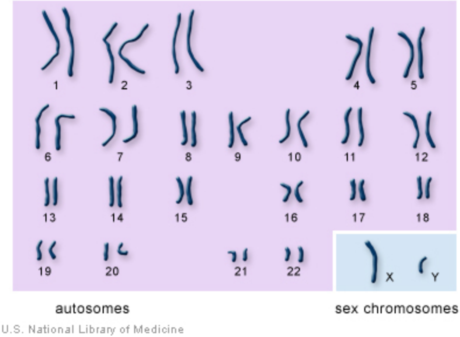
An Integrated Approach for the Assessment of Chromosomal Abnormalities

Ingo Ruczinski

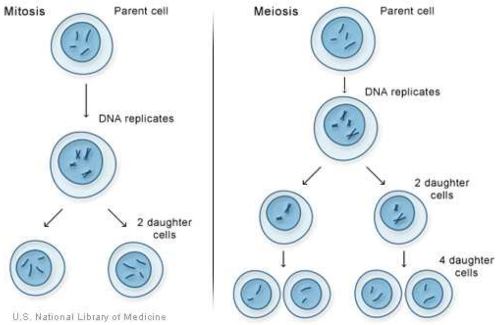
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

June 6, 2007

Karyotypes

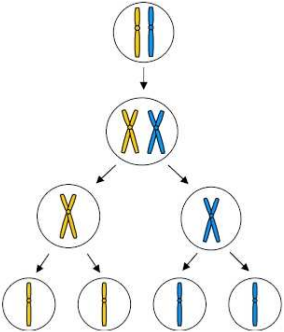

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
autosomes
sex chromosomes
U.S. National Library of Medicine

Mitosis and Meiosis

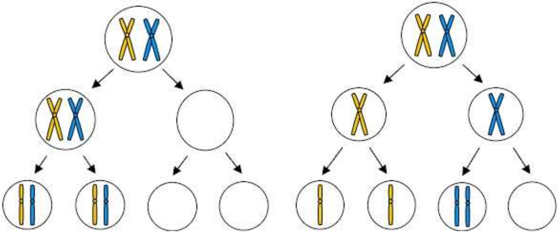


U.S. National Library of Medicine

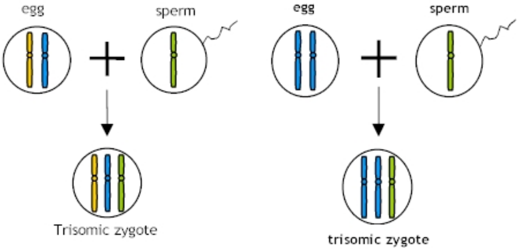
Meiosis



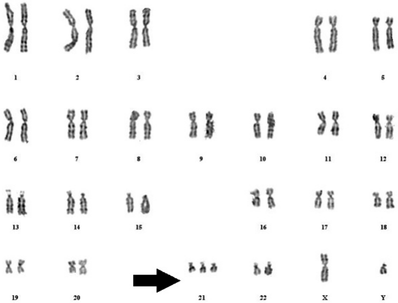
Meiosis Errors



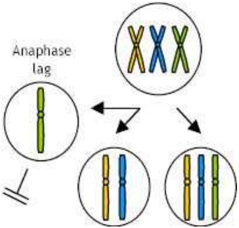
Trisomy

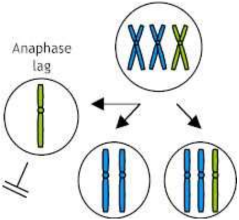


Trisomy

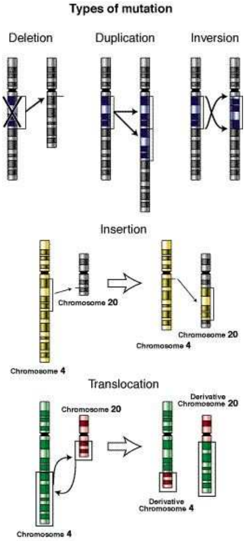


Disomies

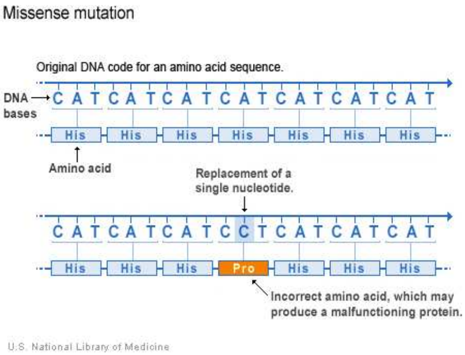




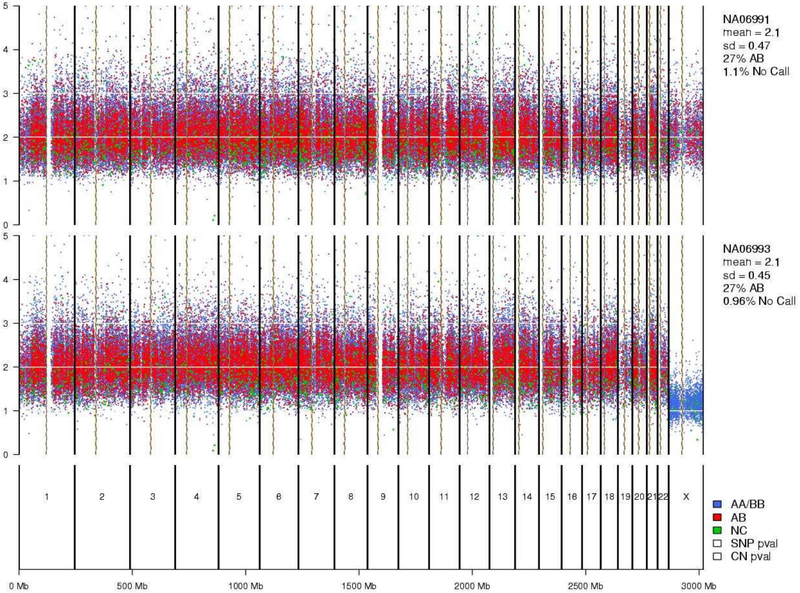
DNA changes


Types of mutation
Deletion
Duplication
Inversion
Insertion
Chromosome 20
Chromosome 4
Chromosome 20
Chromosome 4
Translocation
Chromosome 20
Derivative Chromosome 20
Chromosome 4
Derivative Chromosome 4

DNA changes

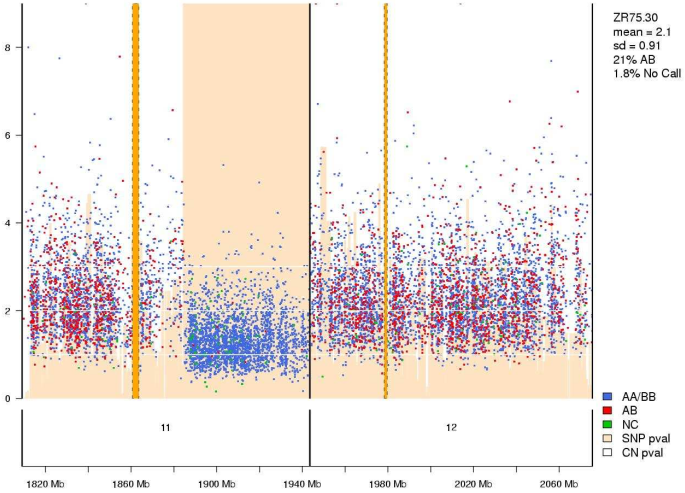

Missense mutation
Original DNA code for an amino acid sequence.
DNA bases
C A T C A T C A T C A T C A T C A T C A T
His His His His His His His
Amino acid
Replacement of a single nucleotide.
C A T C A T C A T C C T C A T C A T C A T
His His His Pro His His His
Incorrect amino acid, which may produce a malfunctioning protein.
U.S. National Library of Medicine

The data

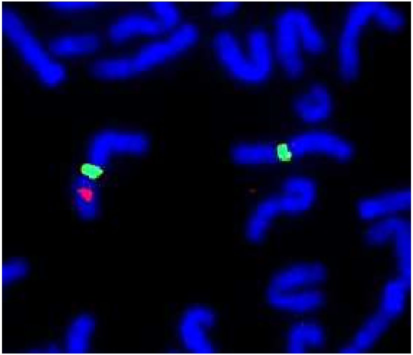


NA06991
mean = 2.1
sd = 0.47
27% AB
1.1% No Call
NA06993
mean = 2.1
sd = 0.45
27% AB
0.96% No Call
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
AA/BB
AB
NC
SNP pval
CN pval
0 Mb
500 Mb
1000 Mb
1500 Mb
2000 Mb
2500 Mb
3000 Mb

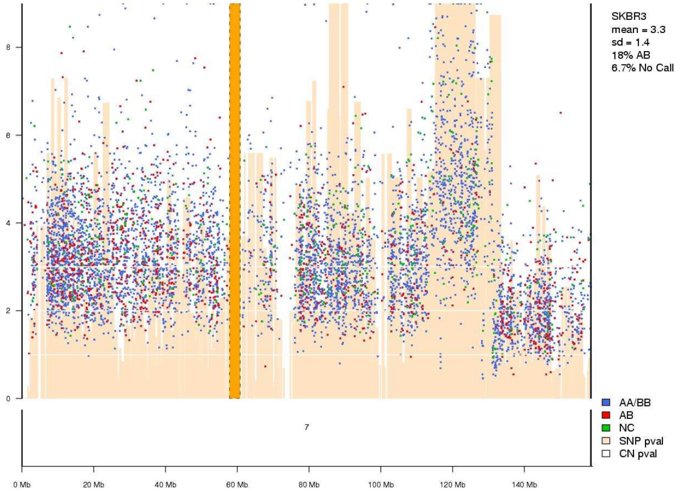
Deletion


ZR75.30
mean = 2.1
sd = 0.91
21% AB
1.8% No Call
0
2
4
6
8
11
12
1820 Mb
1860 Mb
1900 Mb
1940 Mb
1980 Mb
2020 Mb
2060 Mb
🟦 AA/BB
🟥 AB
🟩 NC
□ SNP pval
□ CN pval

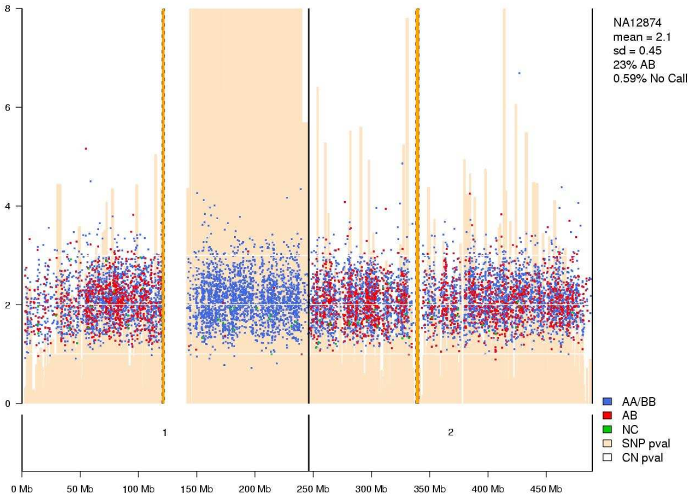
FISH



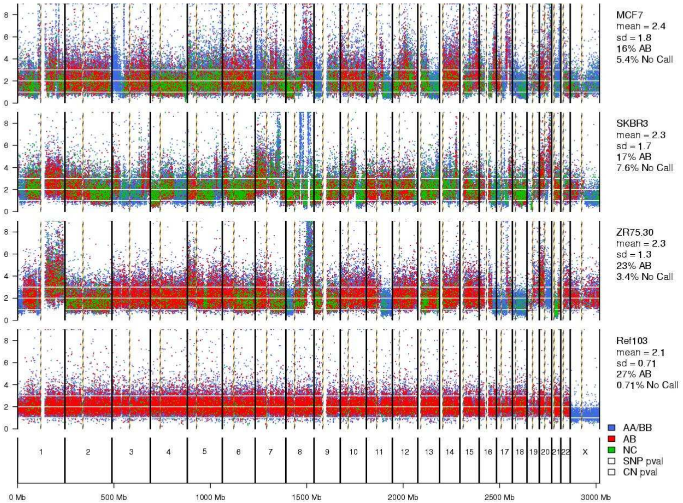
Amplification



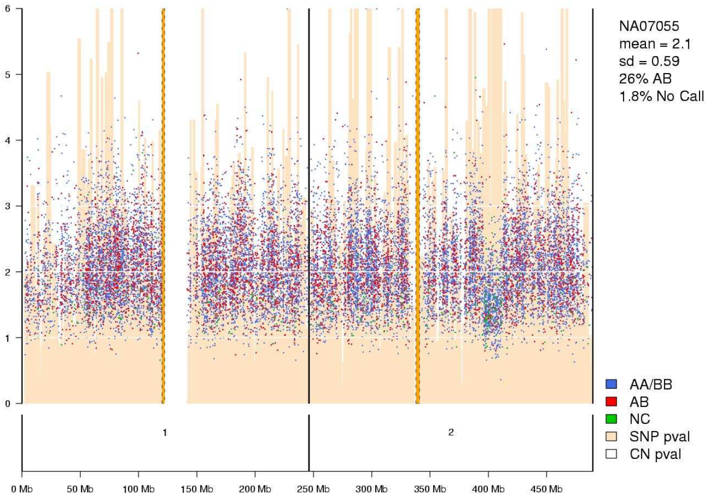
Uniparental Isodisomy



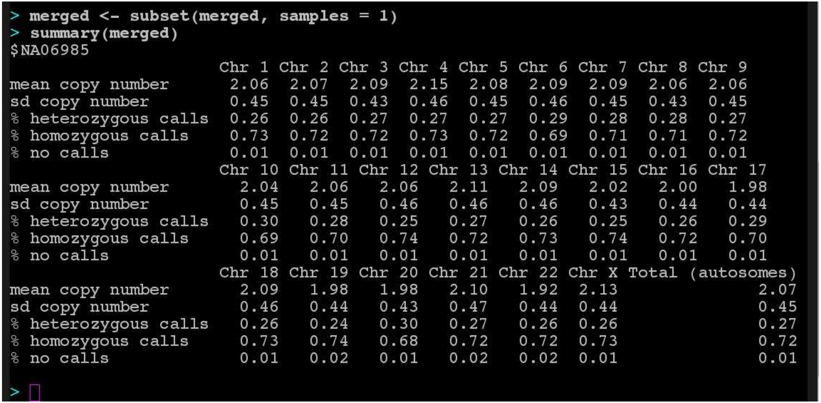
Cancer samples



Mosaicism


6
5
4
3
2
1
0
1
2
0 Mb
50 Mb
100 Mb
150 Mb
200 Mb
250 Mb
300 Mb
350 Mb
400 Mb
450 Mb
NA07055
mean = 2.1
sd = 0.59
26% AB
1.8% No Call
🟦 AA/BB
🟥 AB
🟩 NC
□ SNP pval
□ CN pval

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

>
```

Estimation

1. By SNP:

   Estimate genotype and copy number for each SNP.

2. Within a sample:

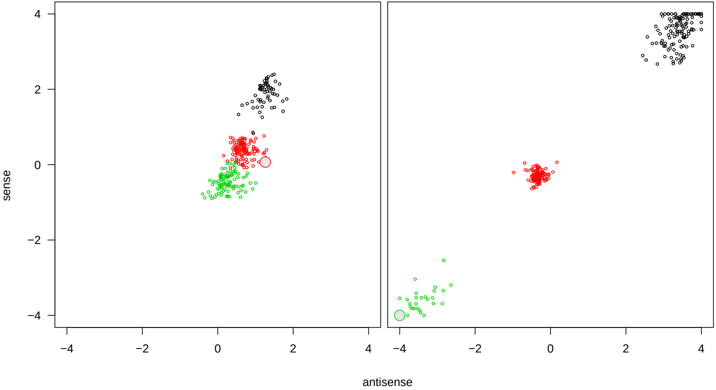
   Borrow strength between SNPs to infer regions of LOH and copy number changes.

3. Between samples:

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

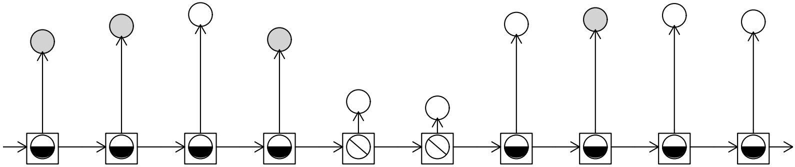
More information

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



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)

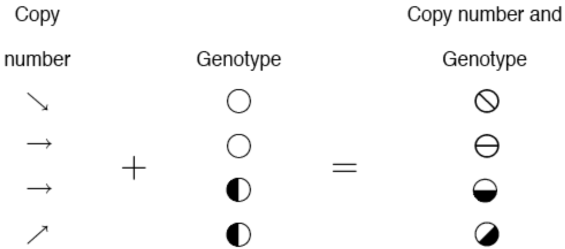
The Vanilla HMM components

- Observations $\widehat{\mathrm{CN}}$ and $\widehat{\mathrm{GT}}$

- Hidden states

- Initial state probability distribution
- Transition probabilities
- Emission probabilities

Hidden states



Copy number

Genotype

Copy number and Genotype

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:

$$
\begin{aligned}
\tau_{\text{◒}|\obslash}(d) &= P\left\{ \text{◒}_{i+1} \mid \obslash_i, d \right\} \\
&= P\left\{ \text{◒}_{i+1}, I \mid \obslash_i, d \right\} + P\left\{ \text{◒}_{i+1}, I^c \mid \obslash_i, d \right\} \\
&= P\left\{ \text{◒}_{i+1} \mid I, \obslash_i, d \right\} \times P\left\{ I \mid \obslash_i, d \right\} + \\
&\quad\; P\left\{ \text{◒}_{i+1} \mid I^c, \obslash_i, d \right\} \times P\left\{ I^c \mid \obslash_i, d \right\} \\
&= P\left\{ \text{◒} \right\} \times \theta(d).
\end{aligned}
$$

Emission probabilities

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

For example:

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

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

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

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

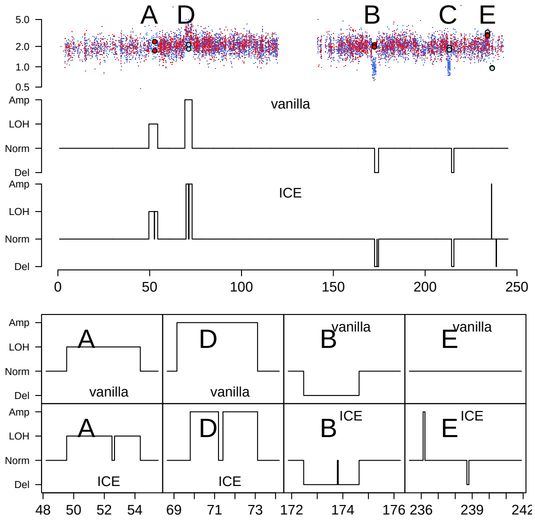
$$\beta_{\bigcirc}\left\{ \widehat{\mathrm{GT}}, \mathrm{S}_{\widehat{\mathrm{GT}}} \right\} \quad = \quad f\left\{ \widehat{\mathrm{GT}} \mid \bigcirc \right\} \times f\left\{ \mathrm{S}_{\widehat{\mathrm{GT}}} \mid \widehat{\mathrm{GT}}, \bigcirc \right\}.$$

Emission Probabilities - Retention

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

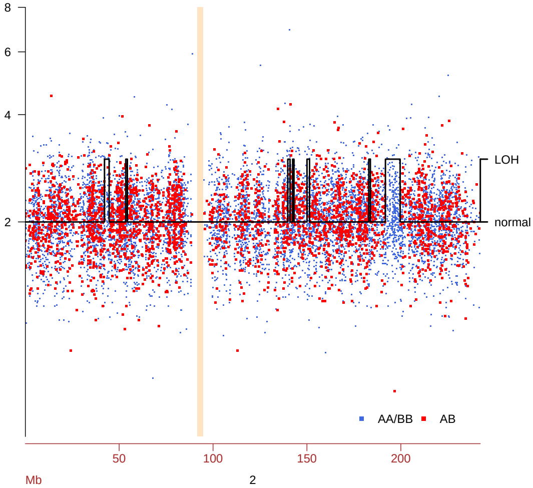
$$
\begin{aligned}
& \beta_{\LEFTcircle}\left\{\widehat{\mathrm{GT}}, \mathrm{S}_{\widehat{\mathrm{GT}}}\right\} \\
= \quad & f\left\{\widehat{\mathrm{GT}} \mid \LEFTcircle\right\} f\left\{\mathrm{S}_{\widehat{\mathrm{GT}}} \mid \widehat{\mathrm{GT}}, \LEFTcircle\right\} \\
= \quad & f\left\{\widehat{\mathrm{GT}}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

Vanilla ICE comparison

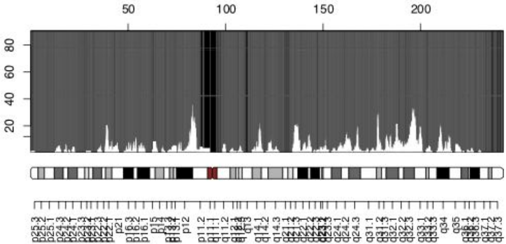


Bioconductor package: ICE

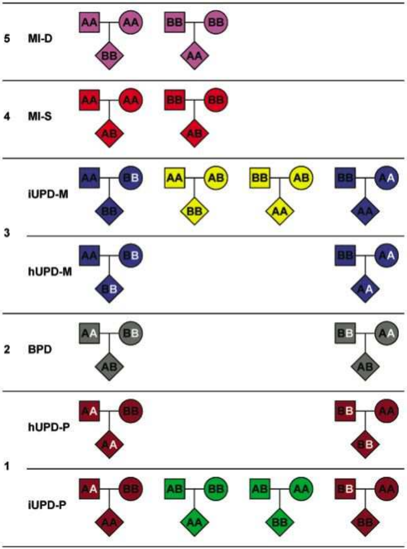
A HapMap sample



Many HapMap samples

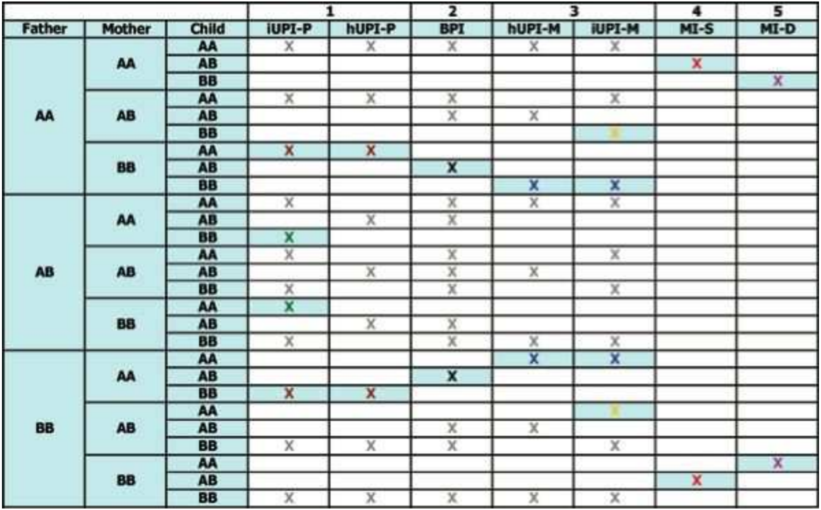


SNP Trio



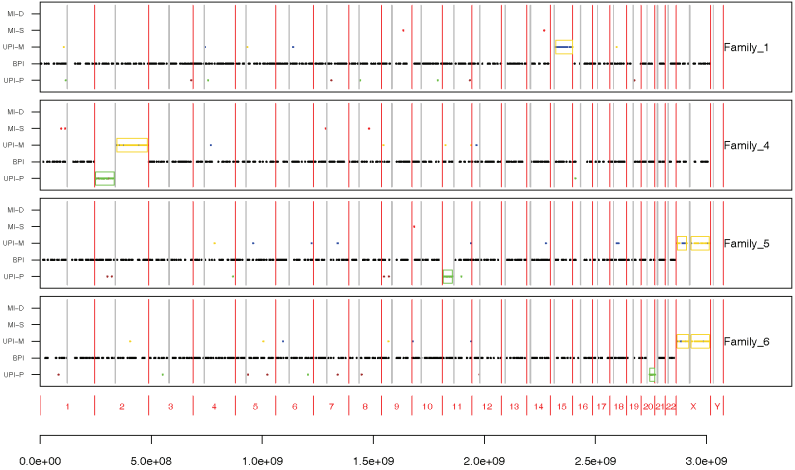
5
MI-D
AA
AA
BB
BB
BB
AA
4
MI-S
AA
AA
AB
BB
BB
AB
3
iUPD-M
AA
BB
BB
AA
AB
BB
BB
AB
AA
BB
AA
AA
hUPD-M
AA
BB
BB
BB
AA
AA
2
BPD
AA
BB
AB
BB
AA
AB
1
hUPD-P
AA
BB
AA
BB
AA
BB
iUPD-P
AA
BB
AA
AB
BB
AA
AB
AA
BB
BB
AA
BB

SNP Trio



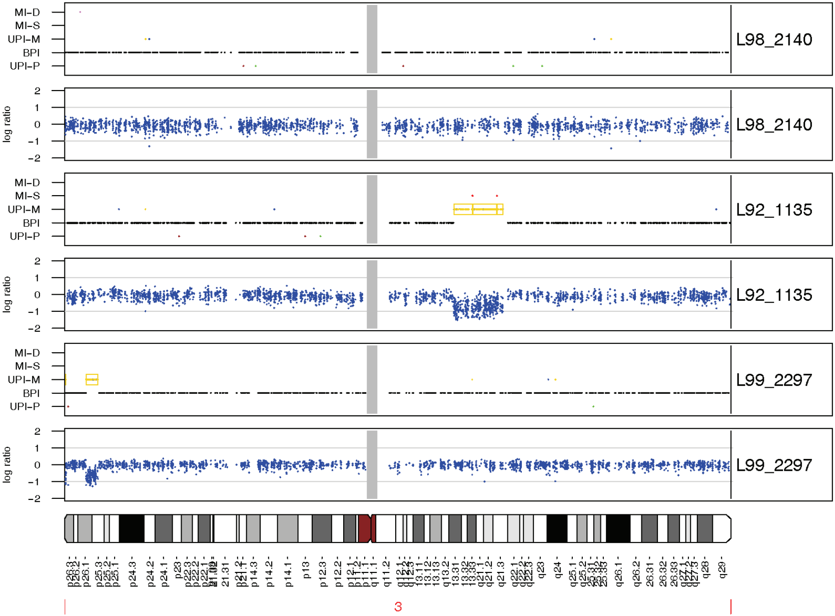
			1		2	3		4	5
Father	Mother	Child	iUPI-P	hUPI-P	BPI	hUPI-M	iUPI-M	MI-S	MI-D
AA	AA	AA	X	X	X	X	X		
		AB						X	
		BB							X
	AB	AA	X	X	X		X		
		AB			X	X			
		BB					X		
	BB	AA	X	X					
		AB			X				
		BB				X	X		
AB	AA	AA	X		X	X	X		
		AB		X	X				
		BB	X						
	AB	AA	X		X		X		
		AB		X	X	X			
		BB	X		X		X		
	BB	AA	X						
		AB		X	X				
		BB	X		X	X	X		
BB	AA	AA				X	X		
		AB			X				
		BB	X	X					
	AB	AA					X		
		AB			X	X			
		BB	X	X	X		X		
	BB	AA							X
		AB						X	
		BB	X	X	X	X	X		

SNP Trio



MI–D
MI–S
UPI–M
BPI
UPI–P
Family_1
Family_4
Family_5
Family_6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
2.5e+09
3.0e+09

SNP Trio



MI-D
MI-S
UPI-M
BPI
UPI-P
L98_2140
log ratio
2
1
0
−1
−2
L98_2140
MI-D
MI-S
UPI-M
BPI
UPI-P
L92_1135
log ratio
2
1
0
−1
−2
L92_1135
MI-D
MI-S
UPI-M
BPI
UPI-P
L99_2297
log ratio
2
1
0
−1
−2
L99_2297
p26.3
p26.2
p26.1
p25.3
p25.2
p25.1
p24.3
p24.2
p24.1
p23
p22.3
p22.2
p22.1
21.31
p21.2
p21.1
p14.3
p14.2
p14.1
p13
p12.3
p12.2
p12.1
p11.2
p11.1
q11.1
q11.2
q12.1
q12.2
q12.3
13.11
13.12
13.13
q13.2
13.31
13.32
13.33
q21.1
q21.2
q21.3
q22.1
q22.2
q22.3
q23
q24
q25.1
q25.2
25.31
25.32
25.33
q26.1
q26.2
26.31
26.32
26.33
q27.1
q27.2
q27.3
q28
q29
3

Probability

$\longrightarrow$ Given the number of SNPs for a particular event region type (M), among all uploaded informative autosomal SNPs (X) of that trio, what is the probability to have the observed number (N) or more informative non-BPI SNPs of that type clustered together solely by chance?

In other words, if the M SNPs of a particular event region type were randomly dispersed among the X informative autosomal SNPs, how probable is it to observe an event of the same or larger magnitude (defined by the number of consecutive SNPs of that event region)?

Probability

- Let p be the probability of a positive outcome in a Bernoulli trial (e.g. a 1 for 0/1 outcomes).
- Assume that we have Z trials.
- Let LS be the length of the largest block of consecutive ones in those Z trials.
- Let $P_Z(LS < N)$ be the probability that LS is smaller than some number N.

Probability

- If $Z < N$, then $P_Z(LS < N) = 1$.
- If $Z = N$, then $P_Z(LS < N) = 1 - p^N$.
- If $Z > N$, then

$$P_Z(LS < N) = \sum_{k=0}^{N-1} p^k (1-p) P_{Z-k-1}(LS < N)$$

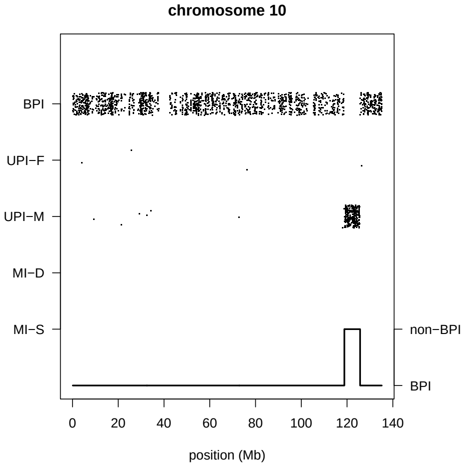
Probability

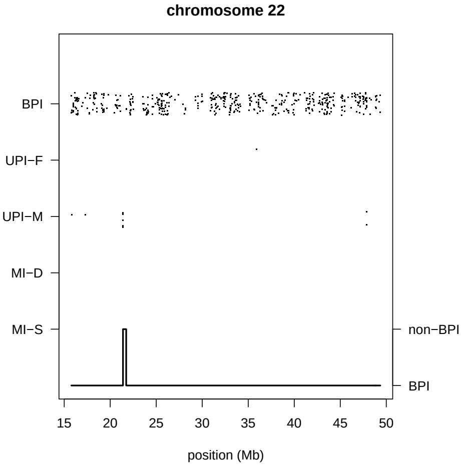
- Tabulate the probabilities $P_1(LS < N), \ldots, P_N(LS < N)$.

- Calculate $P_{N+1}(LS < N), P_{N+2}(LS < N), \ldots, P_X(LS < N)$, iteratively.

- Using $p = M/X$, the probability of having N or more informative SNPs of a particular non-BPI type clustered together solely by chance is equal to $1 - P_X(LS < N)$.

HMM for SNP Trio





Acknowledgments

- Rob Scharpf
- Giovanni Parmigiani
- Rafael Irizarry & Gang

  Benilton Carvalho, Wenyi Wang
- Jonathan Pevsner & Lab

  Nate Miller, Eli Roberson, Jason Ting

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, Ruczinski I (2007).
  A hidden markov model for joint estimation of genotype and copy number in high-throughput SNP chips. *JHU Biostatistics Working papers*, #136.

- 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.