Hidden Markov Models for the Assessment of Chromosomal Alterations using High-throughput SNP Arrays

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

October 4, 2007

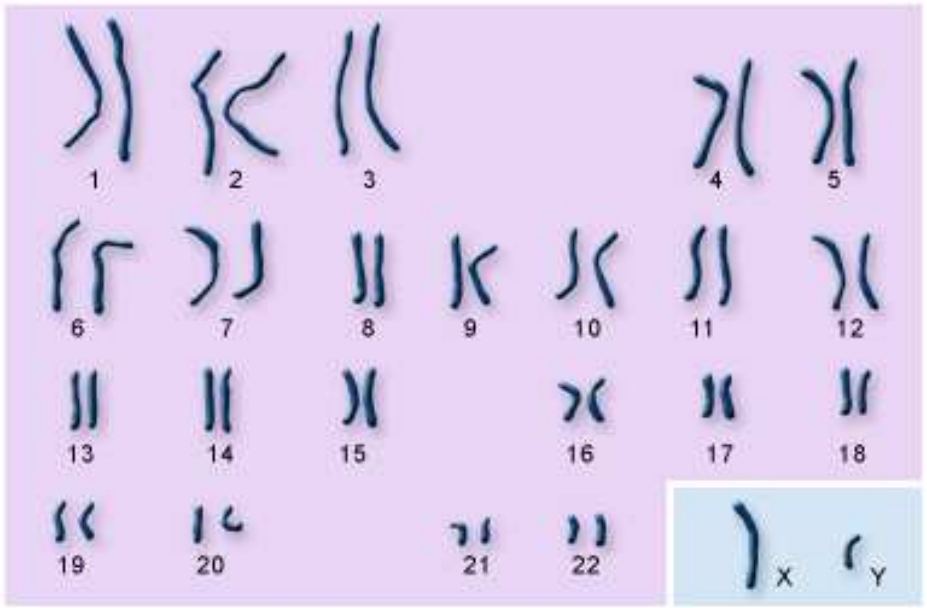
Acknowledgments

- Rob Scharpf
- Giovanni Parmigiani
- Rafael Irizarry

  Benilton Carvalho, Wenyi Wang
- Jonathan Pevsner

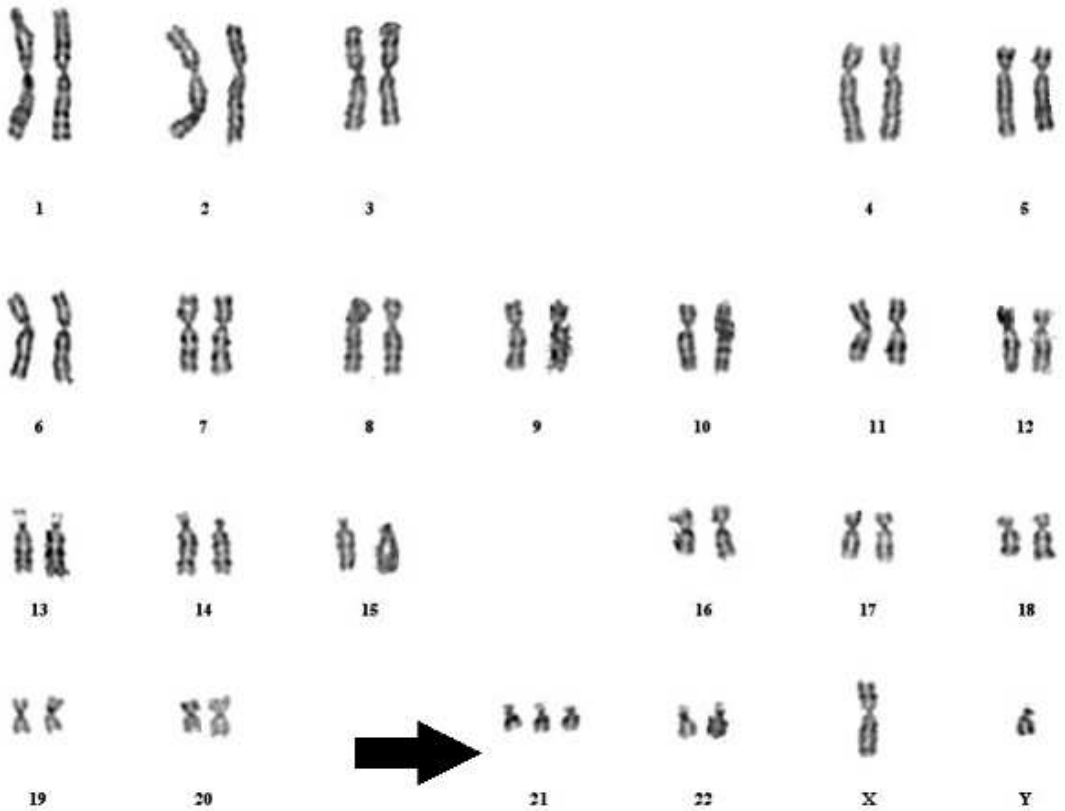
  Nate Miller, Eli Roberson, Jason Ting

Karyotypes

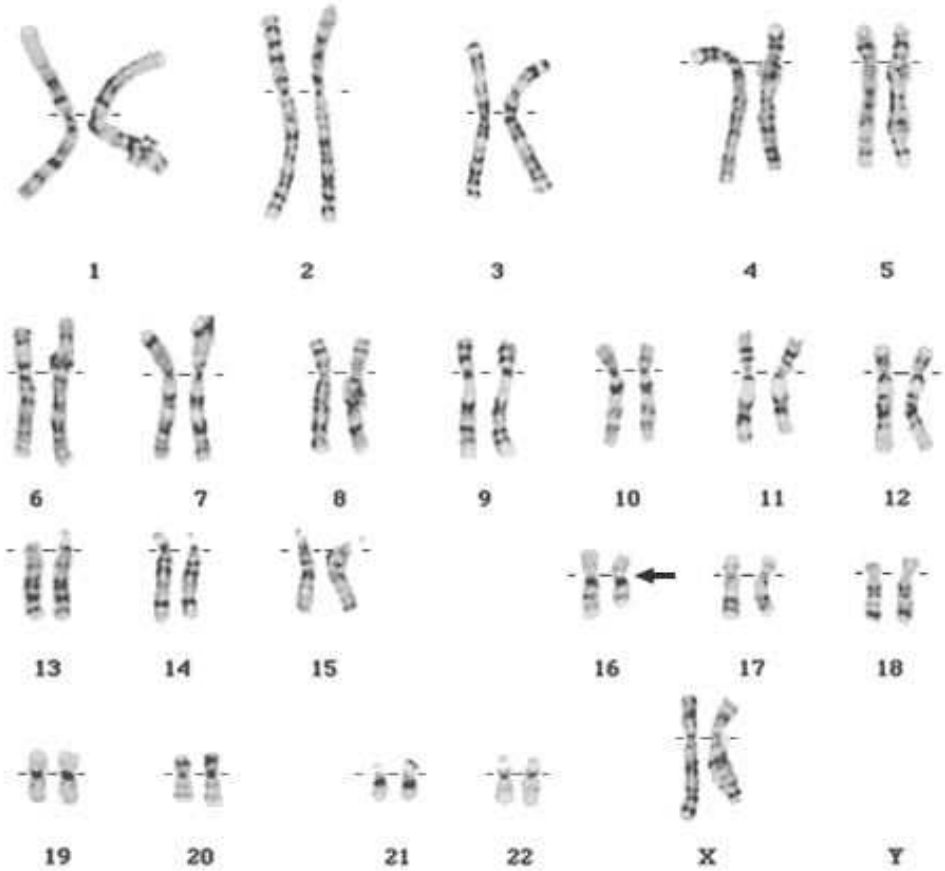


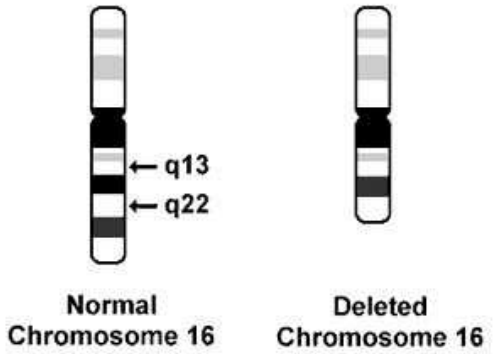
U.S. National Library of Medicine

Trisomy



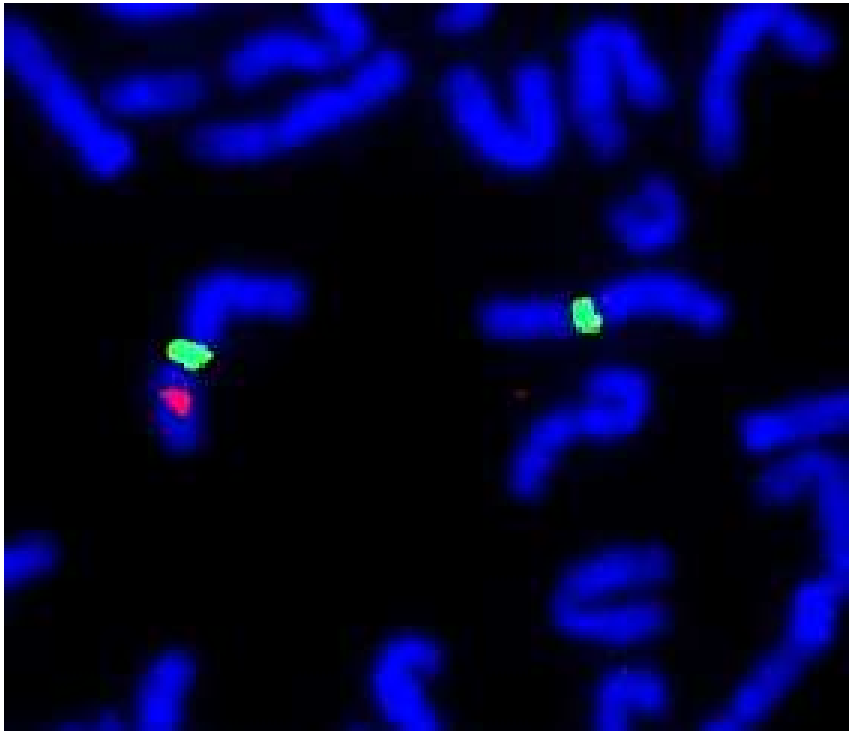
Karyotypes





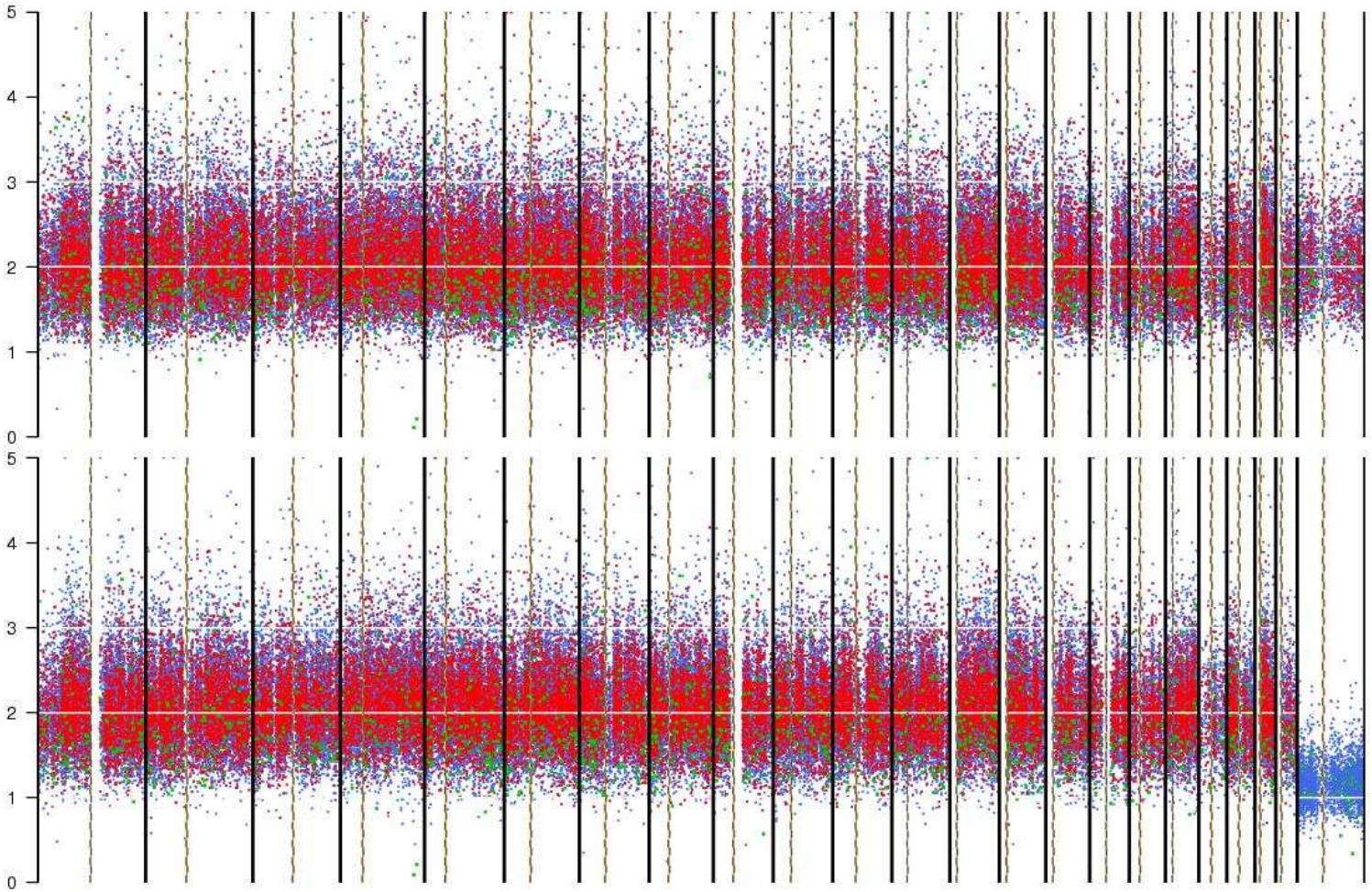
General Cytogenetics Information http://members.aol.com/chrominfo/

FISH

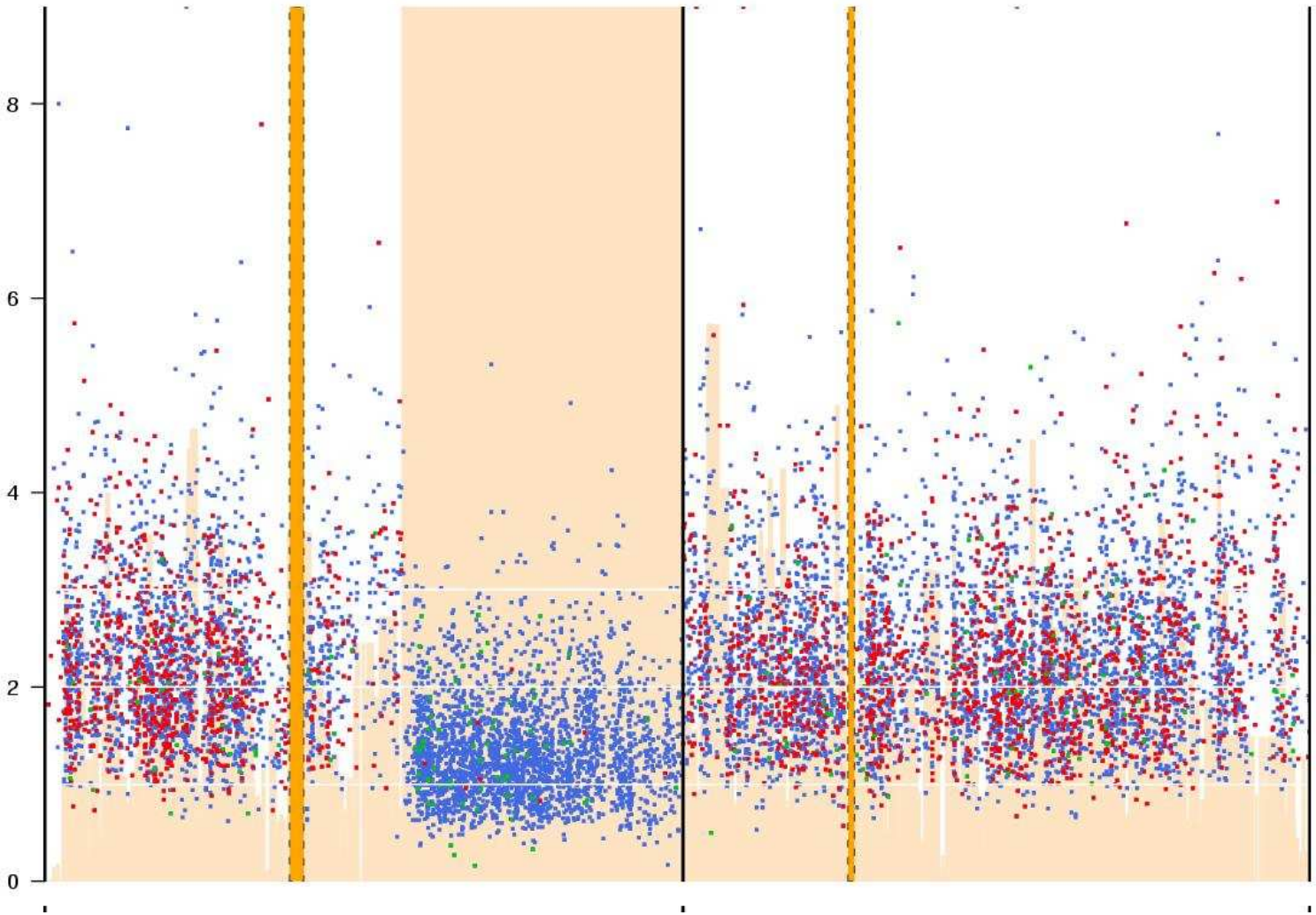


Courtesy of the Pevsner Laboratory

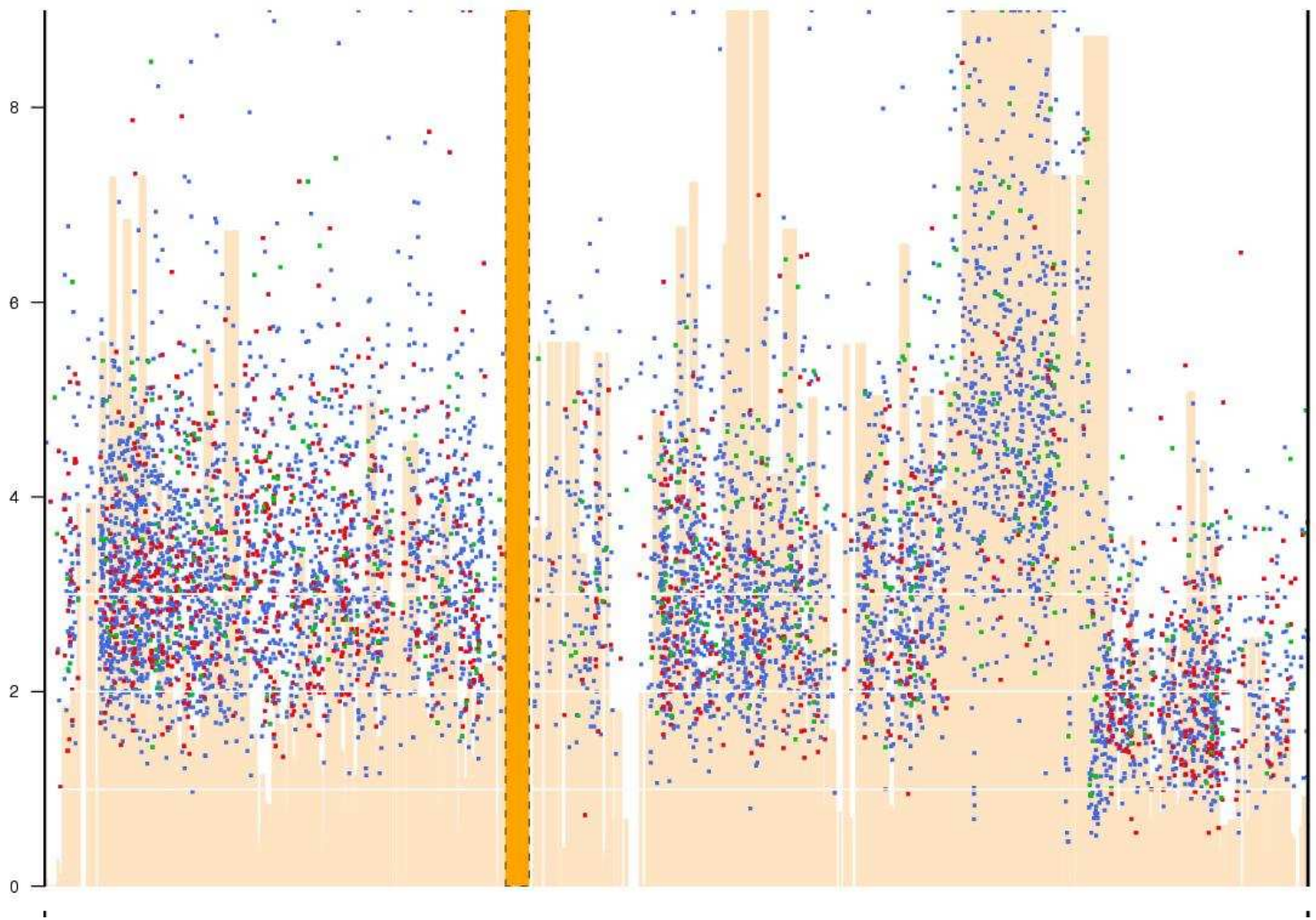
SNP chip data



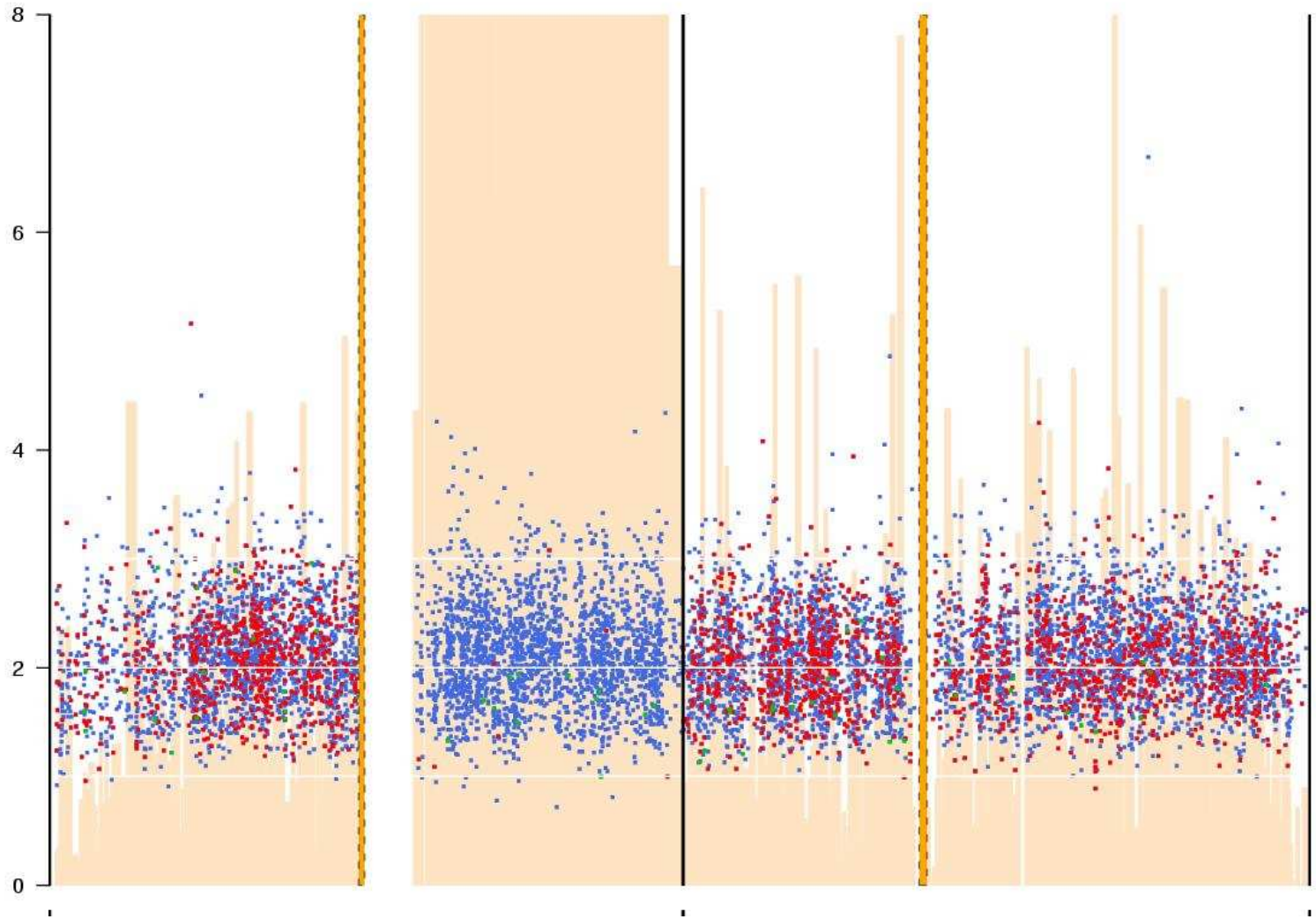
Deletion



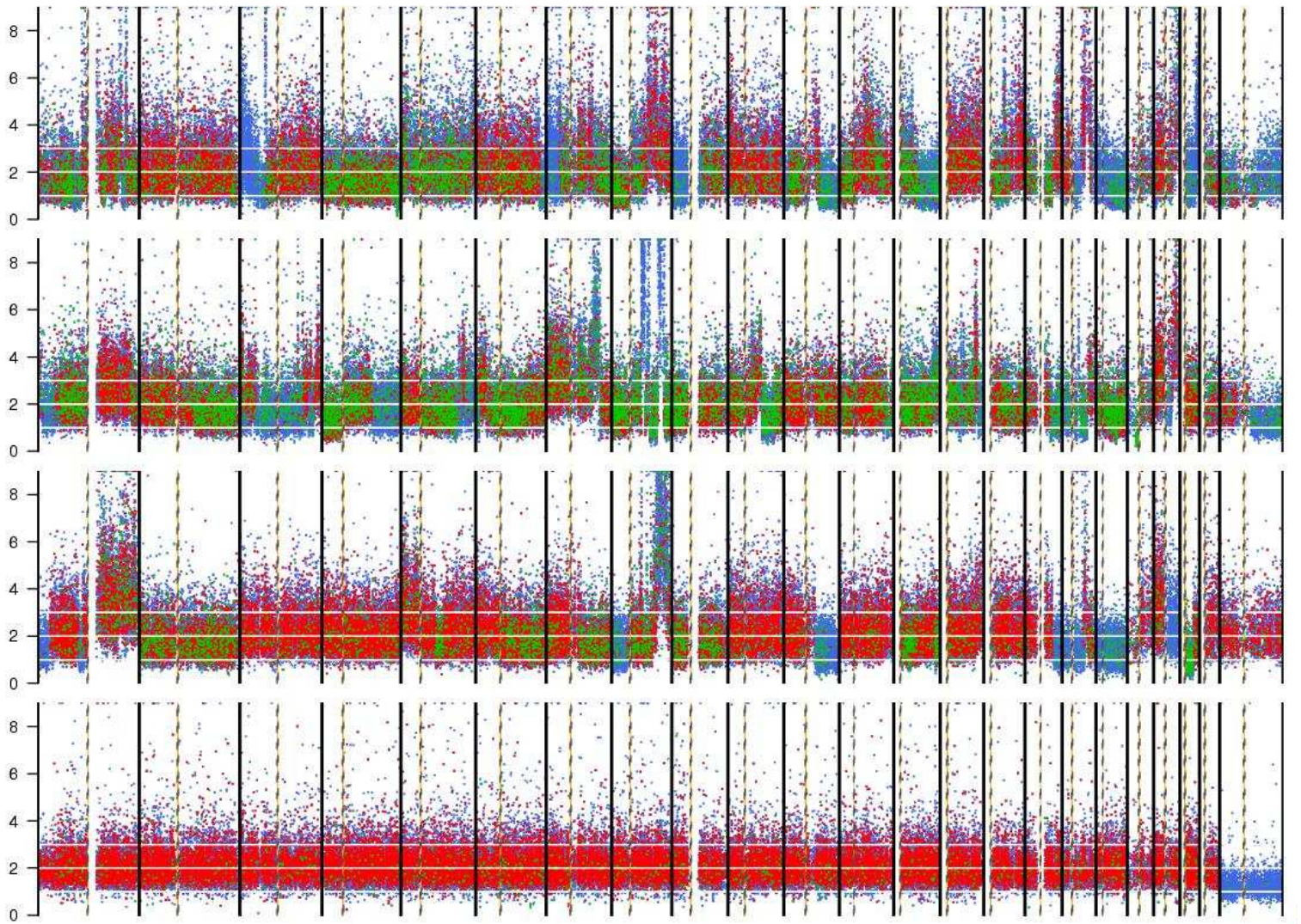
Amplification



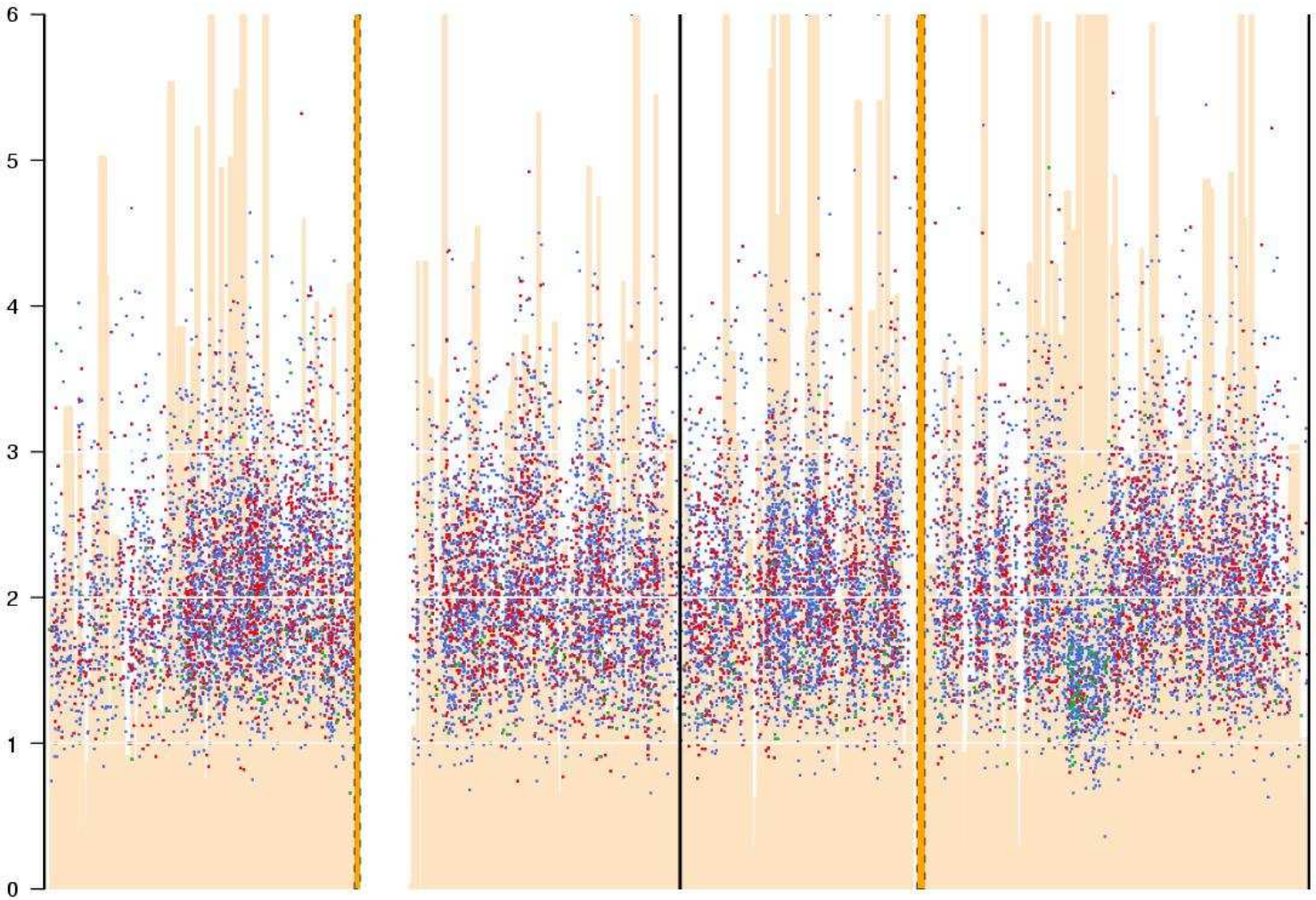
Uniparental Isodisomy



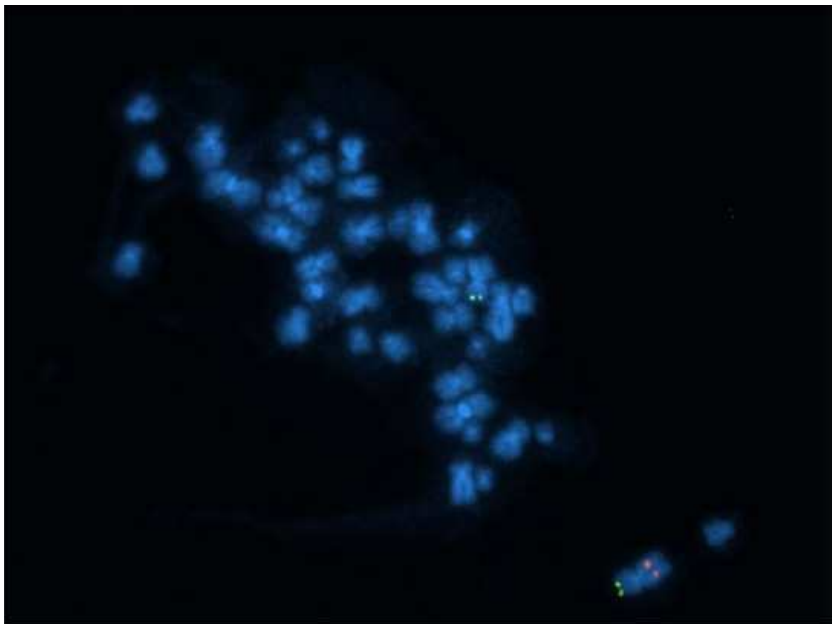
Cancer samples

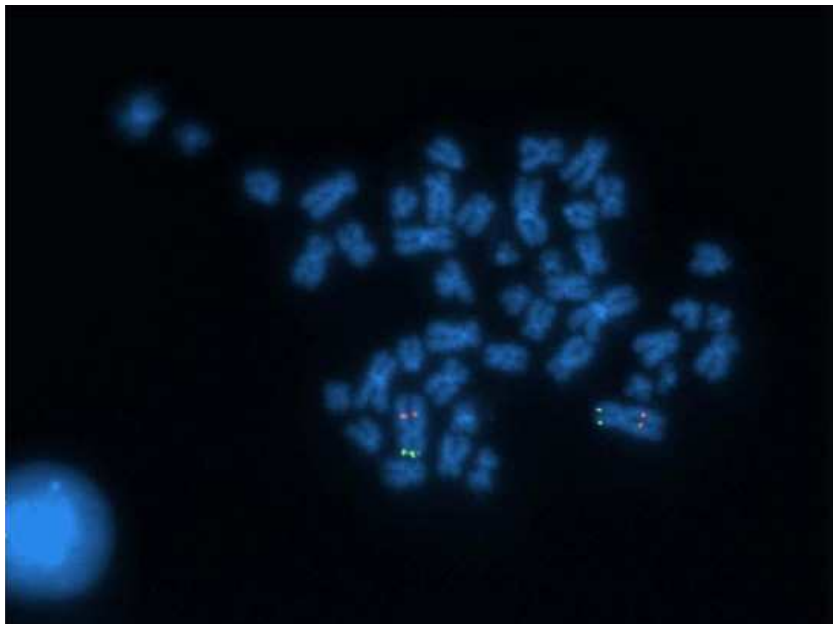


Mosaicism



Mosaicism





Estimation

1. By SNP:

   Estimate genotype and copy number for each SNP.

2. Within a sample:

   Borrow strength between SNPs to infer regions of LOH and copy number changes.

3. Between samples:

   Comparison between normal and disease populations to find chromosomal alterations associated with disease.

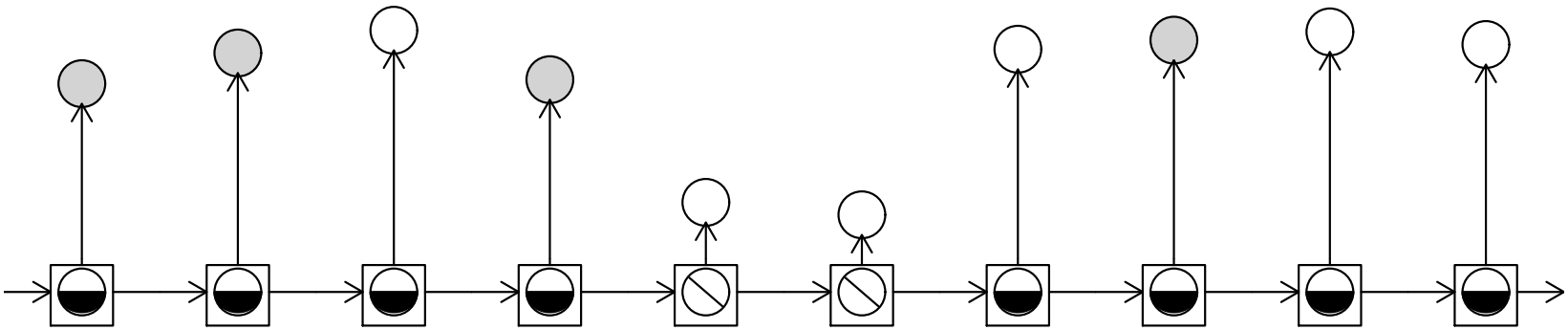
SNPchip S4 classes and methods

```
> merged <- subset(merged, samples = 1)
> summary(merged)
$NA06985
                        Chr 1 Chr 2 Chr 3 Chr 4 Chr 5 Chr 6 Chr 7 Chr 8 Chr 9
mean copy number         2.06  2.07  2.09  2.15  2.08  2.09  2.09  2.06  2.06
sd copy number           0.45  0.45  0.43  0.46  0.45  0.46  0.45  0.43  0.45
% heterozygous calls     0.26  0.26  0.27  0.27  0.27  0.29  0.28  0.28  0.27
% homozygous calls       0.73  0.72  0.72  0.73  0.72  0.69  0.71  0.71  0.72
% no calls               0.01  0.01  0.01  0.01  0.01  0.01  0.01  0.01  0.01
                        Chr 10 Chr 11 Chr 12 Chr 13 Chr 14 Chr 15 Chr 16 Chr 17
mean copy number          2.04   2.06   2.06   2.11   2.09   2.02   2.00   1.98
sd copy number            0.45   0.45   0.46   0.46   0.46   0.43   0.44   0.44
% heterozygous calls      0.30   0.28   0.25   0.27   0.26   0.25   0.26   0.29
% homozygous calls        0.69   0.70   0.74   0.72   0.73   0.74   0.72   0.70
% no calls                0.01   0.01   0.01   0.01   0.01   0.01   0.01   0.01
                        Chr 18 Chr 19 Chr 20 Chr 21 Chr 22 Chr X Total (autosomes)
mean copy number          2.09   1.98   1.98   2.10   1.92  2.13                2.07
sd copy number            0.46   0.44   0.43   0.47   0.44  0.44                0.45
% heterozygous calls      0.26   0.24   0.30   0.27   0.26  0.26                0.27
% homozygous calls        0.73   0.74   0.68   0.72   0.72  0.73                0.72
% no calls                0.01   0.02   0.01   0.02   0.02  0.01                0.01

>
```

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).



Another HMM...?

Novel (and we believe, 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)

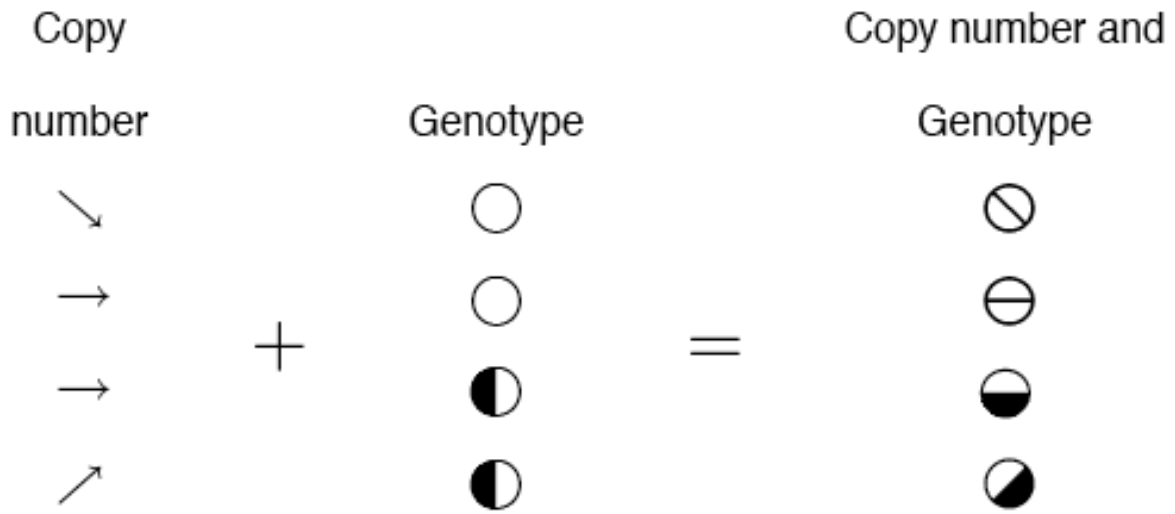
QuantiSNP also models genotype calls and copy number jointly!

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

The Vanilla HMM components

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

Hidden states



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

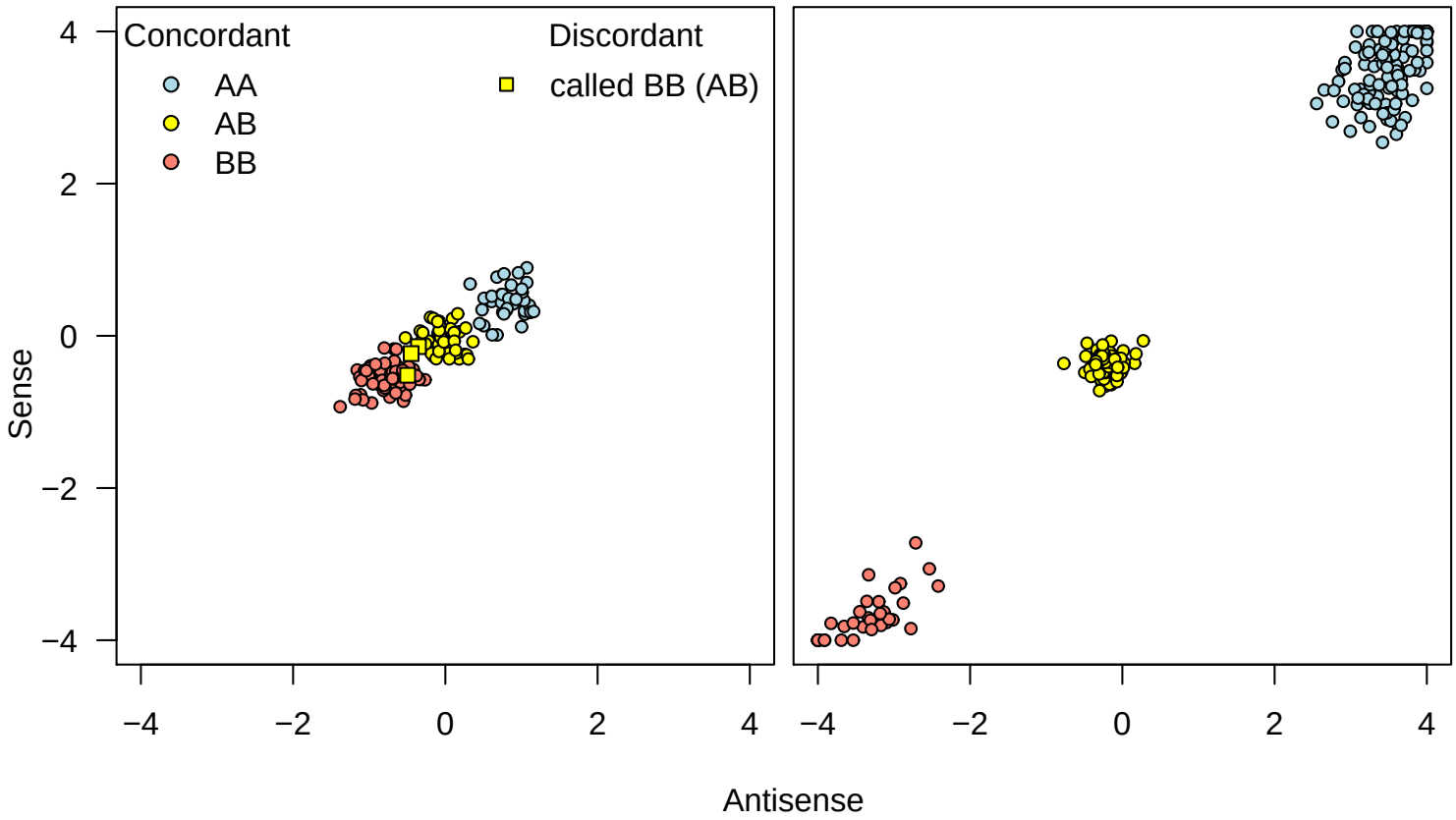
We assume conditional independence between copy number estimates and the genotype calls.

For example:

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

More information

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



Integrating confidence estimates for genotype calls

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

We can estimate from Hapmap the following densities:

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

$\longrightarrow$ Note:

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

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

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

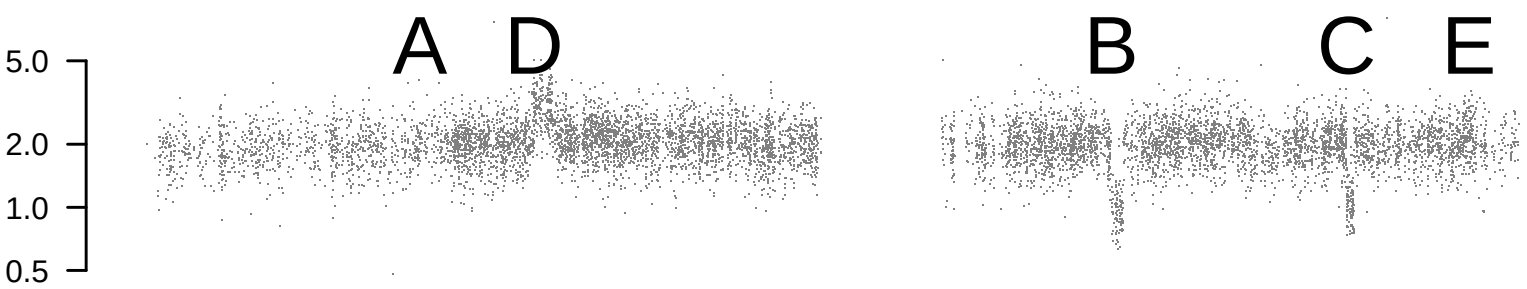
$$\beta_{\bigcirc}\left\{ \widehat{\mathrm{GT}}, \mathrm{S}_{\widehat{\mathrm{GT}}} \right\} = f\left\{ \widehat{\mathrm{GT}} \mid \bigcirc \right\} \times f\left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \widehat{\mathrm{GT}}, \bigcirc \right\}.$$

Emission Probabilities - Retention

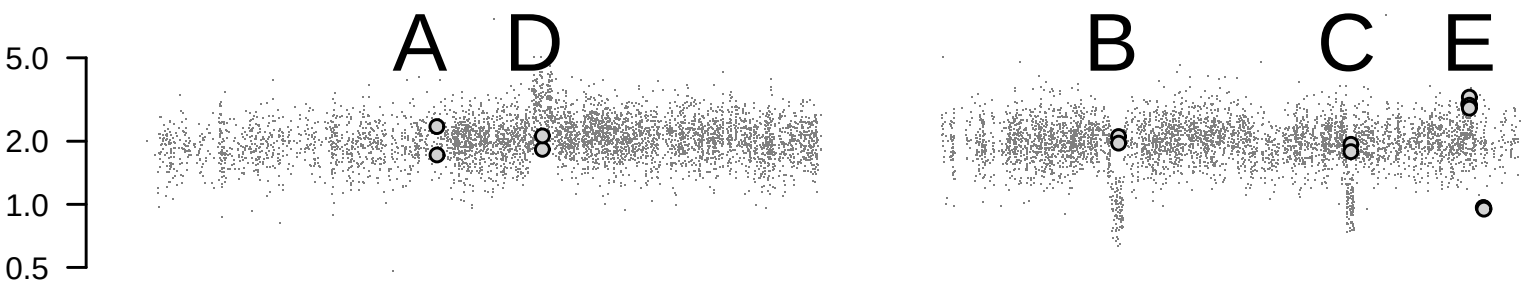
For retention, the true genotype can be HET or HOM:

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

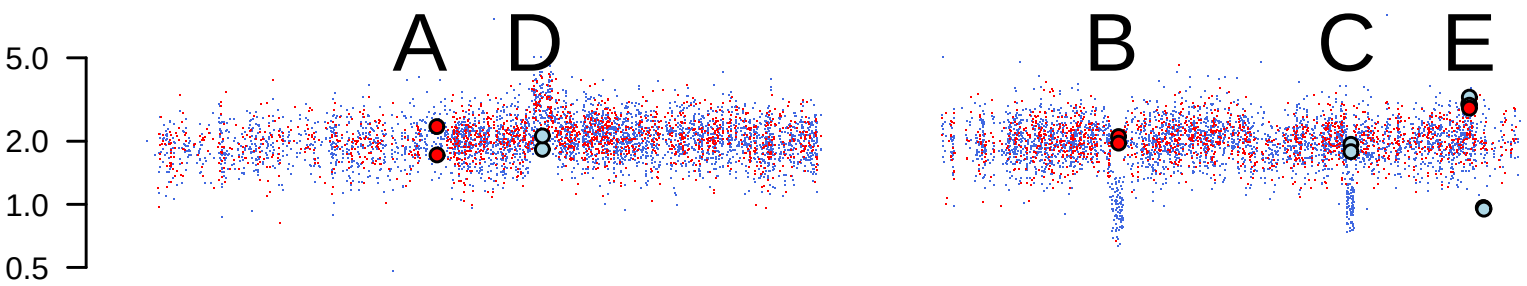
Copy number for chromosome 1



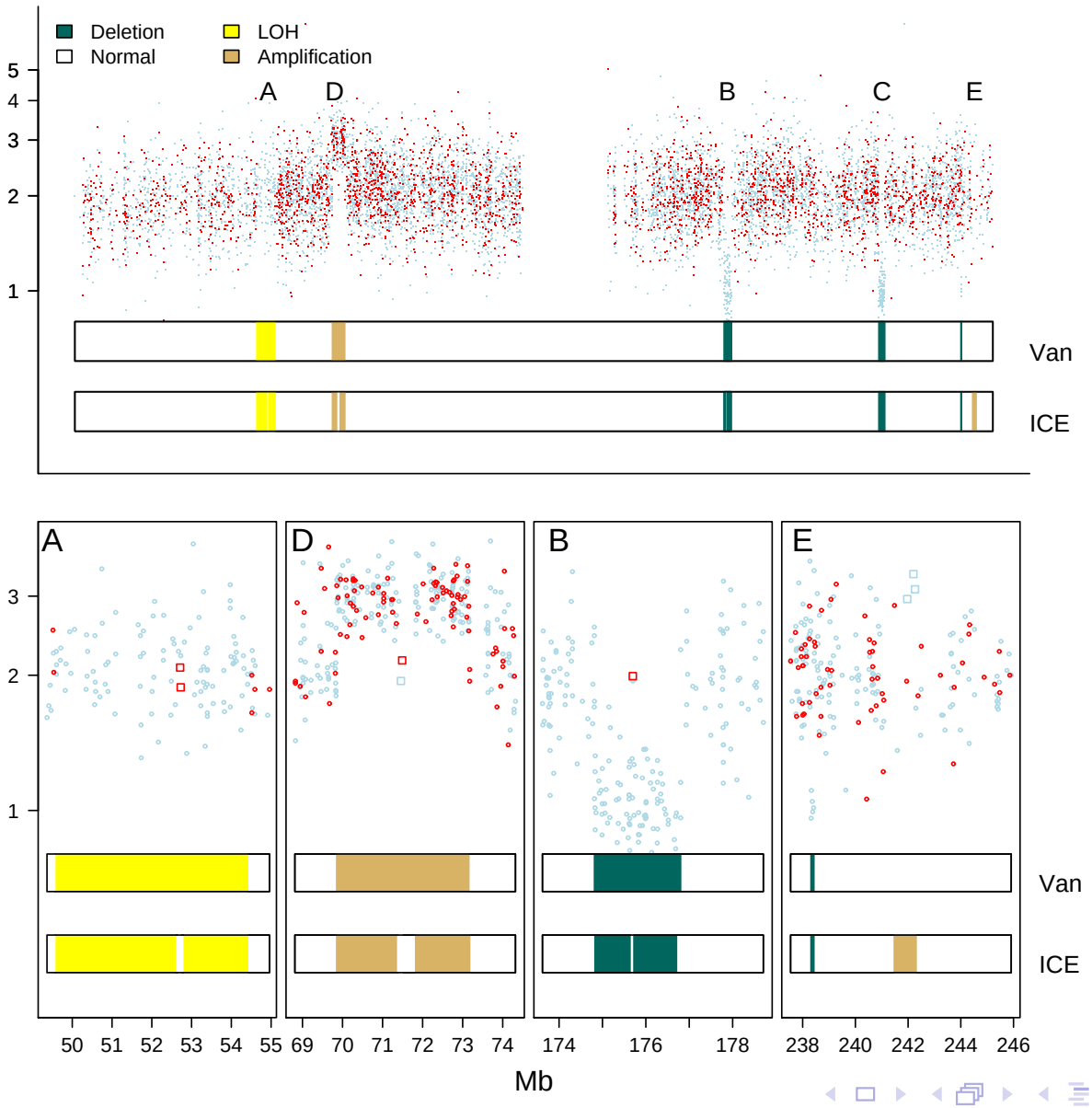
Copy number for chromosome 1



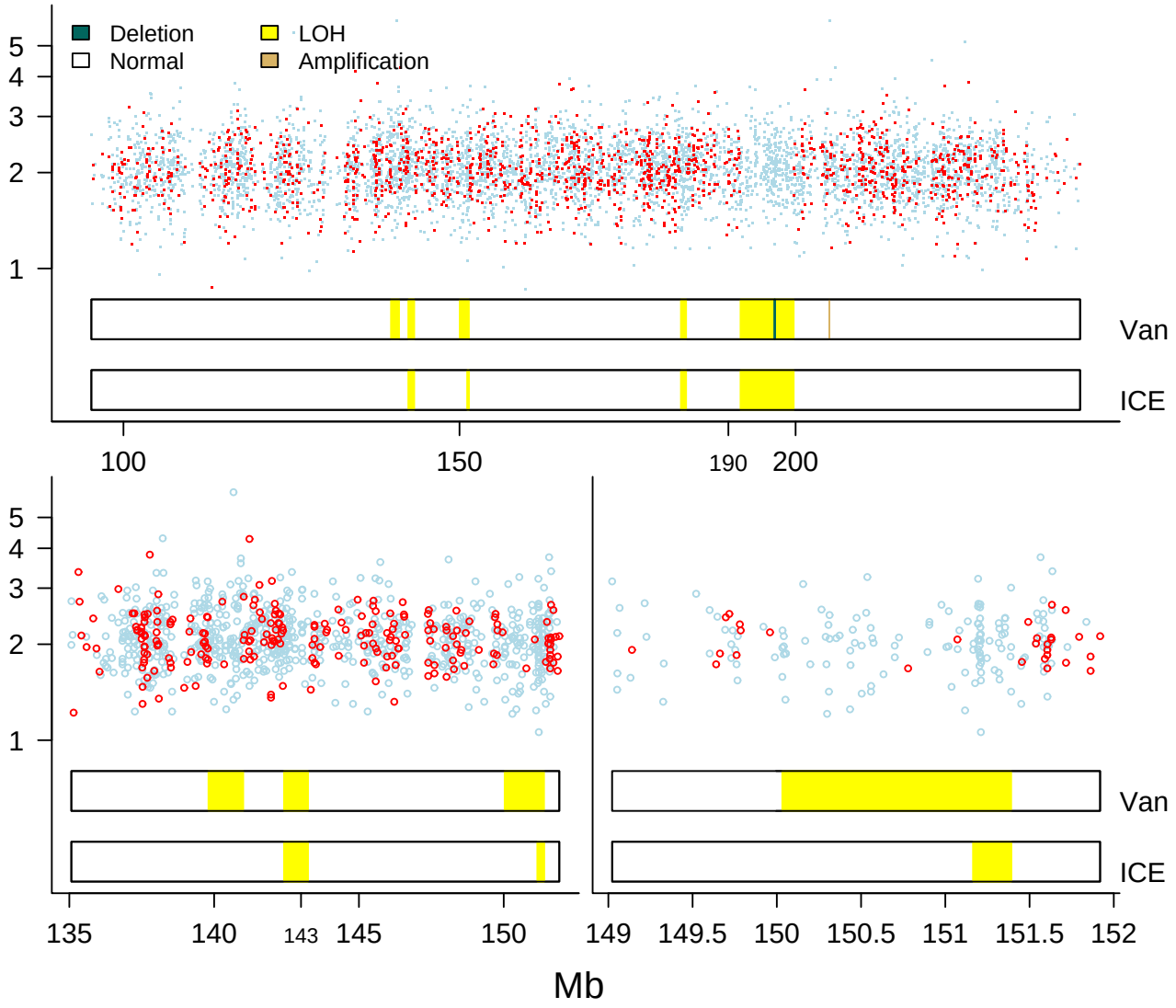
Copy number and genotype calls for chromosome 1



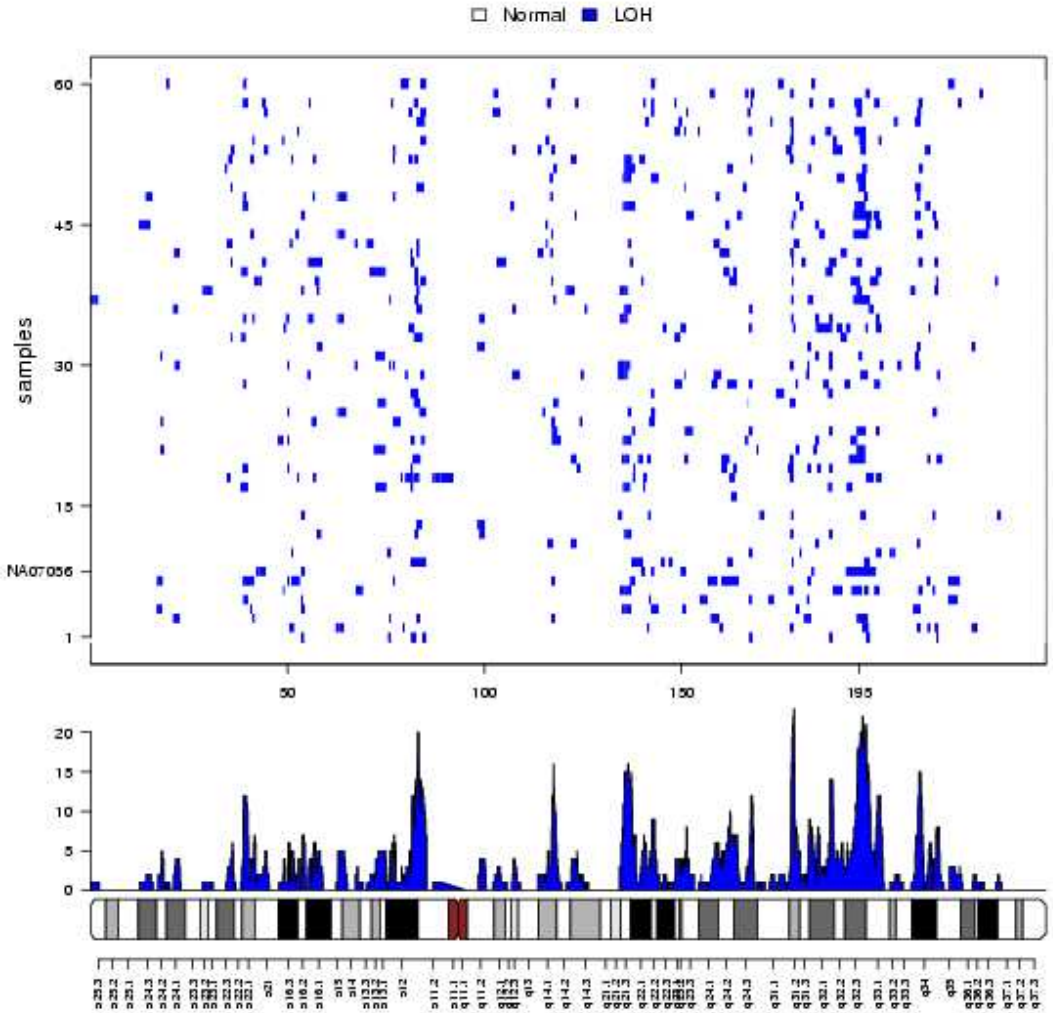
Vanilla and ICE HMMs for genotype and copy number



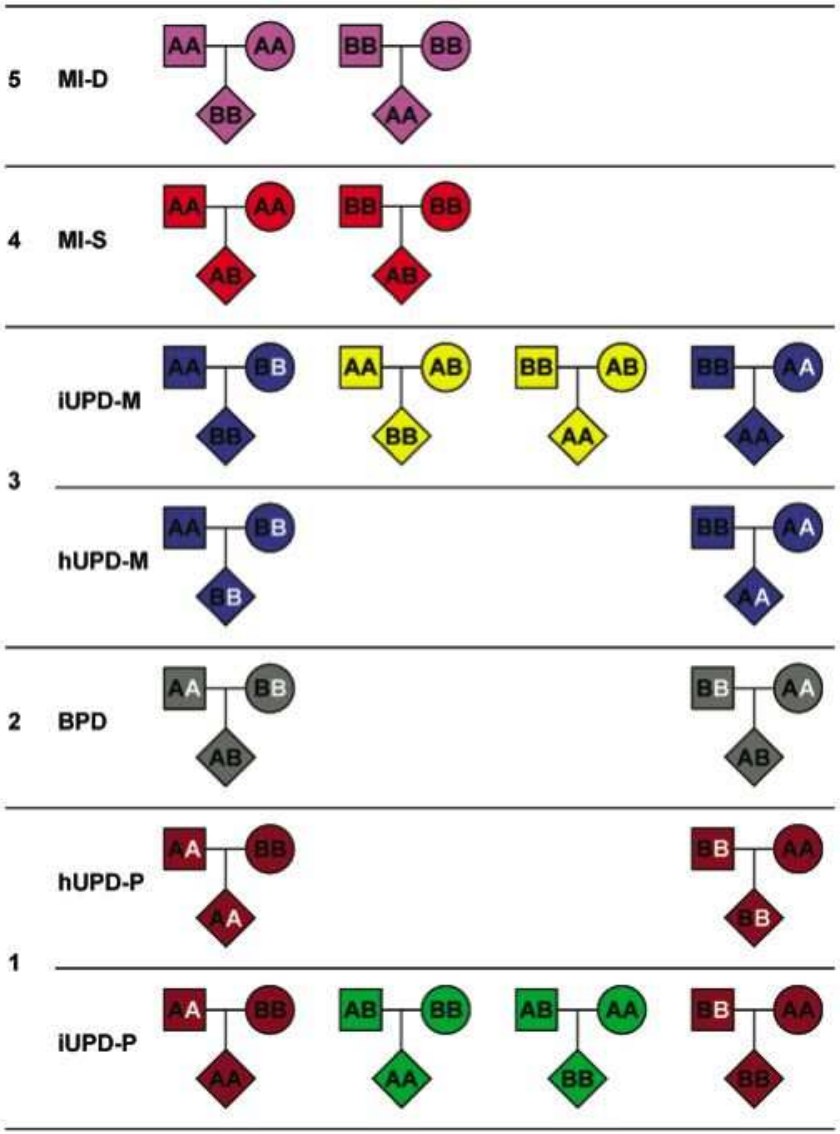
A HapMap sample



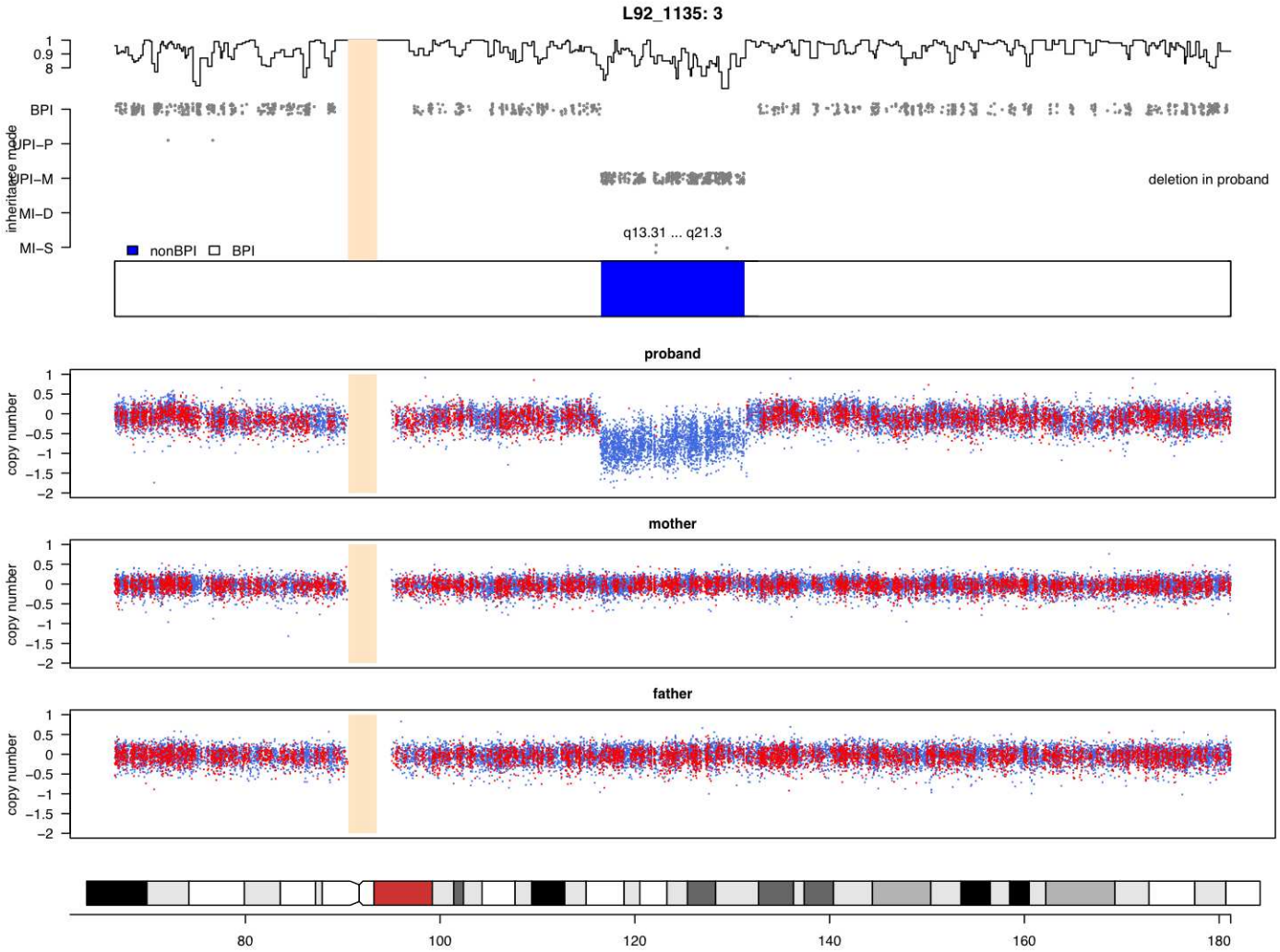
Many HapMap samples



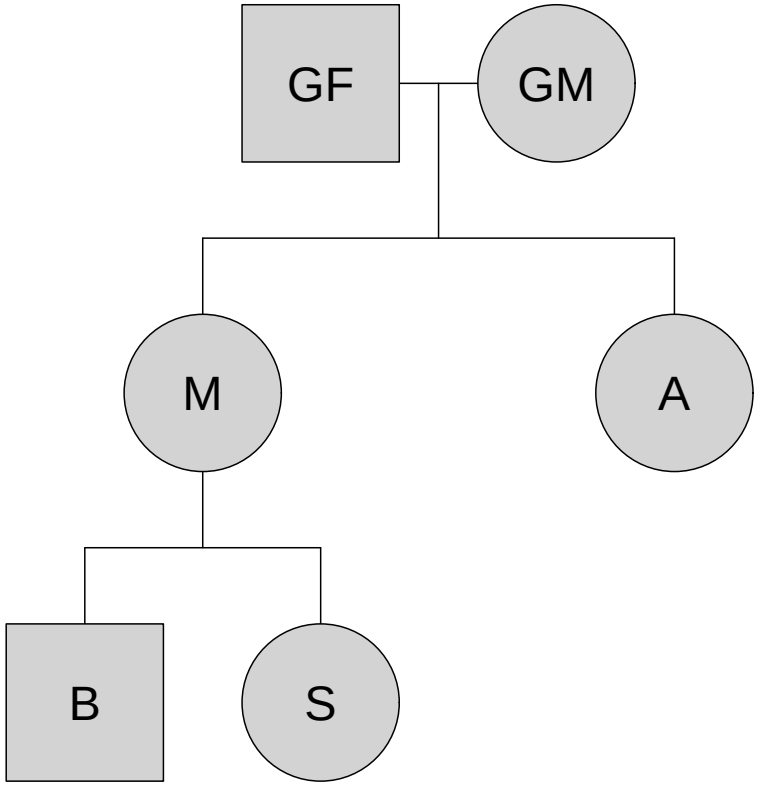
SNP Trio


5
MI-D
AA AA BB
BB BB AA
4
MI-S
AA AA AB
BB BB AB
3
iUPD-M
AA BB BB
AA AB BB
BB AB AA
BB AA AA
hUPD-M
AA BB BB
BB AA AA
2
BPD
AA BB AB
BB AA AB
1
hUPD-P
AA BB AA
BB AA BB
iUPD-P
AA BB AA
AB BB AA
AB AA BB
BB AA BB

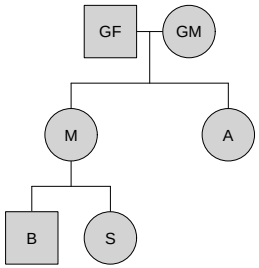
De novo deletion



Homozygous and hemizygous deletions

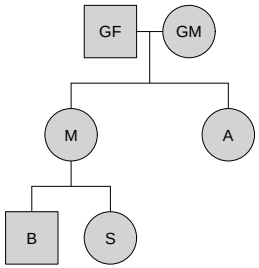


Homozygous and hemizygous deletions



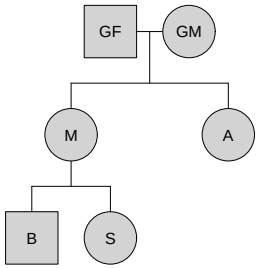
Grandfather		Grandmother		Mother		Aunt		Brother		Sister	
AB	0.01	BB	0.05	AB	-0.11	AB	-0.32	AB	0.06	BB	-0.02
AA	0.27	NC	-5.52	AA	-0.48	AA	-0.45	AB	0.12	BB	-0.42
AB	0.15	NC	-5.04	AA	-0.20	AA	-0.24	AB	-0.09	BB	-0.49
AB	-0.03	NC	-4.59	AA	-0.40	AA	-0.24	AA	0.30	AA	-0.72
BB	0.20	NC	-2.46	NC	-0.38	BB	-0.28	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


Carvalho B, Bengtsson H, Speed TP, Irizarry RA (2007)
Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8(2):485-99.


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 (2007).
Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays *(under revision)*.


Ting JC, Ye Y, Thomas GH, Ruczinski I, Pevsner J (2006).
Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC Bioinformatics*, 7(1):25.


Ting JC, Roberson ED, Miller N, et al, Ruczinski I, Thomas GH, Pevsner J (2007).
Visualization of uniparental inheritance, Mendelian inconsistencies, deletions and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. *Human Mutation*, (in press).


Wang W, Caravalho B, Miller N, Pevsner J, Chakravarti A, Irizarry RA (2006)
Estimating genome-wide copy number using allele specific mixture models. *JHU Biostatistics Working papers*, #122.

http://biostat.jhsph.edu/∼iruczins/