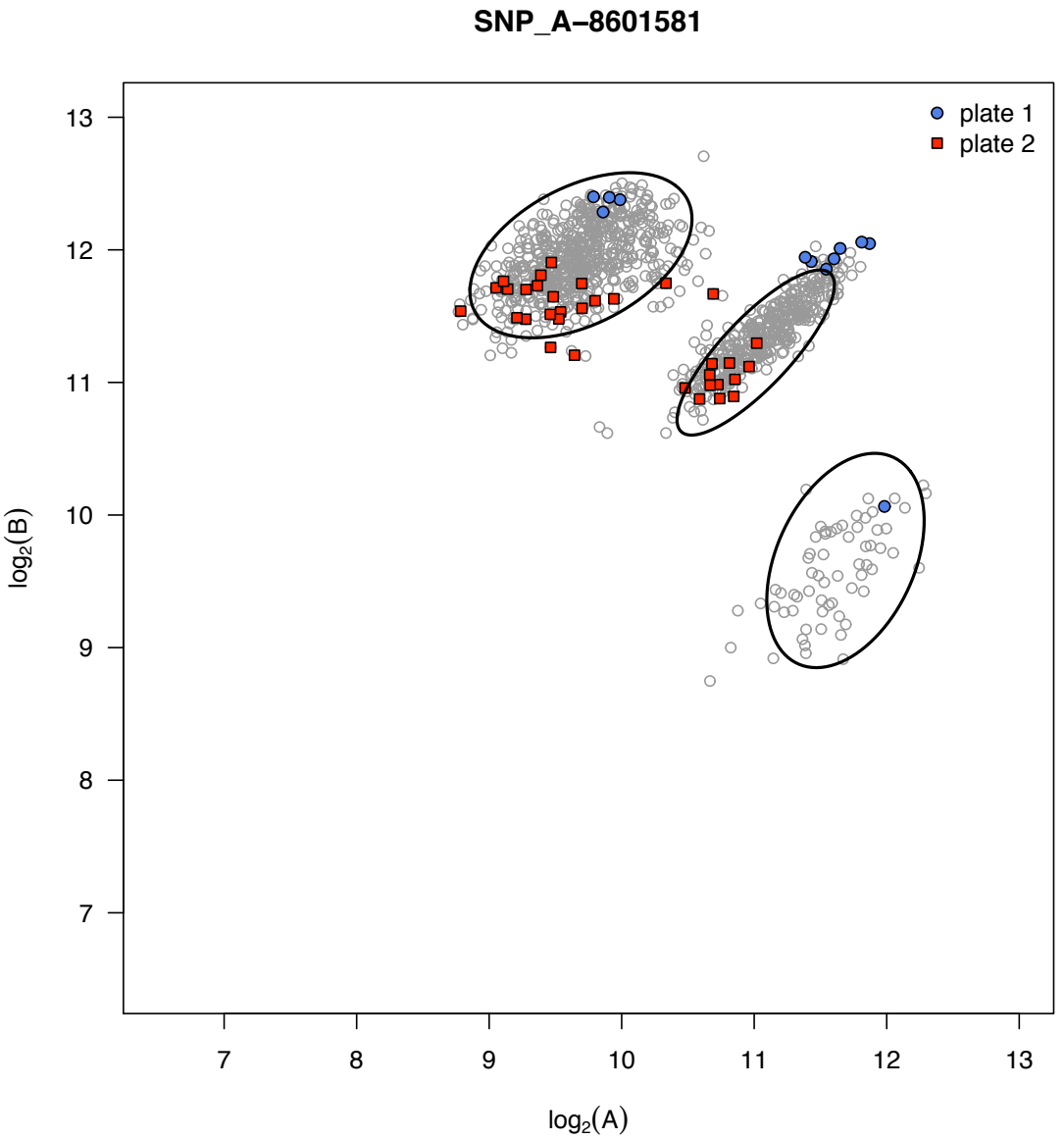
A versus B plots



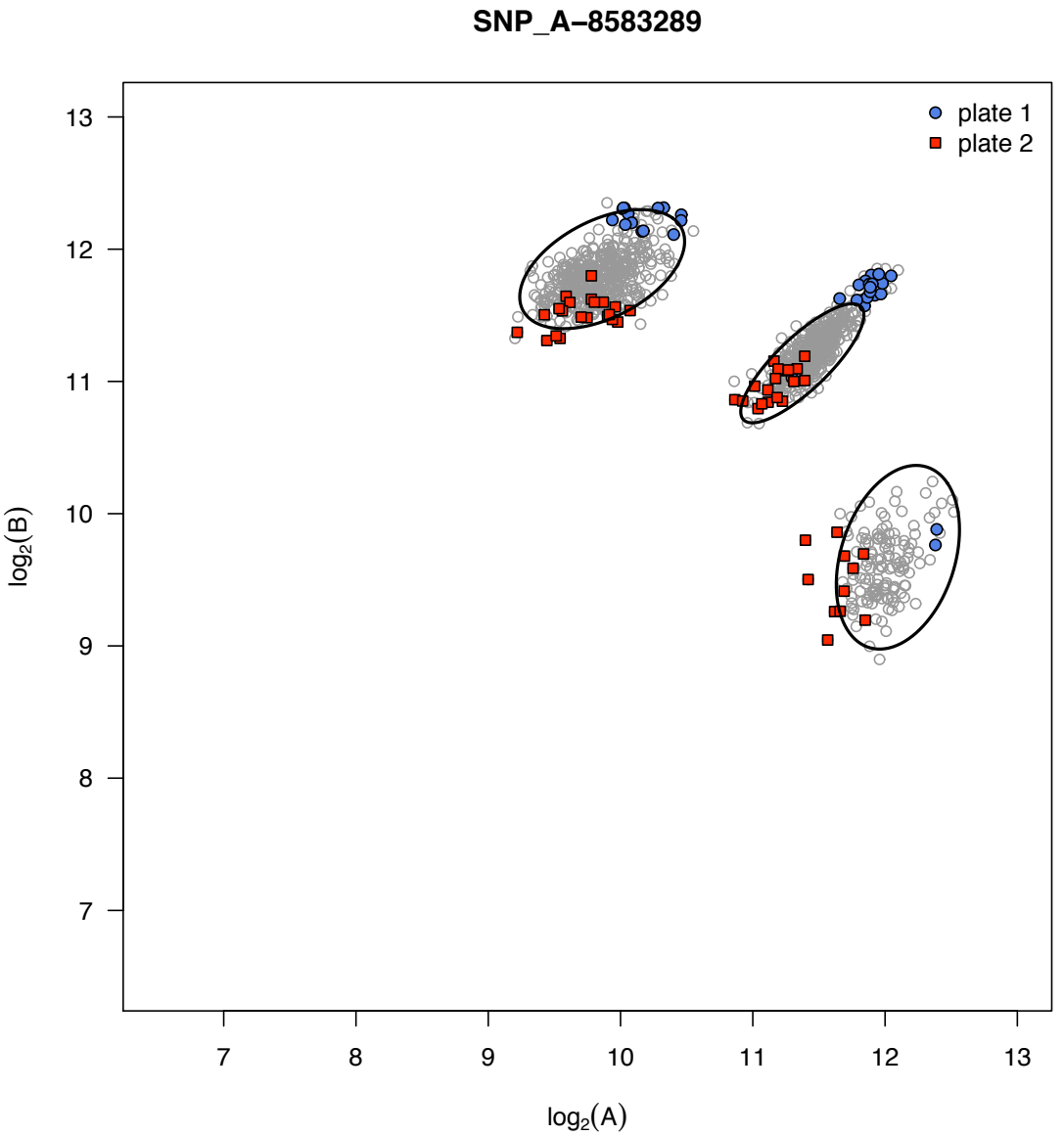
A versus B plots



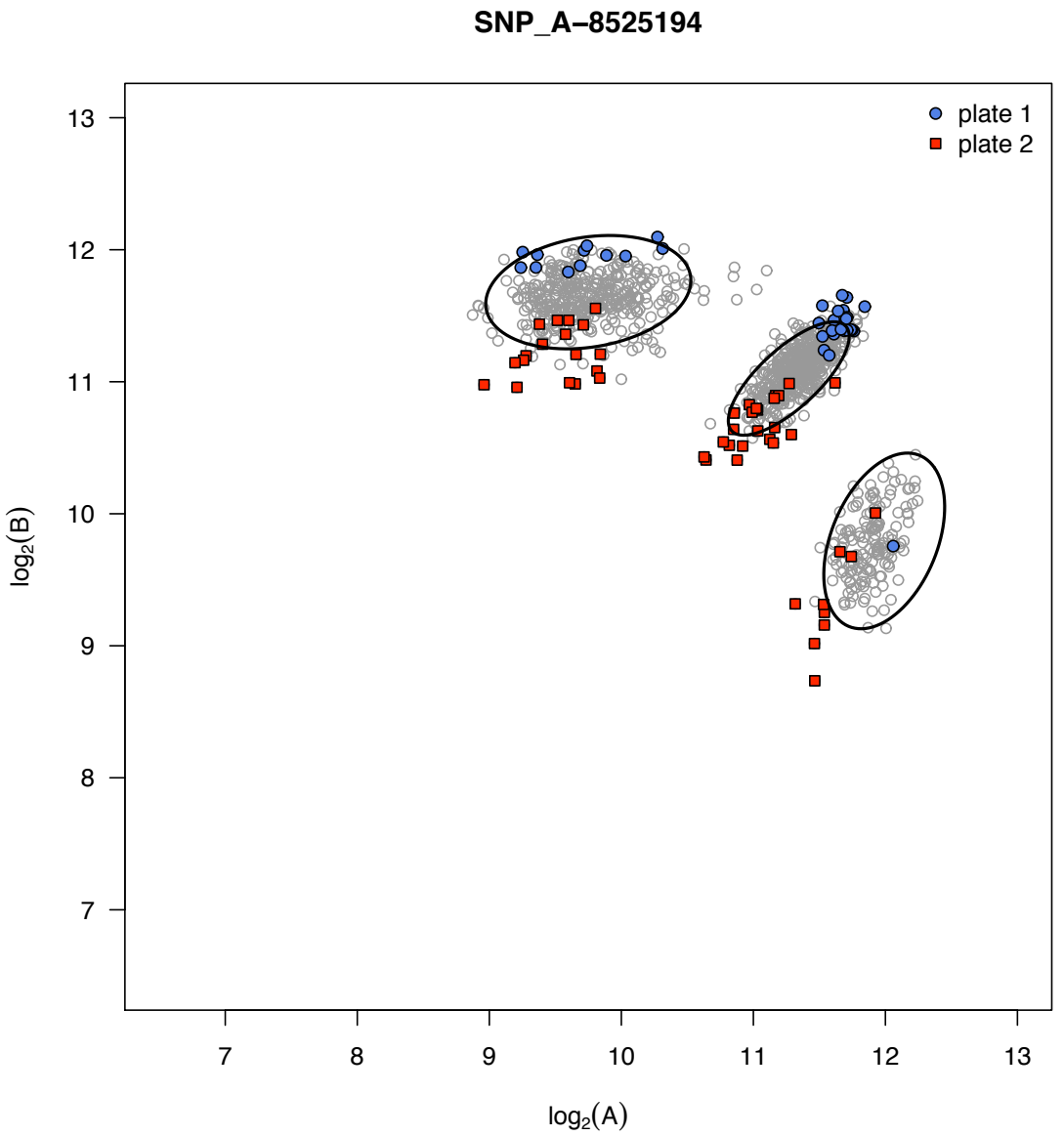
A versus B plots



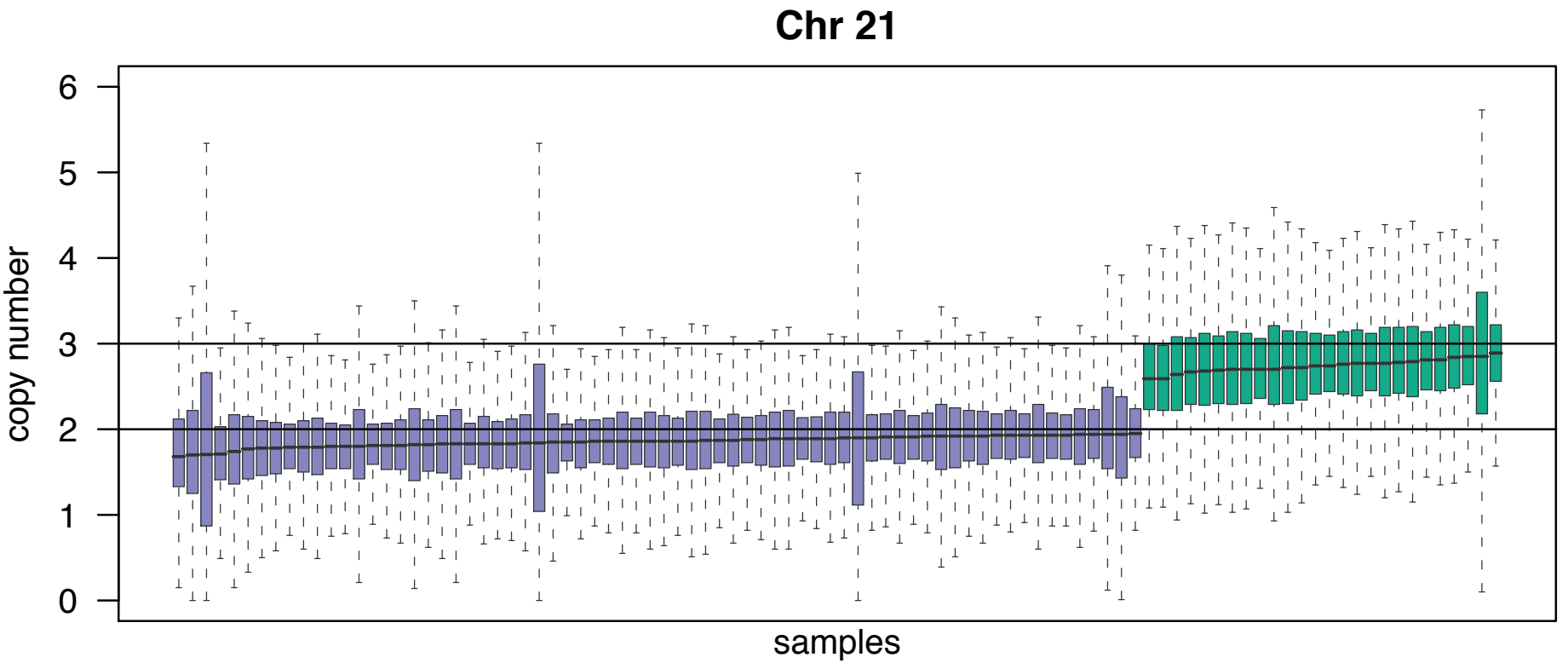
A versus B plots



A versus B plots

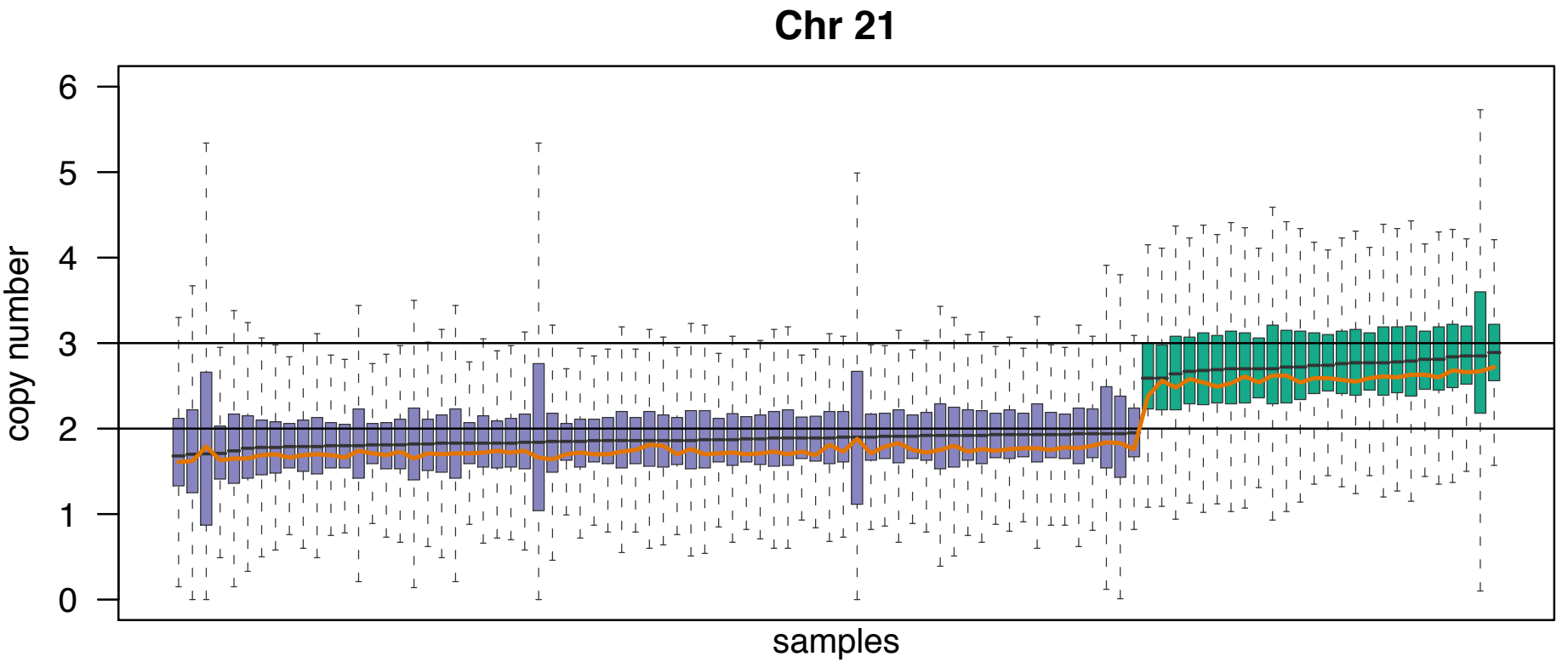


Trisomy 21



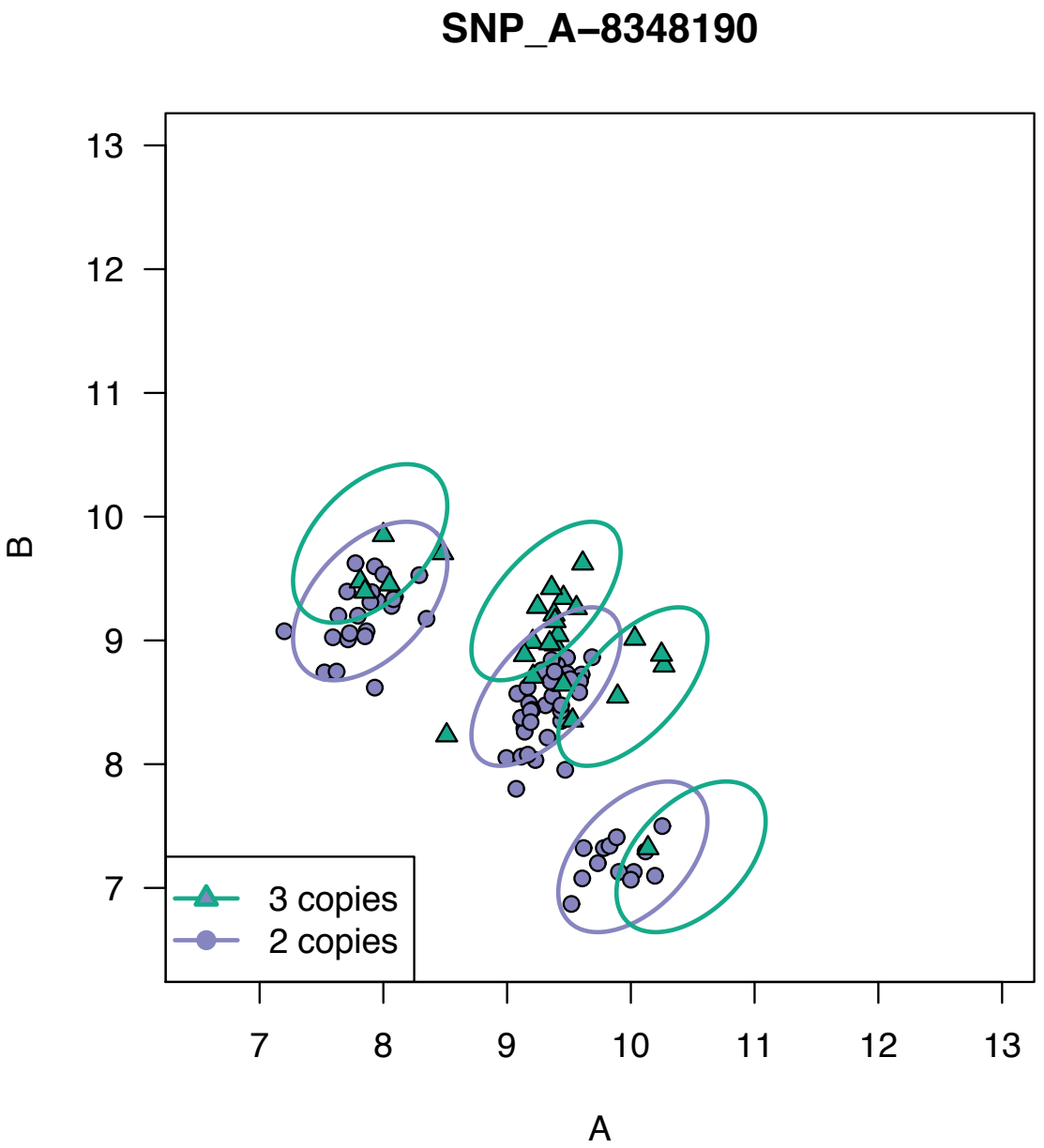
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21

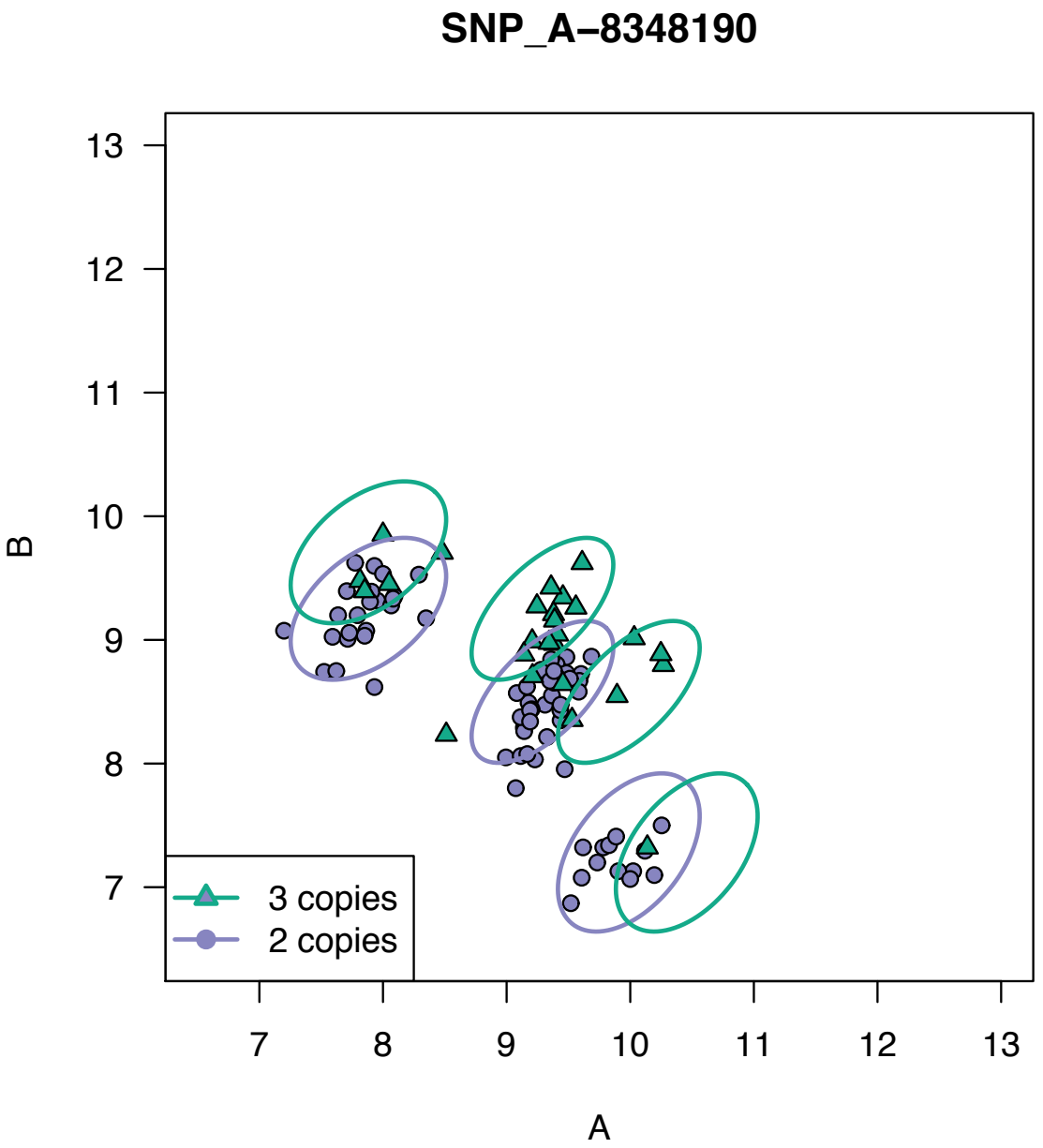


Samples from Aravinda Chakravarti and Betty Doan

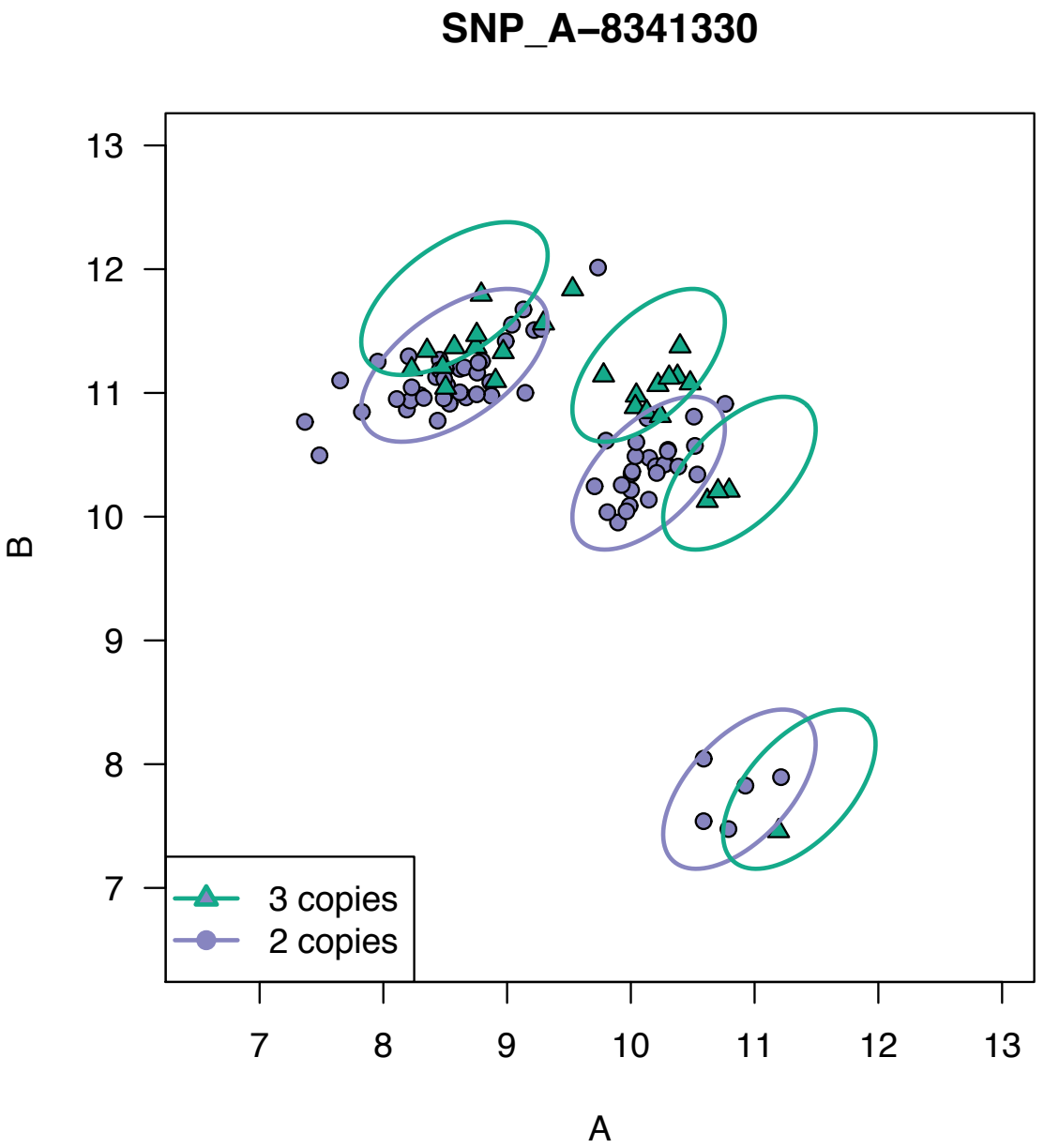
A versus B plots



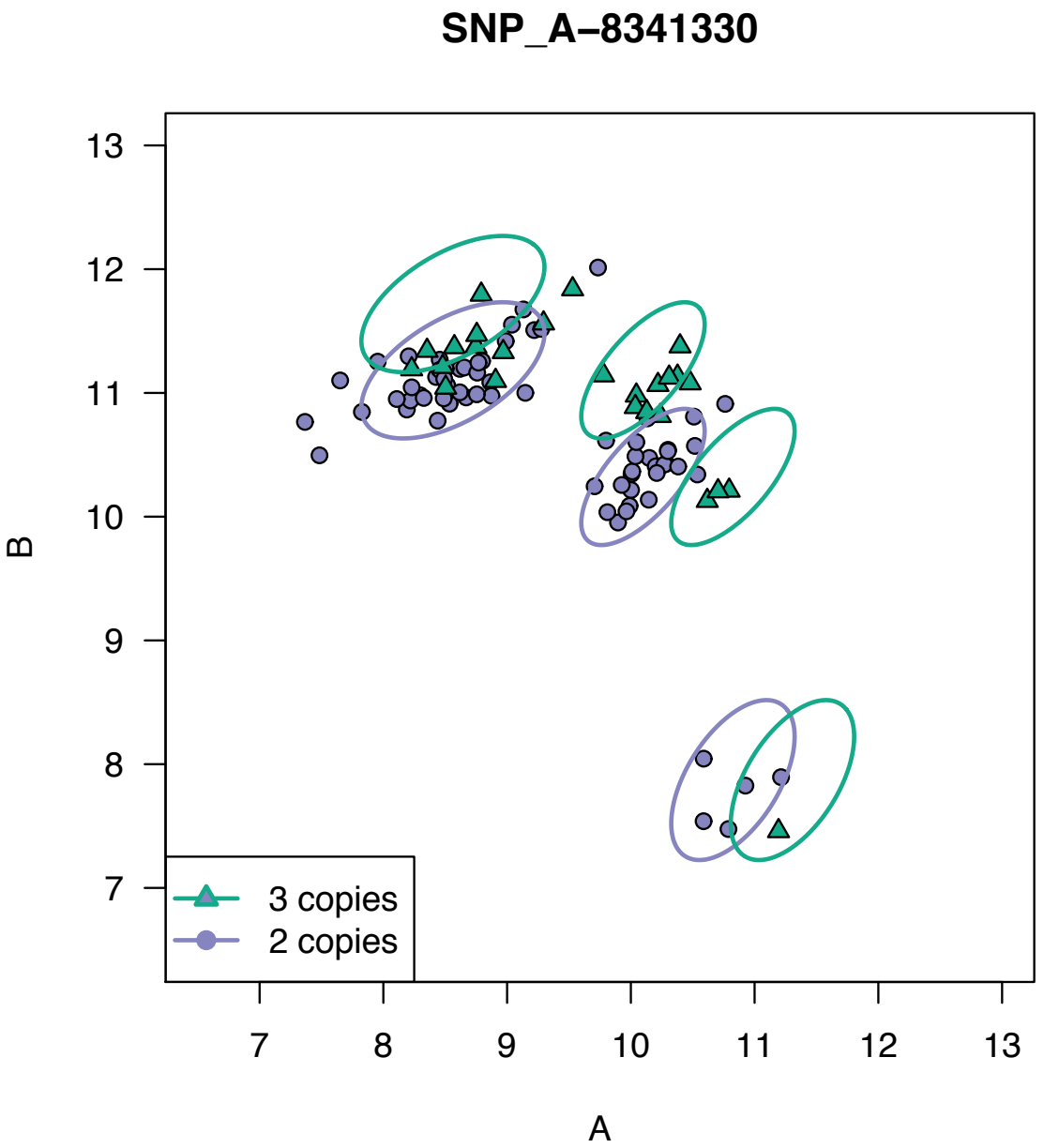
A versus B plots



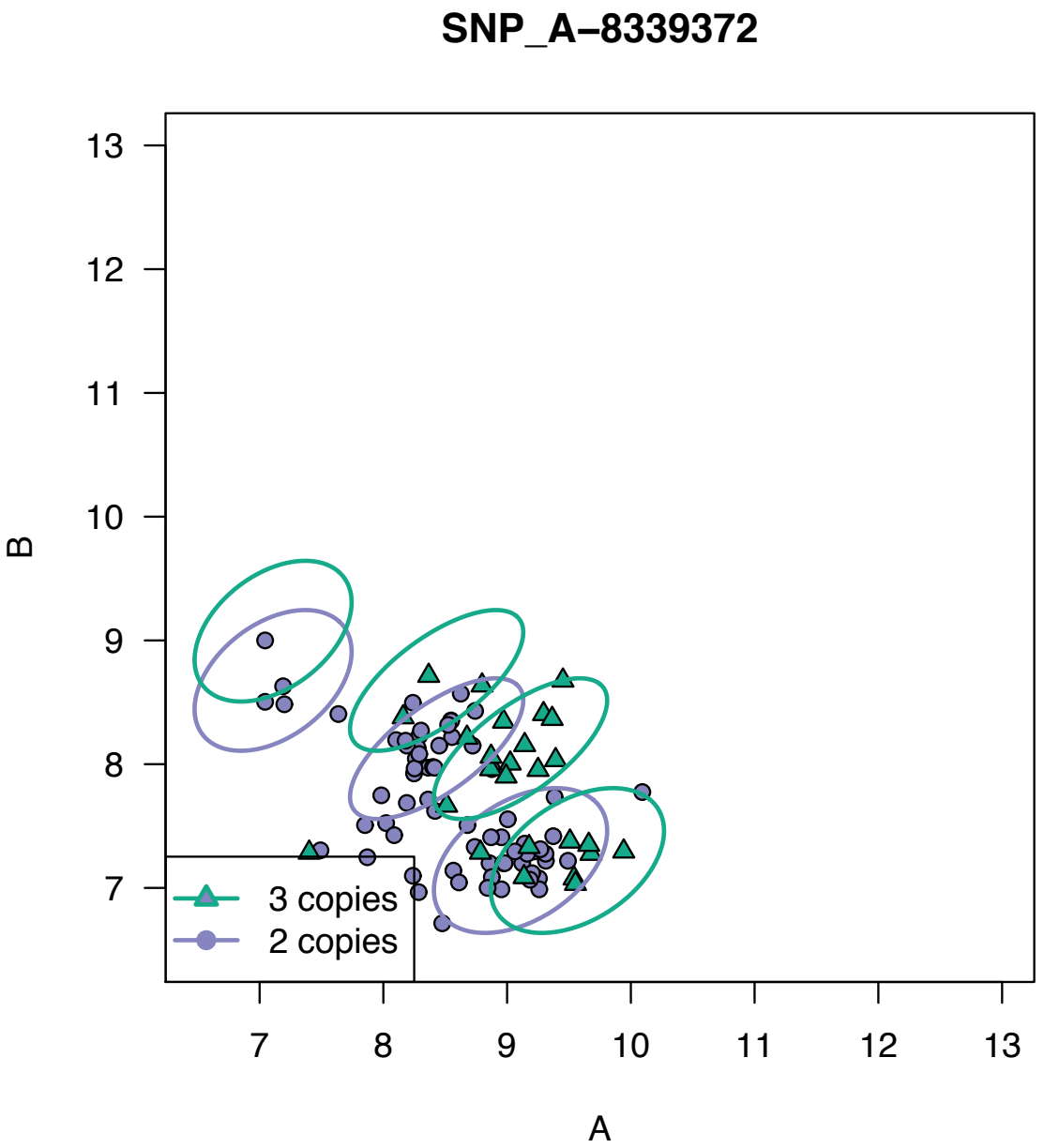
A versus B plots



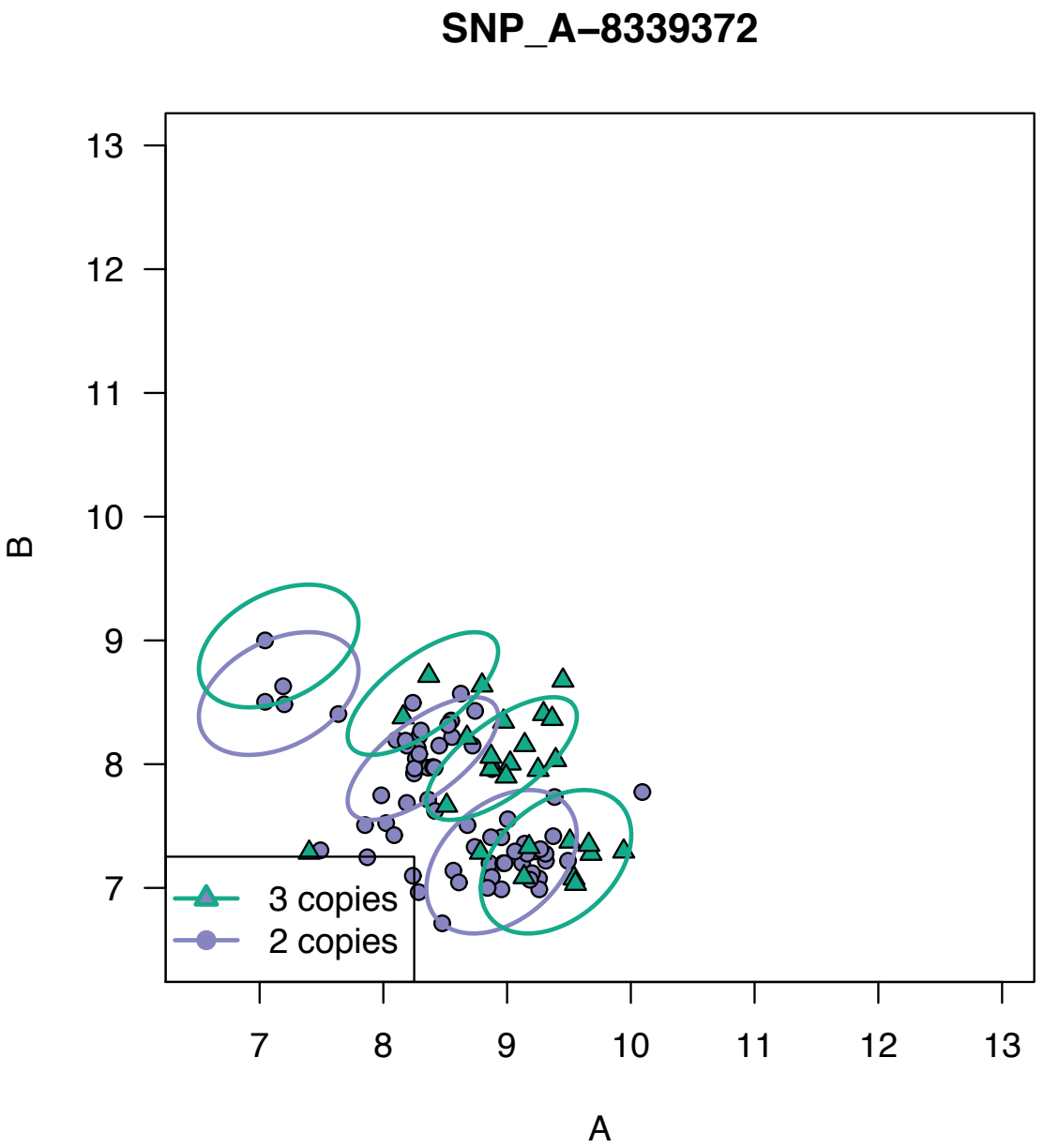
A versus B plots



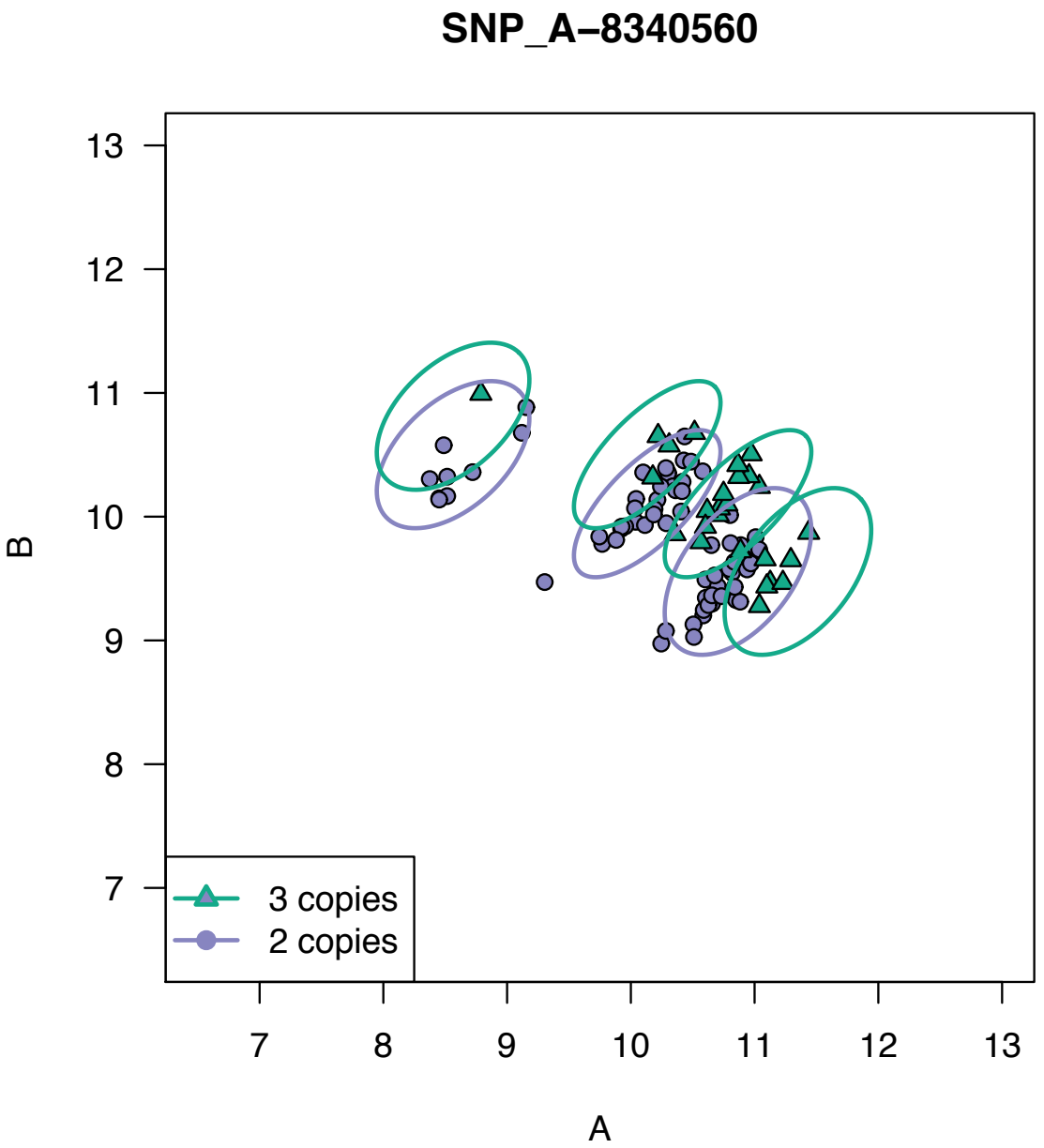
A versus B plots



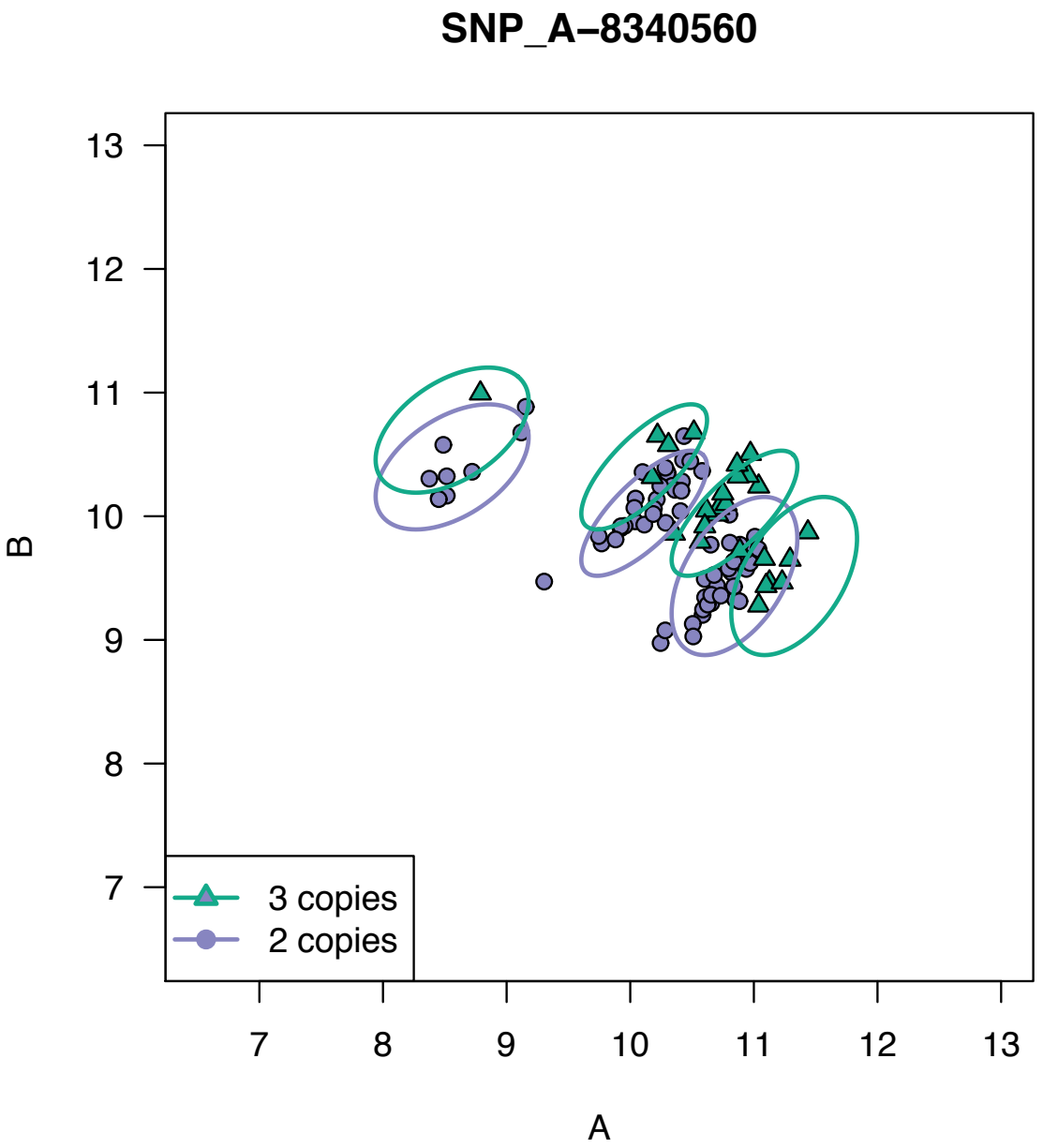
A versus B plots



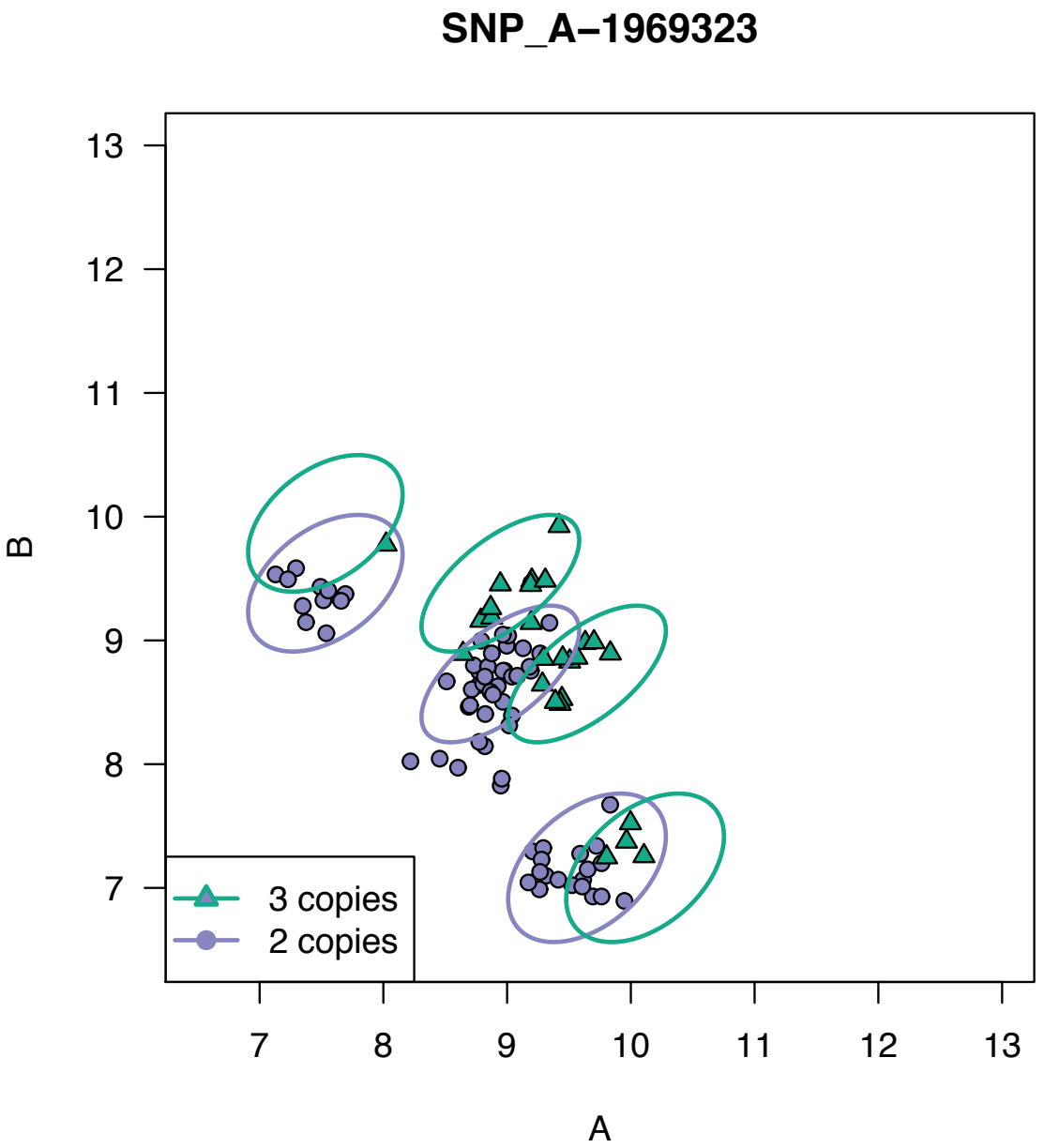
A versus B plots



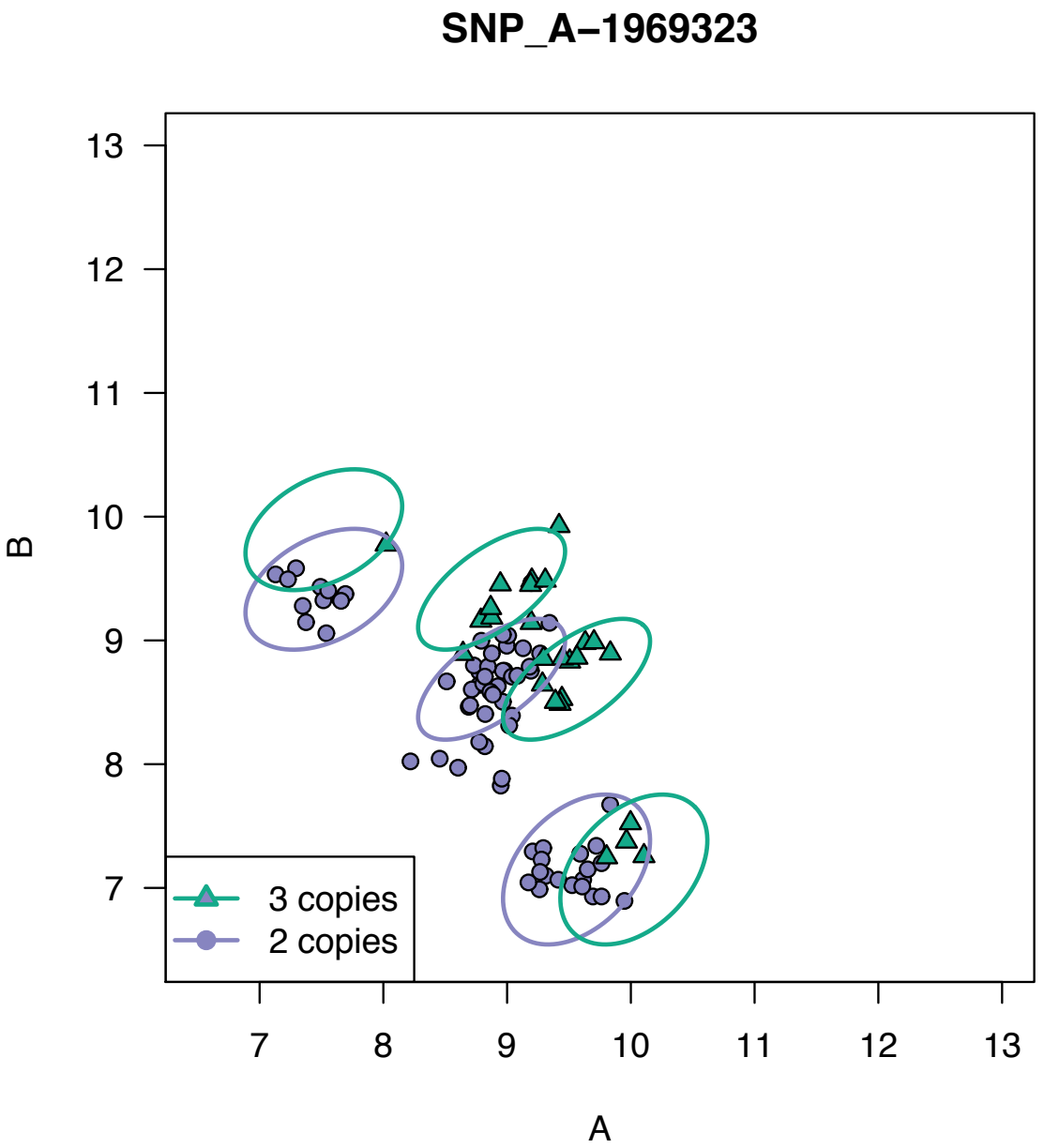
A versus B plots



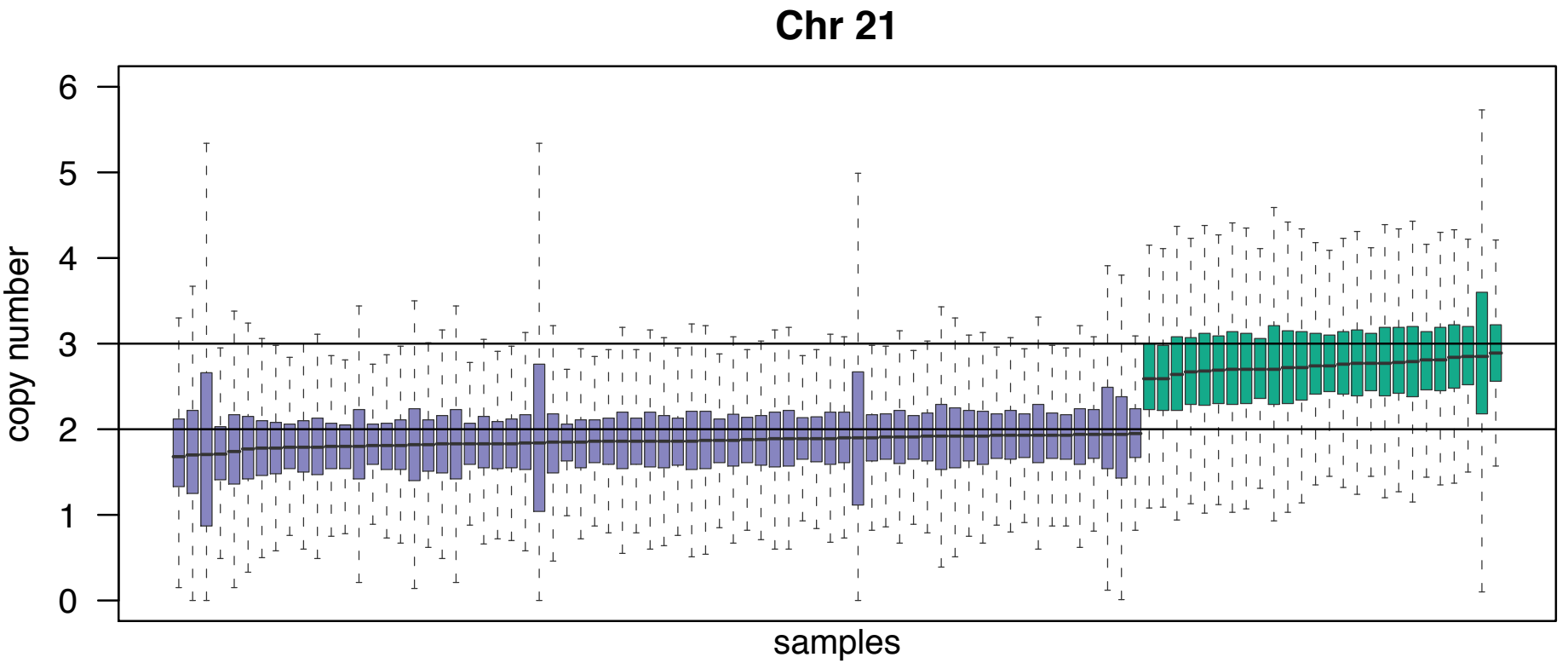
A versus B plots



A versus B plots

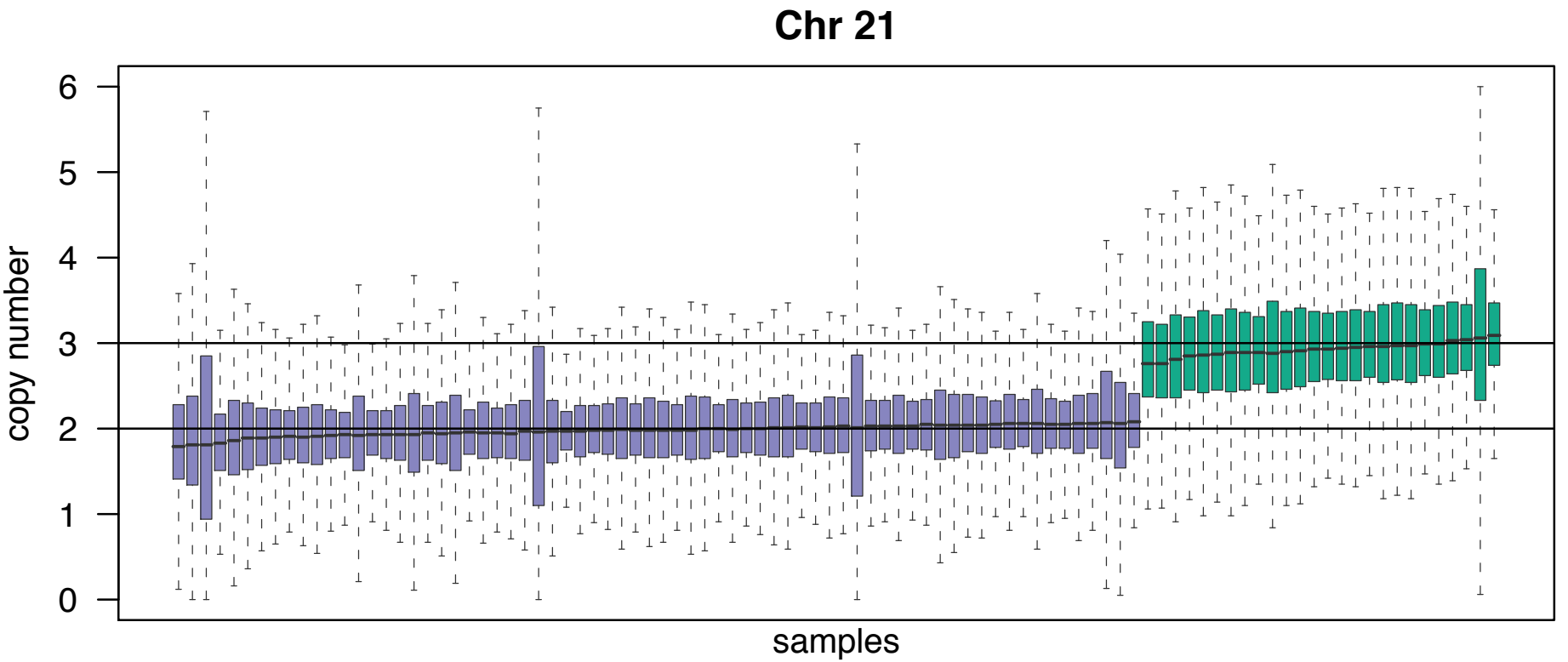


Trisomy 21



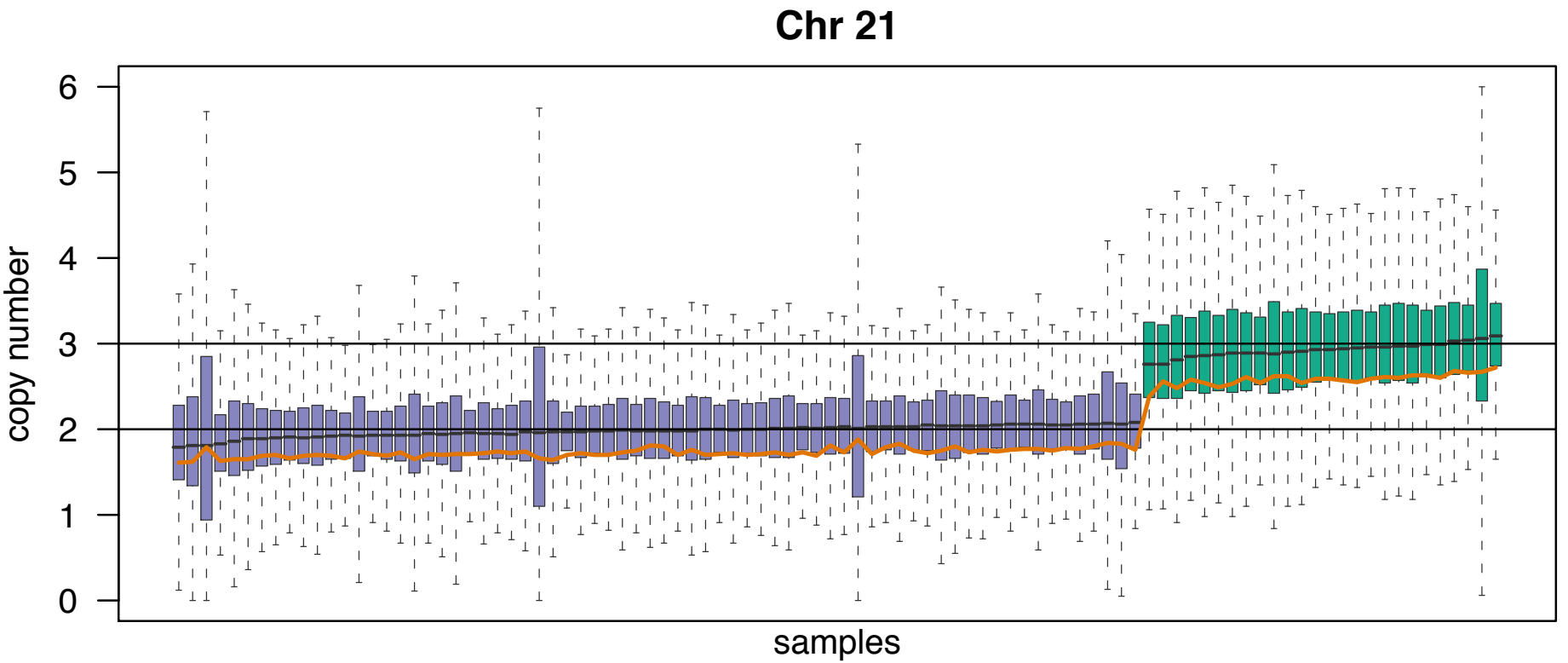
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21



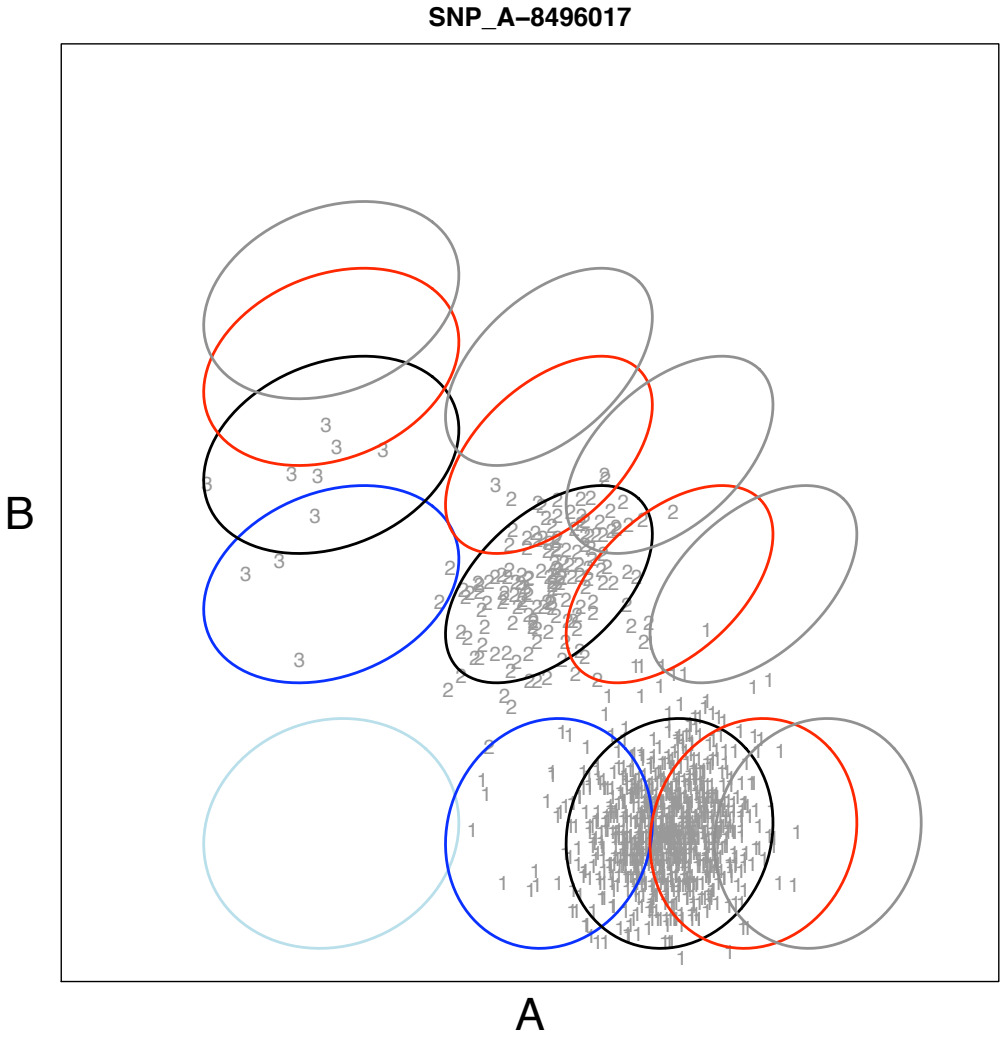
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21

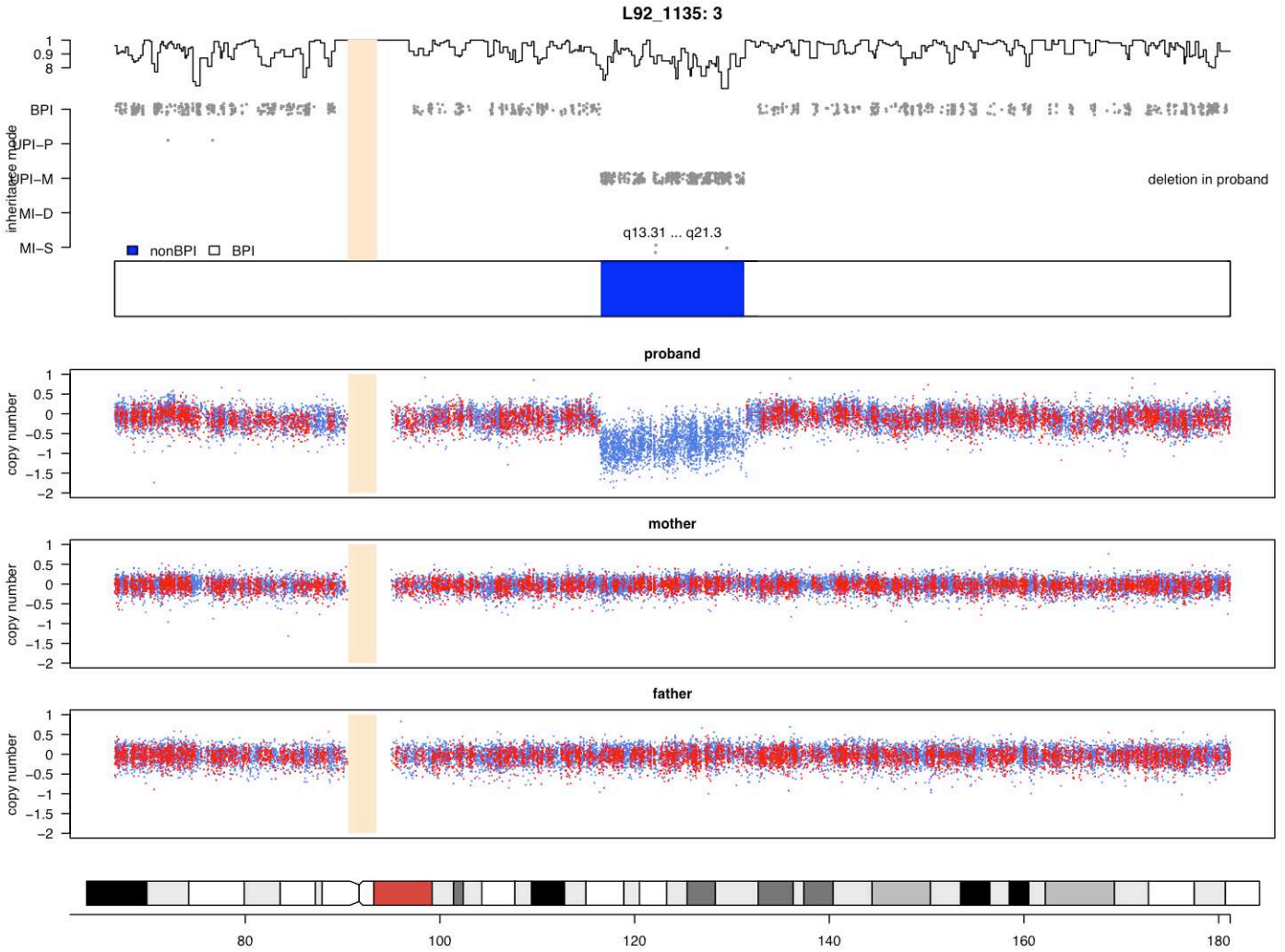


Samples from Aravinda Chakravarti and Betty Doan

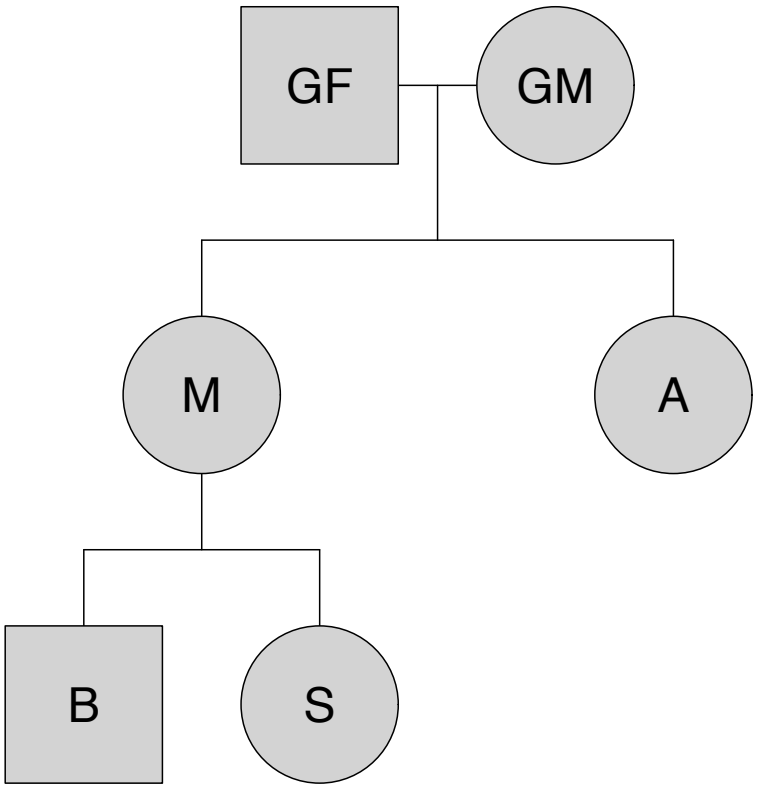
Prediction regions for copy number



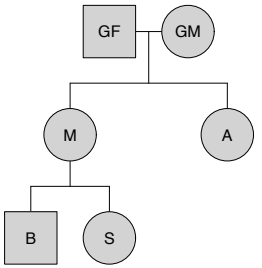
De novo deletion



Homozygous and hemizygous deletions

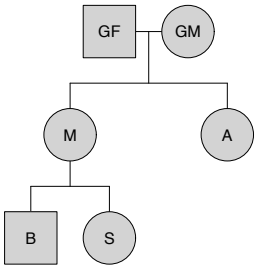


Homozygous and hemizygous deletions



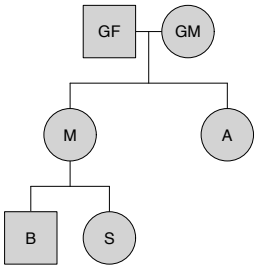
Grandfather		Grandmother		Mother		Aunt		Brother		Sister	
AB	0.01	BB	0.05	AB	-0.11	AB	-0.32	AB	0.06	BB	-0.02
AA	0.27	NC	-5.52	AA	-0.48	AA	-0.45	AB	0.12	BB	-0.42
AB	0.15	NC	-5.04	AA	-0.20	AA	-0.24	AB	-0.09	BB	-0.49
AB	-0.03	NC	-4.59	AA	-0.40	AA	-0.24	AA	0.30	AA	-0.72
BB	0.20	NC	-2.46	NC	-0.38	BB	-0.28	BB	0.22	BB	-0.45
AB	0.03	NC	-6.14	BB	-0.28	BB	-0.42	AB	0.09	AA	-0.70
AB	-0.05	NC	-5.02	BB	-0.17	BB	-0.34	AB	-0.22	AA	-1.06
BB	0.01	NC	-4.04	BB	0.04	BB	-0.68	BB	0.14	NC	-0.98
AB	0.17	NC	-4.06	AA	-0.27	AA	-0.33	AA	-0.03	AA	-0.76
AB	0.01	NC	-4.70	AA	-0.67	AA	-0.52	AA	0.16	AA	-0.80
AB	-0.10	NC	-4.42	BB	-0.25	BB	-0.62	AB	0.13	AA	-0.58
AB	0.01	NC	-8.29	BB	-0.17	BB	-0.15	AB	-0.15	AA	-0.29
BB	0.16	NC	-5.73	BB	-0.64	BB	-0.46	BB	0.10	BB	-0.52
AB	0.06	NC	-7.48	AA	-0.23	AA	-0.33	AB	0.07	BB	-0.47
AA	0.17	NC	-3.70	AA	-0.50	AA	-0.52	AB	-0.06	BB	-0.48
BB	0.02	NC	-5.00	BB	-0.34	BB	-0.45	AB	0.13	AA	-0.55
AA	0.21	NC	-6.10	AA	-0.43	AA	-0.40	AB	0.20	BB	-0.40
BB	0.05	BB	0.11	BB	0.15	BB	0.29	AB	0.13	AB	-0.01

Homozygous and hemizygous deletions



Grandfather		Grandmother		Mother		Aunt		Brother		Sister	
AB	0.01	BB	0.05	AB	-0.11	AB	-0.32	AB	0.06	BB	-0.02
AA	0.27	NC	**-5.52**	AA	**-0.48**	AA	**-0.45**	AB	0.12	BB	**-0.42**
AB	0.15	NC	**-5.04**	AA	**-0.20**	AA	**-0.24**	AB	-0.09	BB	**-0.49**
AB	-0.03	NC	**-4.59**	AA	**-0.40**	AA	**-0.24**	AA	0.30	AA	**-0.72**
BB	0.20	NC	**-2.46**	NC	**-0.38**	BB	**-0.28**	BB	0.22	BB	**-0.45**
AB	0.03	NC	**-6.14**	BB	**-0.28**	BB	**-0.42**	AB	0.09	AA	**-0.70**
AB	-0.05	NC	**-5.02**	BB	-0.17	BB	**-0.34**	AB	-0.22	AA	**-1.06**
BB	0.01	NC	**-4.04**	BB	0.04	BB	**-0.68**	BB	0.14	NC	**-0.98**
AB	0.17	NC	**-4.06**	AA	**-0.27**	AA	**-0.33**	AA	-0.03	AA	**-0.76**
AB	0.01	NC	**-4.70**	AA	**-0.67**	AA	**-0.52**	AA	0.16	AA	**-0.80**
AB	-0.10	NC	**-4.42**	BB	**-0.25**	BB	**-0.62**	AB	0.13	AA	**-0.58**
AB	0.01	NC	**-8.29**	BB	-0.17	BB	-0.15	AB	-0.15	AA	**-0.29**
BB	0.16	NC	**-5.73**	BB	**-0.64**	BB	**-0.46**	BB	0.10	BB	**-0.52**
AB	0.06	NC	**-7.48**	AA	**-0.23**	AA	**-0.33**	AB	0.07	BB	**-0.47**
AA	0.17	NC	**-3.70**	AA	**-0.50**	AA	**-0.52**	AB	-0.06	BB	**-0.48**
BB	0.02	NC	**-5.00**	BB	**-0.34**	BB	**-0.45**	AB	0.13	AA	**-0.55**
AA	0.21	NC	**-6.10**	AA	**-0.43**	AA	**-0.40**	AB	0.20	BB	**-0.40**
BB	0.05	BB	0.11	BB	0.15	BB	0.29	AB	0.13	AB	-0.01

Homozygous and hemizygous deletions



Grandfather		Grandmother		Mother		Aunt		Brother		Sister	
AB	0.01	BB	0.05	AB	-0.11	AB	-0.32	AB	0.06	BB	-0.02
AA	0.27	NC	**-5.52**	AA	**-0.48**	AA	**-0.45**	AB	0.12	BB	**-0.42**
AB	0.15	NC	**-5.04**	AA	**-0.20**	AA	**-0.24**	AB	-0.09	BB	**-0.49**
AB	-0.03	NC	**-4.59**	AA	**-0.40**	AA	**-0.24**	AA	0.30	AA	**-0.72**
BB	0.20	NC	**-2.46**	NC	**-0.38**	BB	**-0.28**	BB	0.22	BB	**-0.45**
AB	0.03	NC	**-6.14**	BB	**-0.28**	BB	**-0.42**	AB	0.09	AA	**-0.70**
AB	-0.05	NC	**-5.02**	BB	-0.17	BB	**-0.34**	AB	-0.22	AA	**-1.06**
BB	0.01	NC	**-4.04**	BB	0.04	BB	**-0.68**	BB	0.14	NC	**-0.98**
AB	0.17	NC	**-4.06**	AA	**-0.27**	AA	**-0.33**	AA	-0.03	AA	**-0.76**
AB	0.01	NC	**-4.70**	AA	**-0.67**	AA	**-0.52**	AA	0.16	AA	**-0.80**
AB	-0.10	NC	**-4.42**	BB	**-0.25**	BB	**-0.62**	AB	0.13	AA	**-0.58**
AB	0.01	NC	**-8.29**	BB	-0.17	BB	-0.15	AB	-0.15	AA	**-0.29**
BB	0.16	NC	**-5.73**	BB	**-0.64**	BB	**-0.46**	BB	0.10	BB	**-0.52**
AB	0.06	NC	**-7.48**	AA	**-0.23**	AA	**-0.33**	AB	0.07	BB	**-0.47**
AA	0.17	NC	**-3.70**	AA	**-0.50**	AA	**-0.52**	AB	-0.06	BB	**-0.48**
BB	0.02	NC	**-5.00**	BB	**-0.34**	BB	**-0.45**	AB	0.13	AA	**-0.55**
AA	0.21	NC	**-6.10**	AA	**-0.43**	AA	**-0.40**	AB	0.20	BB	**-0.40**
BB	0.05	BB	0.11	BB	0.15	BB	0.29	AB	0.13	AB	-0.01

References

Scharpf RB, Ting JC, Pevsner J, Ruczinski I (2007).
SNPchip: R classes and methods for SNP array data.
Bioinformatics, 23(5): 627-8.

Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I (2008).
Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays.
Annals of Applied Statistics, 2(2): 687-713.

Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA (2009).
A multilevel model to assess bacth effects in copy number estimation using SNP arrays.
JHU Department of Biostatistics Working Papers, Working Paper 192.

⟶ Hopefully soon: more on plate effects and the trio HMM method.

http://biostat.jhsph.edu/∼iruczins/