Bioinformatics

Jonathan Pevsner, Ph.D.
Kennedy Krieger Institute/
Johns Hopkins School of Medicine

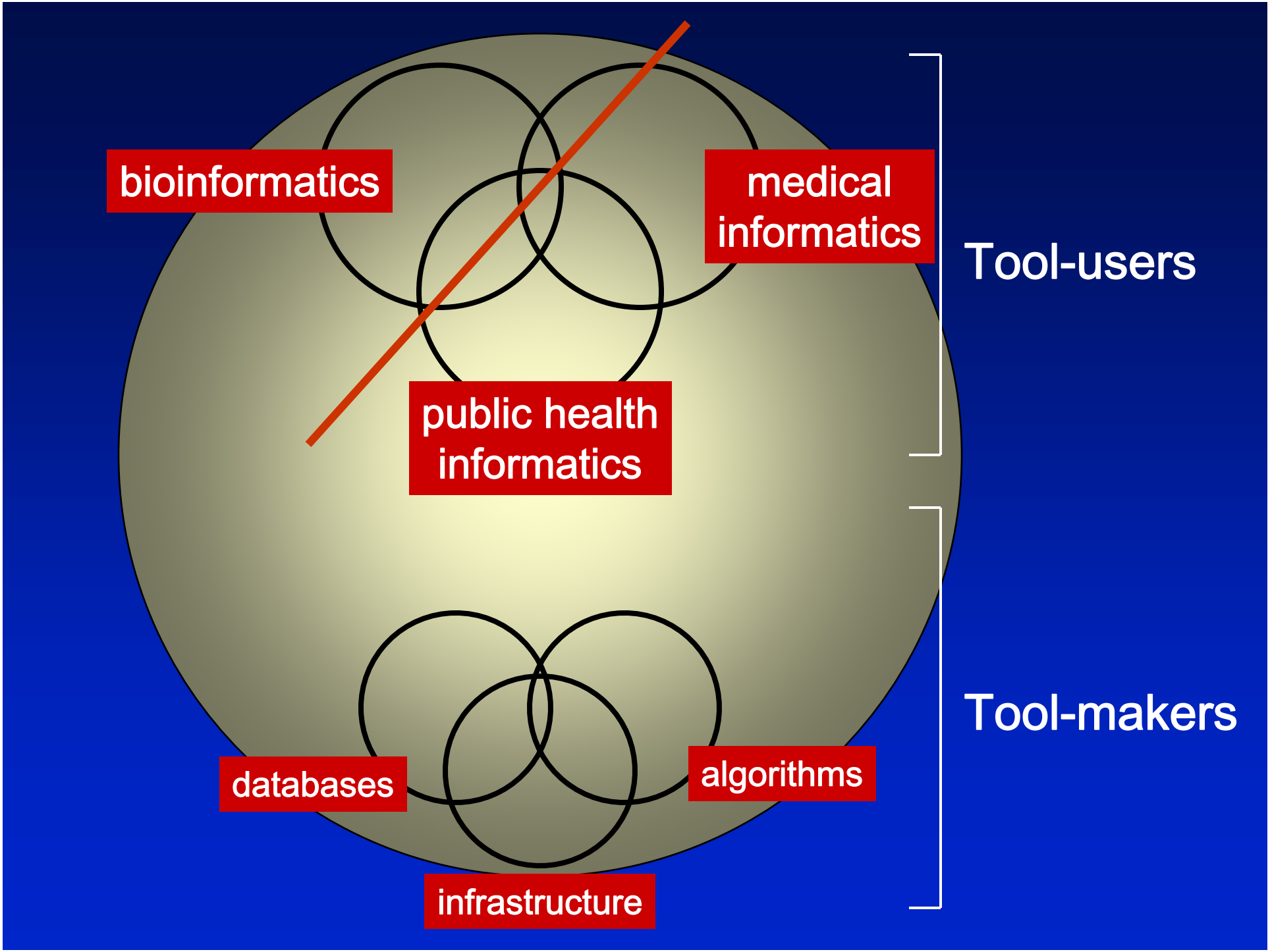
House of Lords Inquiry on Genomic Medicine at NHGRI
June 5, 2008

Definitions of bioinformatics and genomics

- Bioinformatics is the interface of biology and computers. It is the analysis of proteins, genes and genomes using computer algorithms and databases.

- Genomics is the analysis of genomes, including the nature of genetic elements on chromosomes. The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

- Genetics is the study of the origin and expression of individual uniqueness.


bioinformatics
medical informatics
public health informatics
databases
algorithms
infrastructure
Tool-users
Tool-makers

DNA

RNA

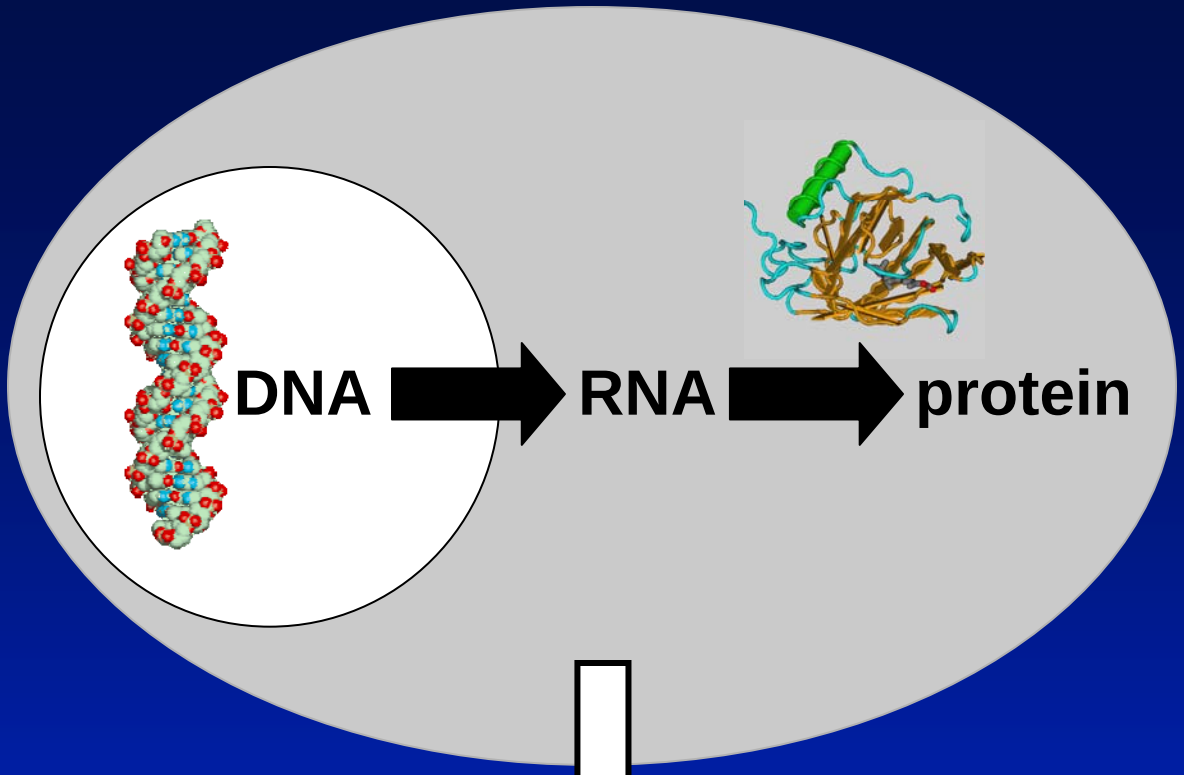
protein

pathway

cell

organism

population



cellular phenotype

clinical phenotype

DNA

RNA

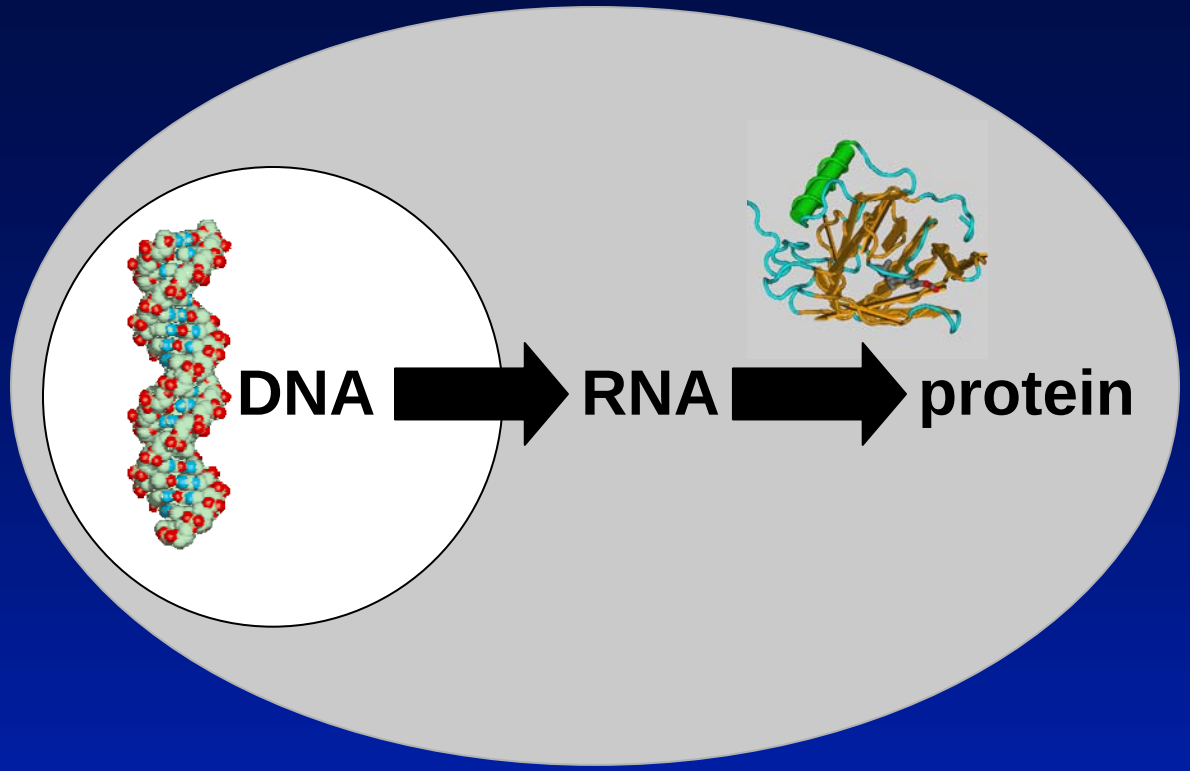
protein

pathway

cell

organism

population



Central dogma of molecular biology:
DNA is transcribed into RNA,
and translated into protein.

Central dogma of bioinformatics/genomics:
the genome is transcribed into the transcriptome,
and translated into the proteome.

DNA

RNA

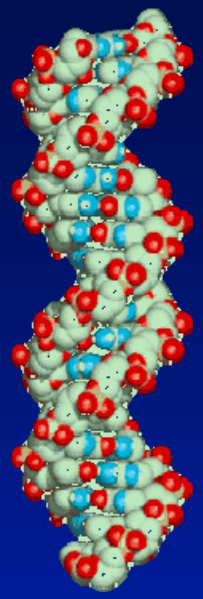
protein

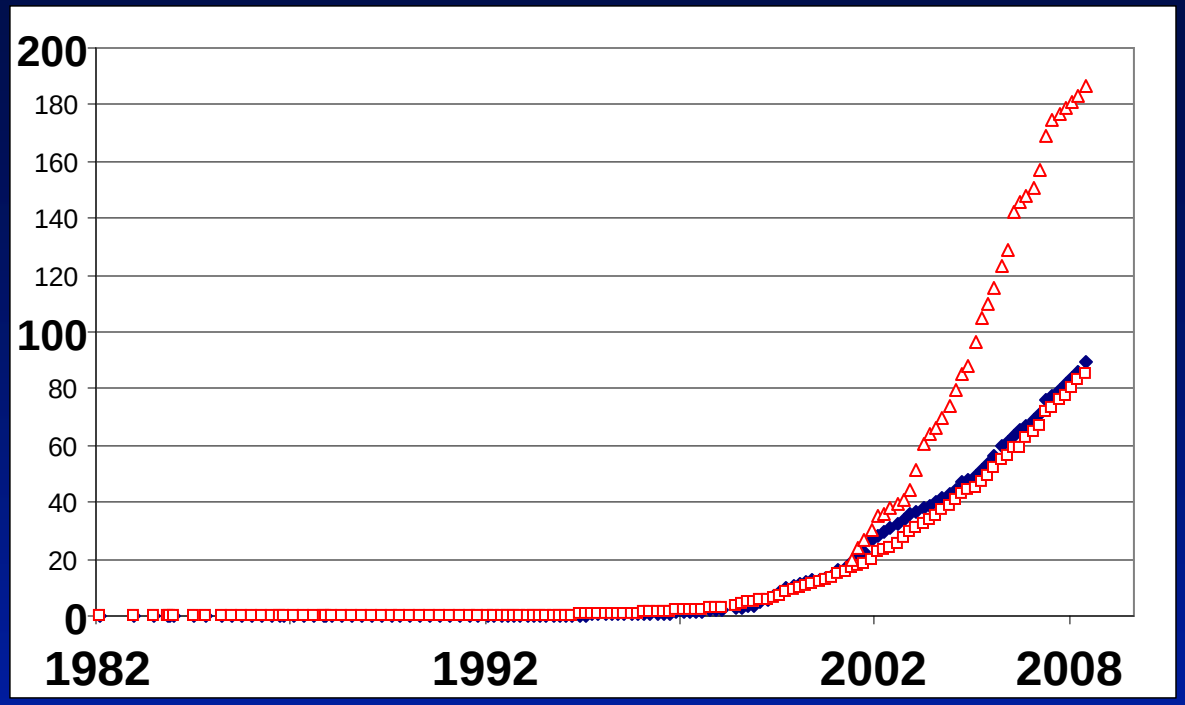
pathway

cell

organism

population





Over 180 billion base pairs of DNA have now been sequenced, from >165,000 organisms.

DNA

RNA

protein

pathway

cell

organism

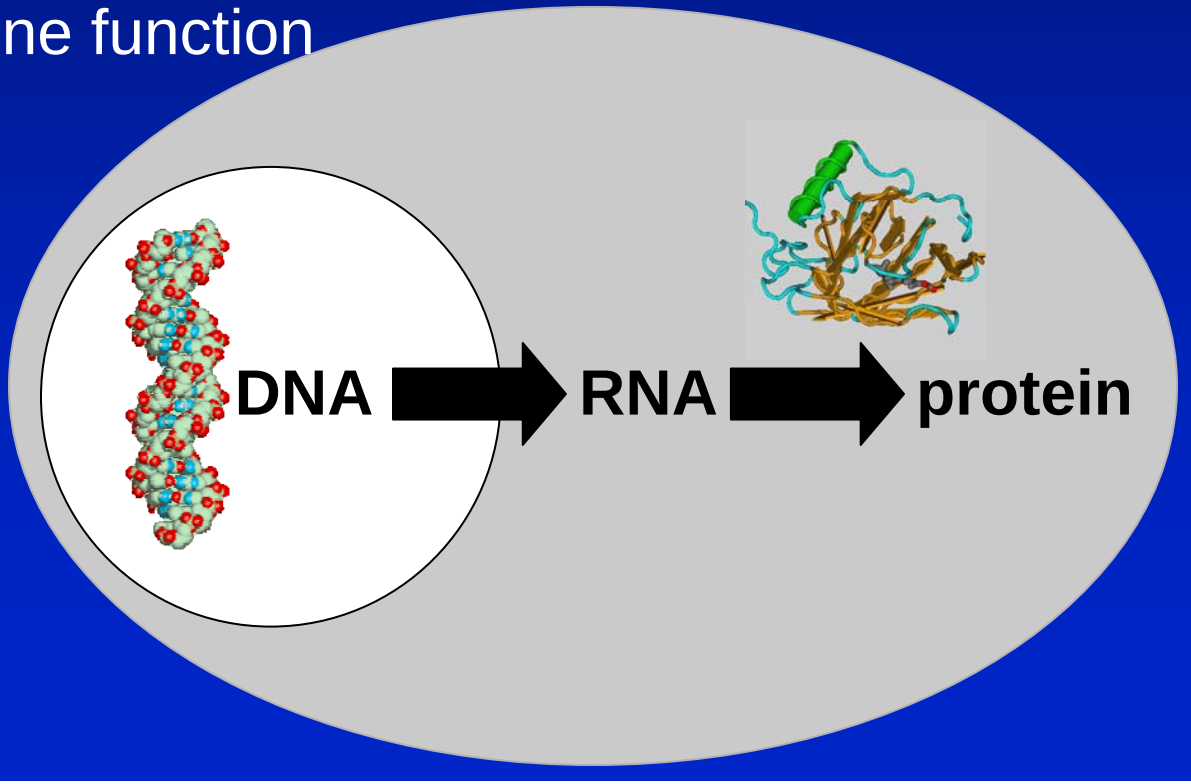
population

Sequence analysis

- Pairwise alignment
- Multiple sequence alignment
- Phylogeny
- Database searching (e.g. BLAST)

Functional genomics

- RNA studies; gene expression profiling
- Proteomics; protein structure
- Gene function



DNA

RNA

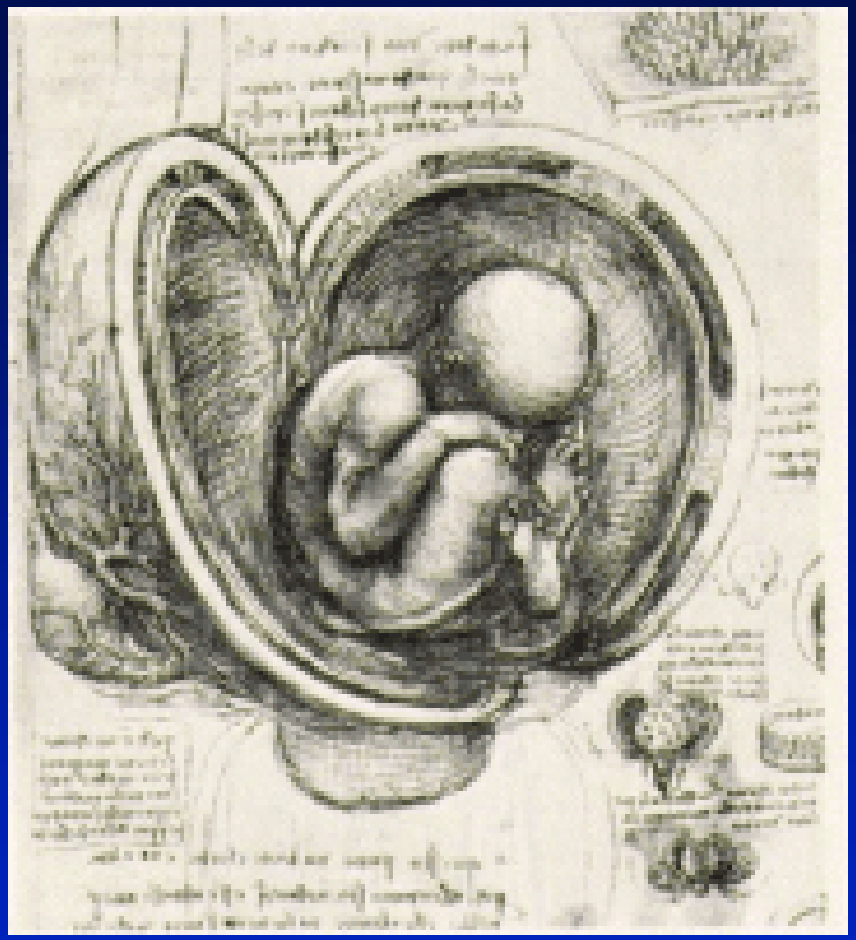
protein

pathway

cell

organism

population



Time of development

Body region, physiology, pharmacology, pathology

DNA

RNA

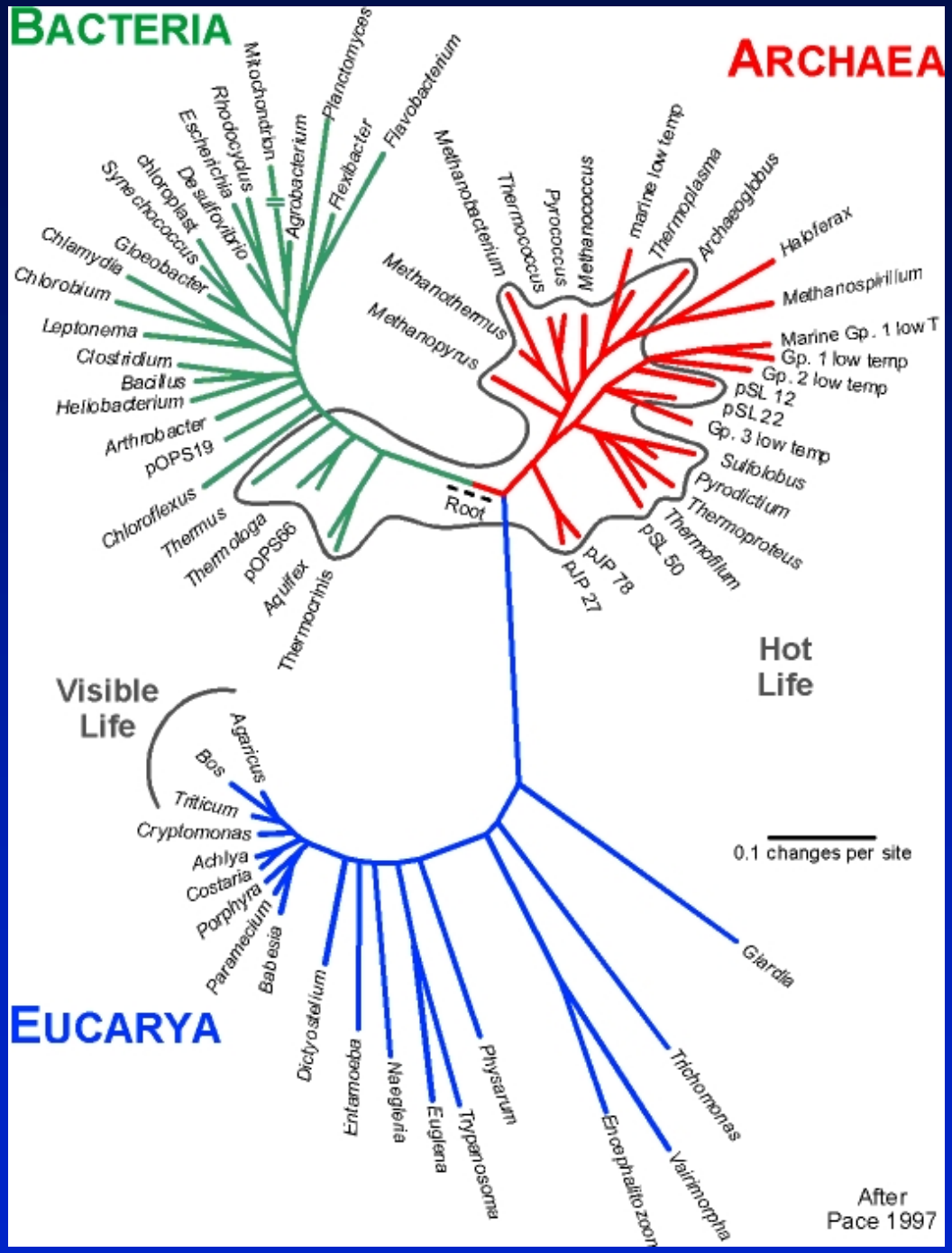
protein

pathway

cell

organism

population



DNA

RNA

protein

pathway

cell

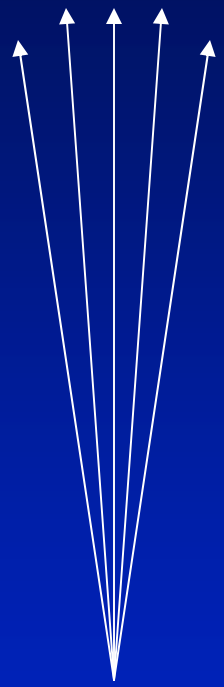
organism

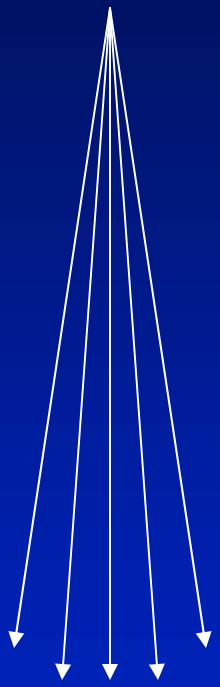
population

DNA database (e.g. GenBank/EMBL)

multiple genes

one gene





one disease

multiple diseases

Disease database (e.g. OMIM)

DNA

RNA

protein

pathway

cell

organism

population

DNA database (e.g. GenBank/EMBL)

Challenges in creating a disease database

- gene-centric versus disease-centric: different concepts apply at these levels
- complexity of disease mechanisms not readily captured
- the connection between a gene and a disease is often obscure; some molecular changes causal, others reactive (secondary)
- false positive and false negative error rates difficult to estimate

Disease database (e.g. OMIM)

DNA

RNA

protein

pathway

cell

organism

population

Phenotype

We see 500 inpatients and 13,000 outpatients per year at the Kennedy Krieger Institute. Why do children engage in self-injurious behavior? In many cases, there are chromosomal insults.

DNA

RNA

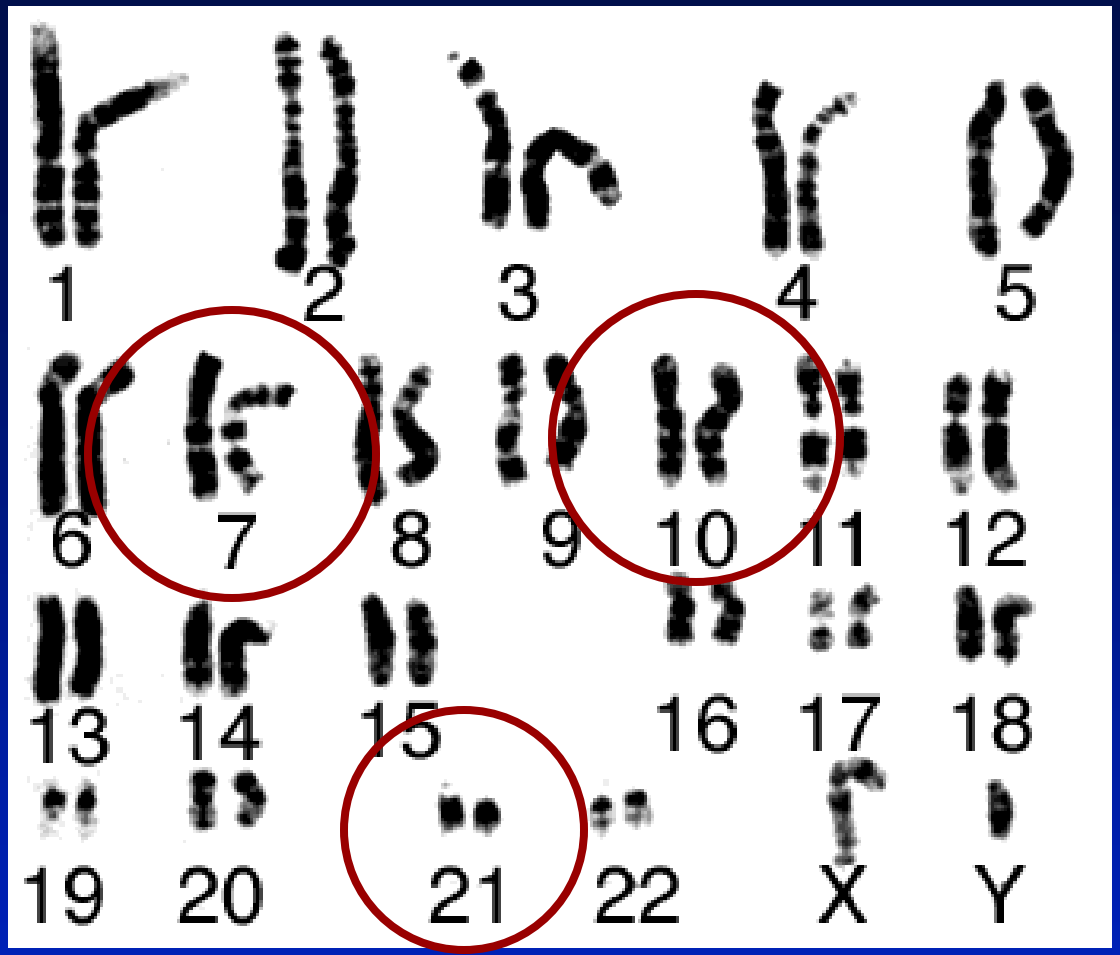
protein

pathway

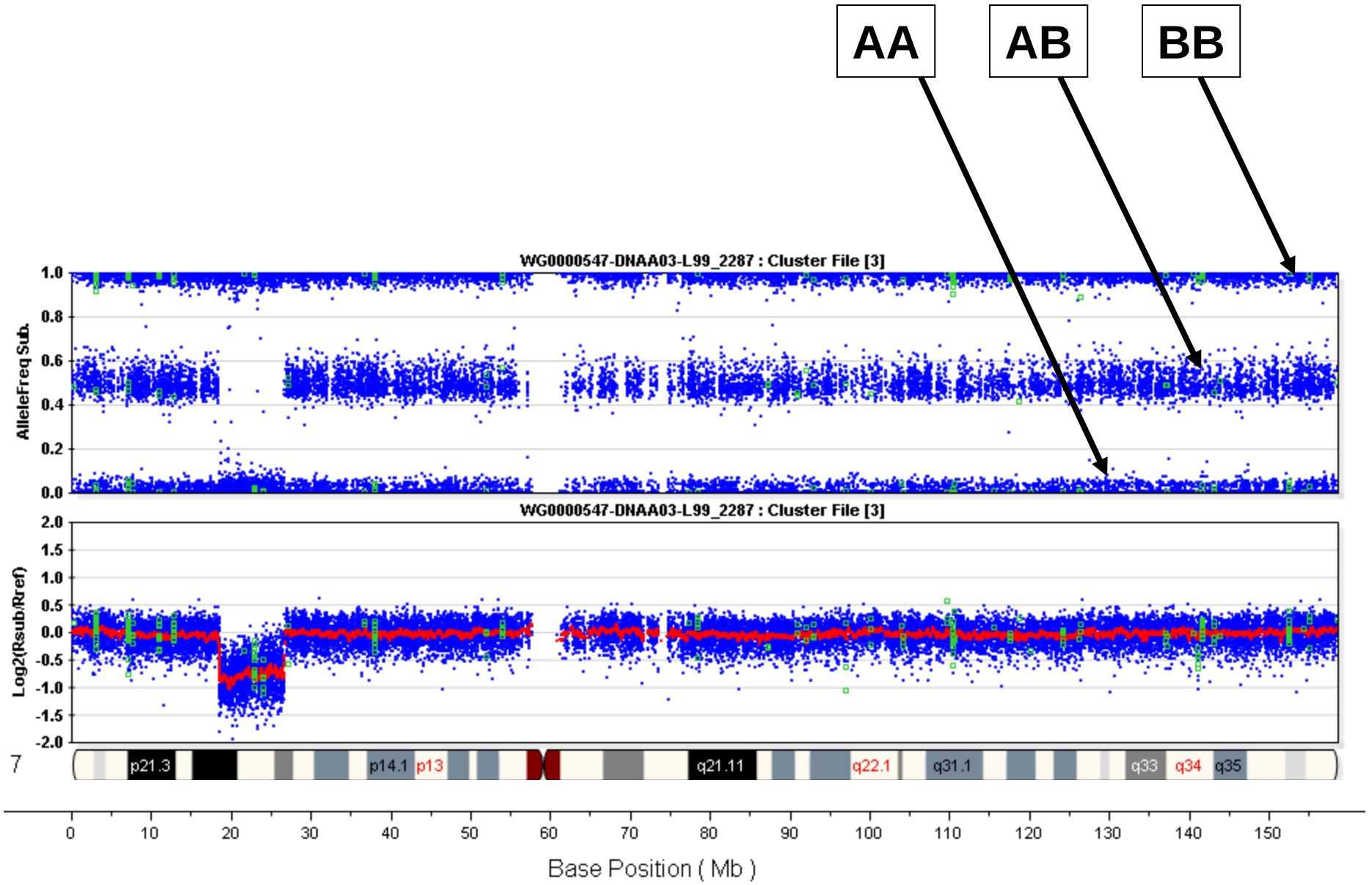
cell

organism

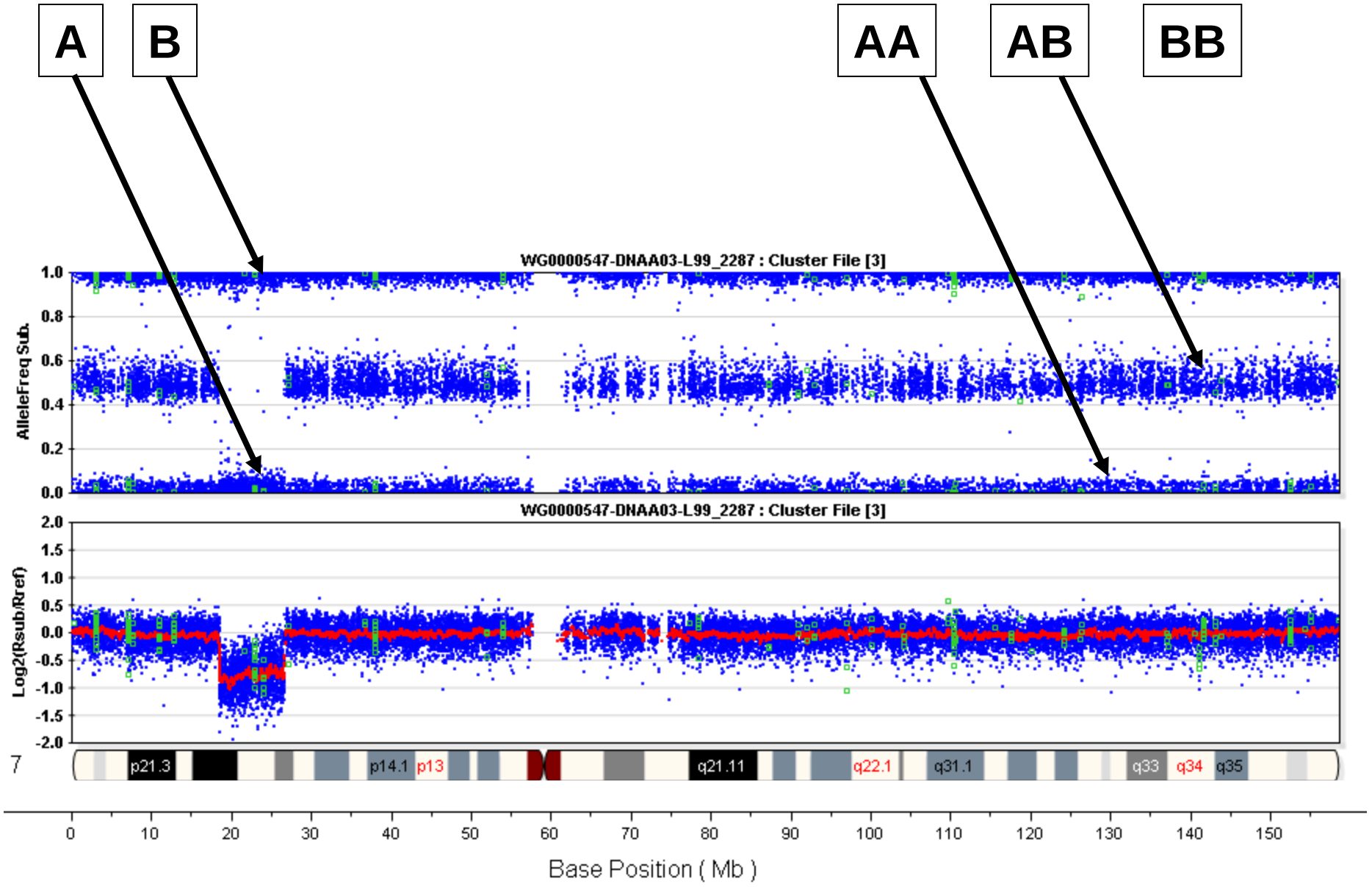
population



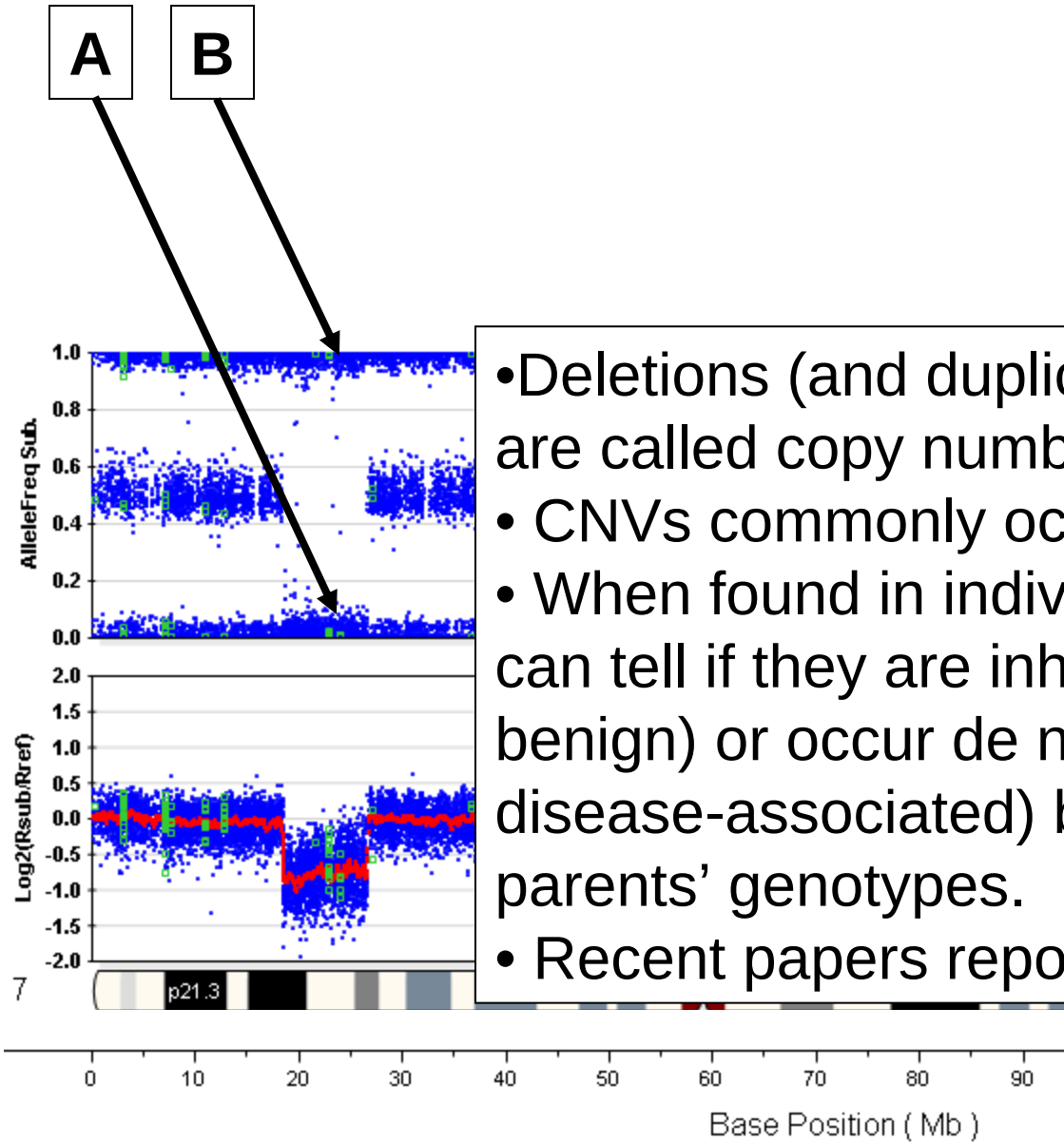
Single nucleotide polymorphisms (SNPs) to investigate chromosomes: A case of 7p deletion



A case of 7p deletion

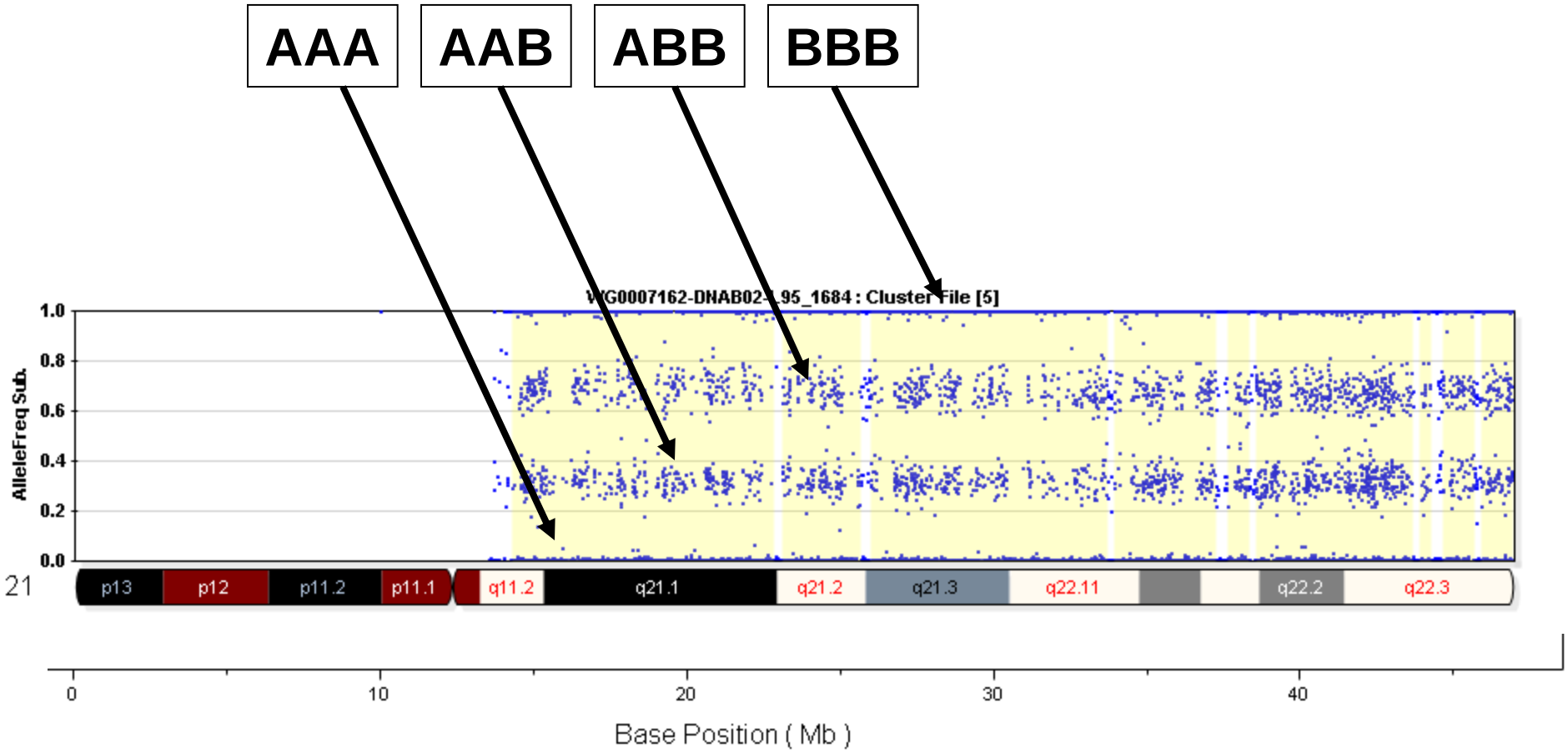


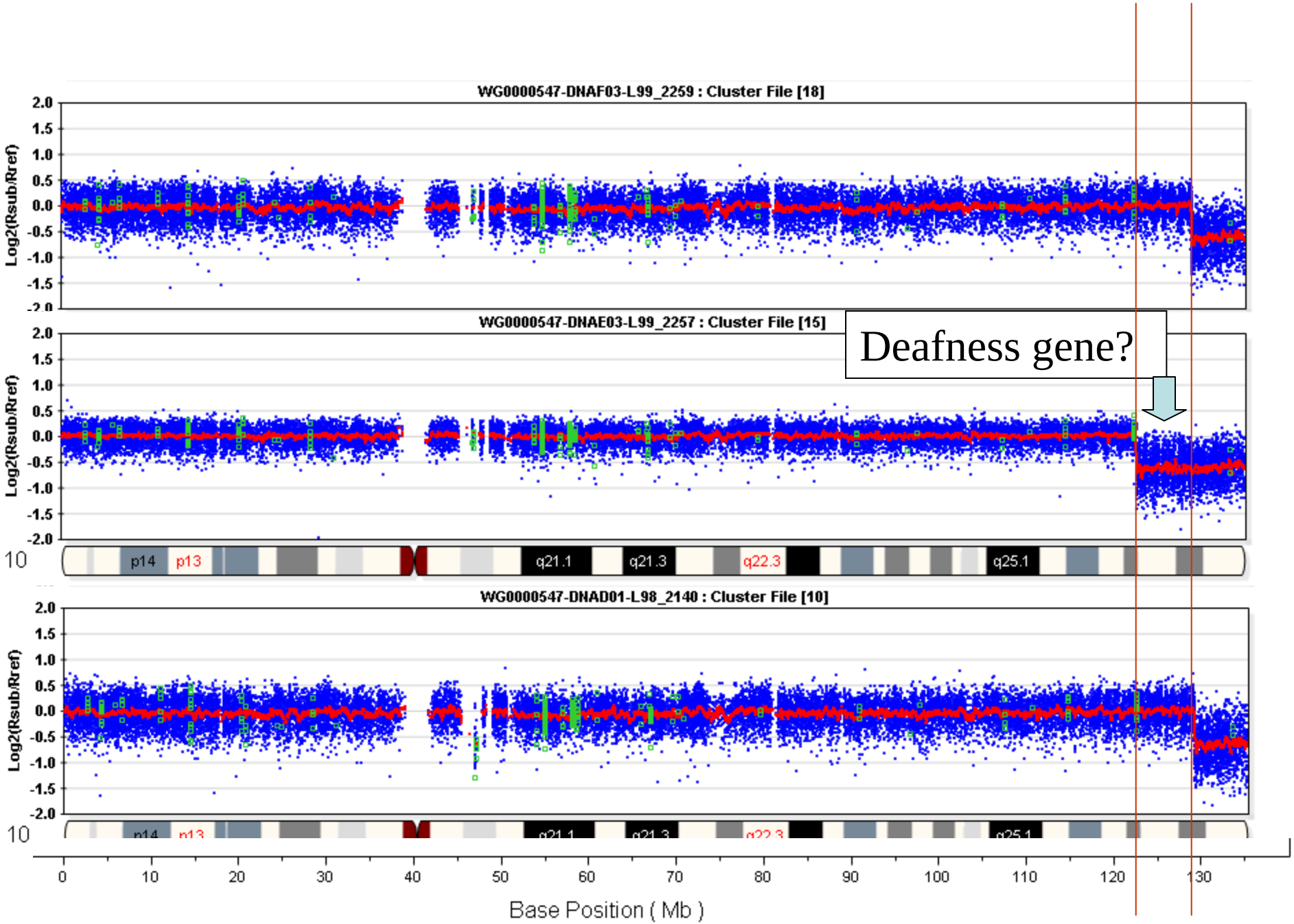
A case of 7p deletion



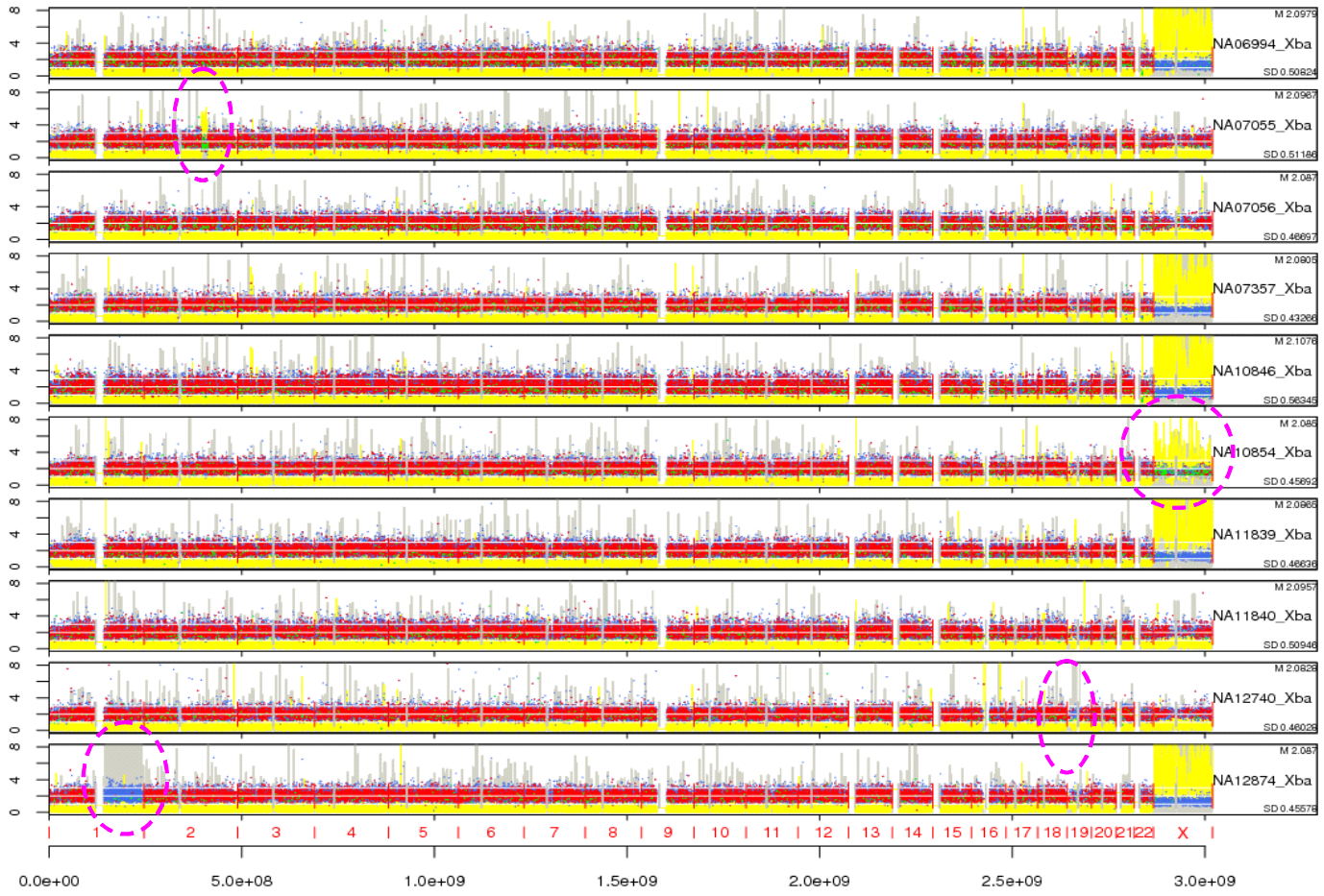
- Deletions (and duplications) such as these are called copy number variants (CNVs).
- CNVs commonly occur in normal individuals.
- When found in individuals with disease, we can tell if they are inherited (likely to be benign) or occur de novo (more likely to be disease-associated) by comparison to the parents' genotypes.
- Recent papers report many CNVs in autism.

A case of trisomy 21 (Down syndrome)

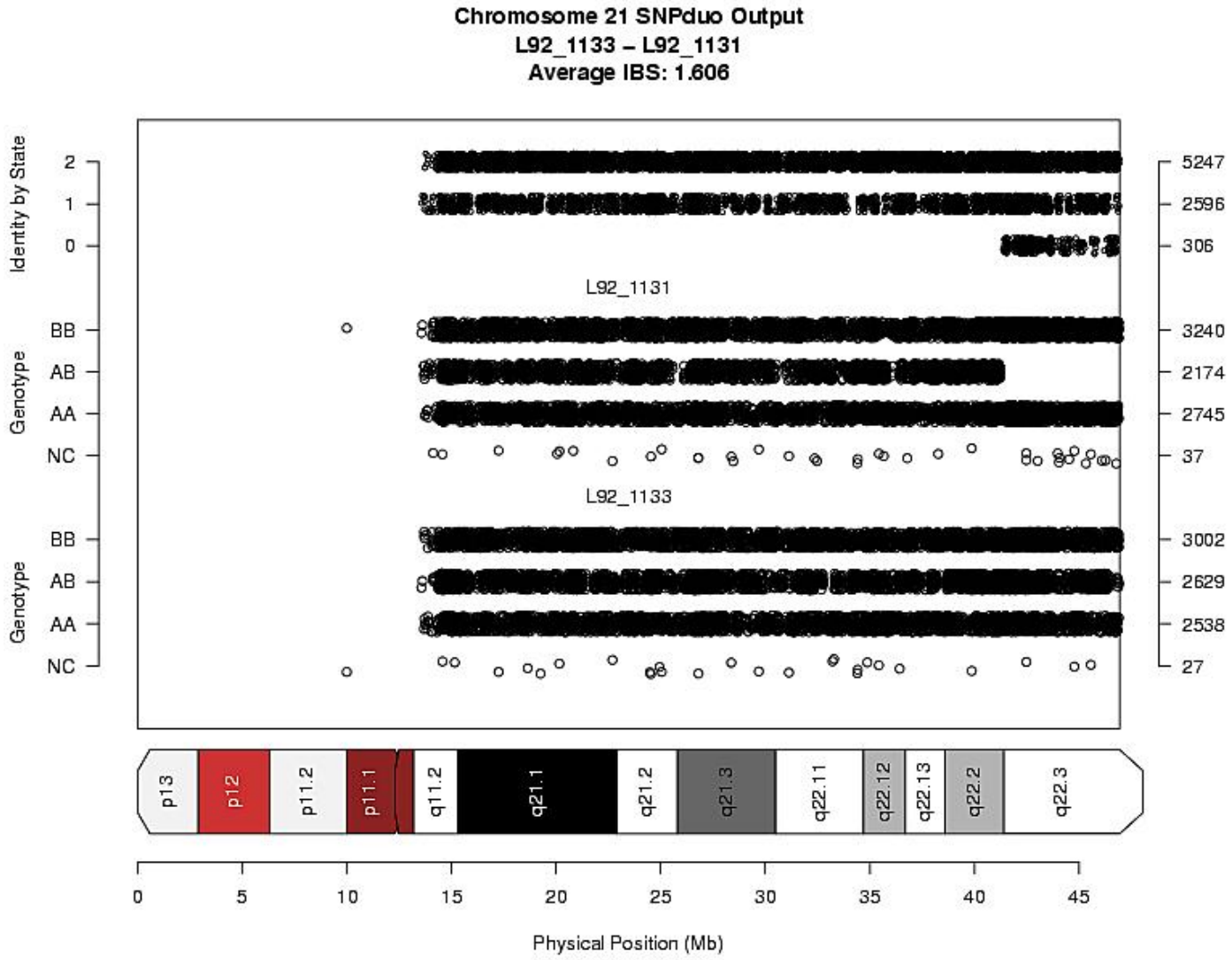


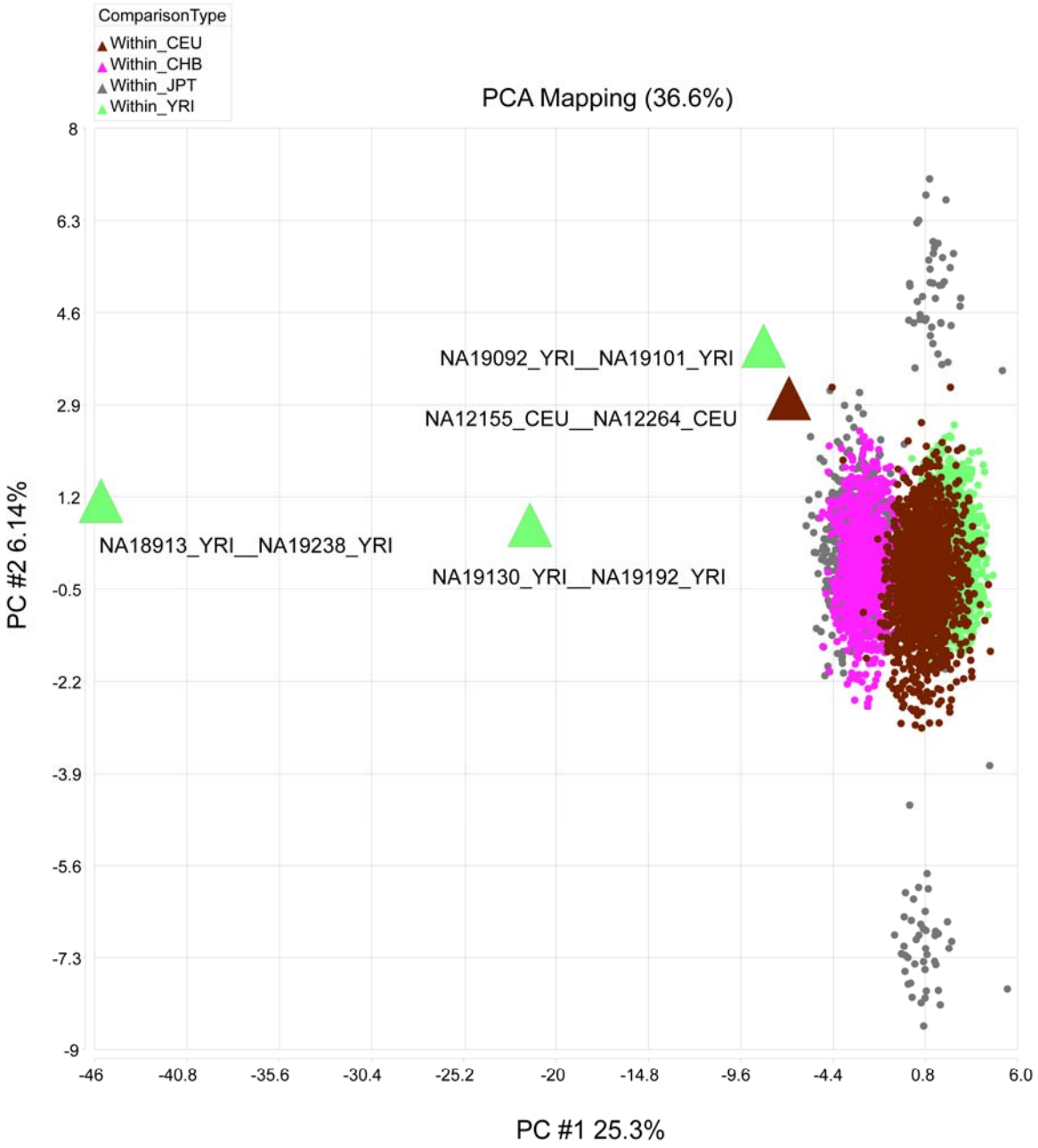

WG0000547-DNAF03-L99_2259 : Cluster File [18]
WG0000547-DNAE03-L99_2257 : Cluster File [15]
WG0000547-DNAD01-L98_2140 : Cluster File [10]
Log2(Rsub/Rref)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Deafness gene?
10
p14
p13
q21.1
q21.3
q22.3
q25.1
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Base Position (Mb)

SNPscan: identifying chromosomal abnormalities in SNP data

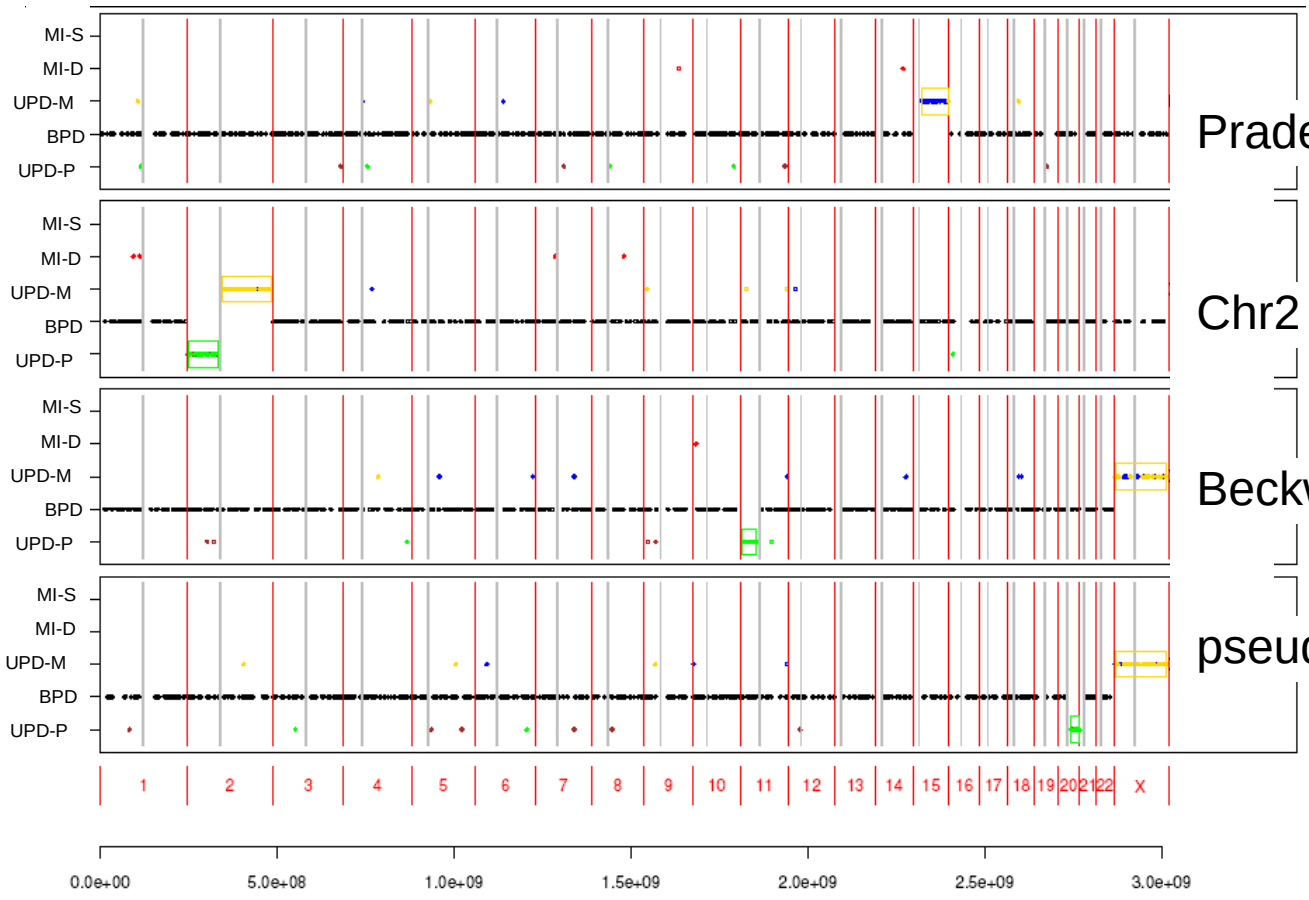


SNPduo: pairwise SNP data analyses




ComparisonType
Within_CEU
Within_CHB
Within_JPT
Within_YRI
PCA Mapping (36.6%)
NA19092_YRI__NA19101_YRI
NA12155_CEU__NA12264_CEU
NA18913_YRI__NA19238_YRI
NA19130_YRI__NA19192_YRI
PC #2 6.14%
PC #1 25.3%
8
6.3
4.6
2.9
1.2
-0.5
-2.2
-3.9
-5.6
-7.3
-9
-46
-40.8
-35.6
-30.4
-25.2
-20
-14.8
-9.6
-4.4
0.8
6.0

SNPtrio: identifying chromosomal abnormalities in SNP data



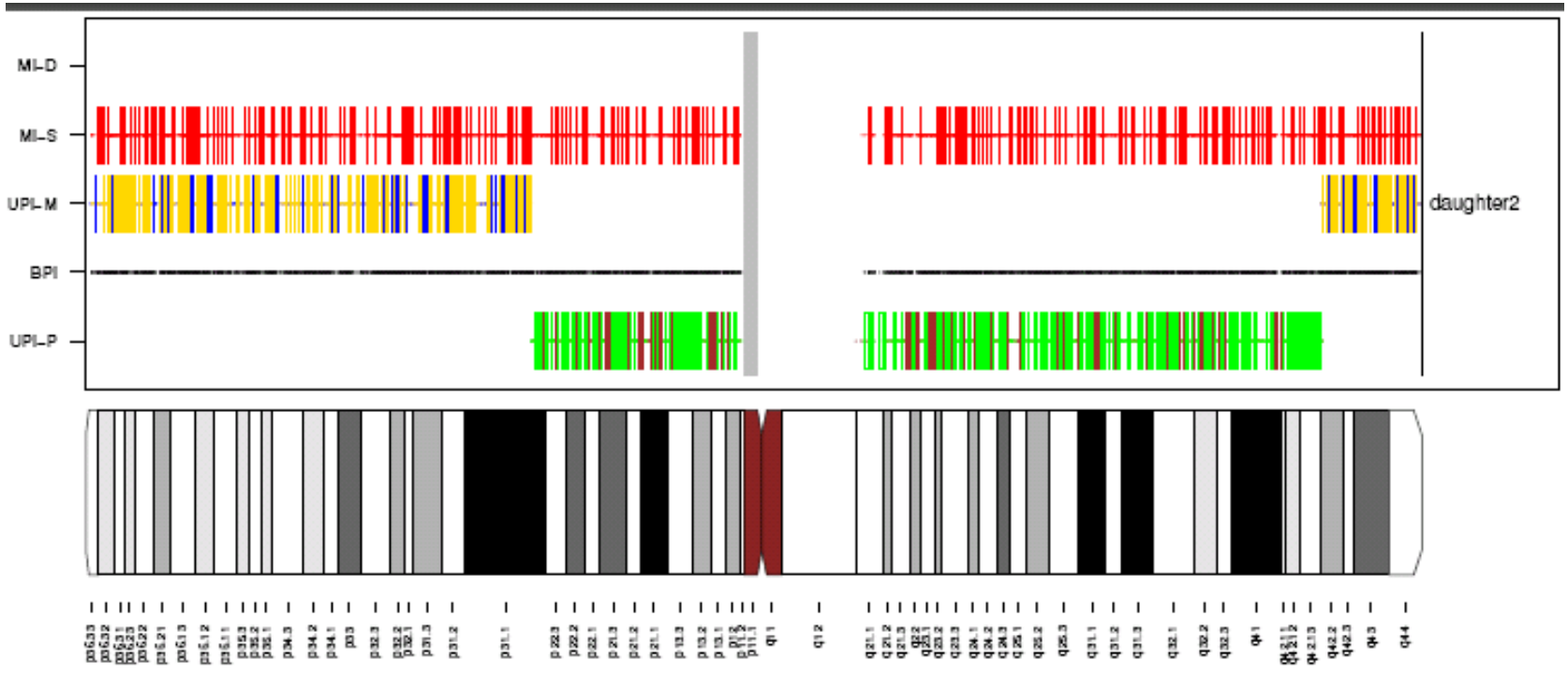
Prader-Willi

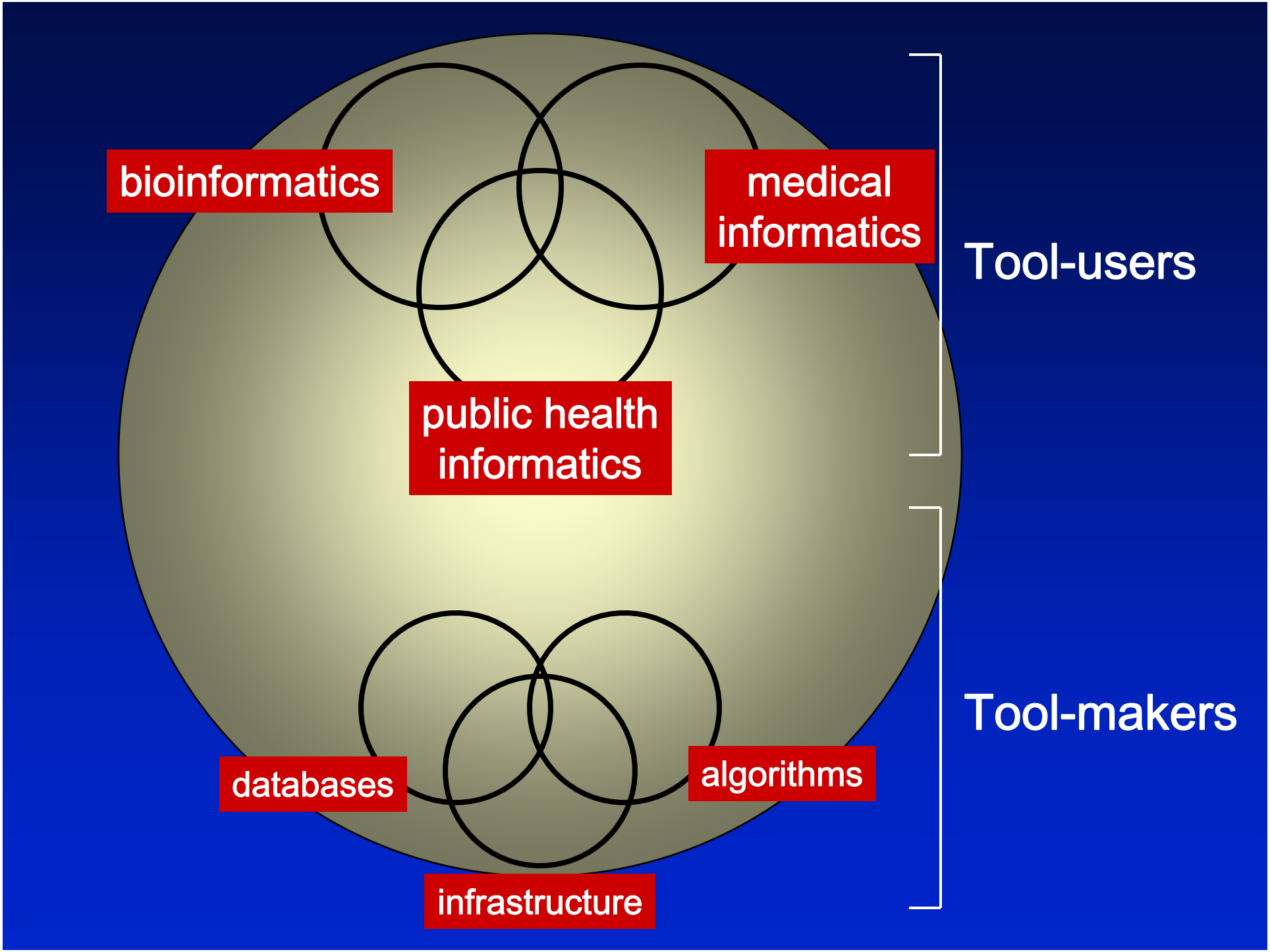
Chr2 abnormality

Beckwith-Wiedemann

pseudohypoparathyroidism

PediSNP: identifying meiotic crossovers in SNP data




bioinformatics
medical informatics
public health informatics
databases
algorithms
infrastructure
Tool-users
Tool-makers

Conclusions: barriers and opportunities

Bioinformatics is a multidisciplinary field that serves biology and medicine. There is a need for computer programmers, biologists, clinicians, and biostatisticians.

For both research and healthcare applications, understanding the relationship between genotype and phenotype presents great challenges.

-- how to visualize large data sets
-- how to find relevant relationships
-- how to facilitate interoperability, extensibility, growth