Bioinformatics

Some selected examples . . . and a bit of an overview

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

July 19, 2007 @ EnviroHealth Connections

Bioinformatics and Computational Biology

Wikipedia:

Bioinformatics and computational biology involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level.

. . .

Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.

Bioinformatics and Computational Biology

Wikipedia:

The terms bioinformatics and computational biology are often used interchangeably. However bioinformatics more properly refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems inspired from the management and analysis of biological data. Computational biology, on the other hand, refers to hypothesis-driven investigation of a specific biological problem using computers, carried out with experimental or simulated data, with the primary goal of discovery and the advancement of biological knowledge.

Bioinformatics and Computational Biology

NIH definition of Bioinformatics and Computational Biology:

Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science.

. . .

Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.

Bioinformatics and Computational Biology

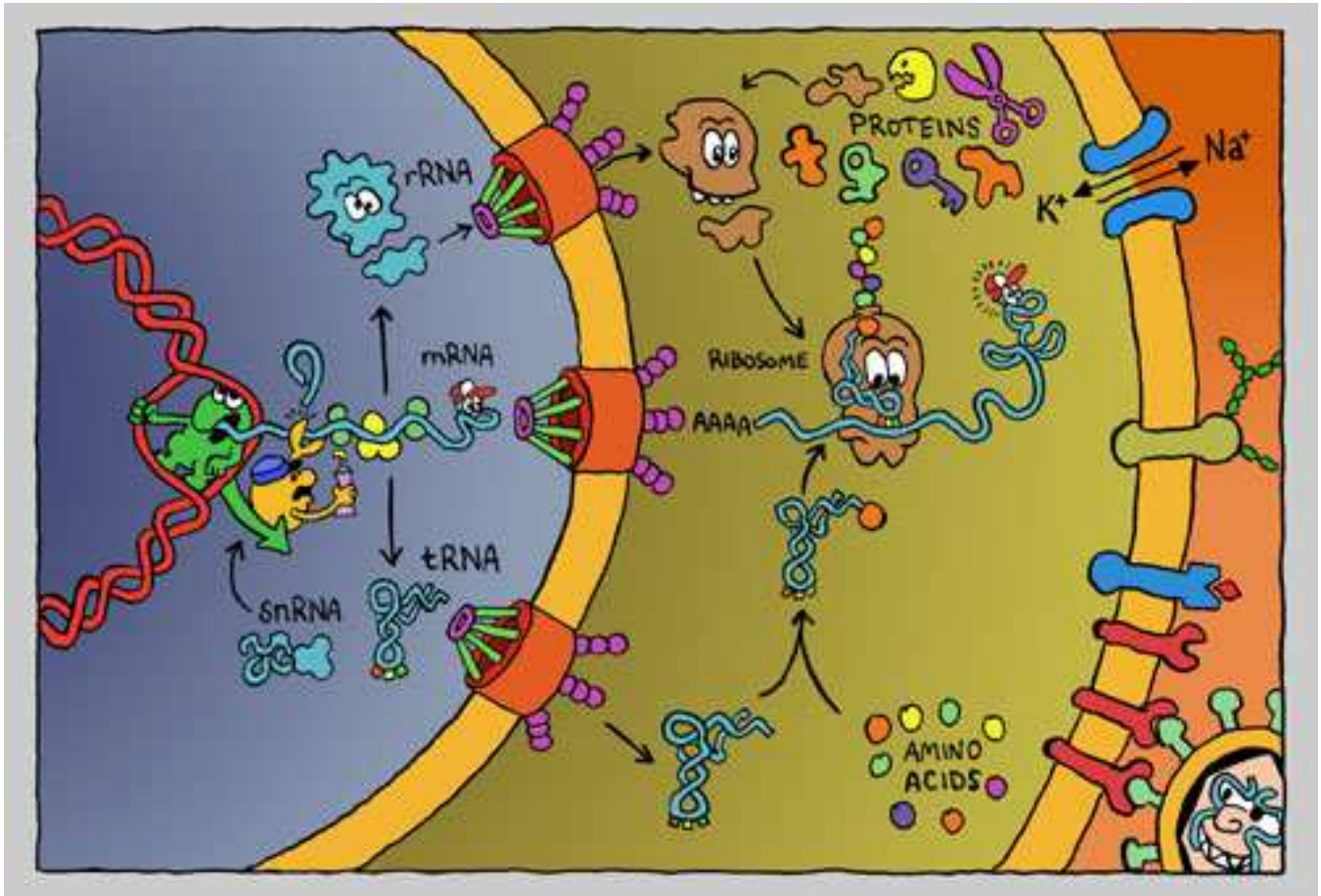
NIH definition of Bioinformatics and Computational Biology:

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

The central dogma of biology



Drawn by Ebbe Sloth Andersen http://old.mb.au.dk/

Topics

DNA $\longrightarrow$ RNA $\longrightarrow$ Protein

DNA Sequence analysis, genome annotation, evolutionary biology, phylogeny, DNA alterations, comparative genomics, SNP association studies

RNA Analysis of gene expression and regulation

Proteins Analysis of protein expression, protein-protein docking, prediction of protein structure

Systems Biology:
modeling biological systems, gene/protein networks

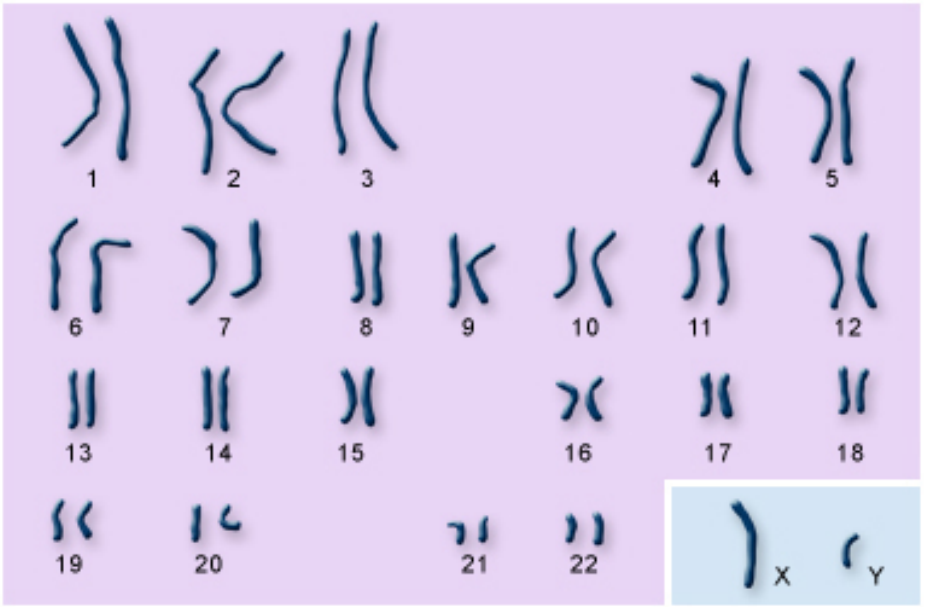
Some selected examples

1. Chromosomal alterations
2. Protein structure prediction
3. 2D gel electrophoresis

Some selected examples

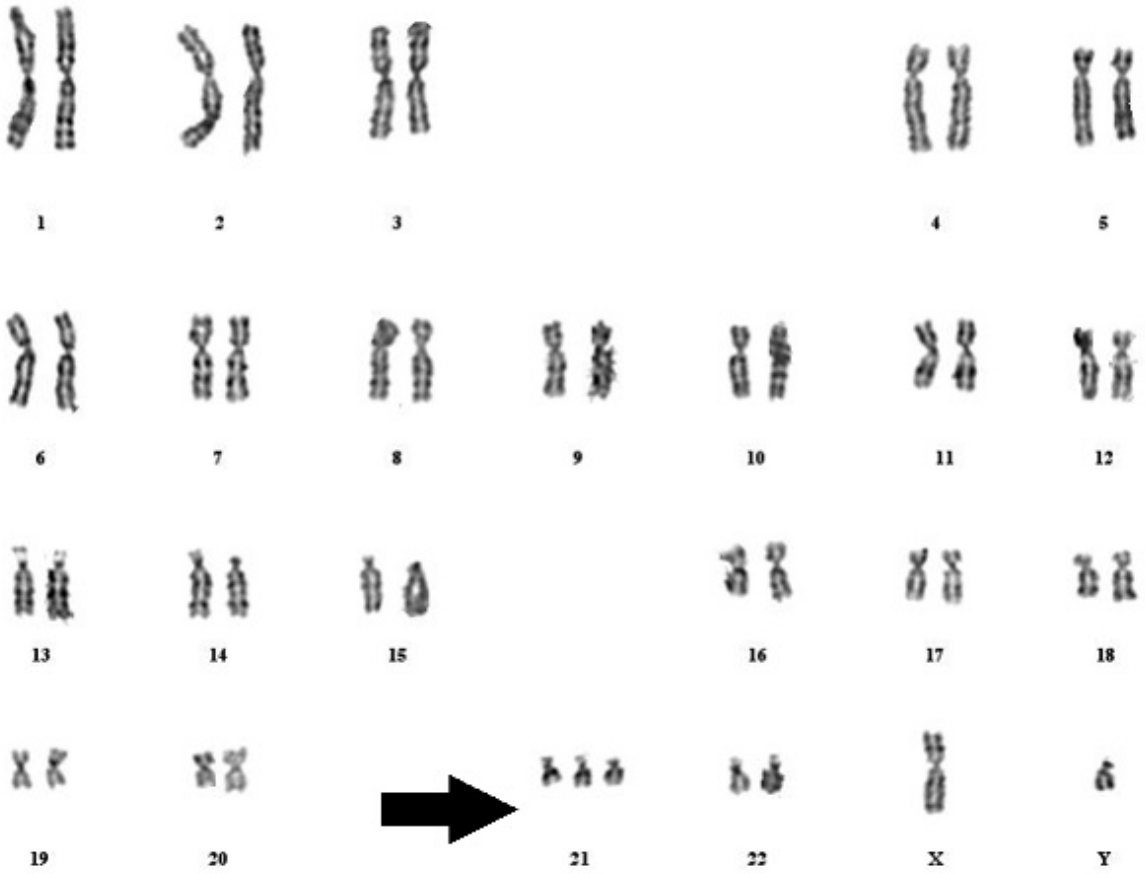
1. Chromosomal alterations
2. Protein structure prediction
3. 2D gel electrophoresis

Karyotypes



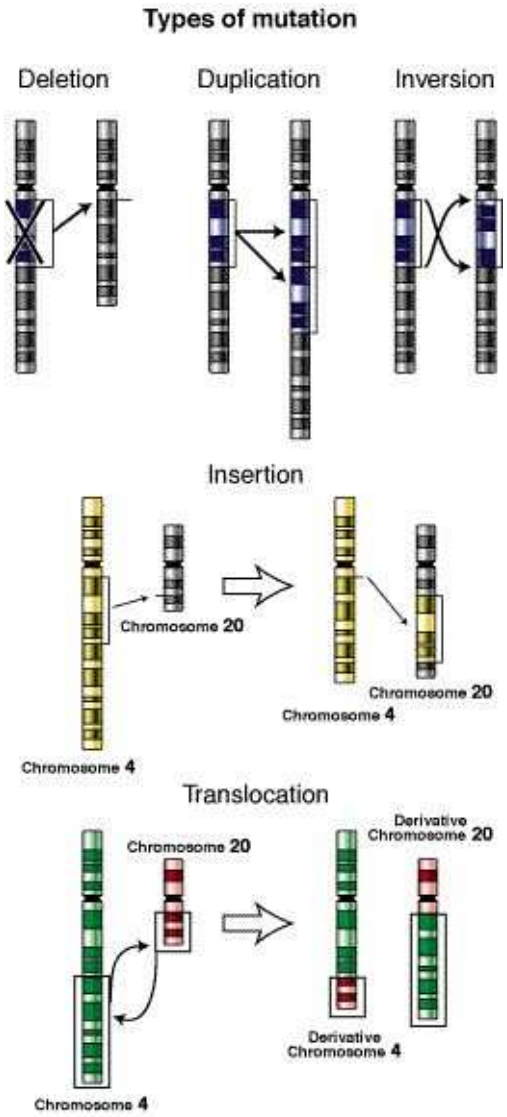
U.S. National Library of Medicine

Trisomy

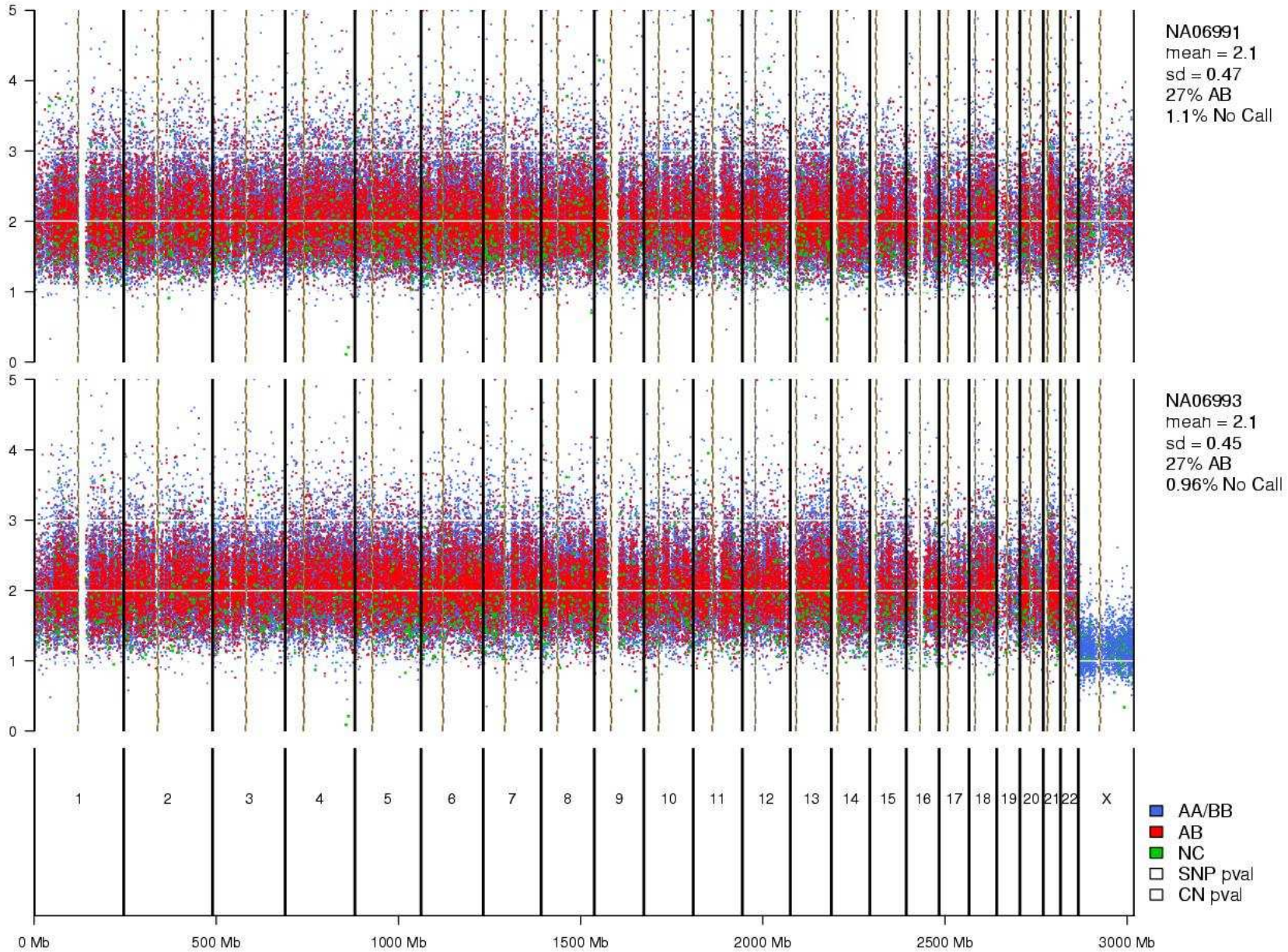


http://www.medgen.ubc.ca

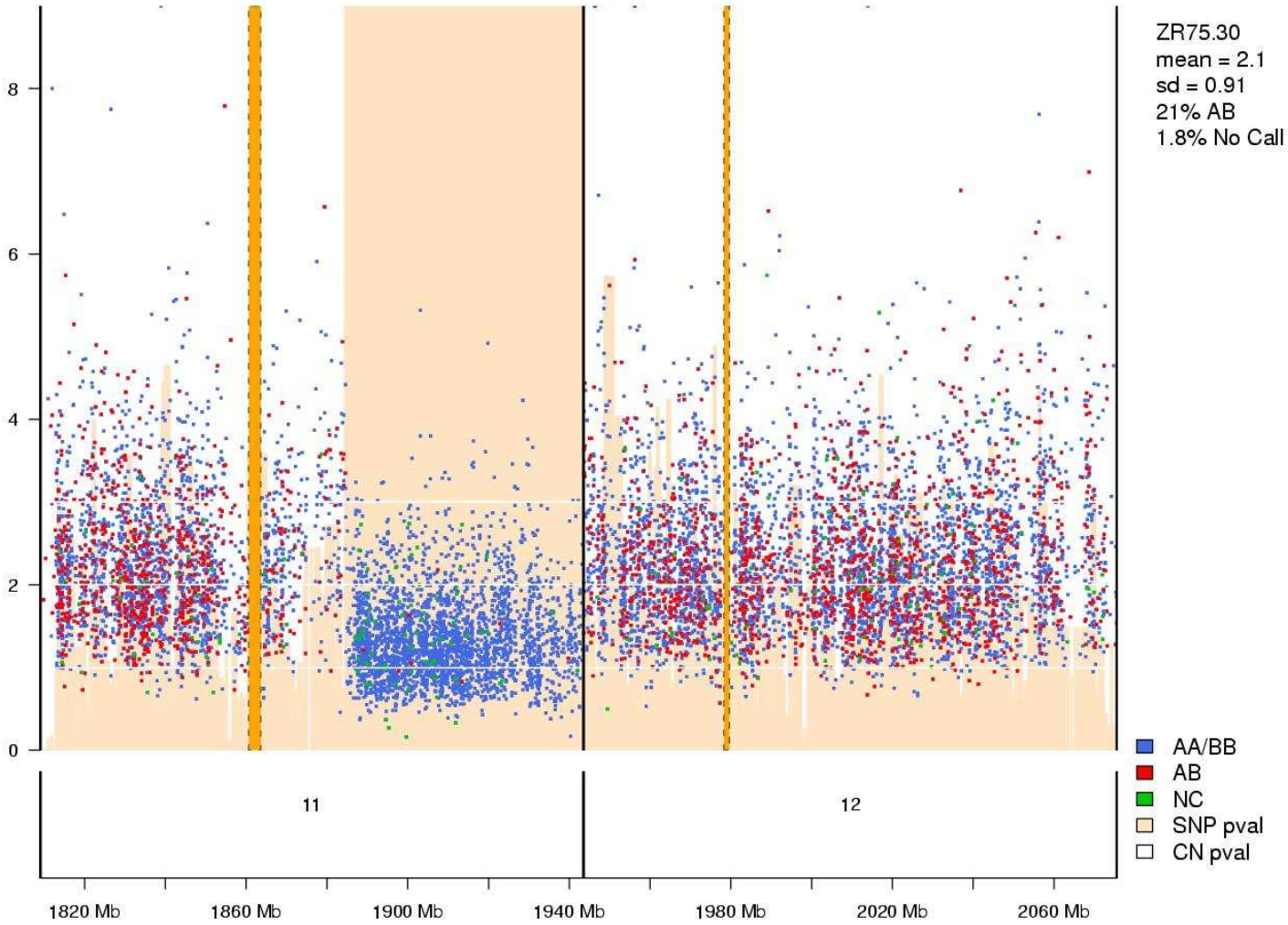
DNA changes


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

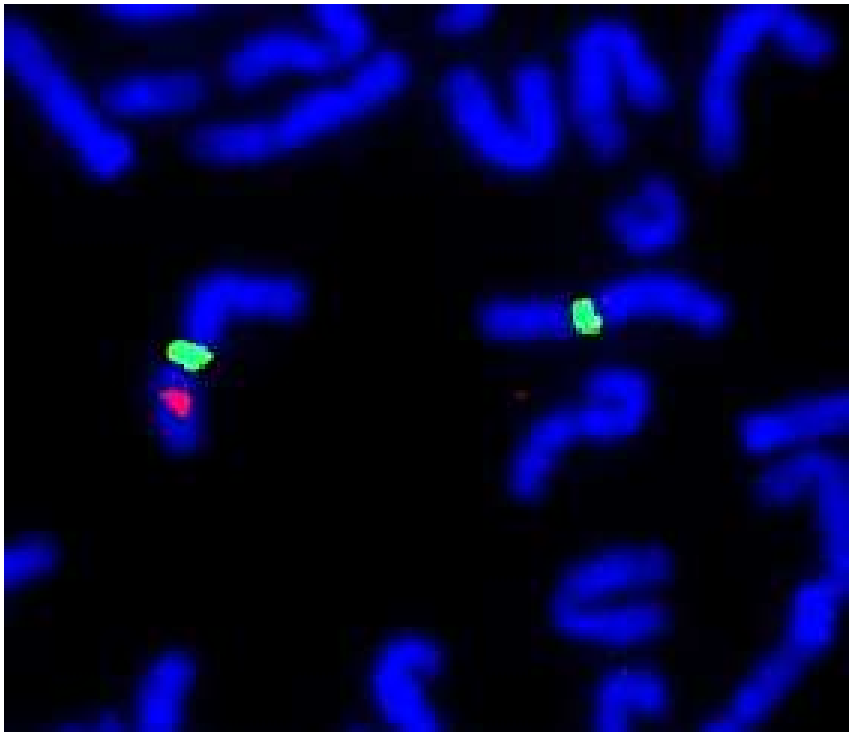
The data


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
0 Mb
500 Mb
1000 Mb
1500 Mb
2000 Mb
2500 Mb
3000 Mb
AA/BB
AB
NC
SNP pval
CN pval

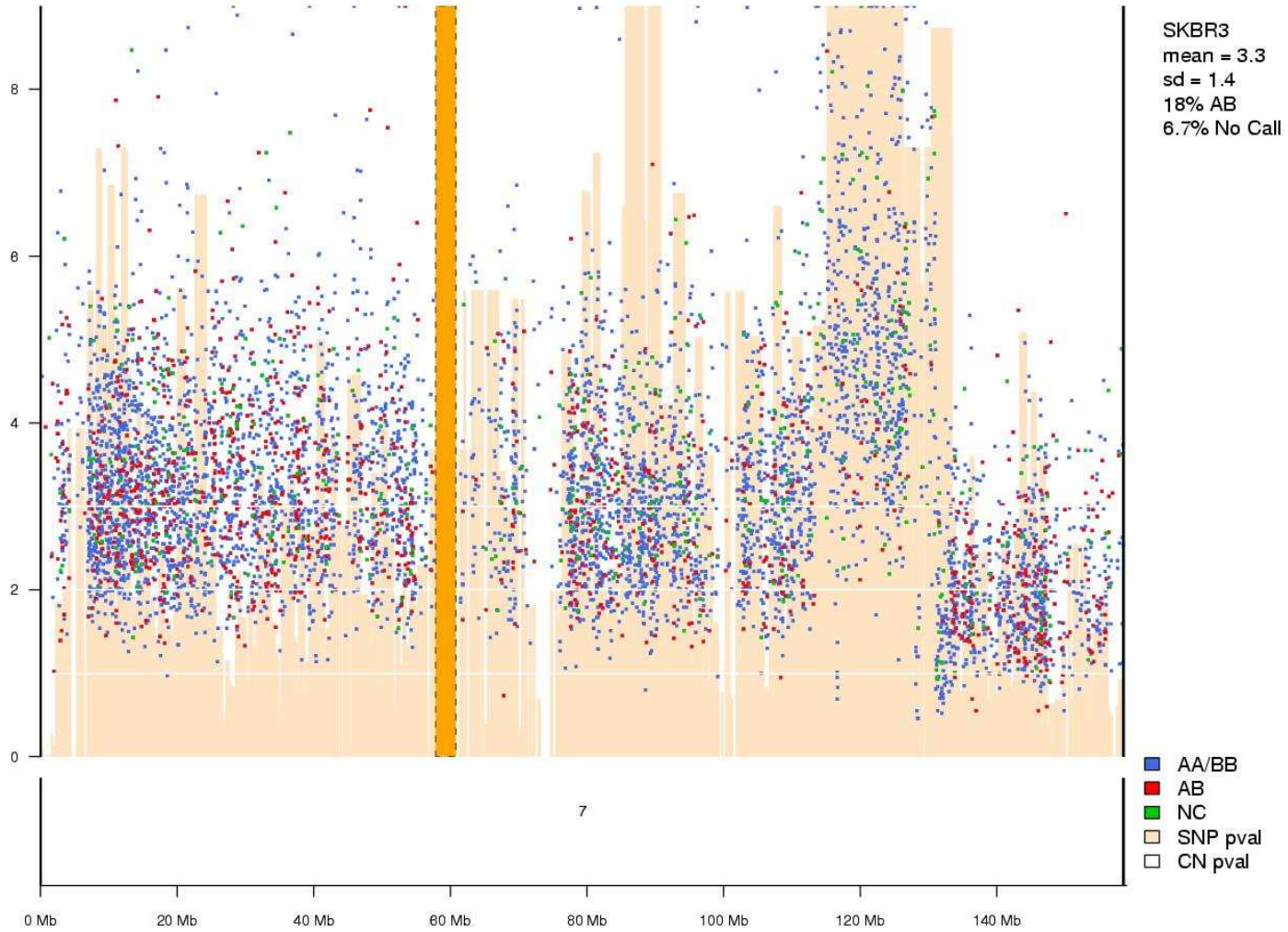
Deletion


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

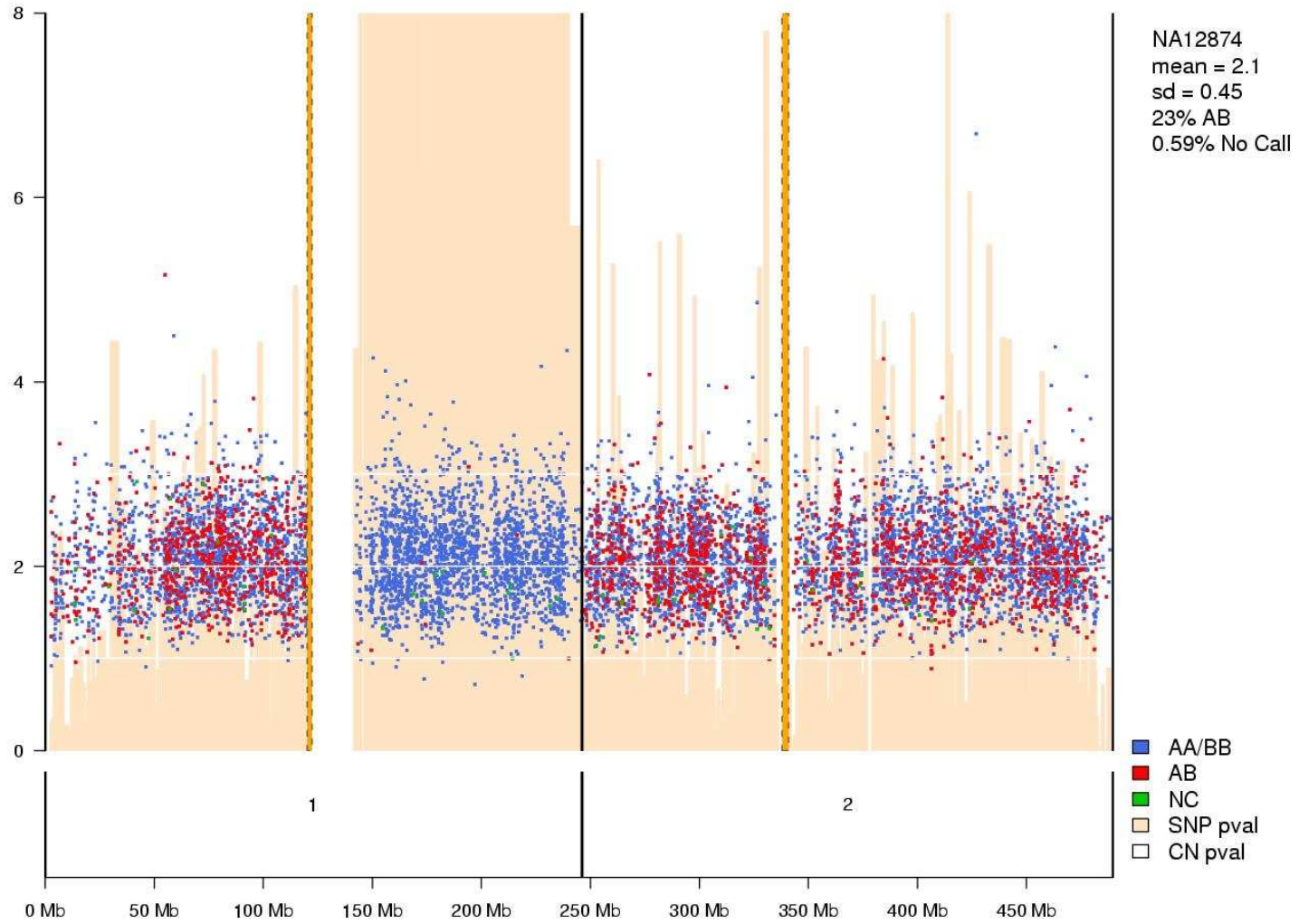
FISH



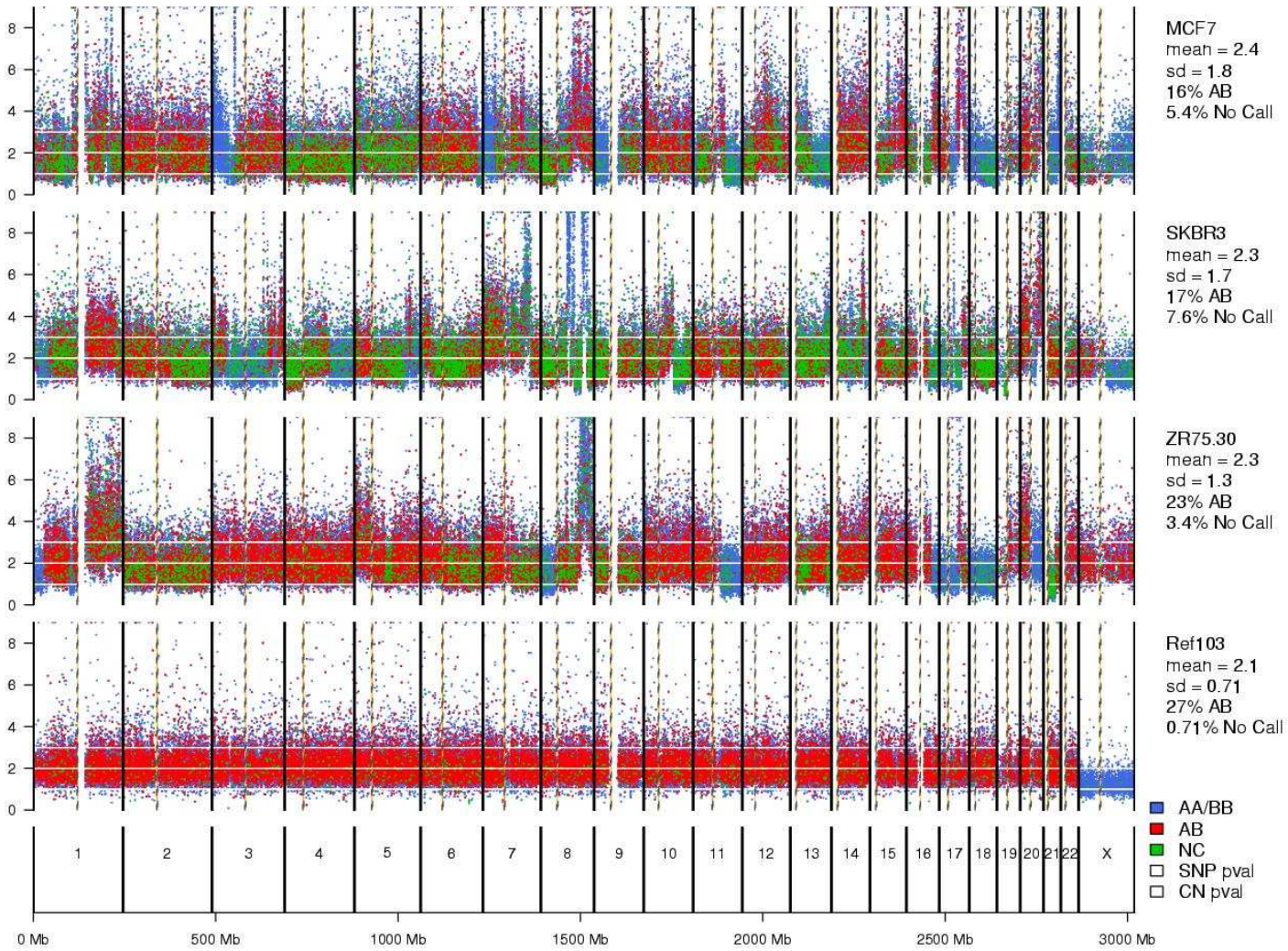
Amplification



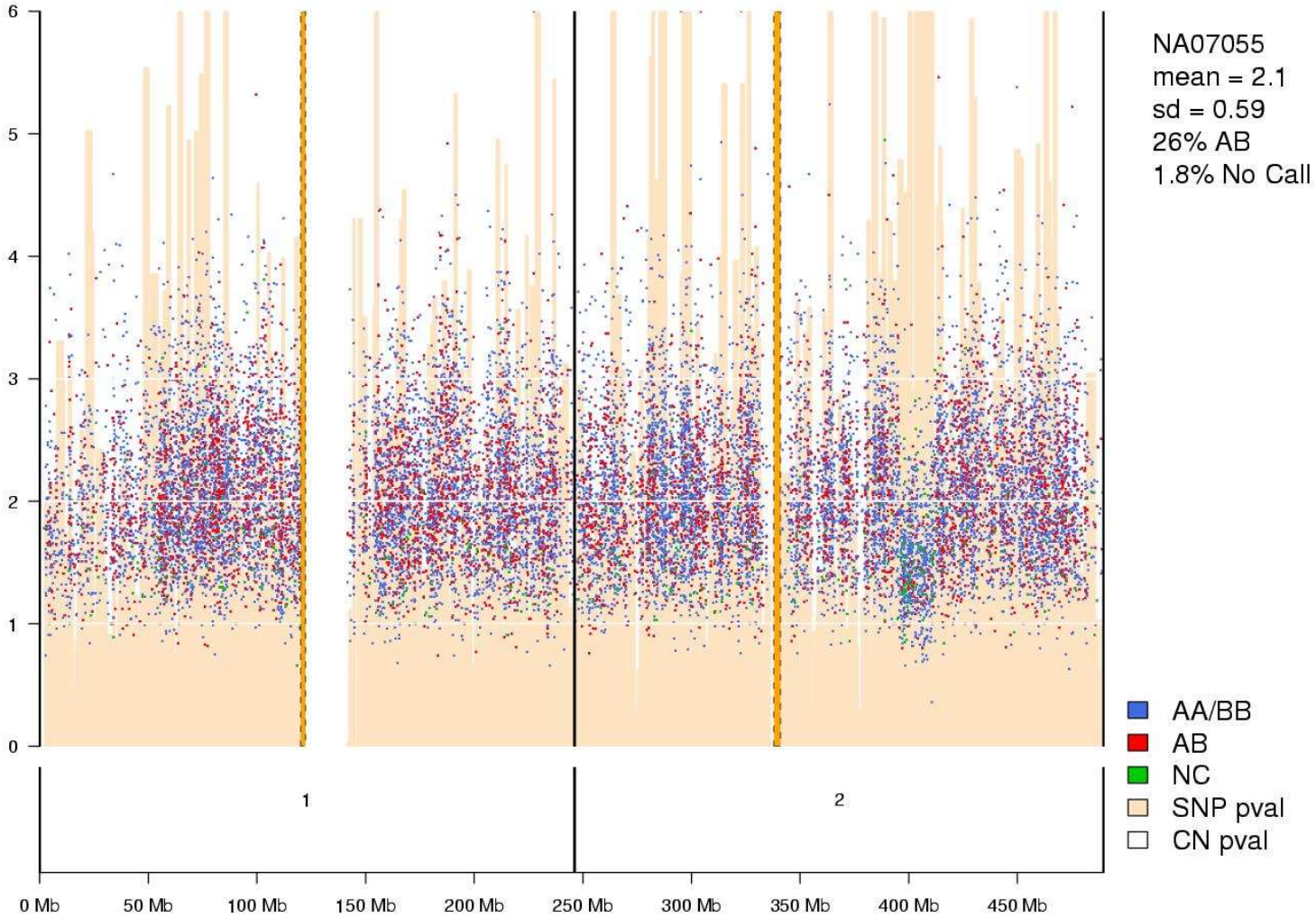
Uniparental Isodisomy



Cancer samples



Mosaicism


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

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.

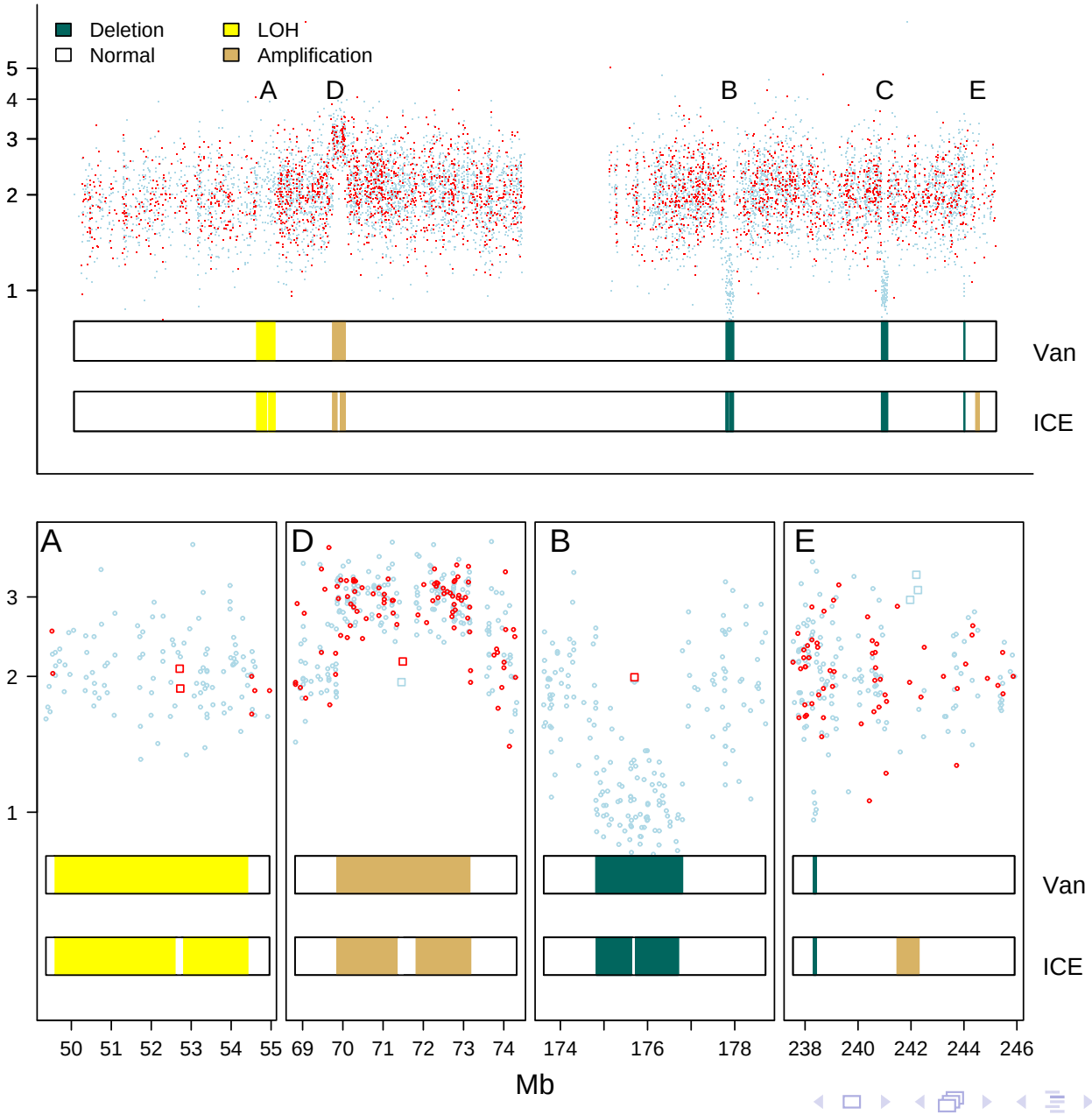
2. Within a sample:

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

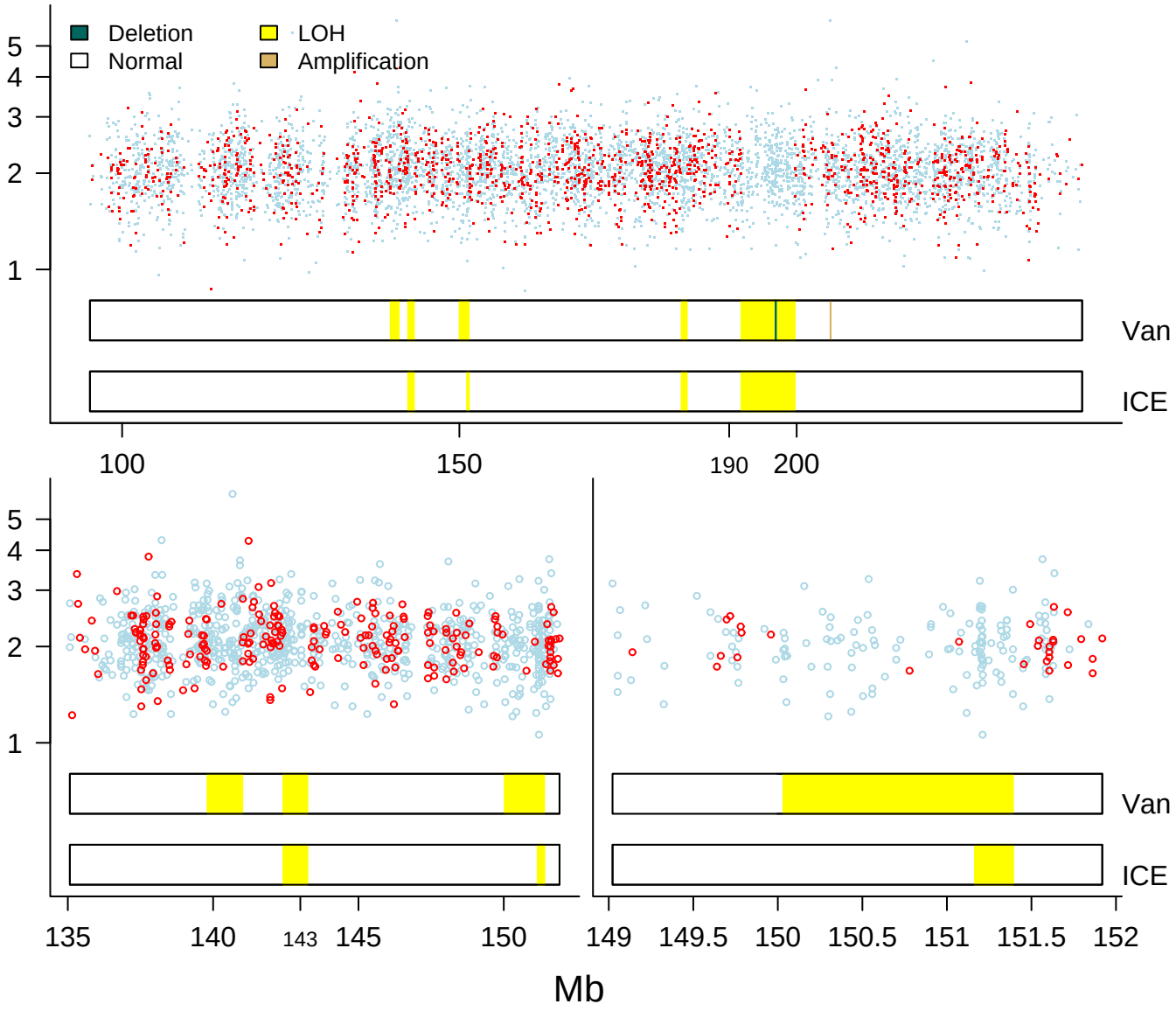
3. Between samples:

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

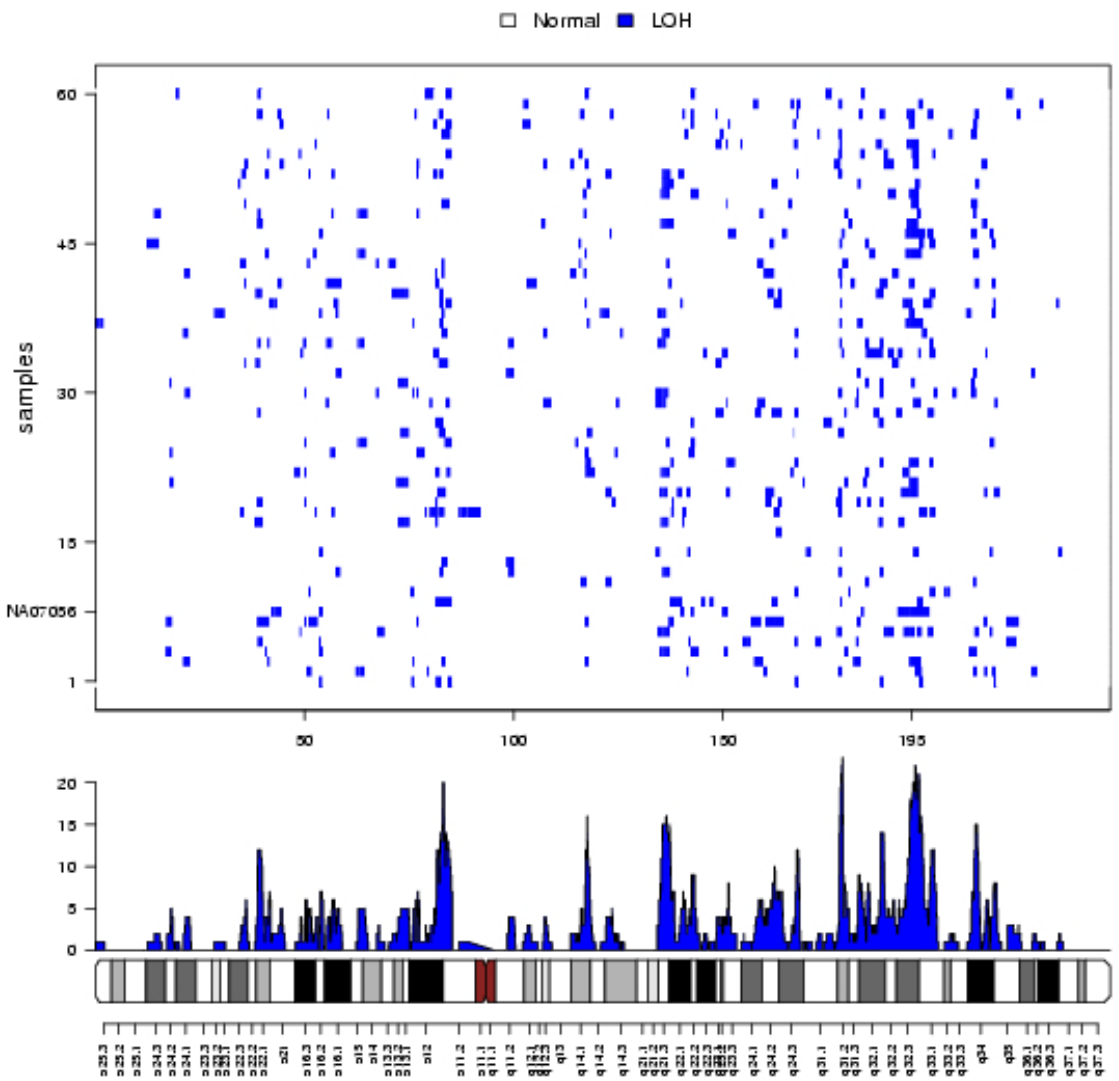
Vanilla ICE



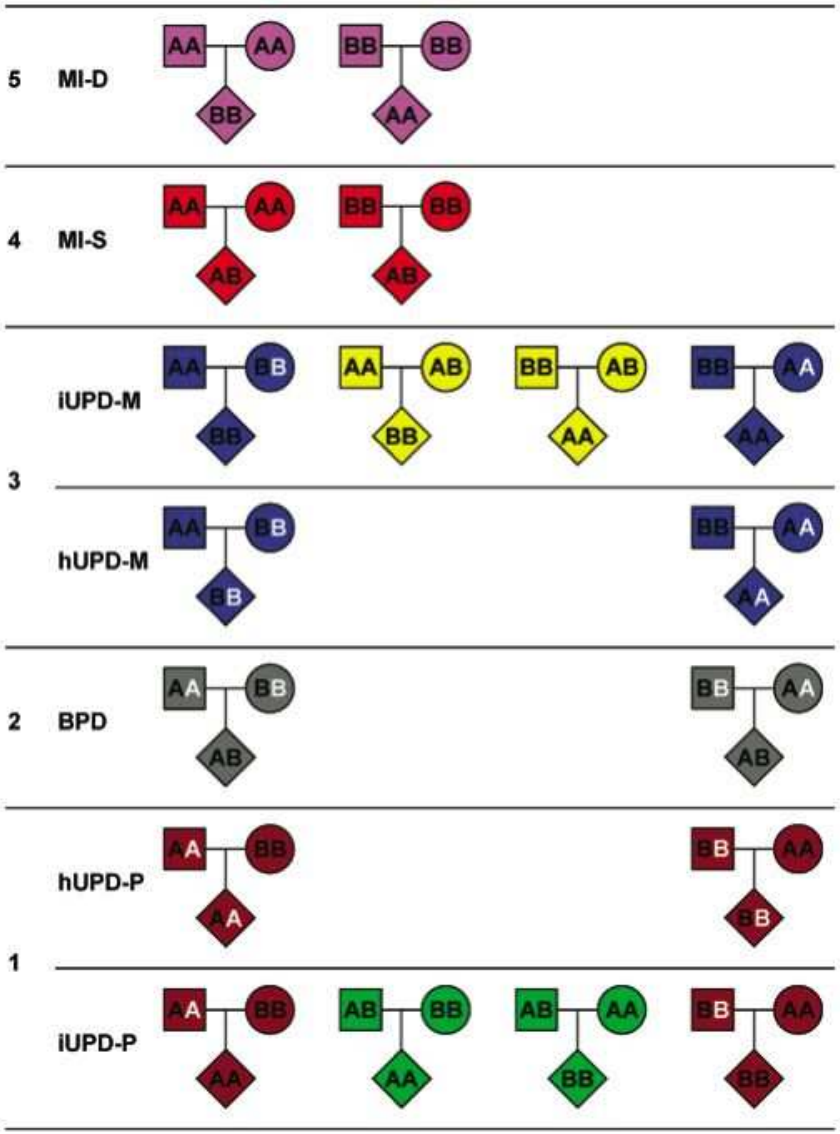
A HapMap sample



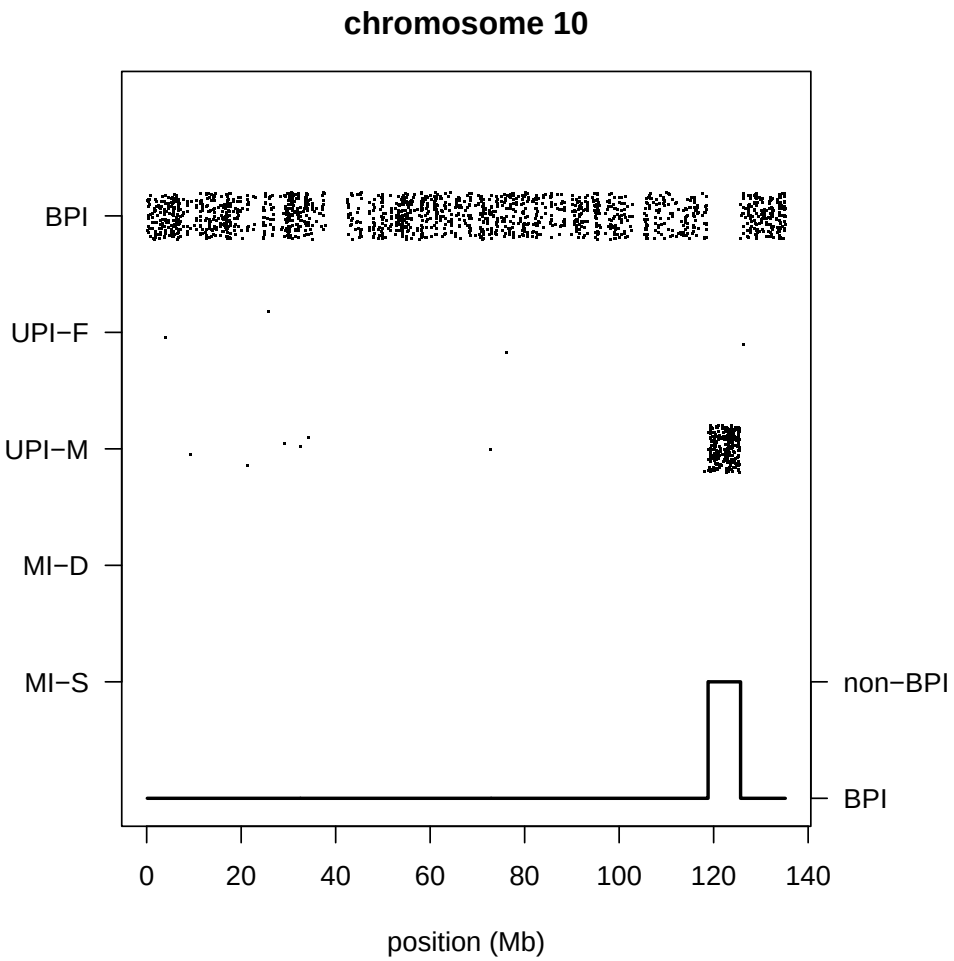
Many HapMap samples

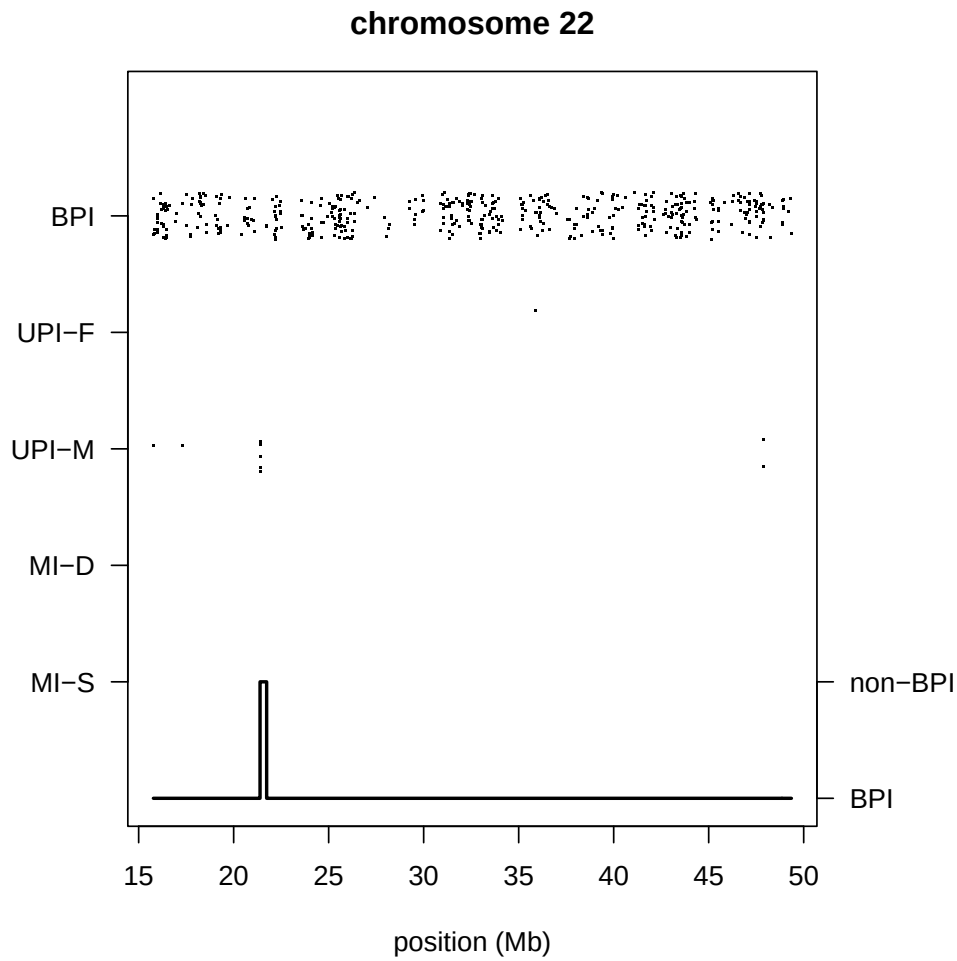


SNP Trio

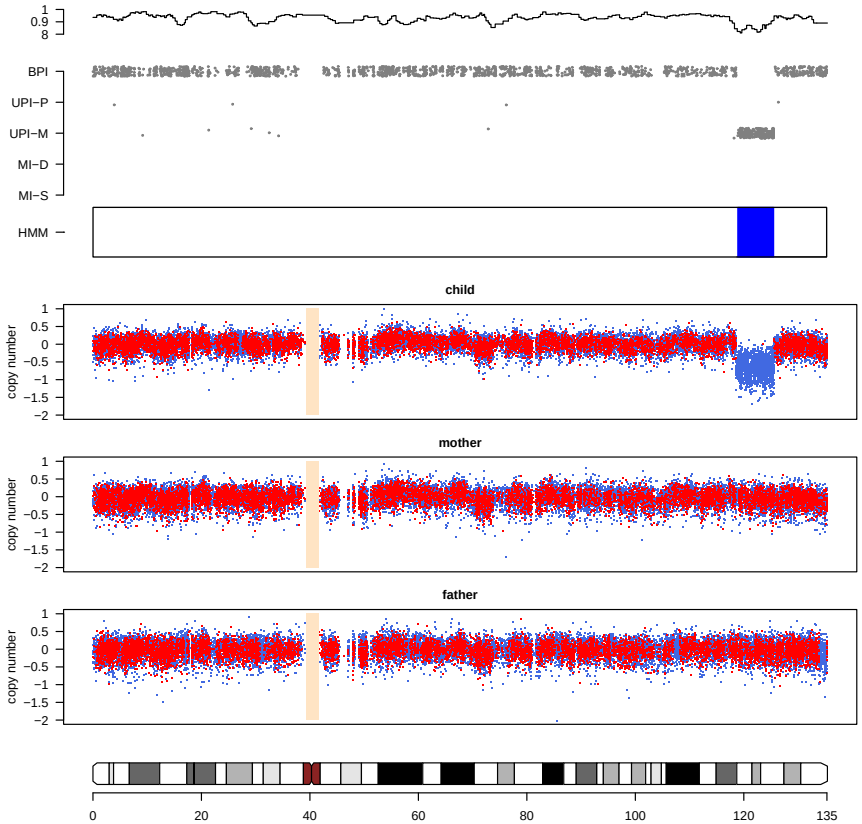

5
MI-D
AA AA BB
BB BB AA
4
MI-S
AA AA AB
BB BB AB
3
iUPD-M
AA BB BB
AA AB BB
BB AB AA
BB AA AA
hUPD-M
AA BB BB
BB AA AA
2
BPD
AA BB AB
BB AA AB
1
hUPD-P
AA BB AA
BB AA BB
iUPD-P
AA BB AA
AB BB AA
AB AA BB
BB AA BB

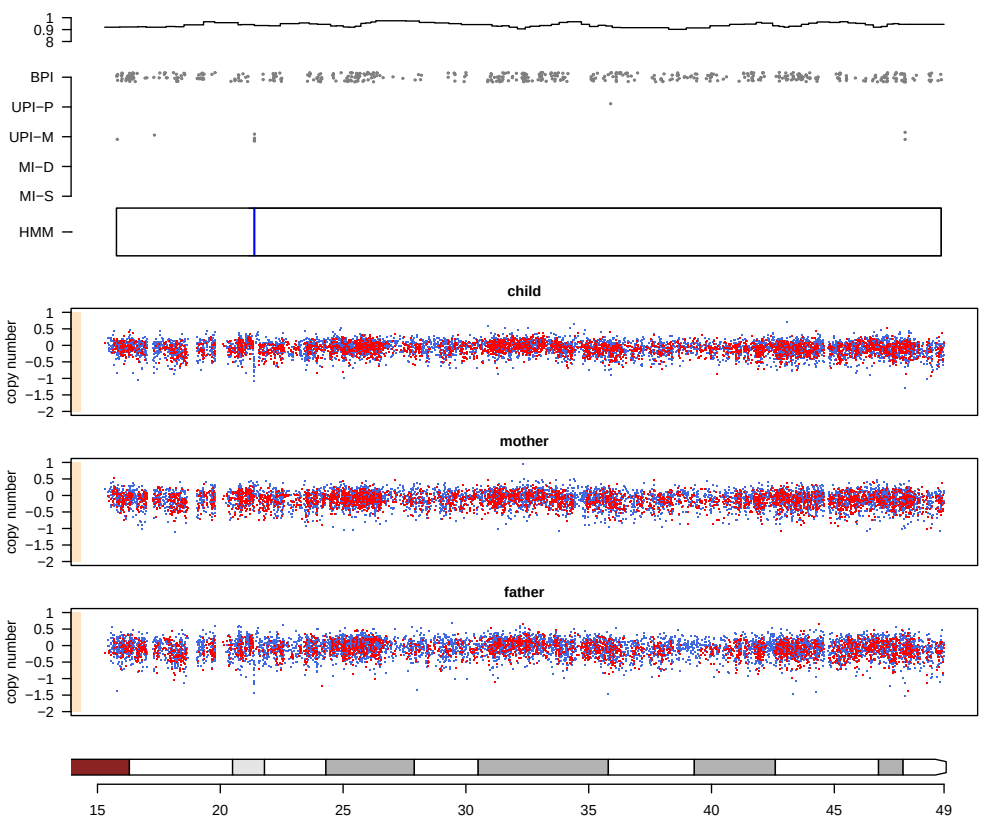
HMM for SNP Trio





HMM for SNP Trio

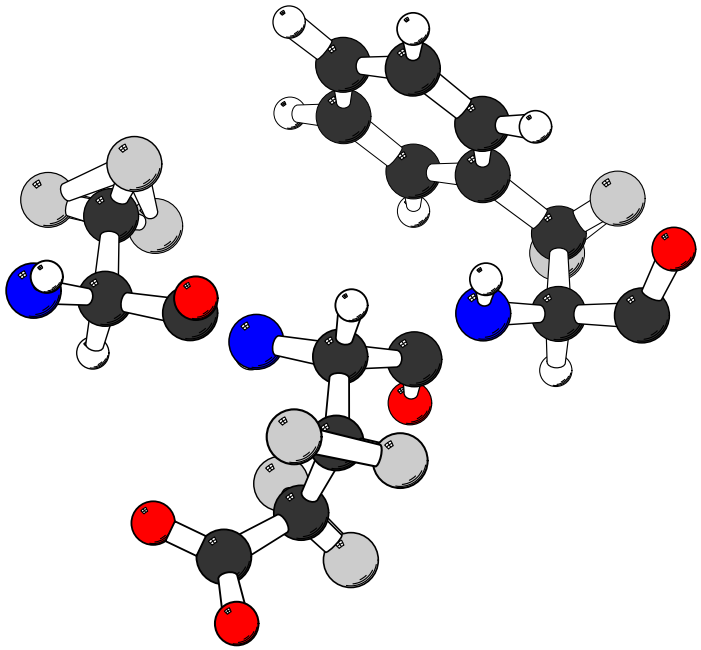


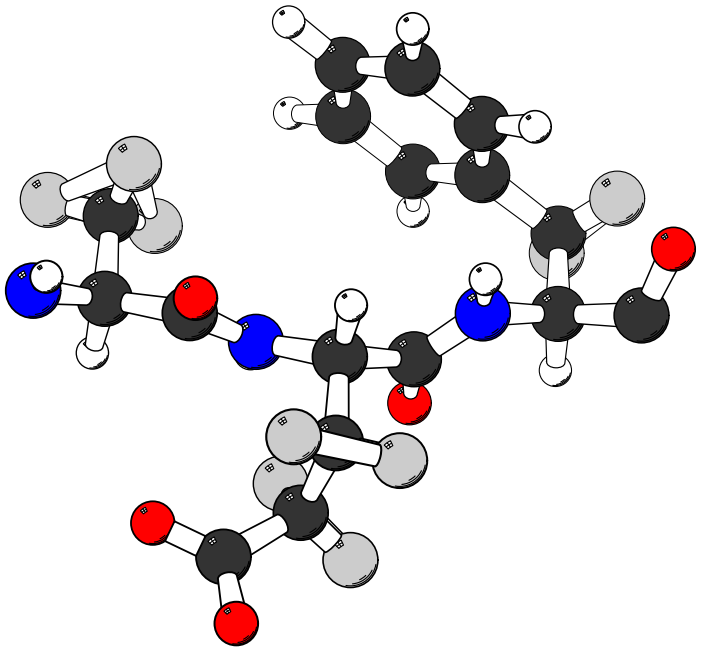


Some selected examples

1. Chromosomal alterations
2. Protein structure prediction
3. 2D gel electrophoresis

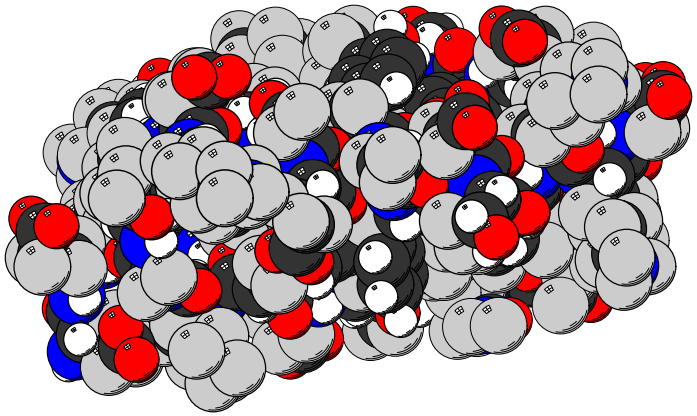
Proteins

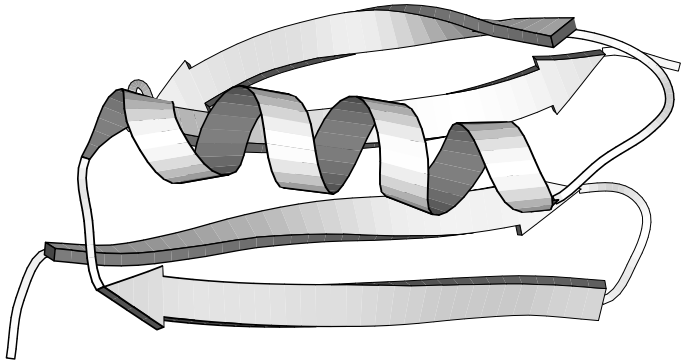




⟶ Amino acids are the building blocks of proteins.

Proteins

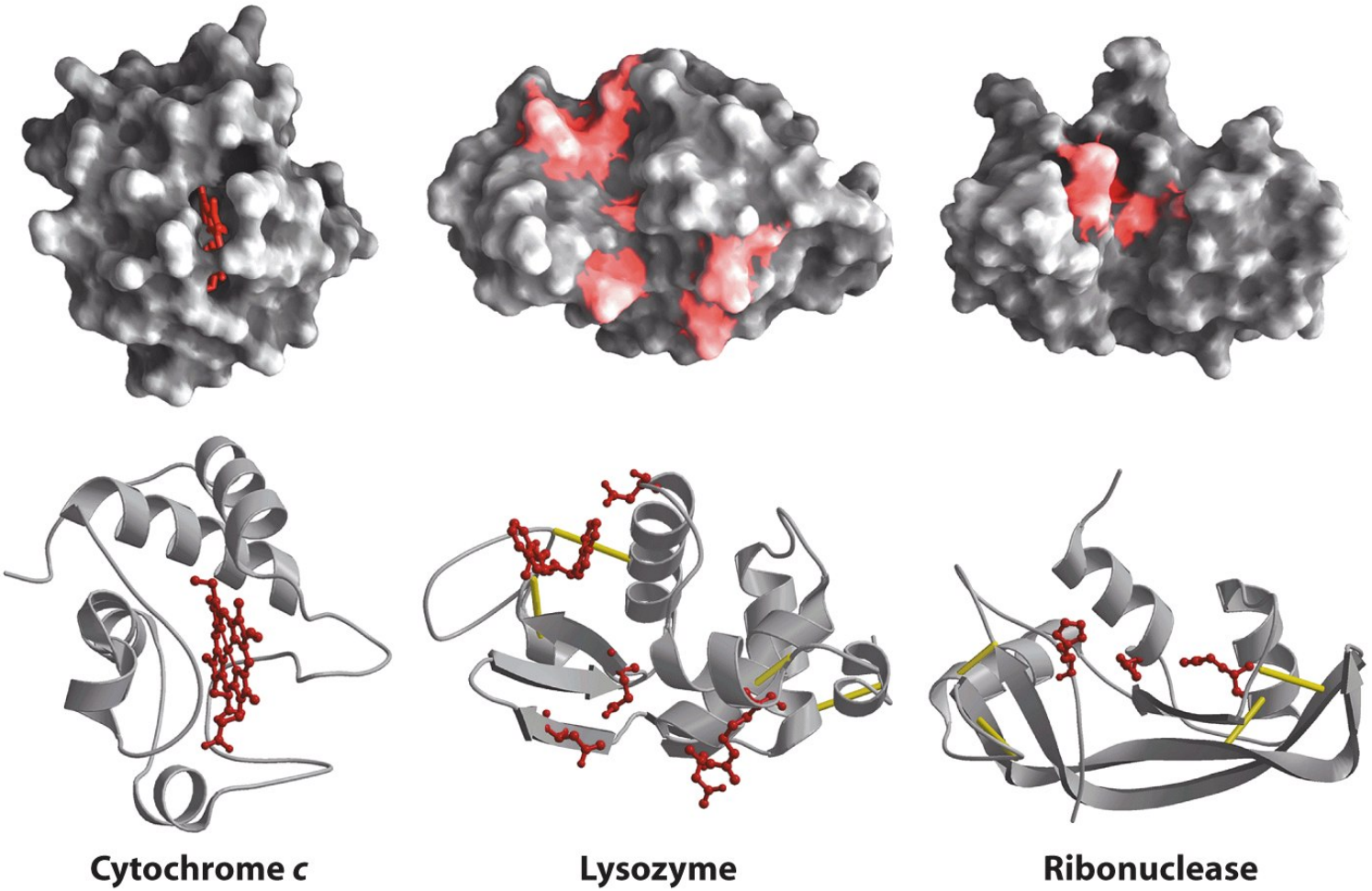




$\longrightarrow$ Both figures show the same protein (the bacterial protein L).

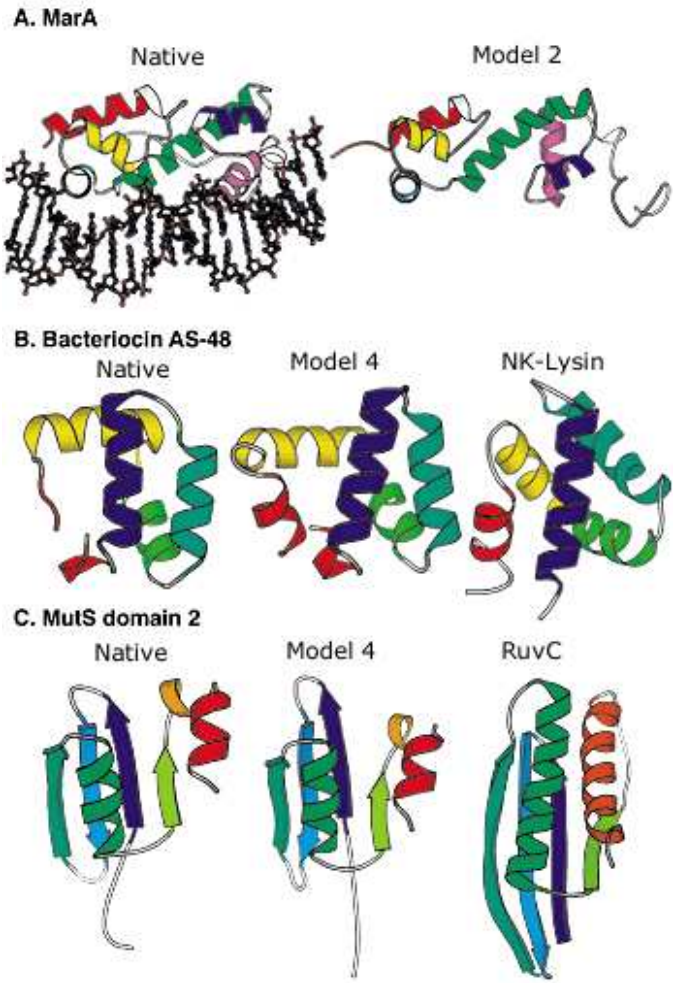
The right figure also highlights the secondary structure elements.

Proteins

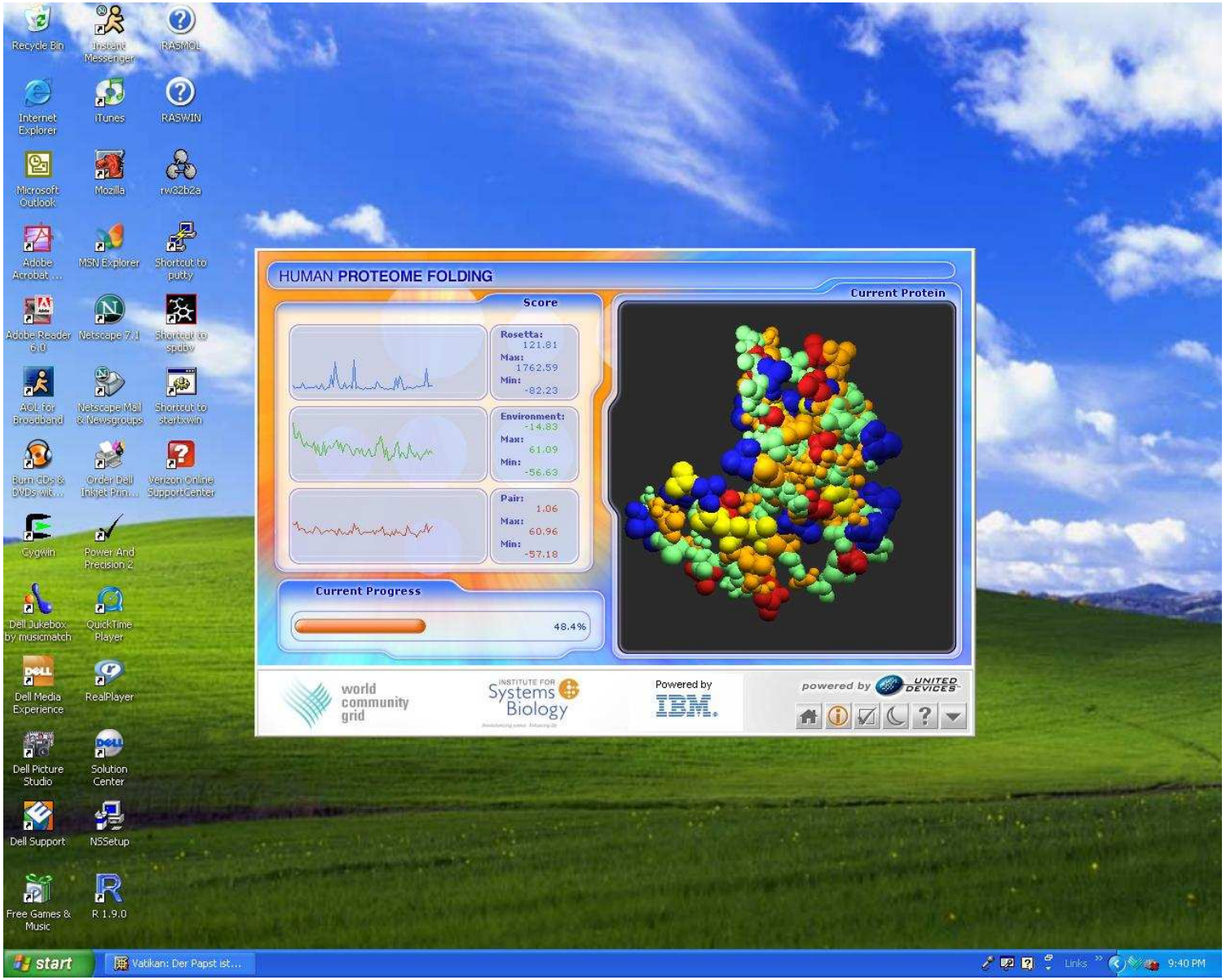


From Lehninger, *Principles of Biochemistry*

Functional Annotation



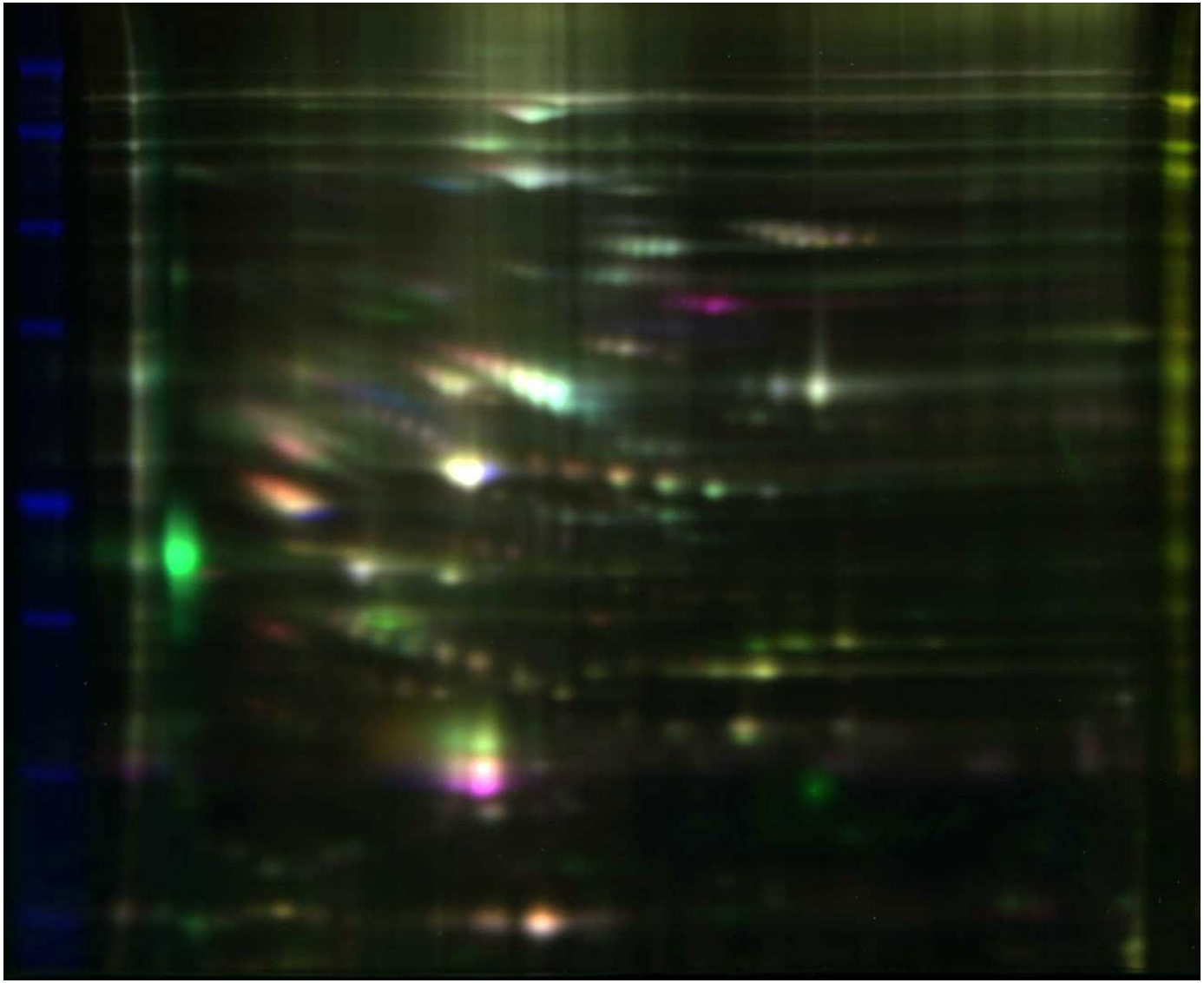
Genome Wide Annotation



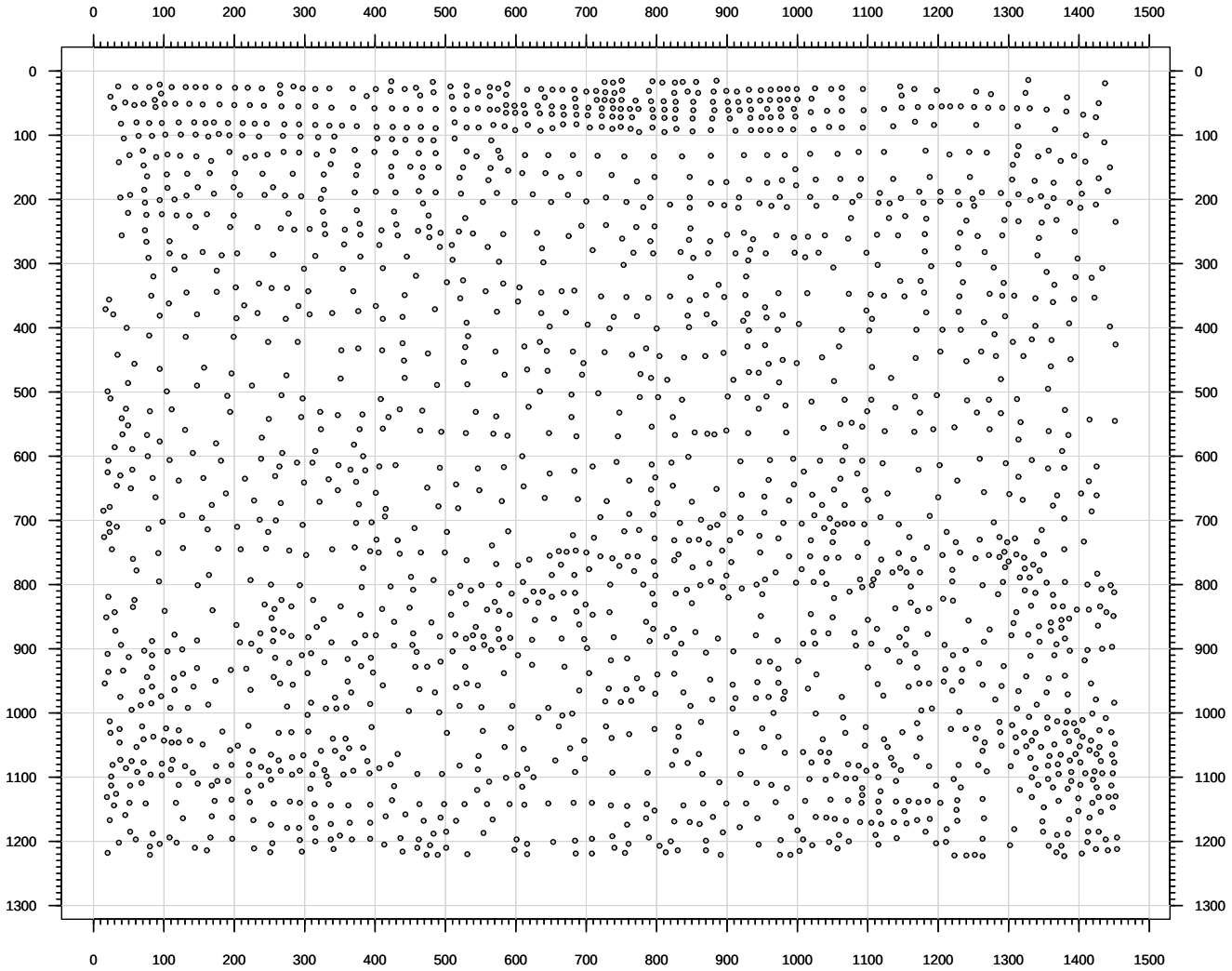
Some selected examples

1. Chromosomal alterations
2. Protein structure prediction
3. 2D gel electrophoresis

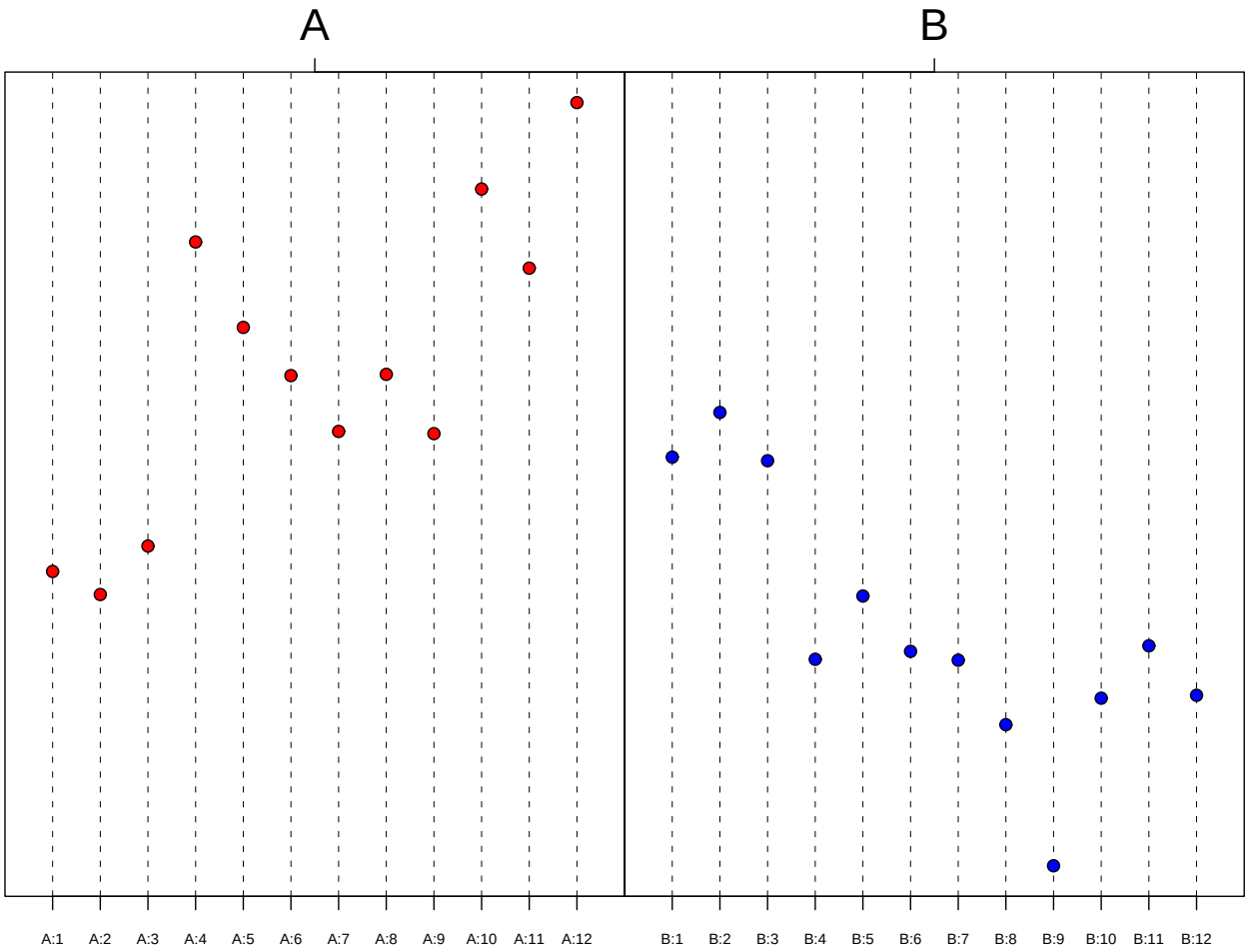
2D Gel Electrophoresis



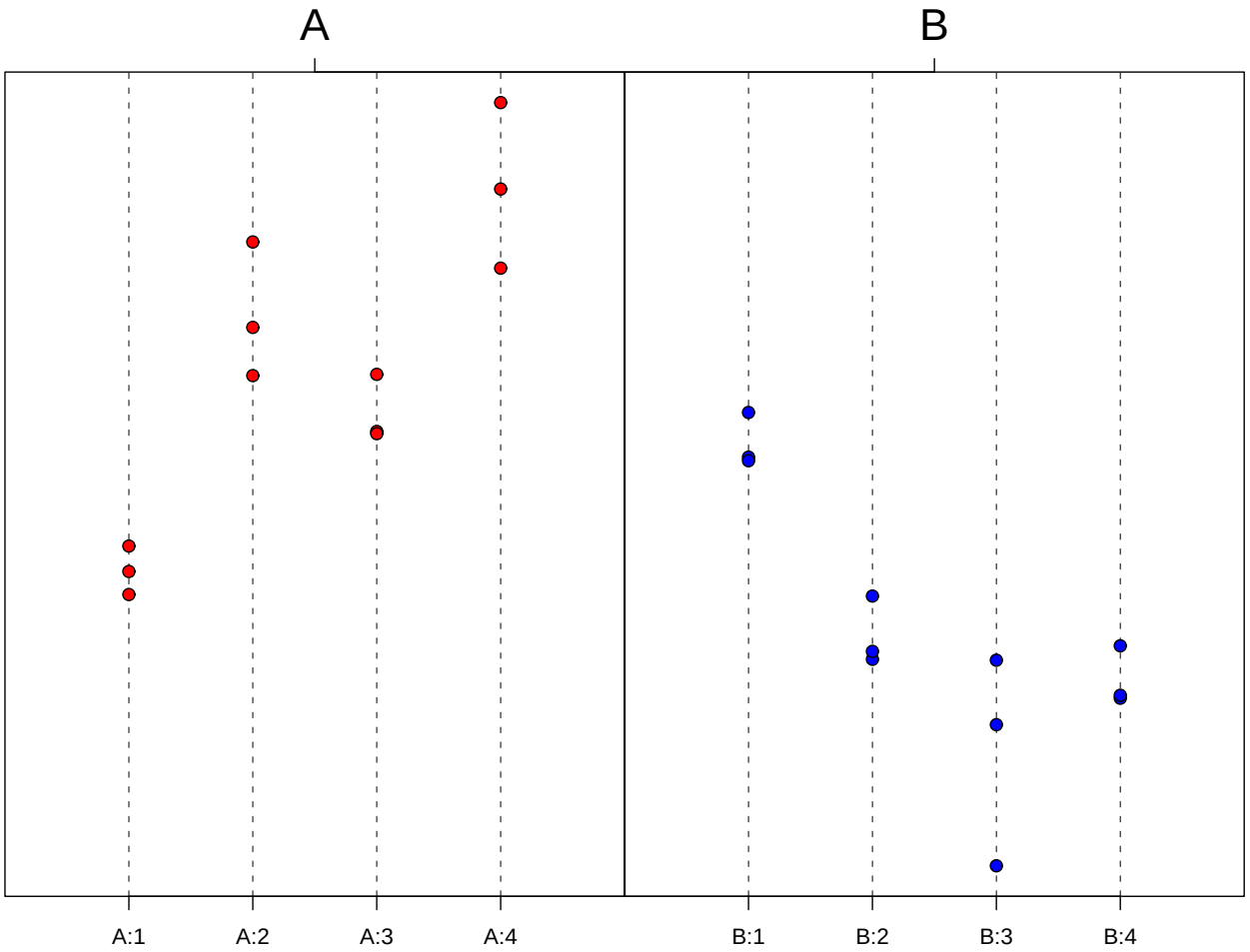
2D Gel Electrophoresis



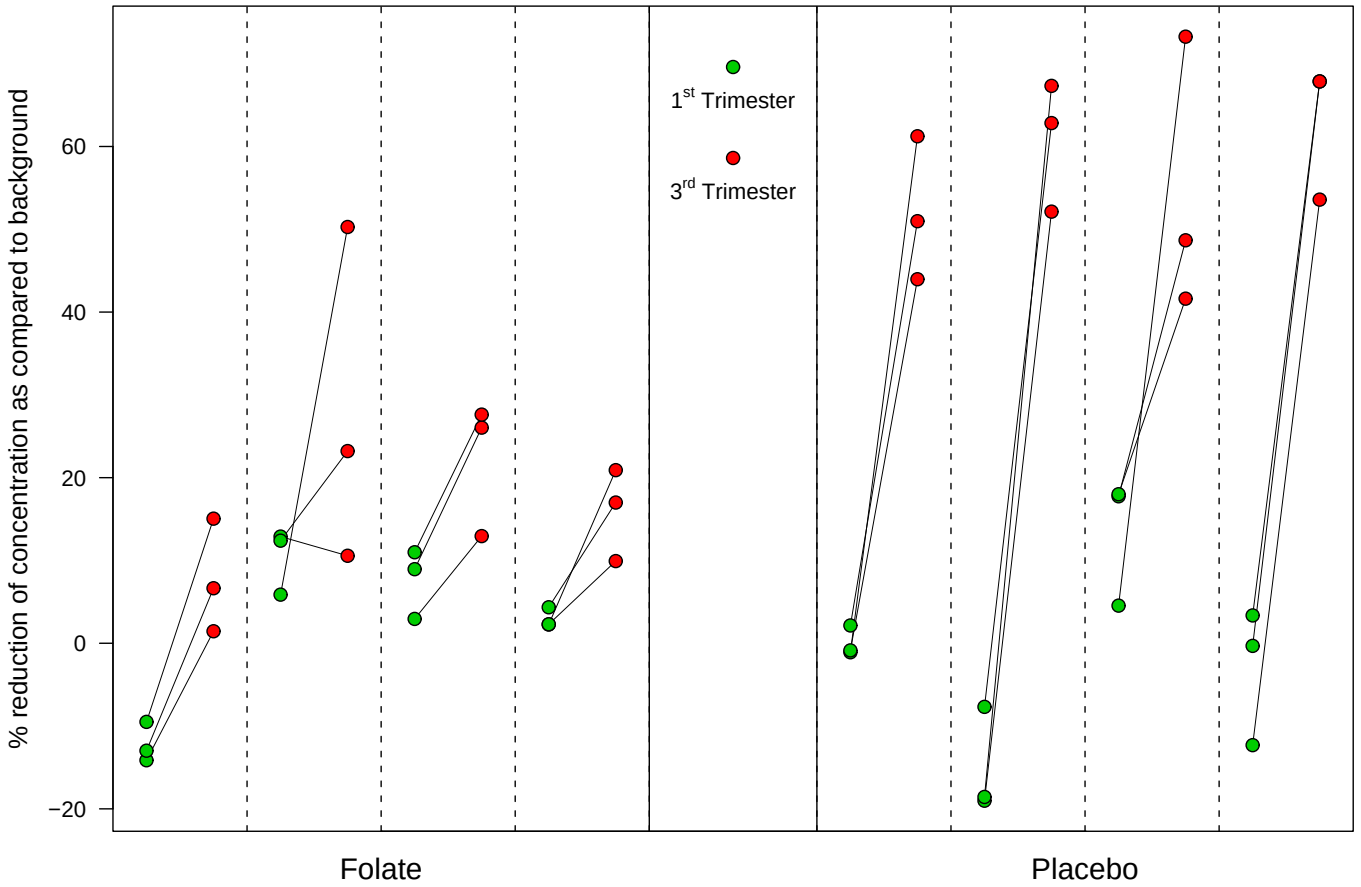
2D Gel Electrophoresis



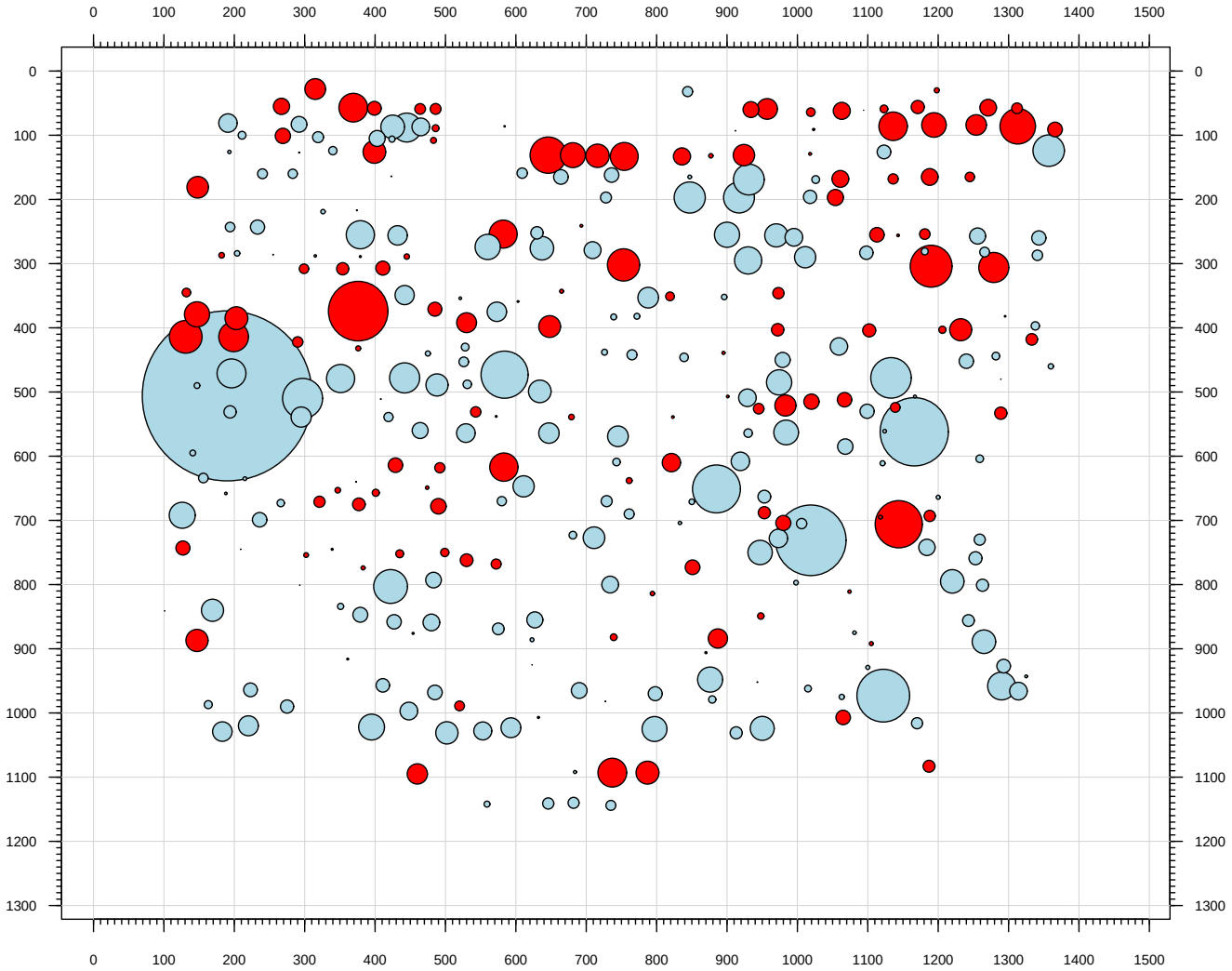
2D Gel Electrophoresis



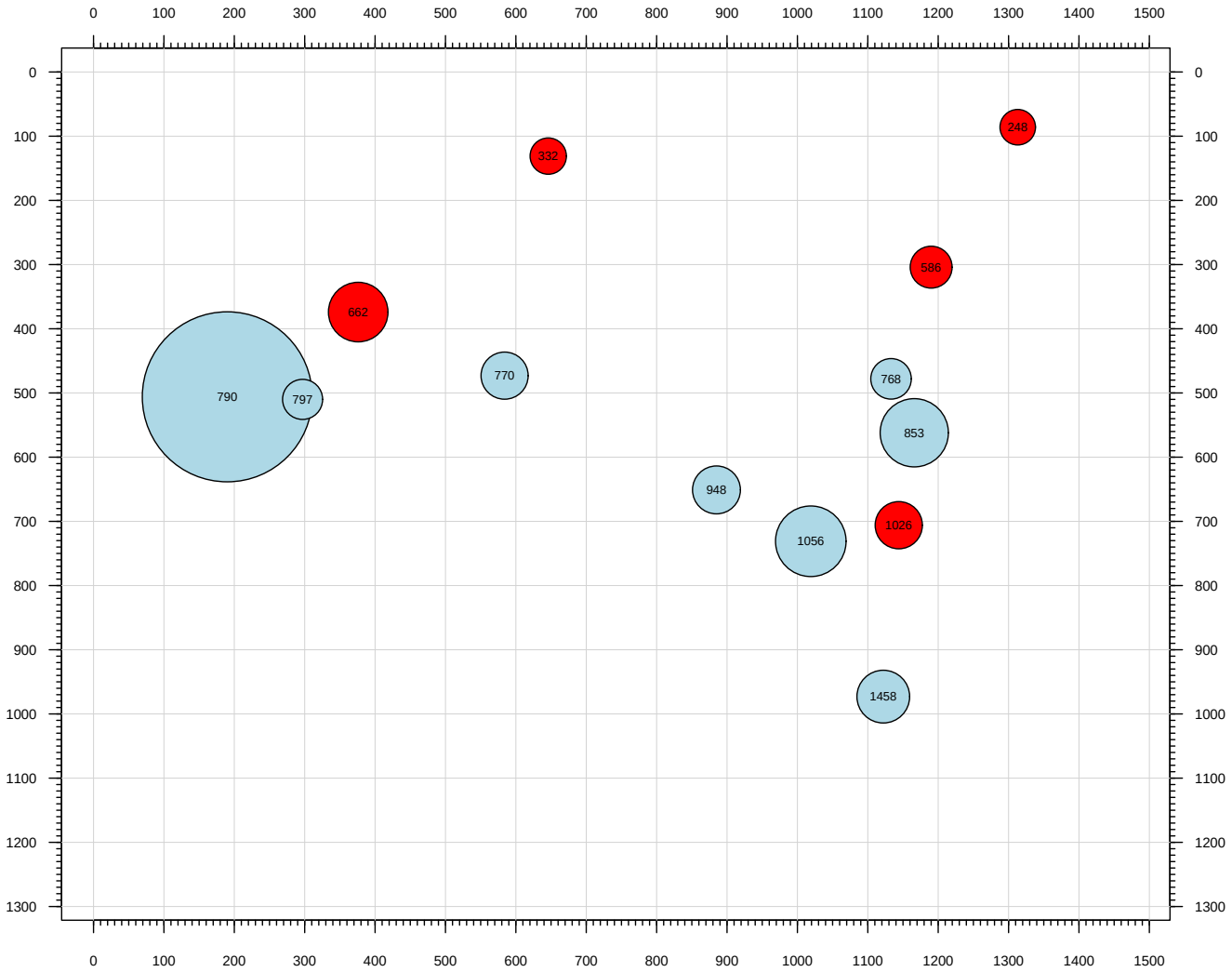
2D Gel Electrophoresis



2D Gel Electrophoresis



2D Gel Electrophoresis



http://biostat.jhsph.edu/∼iruczins/