Visualizing and Analyzing High Density SNP Data with SNPscan

Ingo Ruczinski and Robert Scharpf
in collaboration with Jonathan Pevsner and Jason Ting

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

ENAR, 2006

Outline

1. Background
   - SNPscan: the Biologists
   - The Biology
2. Statistical challenges
3. The R package
   - Bioconductor

Outline

1. Background
   - SNPscan: the Biologists
   - The Biology
2. Statistical challenges
3. The R package
   - Bioconductor

SNPscan web tool

SNPscan at the Pevsnerlab

http://pevsnerlab.kennedykrieger.org/snpscan_01_snpscan.htm

Ingo's Pond | News | Links | Running | Science | Travel

SNPscan

Home | DRAGON | SNOMAD | SNPscan | Bioinformatics | Bioinformatics Course | Microarrays | Lead Poisoning
Autism and Rett Syndrome | Down Syndrome | Leonardo da Vinci | Publications | Lab Personnel & Contact Info

Welcome to SNPscan. Your input file should be a text output file exported from Affymetrix's Copy Number Analysis Tool (CNAT) v2.0 or later, with the first line ("Copy Number Analysis Tool" form CNAT 2.0/2.1 or "Copy Number Viewer" from CNAT 3.0) been deleted. If you use CNAT 3.0 or GTYPE 4.0, do not use any file saved in the ".cnt" file format as your input to SNPscan. Select "Export", "Table" form GTYPE's CNAT Viewer menu bar to save you SNP data as a tab-delimited file. The first line of your input file shall be ColumnNanes (column headers), followed by Data lines, one line per SNP. The first field of each Data lines shall be a serial number for that SNP. SNPscan expects the ColumnNames to be in the exact format exported by the Copy Number Analysis Tool (XXX_Call, XXX_SPA_CN, etc.) Visit the sample data page to see examples of the appropriate formats.

The current file size limitation is 40 MB per upload. It's enough for 10 100K SNP cases. You can increase this limitation by installing SNPscan locally (instructions).

The Copy Number of each SNP will be plotted as dots against chromosomal positions, color coded according to each SNP's call: AA, BB, AB, NoCall. In addition, LOH and p-value can also be plotted, as gray bars and yellow lines, respectively.

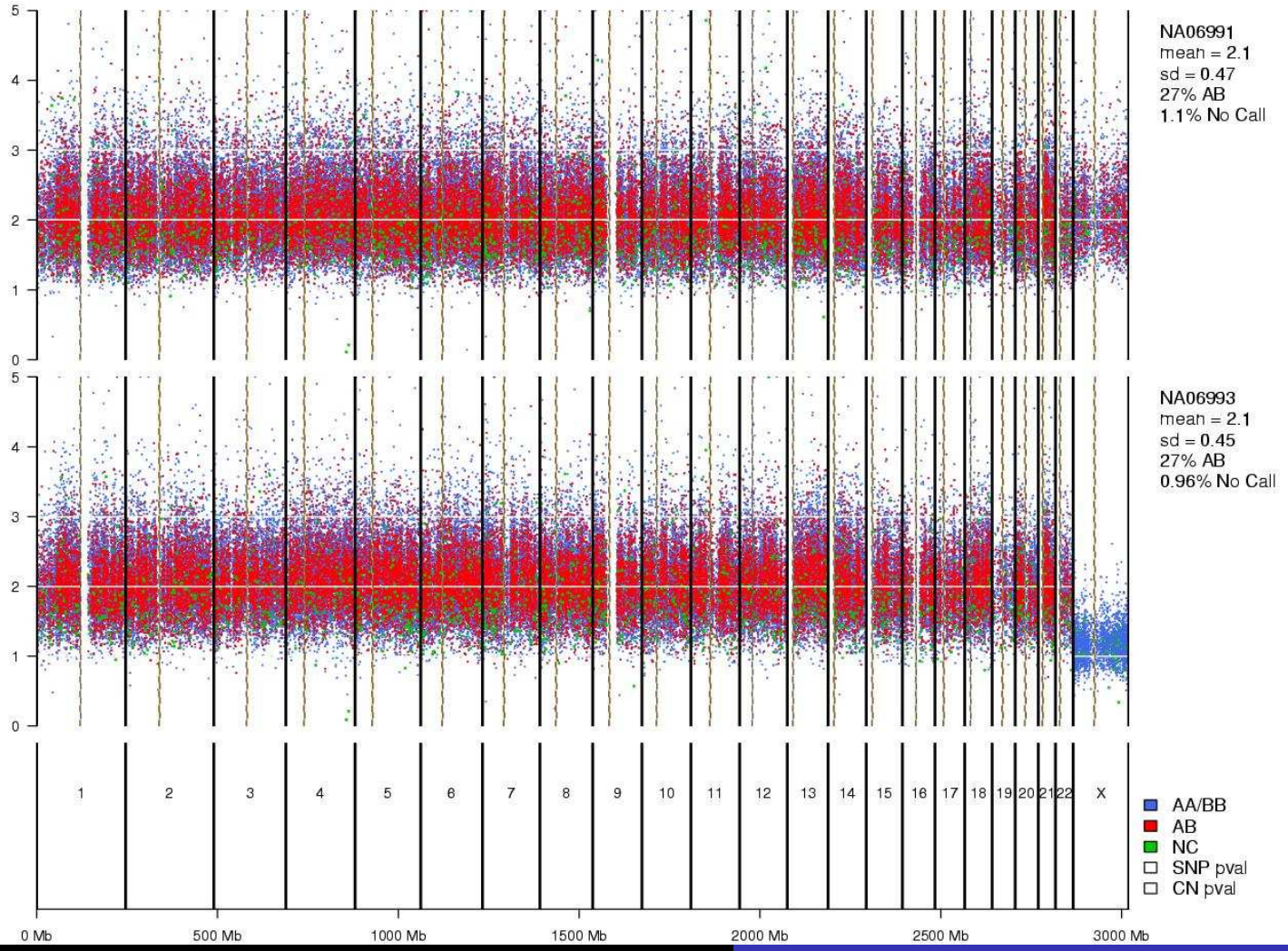
Step 1	Please choose a file to upload: (If you want to try a sample file, visit the sample data page, or try a 10K dataset by clicking here and saving as a text file to your hard drive.)	Choose File no file selected
Step 2	Enter the file name of your selection here: (For example, for file C:\Documents and Settings\user\Desktop\data.txt, enter data.txt. This step is required, and it allows SNPscan to work with a variety of different browsers and operating systems.)	
Step 3	The height of plot in inches:	8
Step 4	The width of plot in inches:	11
Step 5	The width ratio of label vs. plot:	0.12
Step 6	Upper limit for the Y-axis (Copy Number): (Leave this blank to plot the whole range; for most applications, try the default value 8)	8
Step 7	Specify chromosomes: (For most applications, leave this blank to select all chromosomes as a default. You can enter any chromosome numbers separated by spaces [e.g. 1 9 21], or a single chromosome group letter [A-G,X,Y]. Please use capital letters (e.g. X not x). Enter asterisk symbol '*' for all autosomes.)	
Step 8	Plot LOH? (For most applications, select Yes.)	No Yes
Step 9	Plot *p* value? (Choose SPA to use single point p_values in your plot. The regional smoothed p_value is called CPA for 10K data, and GSA for 50K and 100K data. Reginal smoothed p_value is recommended. If your dataset is noisy [high variance], turn this off so that you can see your results more easily. Your plot will not be completed if you selected a wrong type here.)	No SPA_pValue (10K, 50K, 100K) CPA_pValue (10K only) GSA_pValue (50K & 100K)
Step 10	Plot comparison of paired data? Homoz. unchanged: gray; Heteroz. unchanged: gray; Homoz. to Heteroz.: green; Heteroz. to Homoz.: red; NoCall in #1: none; NoCall in #2: none; NoCall in both: none; Diploid Switch: black (If your data file includes multiple SNP data sets, each pair of datasets will be analyzed with an output of two tracks plus a third track showing the differences as specified above. Step 10 is typically used for paired tumor versus normal sample comparisons.)	No Yes
Step 11	Your data is based on which NCBI Build? (Build 34 is also known as "July 2003", or "hg16" in UCSC Genome Browser. Build 35 is known as "May 2004", or "hg17", respectively. As of December 2005, your Affymetrix SNP data is most likely to be based on Build 34. Affy's CNAT 3.0 uses Build 35.)	Build 34 Build 35
Step 12	Click on the Submit button: Depending on the size of your file, the work load of the server and the traffic load of the network, it may take SNPscan some time before your results are sent back to you. Please be patient!	Submit

Outline

1. Background
   - SNPscan: the Biologists
   - The Biology
2. Statistical challenges
3. The R package
   - Bioconductor

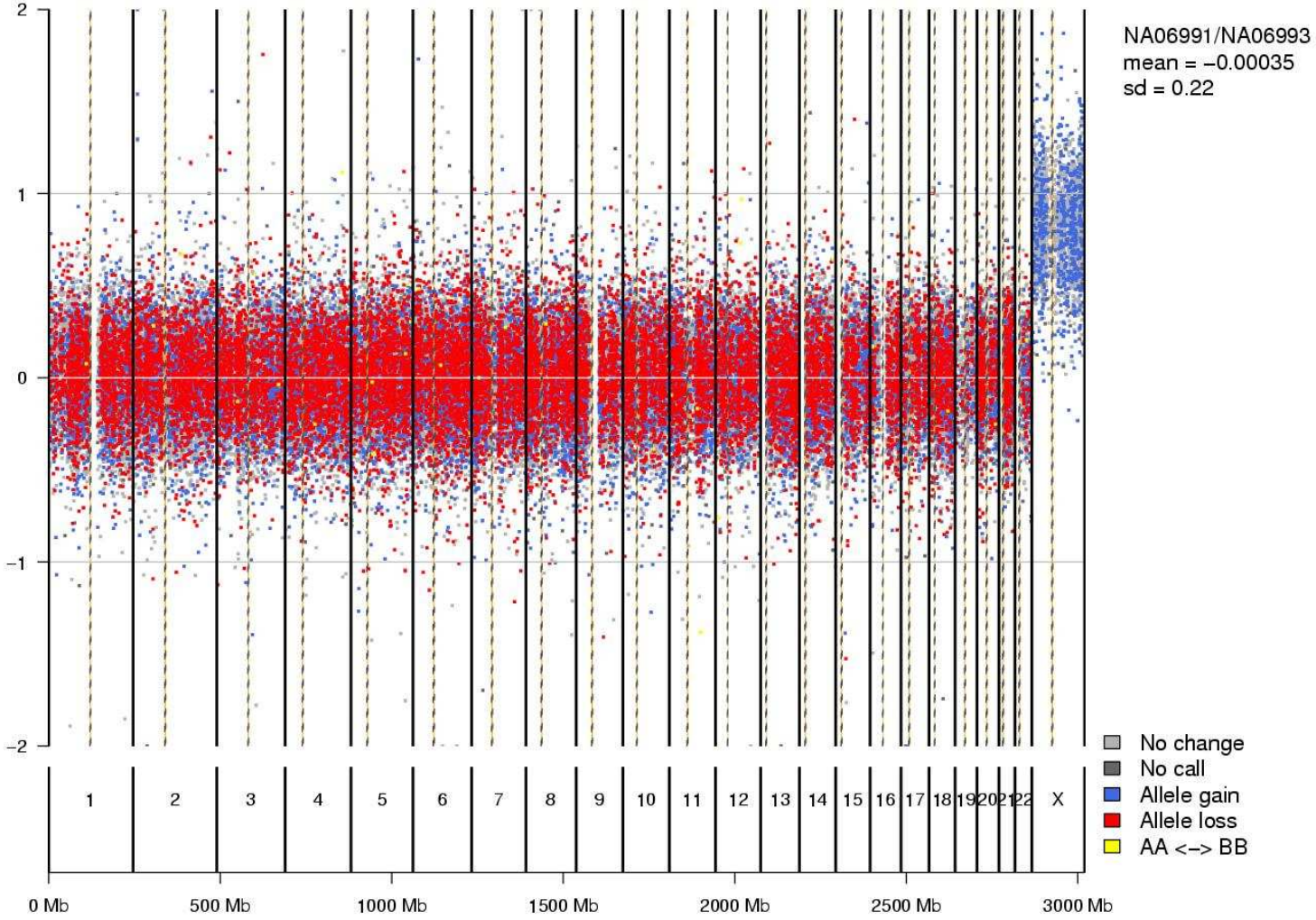
Normal male and female

> plot(1:23, normalFM)



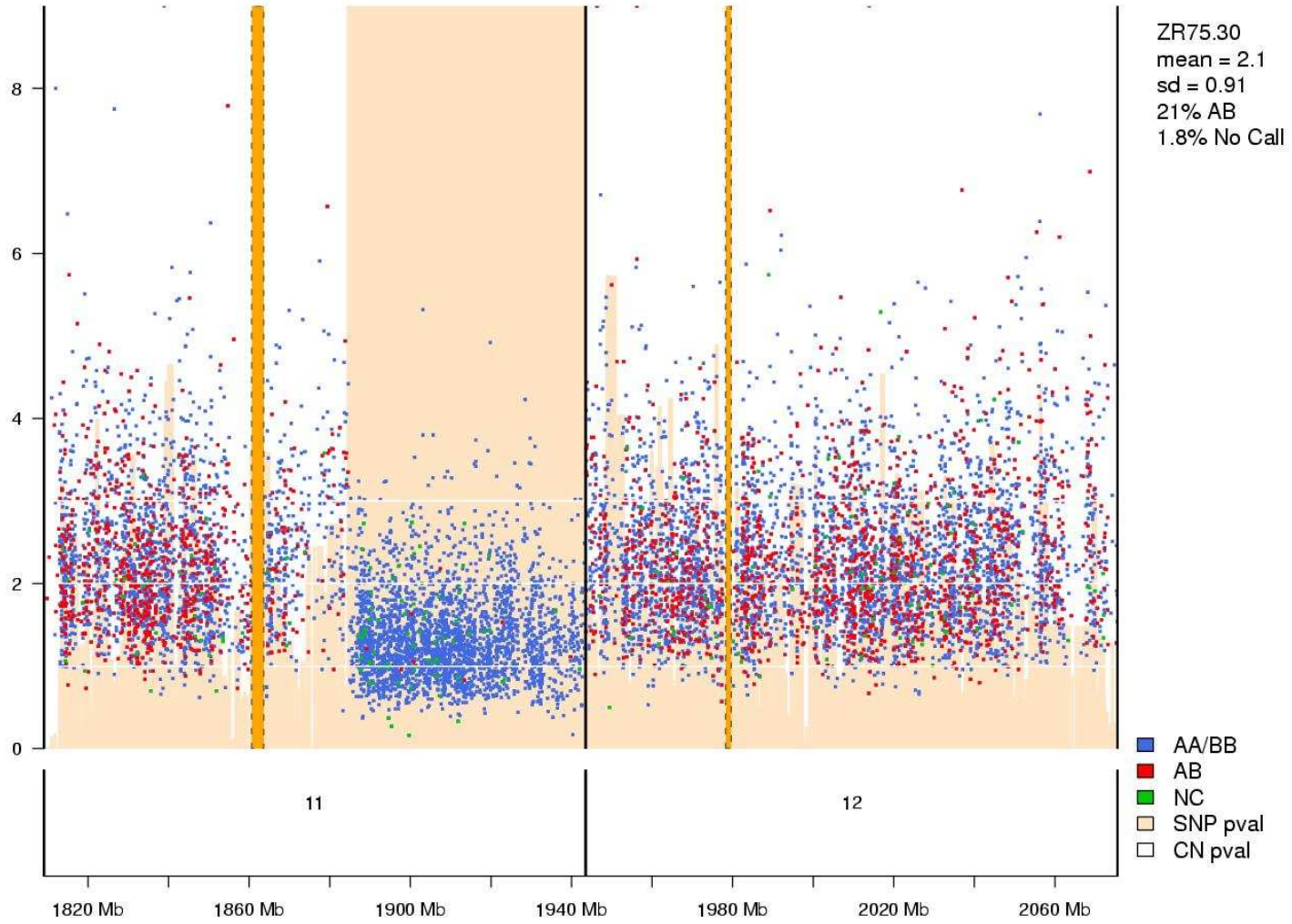
Paired plot

```
> plot(1:23, normalFM, paired = TRUE)
```



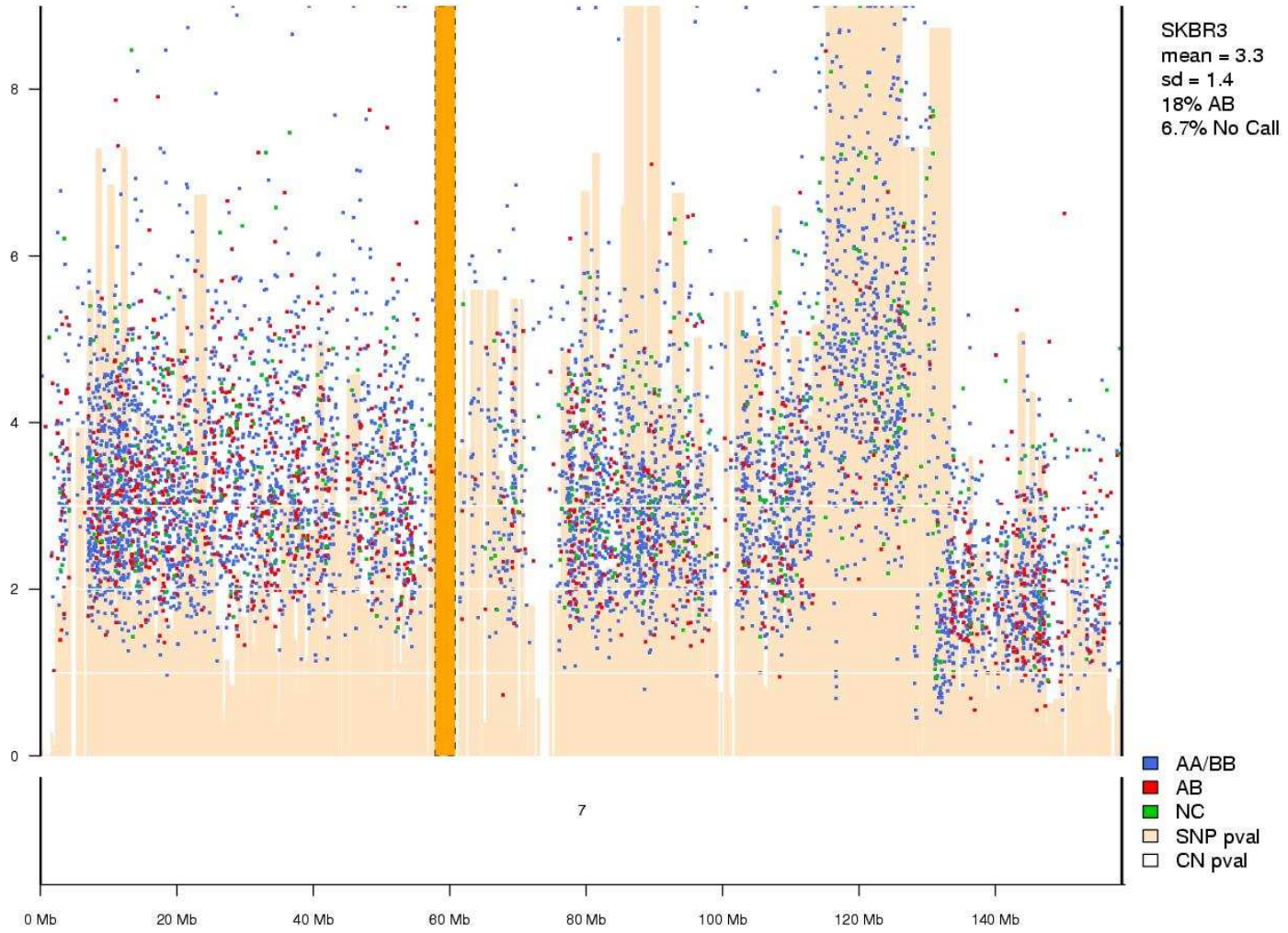
Zoom-feature: deletion

```
> plot(11:12, cancer)
```



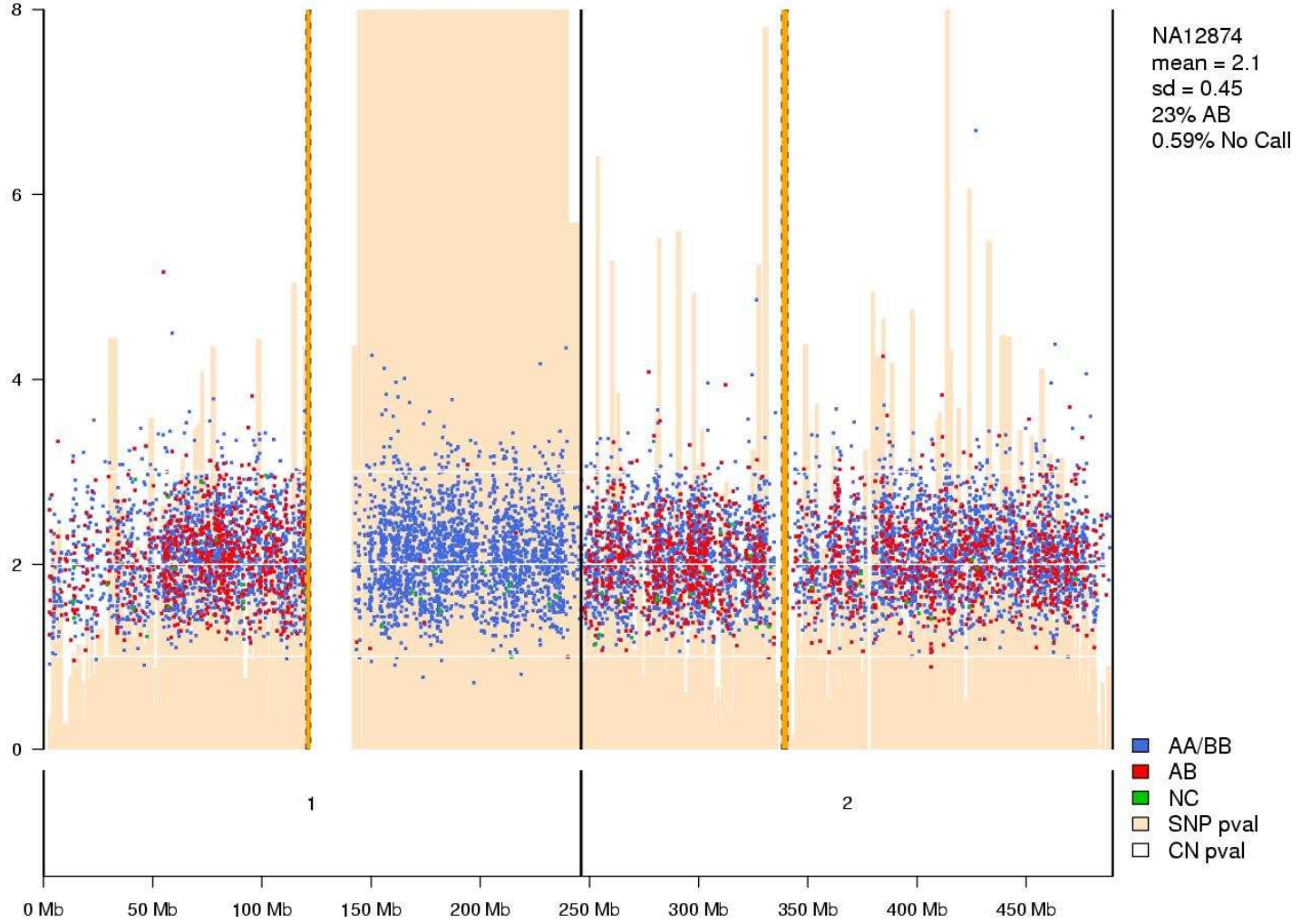
Amplification

```
> plot(11:12, cancer)
```



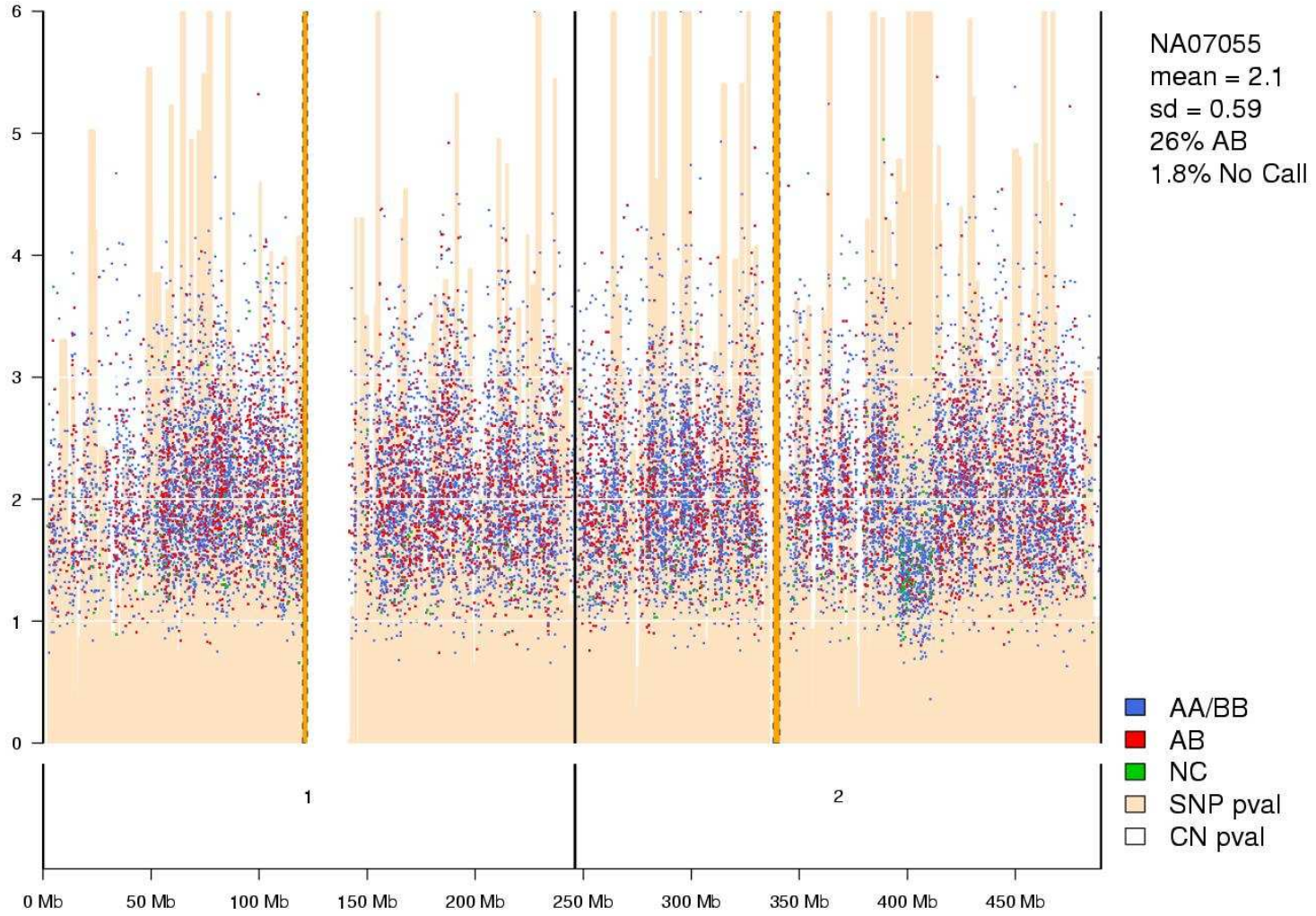
Uniparental Isodisomy

plot(1:2, normal)



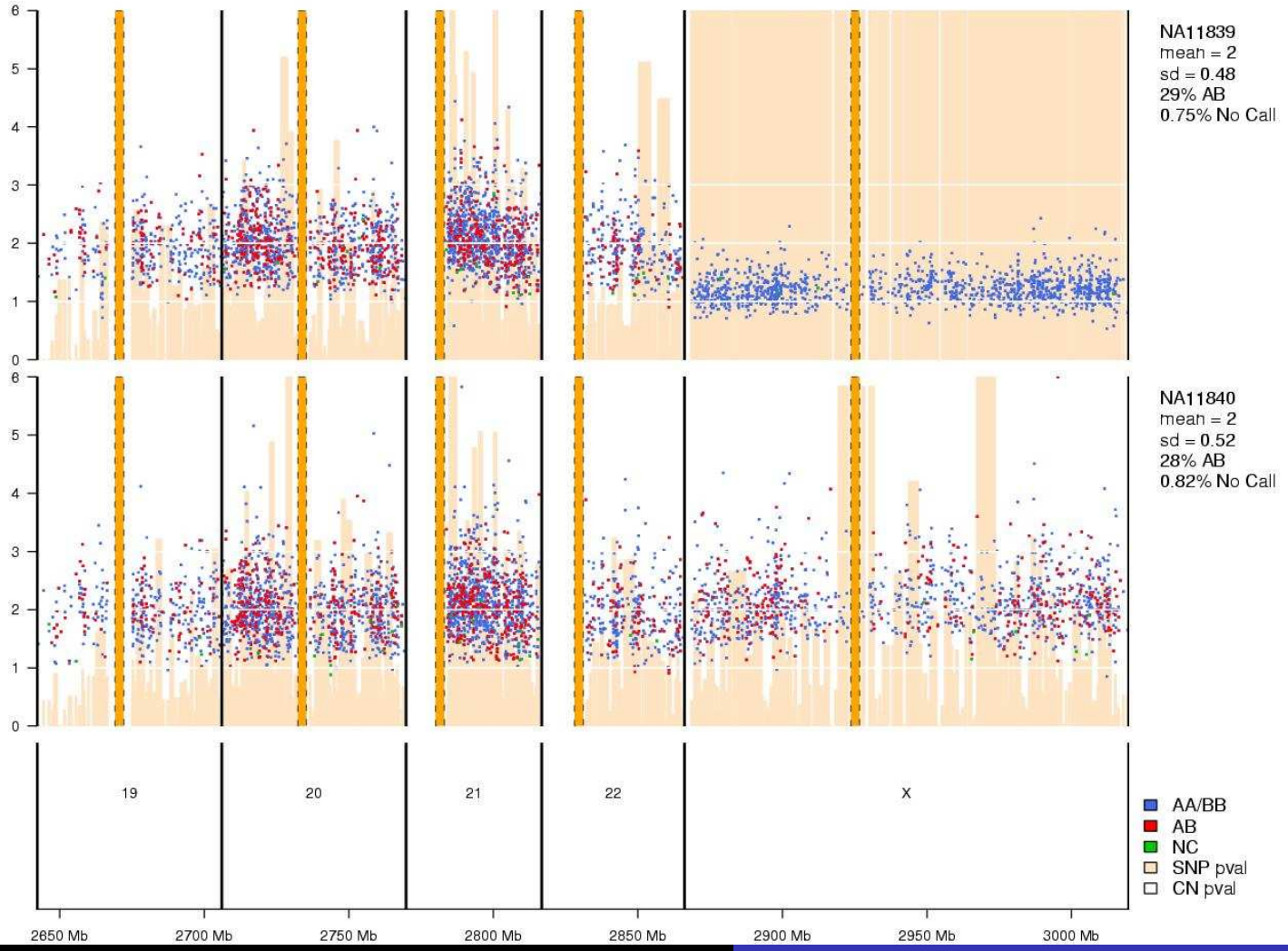
Mosaicism

```
> plot(1:2, mosaicExample)
```



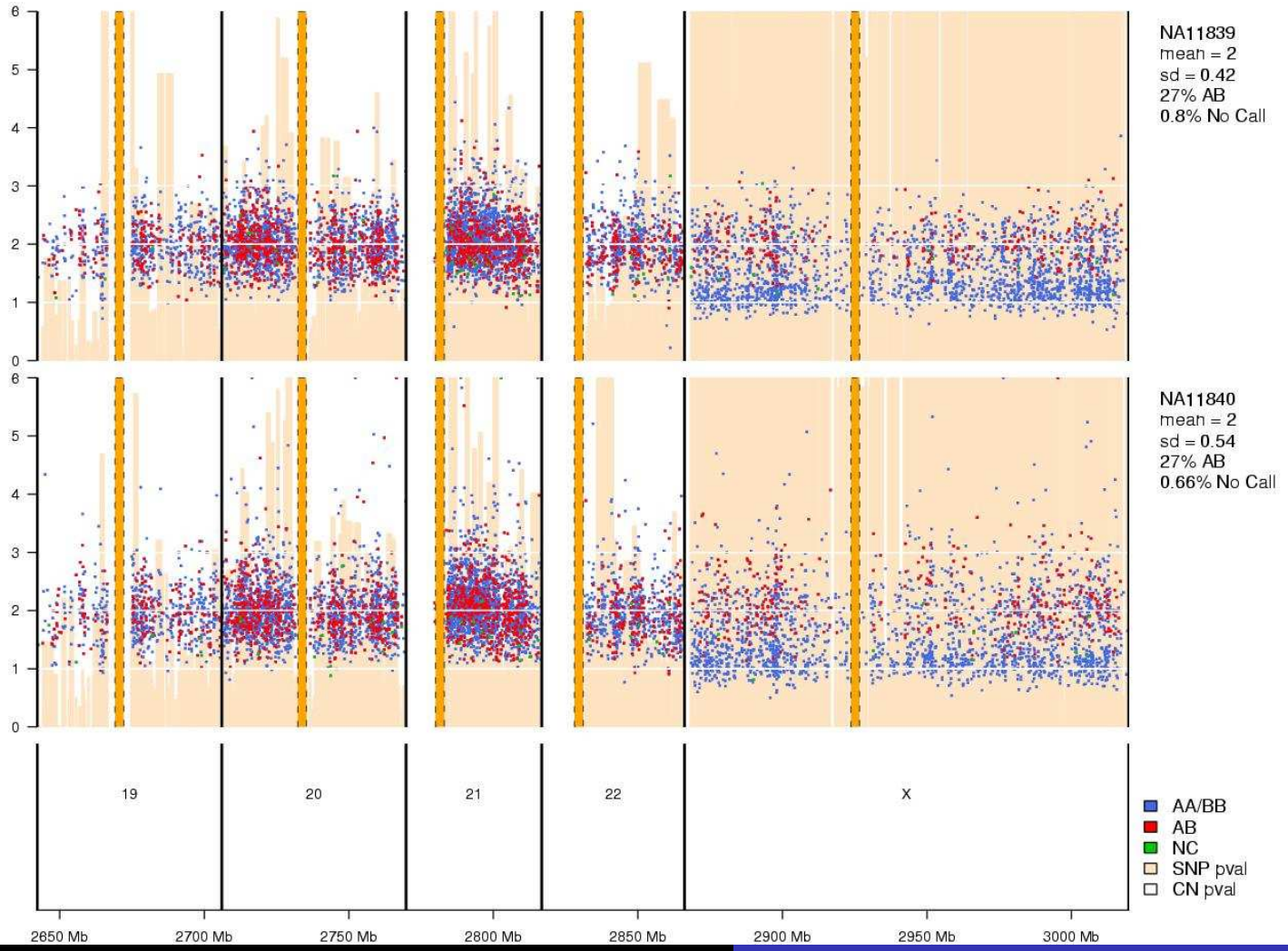
Mislabeled Samples

```
> plot(1:23, hind)
```



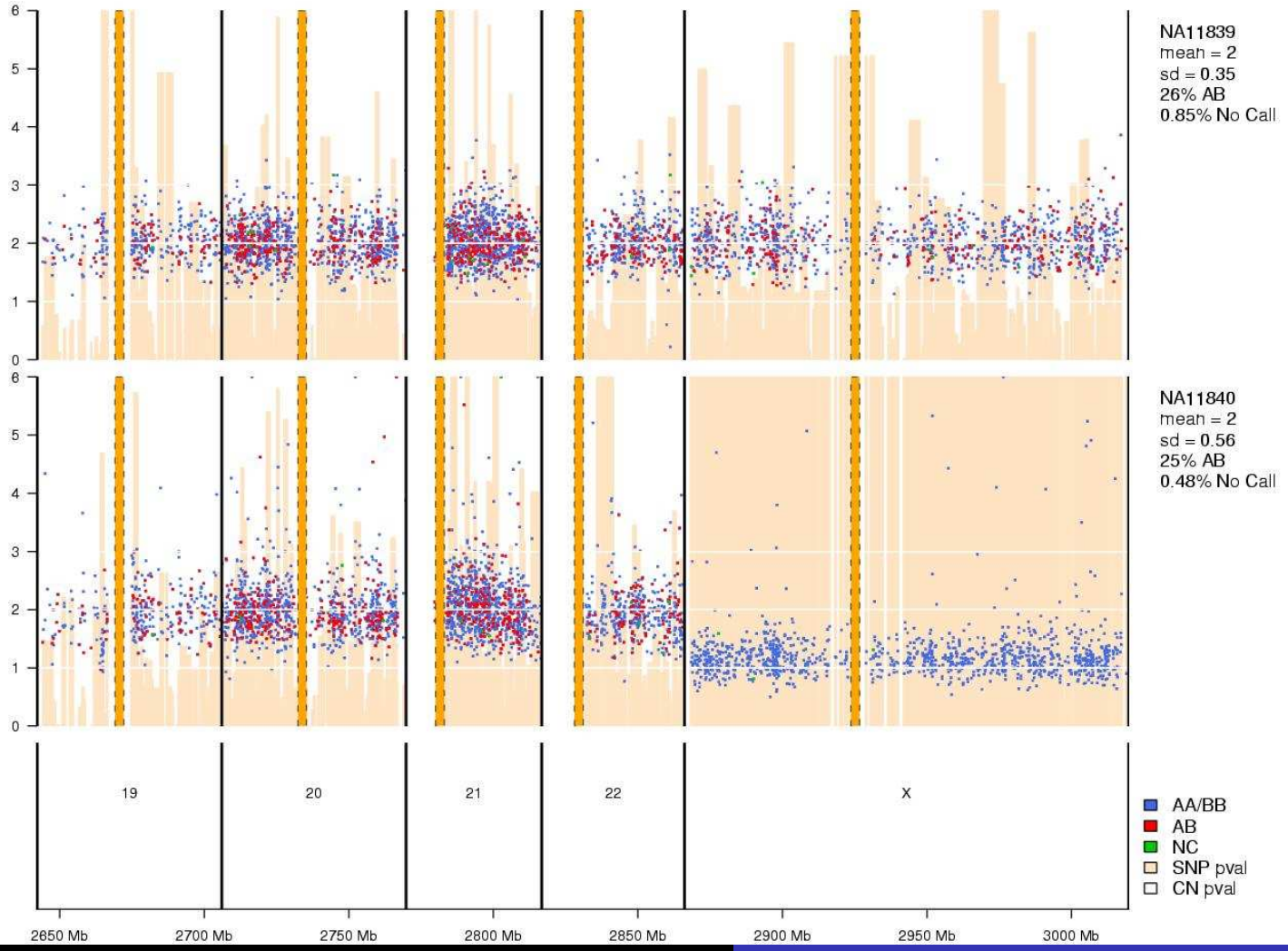
Mislabeled Samples

```
> plot(1:23, mergedXbaHind)
```



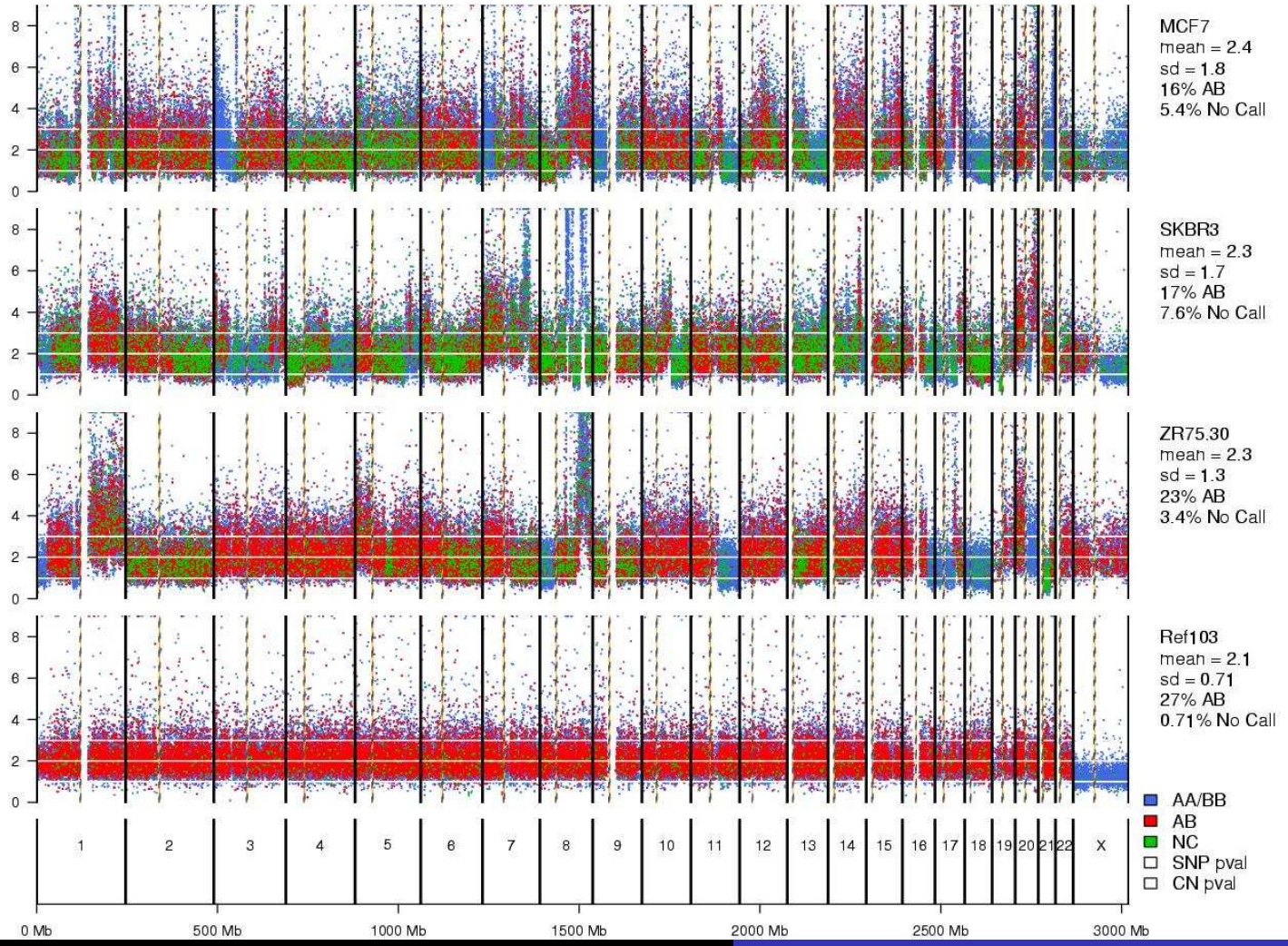
Mislabeled Samples

```
> plot(1:23, xba)
```

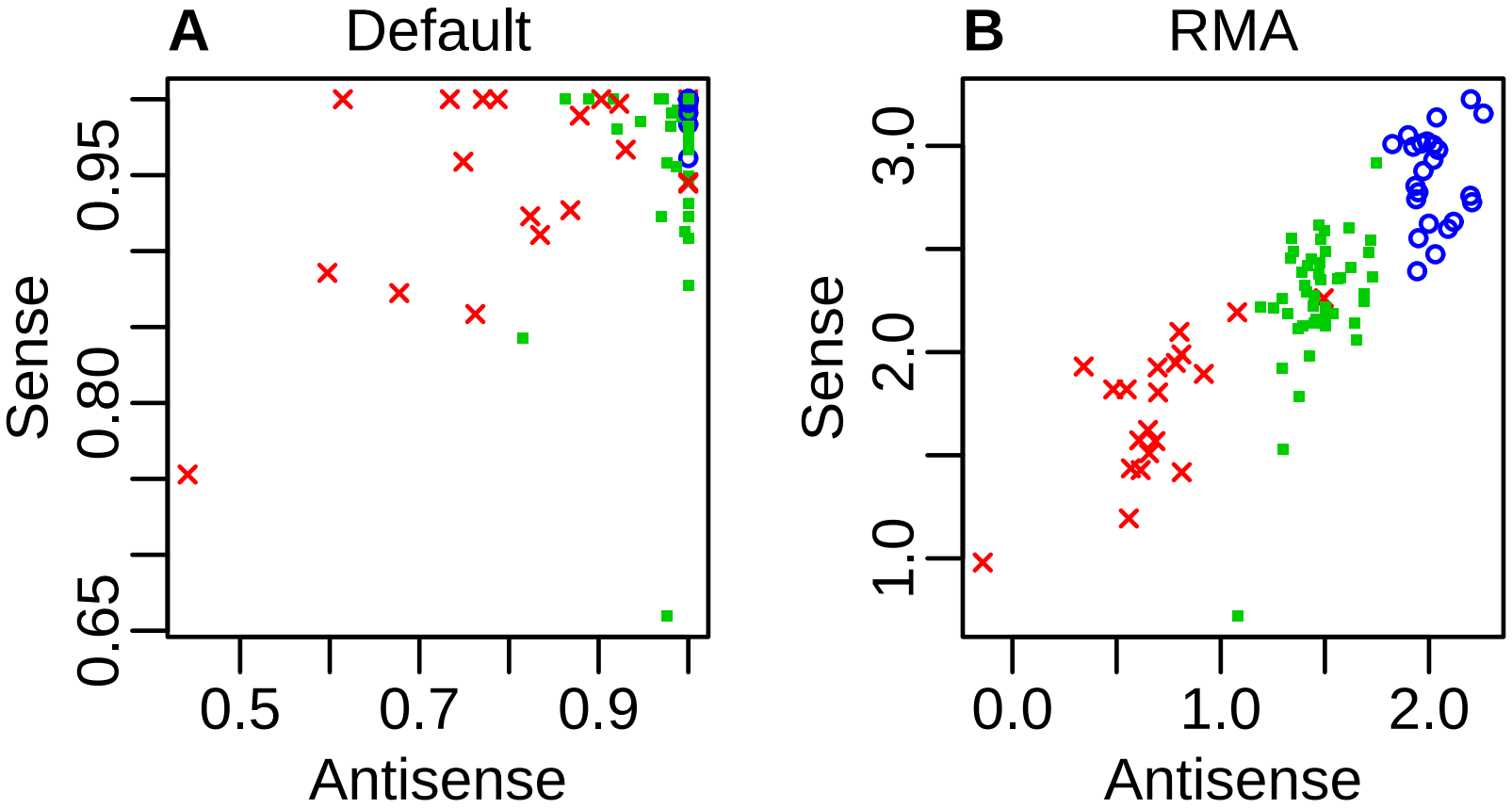


Genome-wide view of cancer samples

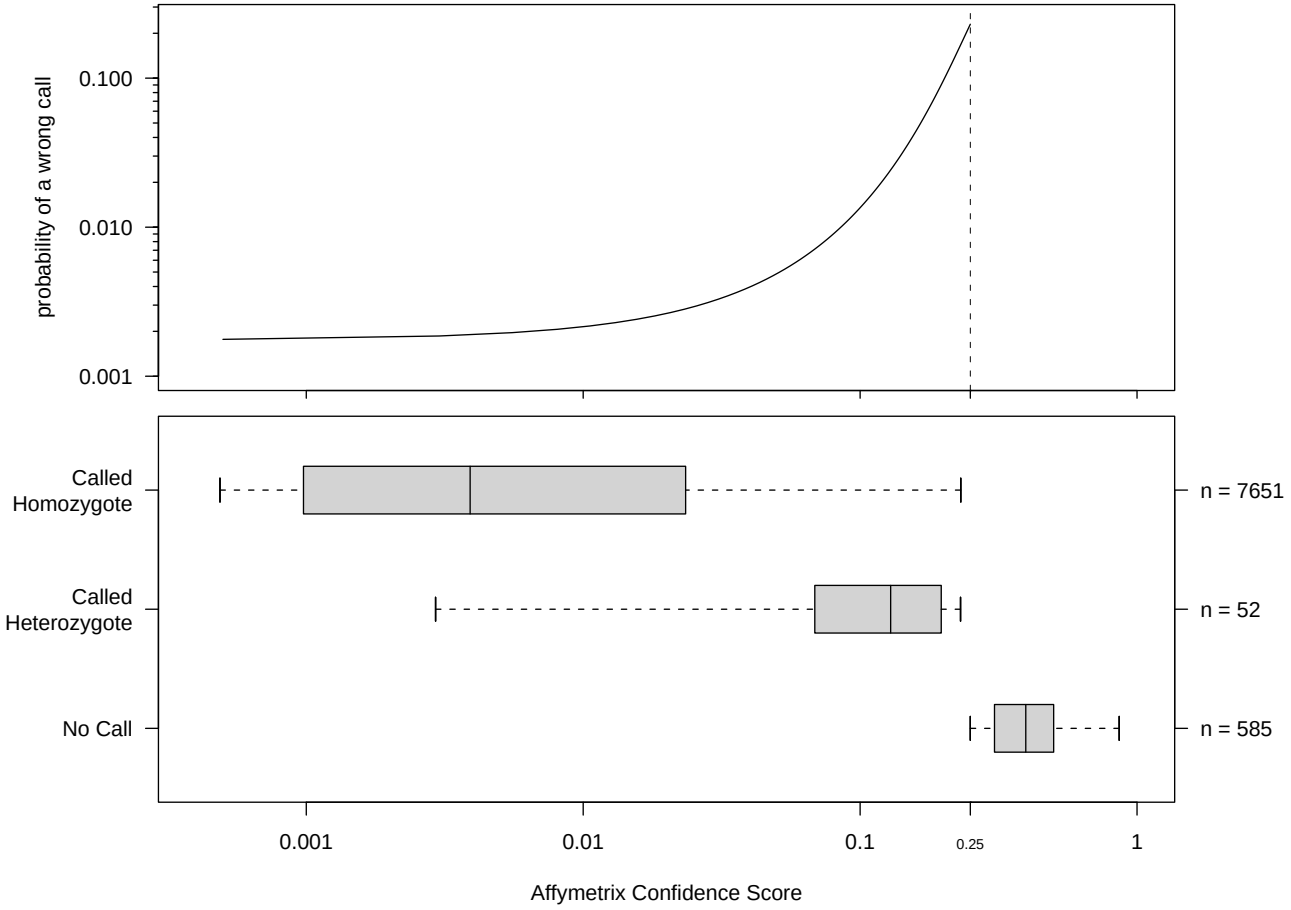
```
> plot(1:23, cancer)
```



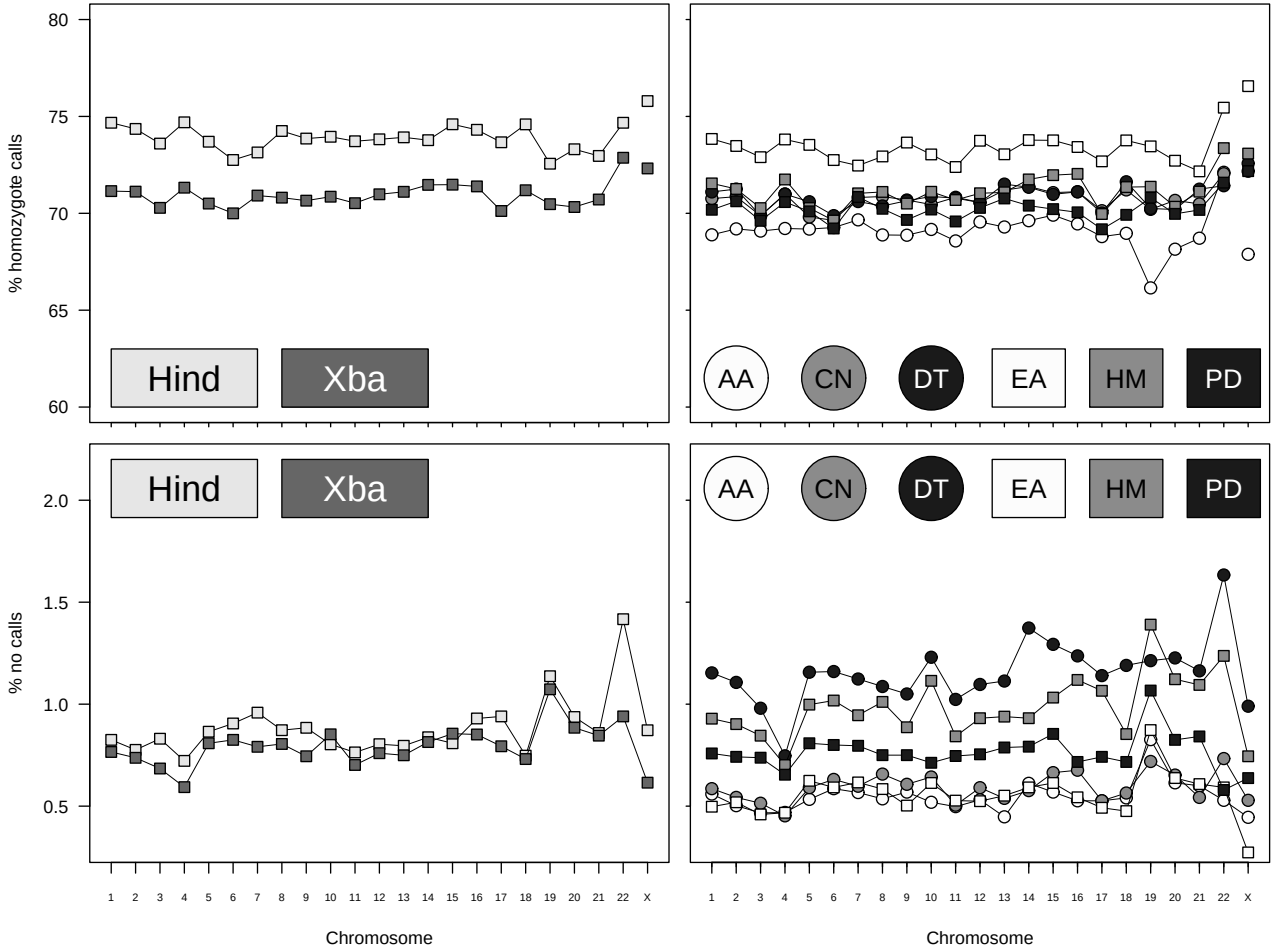
"Who cares? You've made a nice-looking plot that anyone can do with a tab-delimited text file." - Rafael Irizarry



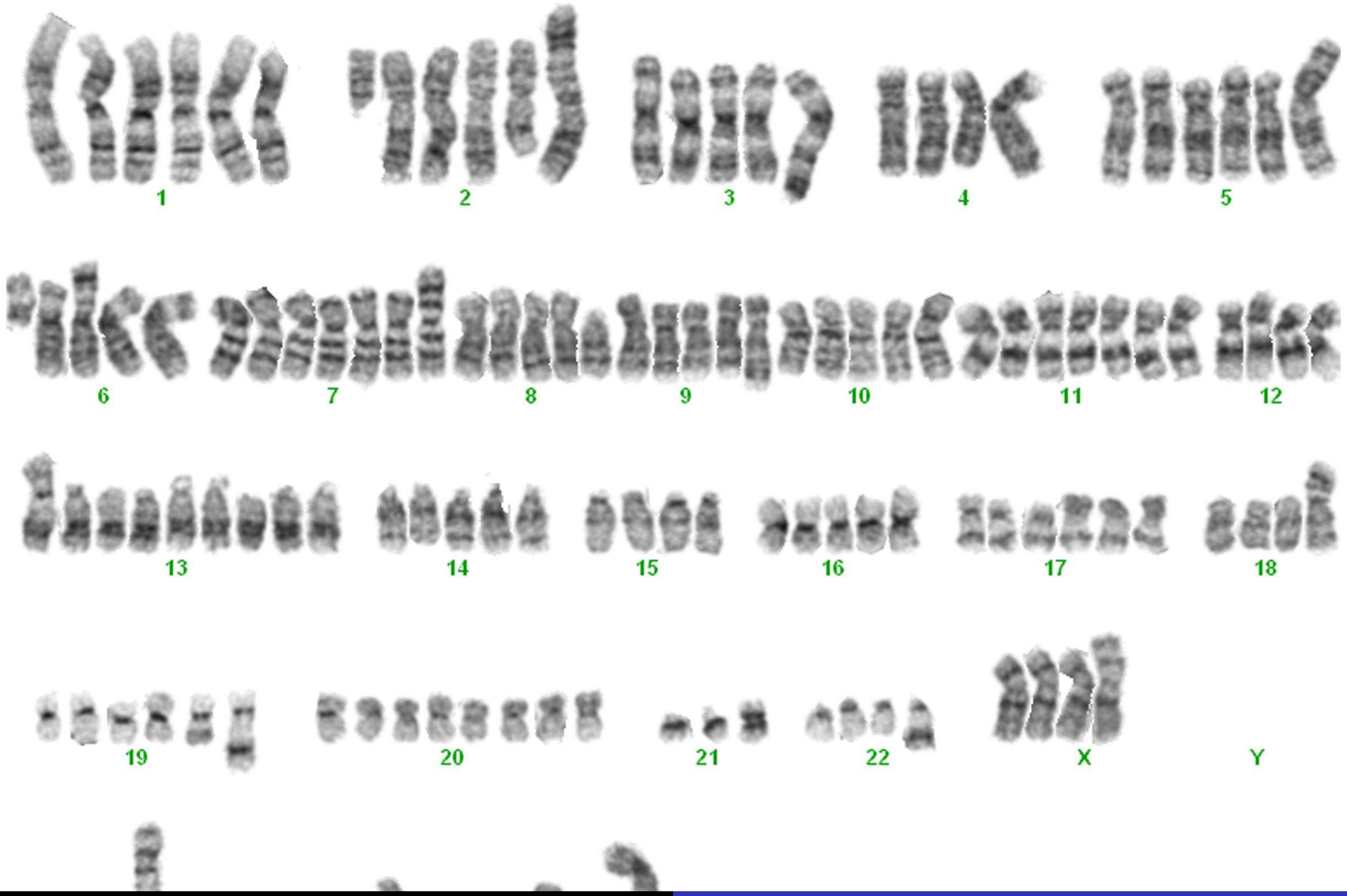
Confidence scores



Chip/Racial differences



Non-diploid samples



Outline

1. Background
   - SNPscan: the Biologists
   - The Biology
2. Statistical challenges
3. The R package
   - Bioconductor

Utilizing bioconductor infrastructure

```
- attr(*, "assayData")=List of 2
 ..$ copyNumber: num [1:115571, 1:10] 2.18 1.27 0.9
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:115571] "SNP_A-1718890" "SNP_A-
 .. .. ..$ : chr [1:10] "NA06985" "NA06991" "NA0699
 ..$ snpCall   :`data.frame':  115571 obs. of  10 v
 .. ..$ NA06985: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA06991: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA06993: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA06994: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA07000: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA07019: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA07022: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA07029: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA07034: Factor w/ 4 levels "AA","AB","BB",
 .. ..$ NA07048: Factor w/ 4 levels "AA","AB","BB",
- attr(*, "sampleNames")= chr [1:10] "NA06985" "NA0
```

summary()

```
> 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

>
```

Summary

- Genome-wide and chromosome-specific visualization tools are available
- Take advantage of the Bioconductor infrastructure
- Reasons to use eSet classes:
  - easy
  - methods for the class work regardless of how the pre-processing was performed
  - facilitates collaborative research
  - reproducible research

Acknowledgments

- Benilton Carvalho
- David Dezentje
- Seth Falcon
- Rafael Irizarry
- Giovanni Parmigiani