

# Microarray Data Analysis

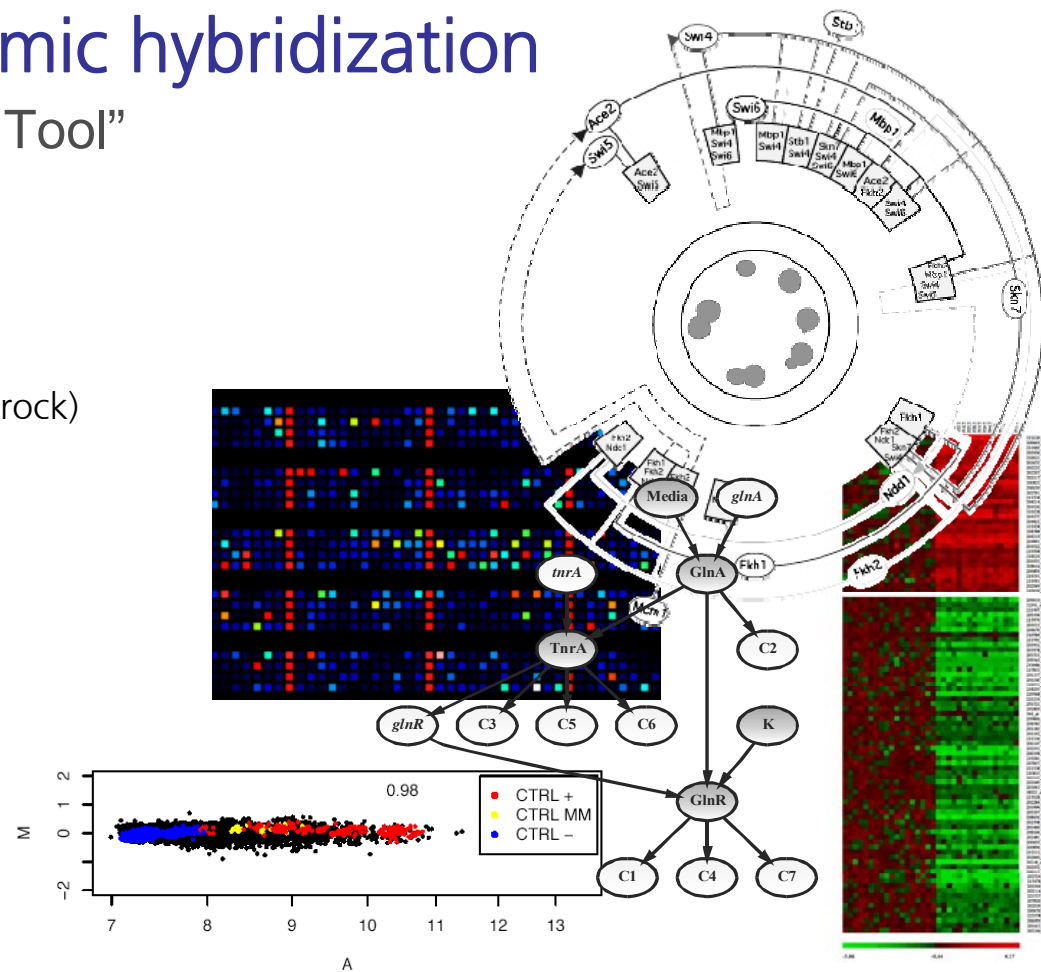
—Course 27621, DTU 2008

## Comparative genomic hybridization

“Poor Man’s Sequencing Tool”

*Carsten Friis*

(...with several slides from H. Willenbrock)



# Outline

---

Introduction to comparative genomic hybridization (CGH) and array CGH

Data analysis approaches

- Breakpoint detection
- Loss and gain analysis

Real data example: Comparative genomic profiling of bacterial strains

---

Introduction to comparative genomic hybridization (CGH) and array CGH

Data analysis approaches

- Breakpoint detection
- Loss and gain analysis

Real data example: Comparative genomic profiling of bacterial strains

---

# Comparative Genomic Hybridization

---

## Study types :

- Gain or loss of genetic material
- To find variations in the genetic material

## Purposes:

- Study of chromosomal aberrations often found in cancer and developmental abnormalities
- Study of variations in the baseline sequence in a microbial population (microbial comparative genomics)

# Genetic Alterations and Disease

---

A Variety of Genetic Alterations Underlie Developmental Abnormalities and Disease

Inappropriate gene activation or inactivation can be caused by:

- Mutation
- Epigenetic gene silencing (e.g. addition of methyl groups)
- Reciprocal translocation (exchange of fragments between two non-homologous chromosomes)
- **Gain or loss of genetic material**

*Any of the above may lead to an oncogene activation or to inactivation of a tumor suppressor*

---

# Microarrays for copy number analysis

---

BAC arrays

Affymetrix SNP chip (500 K)

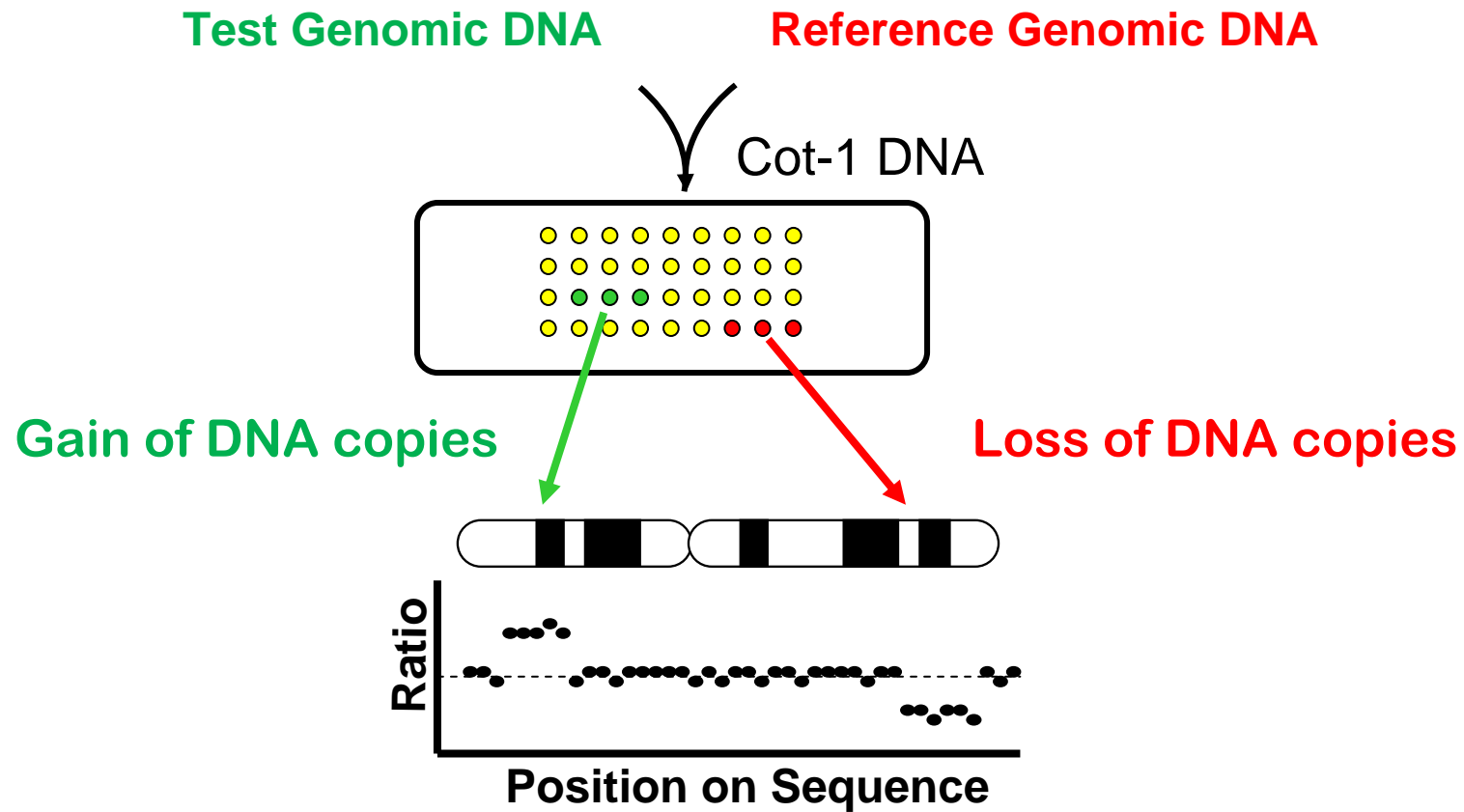
Representational oligonucleotide microarray analysis  
(ROMA)

Whole genome tiling arrays

Own design (NimbleGen/NimbleExpress)

---

## Array CGH Maps DNA Copy Number Alterations to Positions in the Genome



# Advantages over Expression Arrays

---

Hybridization of DNA to microarray (DNA is much more stable)

Little normalization is necessary

Use of spatial coherence in the analysis

Only 1 sample is necessary to draw conclusions

- it is still necessary with biological replicates to be able to draw general conclusions regarding a certain biological subtype

Results may be easier interpretable and correlated with sample phenotypes

---

# Outline

---

Introduction to comparative genomic hybridization (CGH) and array CGH

## Data analysis approaches

- Breakpoint detection
- Loss and gain analysis

Real data example: Comparative genomic profiling of bacterial strains

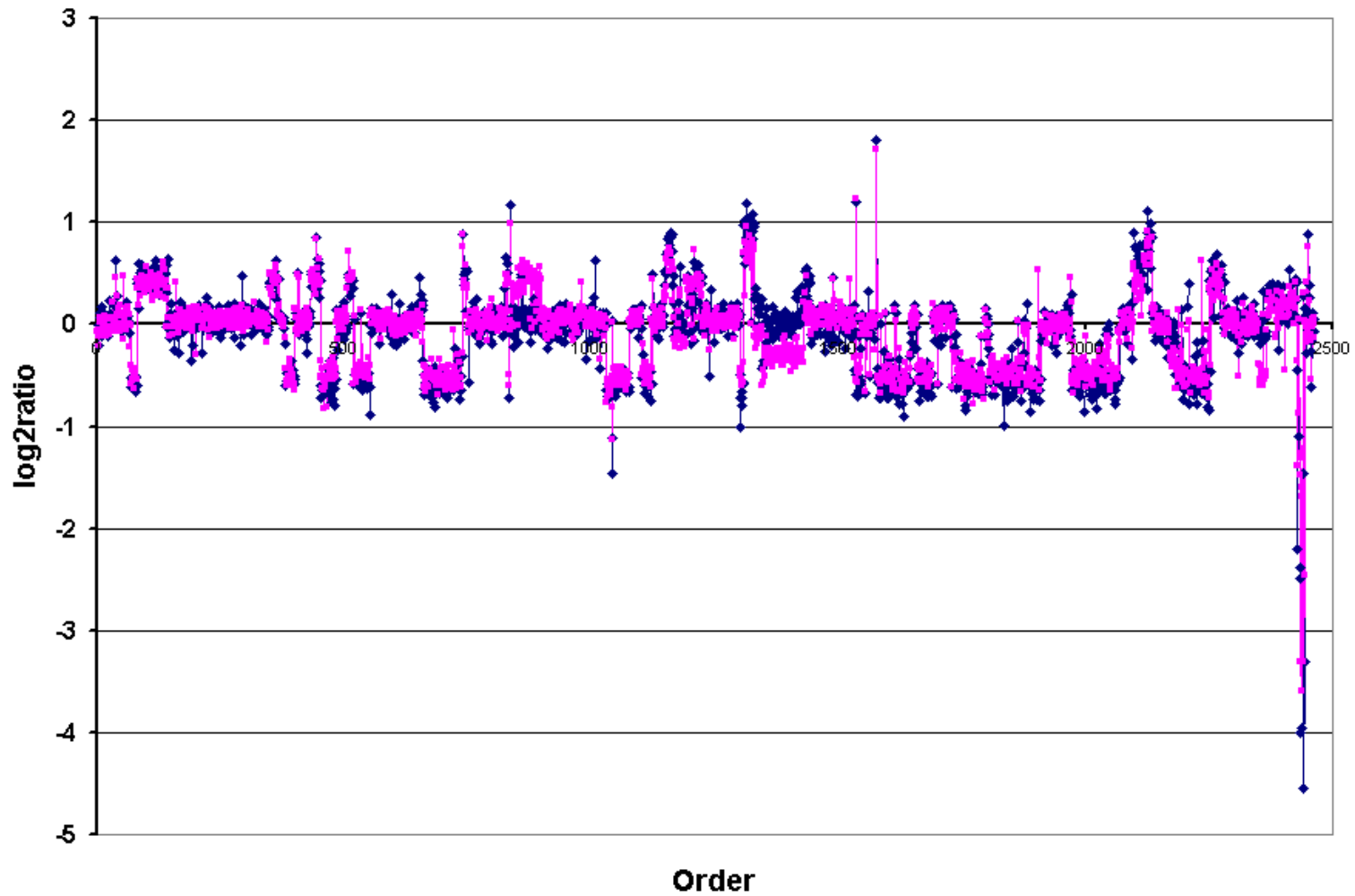
---

**Goal:** To partition the clones into sets with the same copy number and to characterize the genomic segments in terms of copy number.

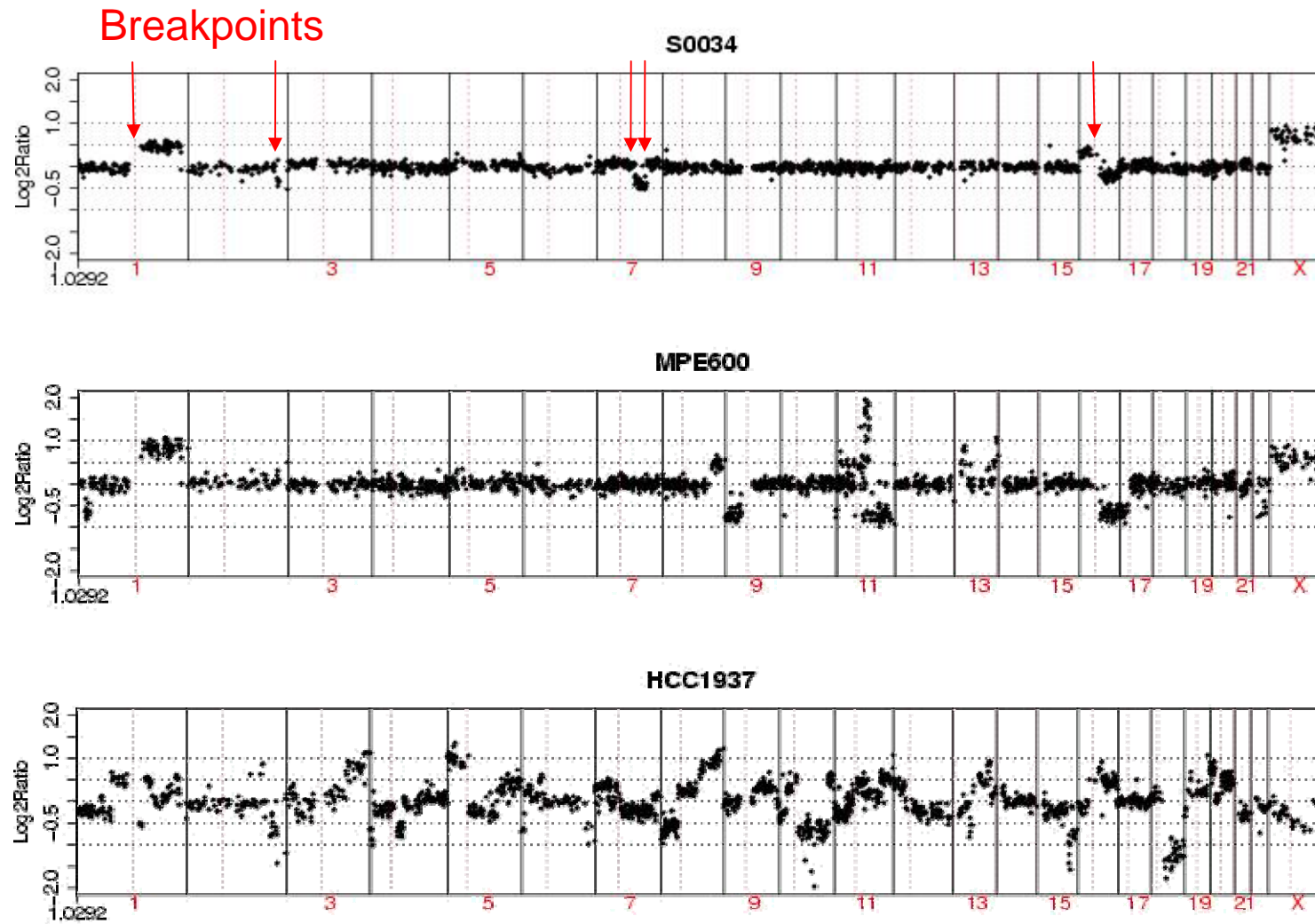
***Biological model:*** genomic rearrangements lead to gains or losses

- Sizable contiguous parts of the genome, possibly spanning entire chromosomes
  - Or, alternatively, to focal high-level amplifications
-

# Copy Number Profiles of a Tiling Array

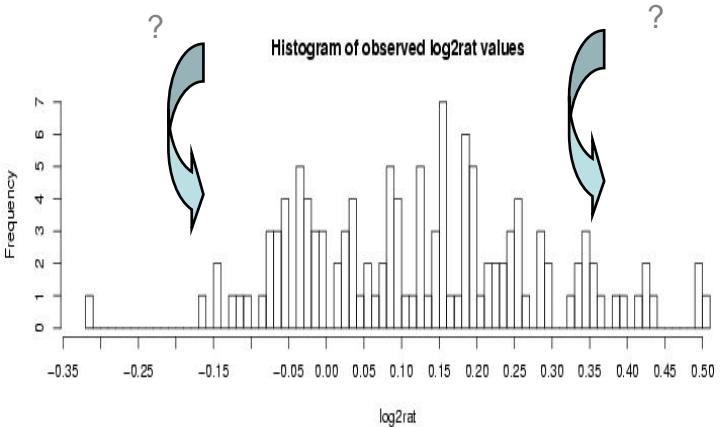
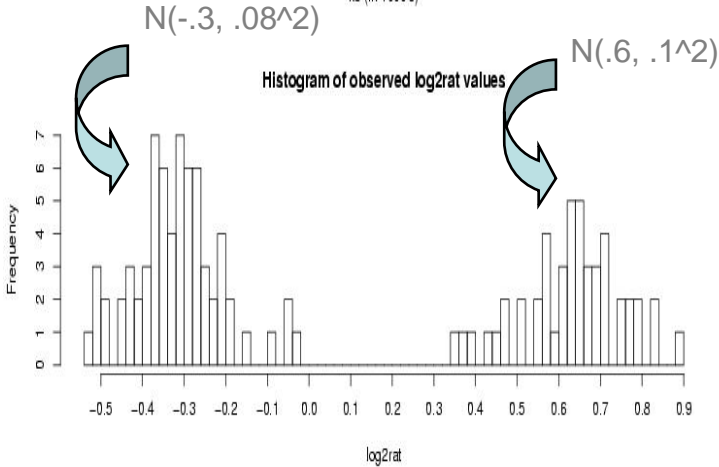
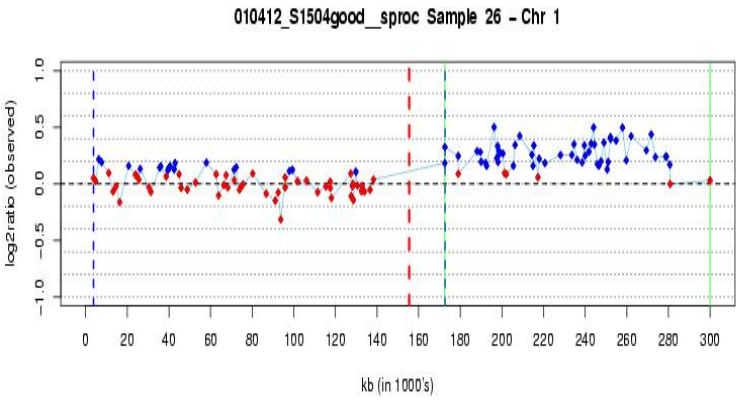
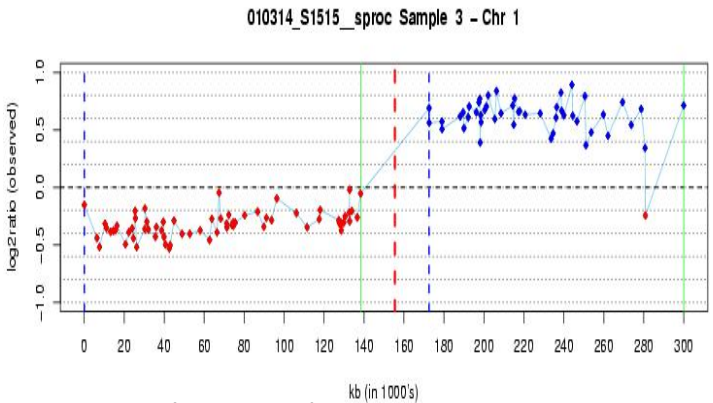


# Varying genomic complexity



# Observed clone value and spatial coherence

Useful to make use of the physical dependence of the nearby clones, which translates into copy number dependence



## **HMM:** Hidden Markov Model (aCGH package)

- Fit HMMs in which any state is reachable from any other state (Fridlyand et al, JMVA, 2004).

## **CBS:** Circular binary segmentation (DNAcopy package)

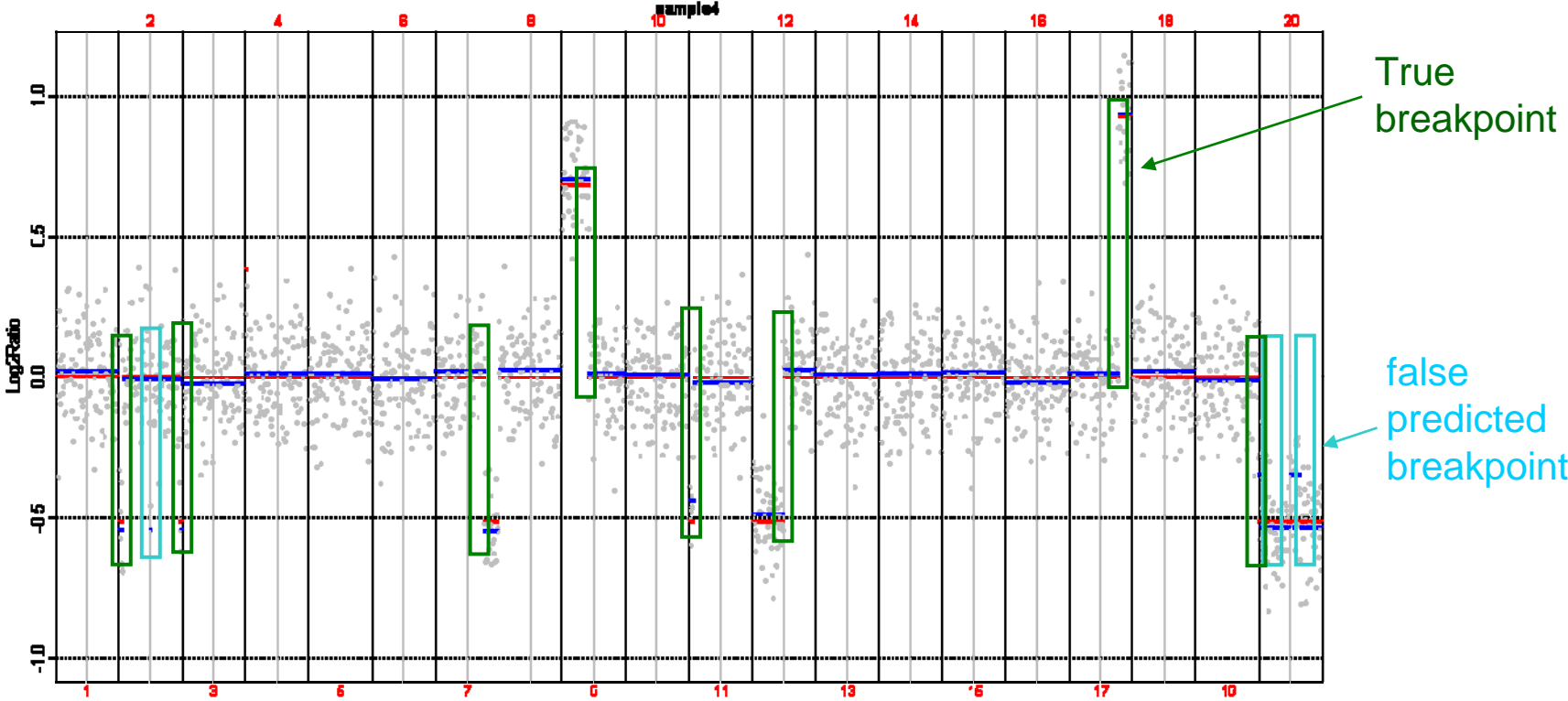
- Tertiary splits of the chromosomes into contiguous regions of equal copy number and assesses significance of the proposed splits by using a permutation reference distribution (Olshen et al, Biostatistics, 2004).

## **GLAD:** Gain and Loss Analysis of DNA (GLAD package)

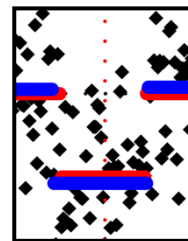
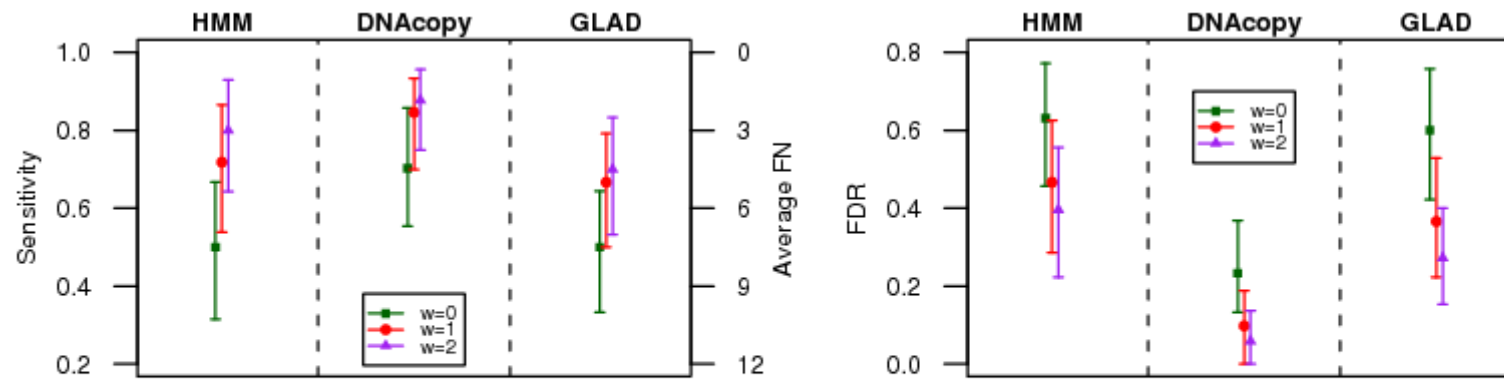
- Detects chromosomal breakpoints by estimating a piecewise constant function that is based on adaptive weights smoothing (Hupe et al, Bioinformatics, 2004).

# Comparison Scheme

Use of simulated data, where the truth is known and the noise is controlled



# Breakpoint Detection Accuracy



# Conclusions so far

---

## Signal2noise:

- **CBS** consistently the best performance
  - **HMM** has the highest False Discovery Rate
  - **GLAD** is least sensitive
-

# Outline

---

Introduction to comparative genomic hybridization (CGH) and array CGH

## Data analysis approaches

- Breakpoint detection
- Loss and gain analysis
- Application of segmentation to testing

Real data example: Comparative genomic profiling of bacterial strains

---

# Merging segments

---

**Note:** that all procedures operate on individual chromosomes, therefore resulting in a large number of segments with mean values close to each other

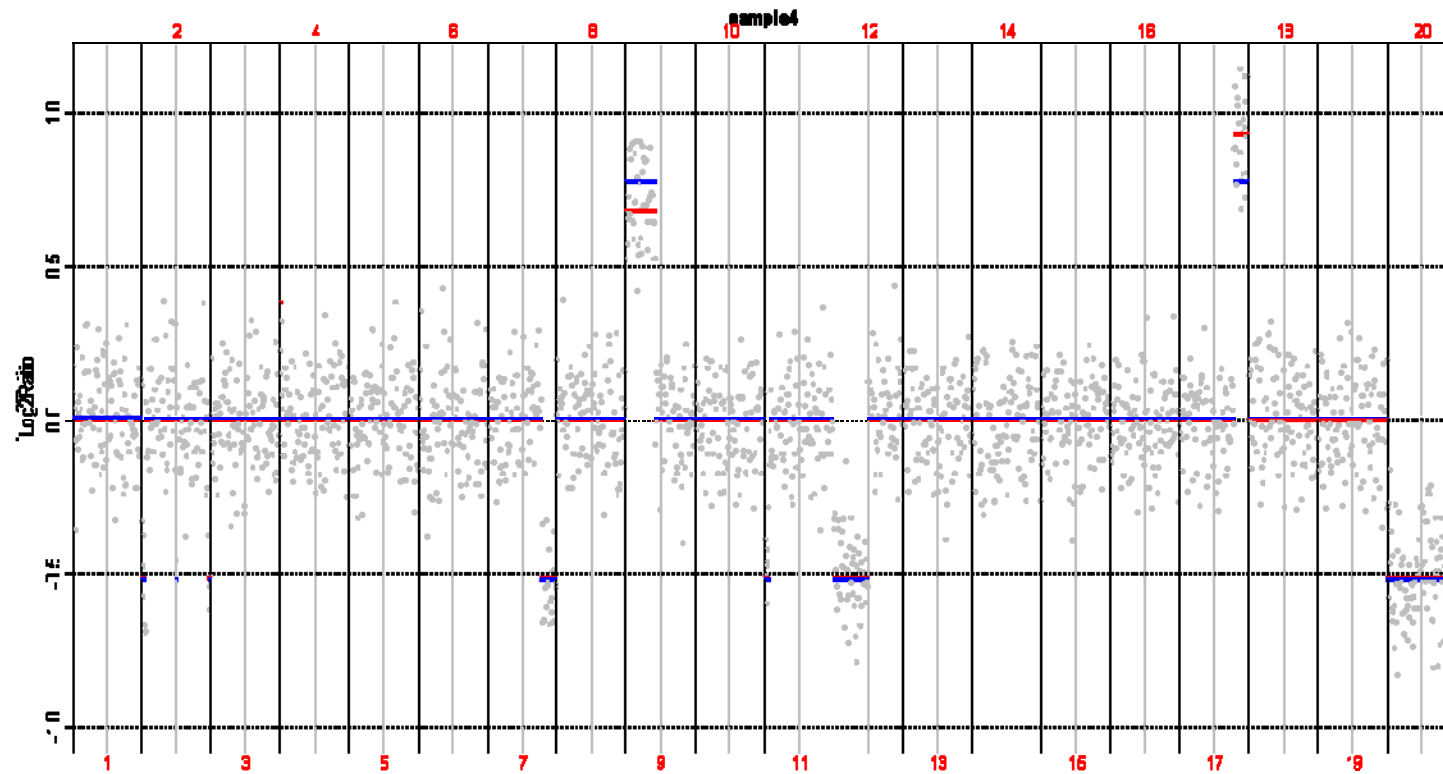
*Additional Challenge:* reduce number of segments by merging the ones that are likely to correspond to the same copy number

This will facilitate inference of altered regions

---

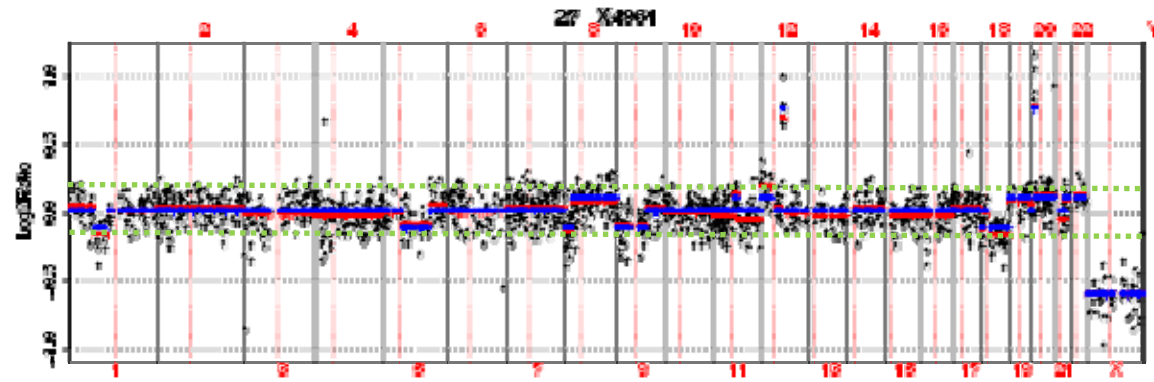
# Merging

For estimating actual copy number levels from segmentations

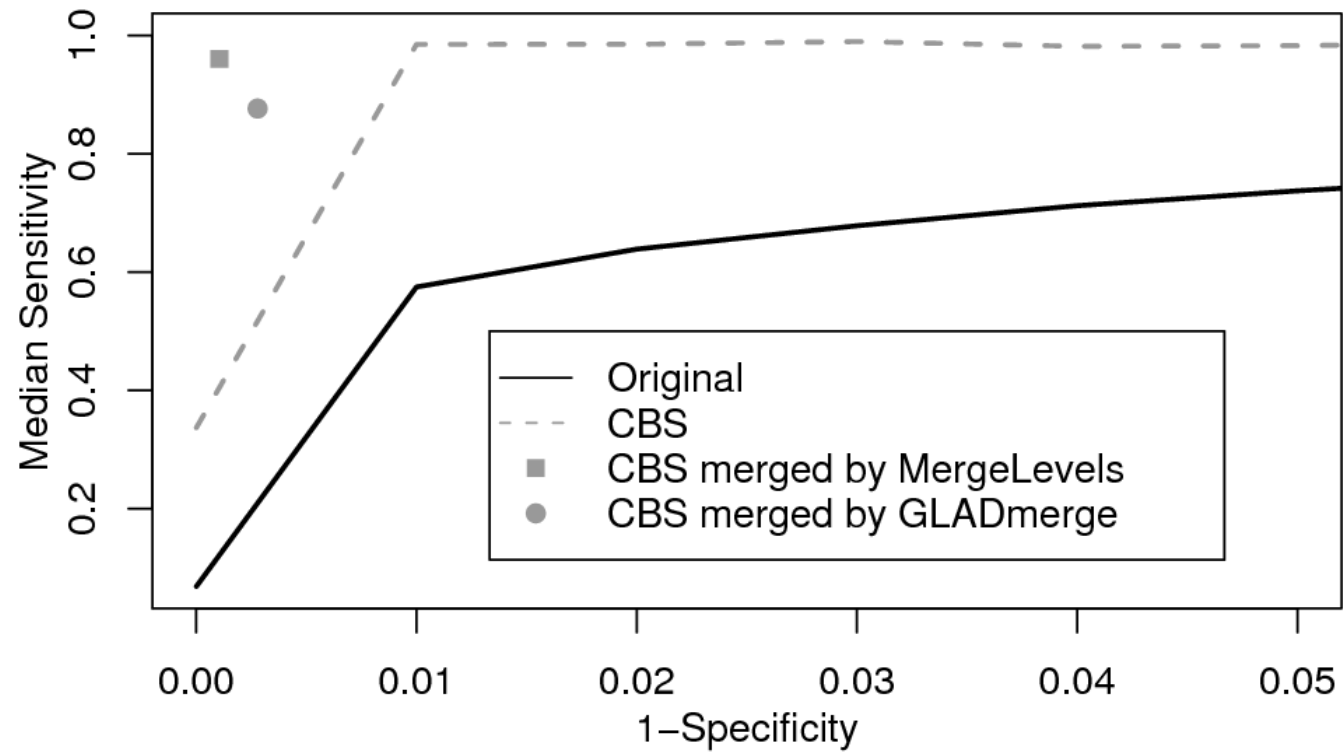


# Segmentation and Merging

---



## Identification of copy number alterations for varying thresholds



# Outline

---

Introduction to comparative genomic hybridization (CGH) and array CGH

Data analysis approaches

- Breakpoint detection
- Loss and gain analysis

**Real data example: Comparative genomic profiling of bacterial strains**

---

## Real Data Example:

---

Comparative genomic profiling of several *Escherichia coli* strains

The microarray design included probes for:

- 7 known *E. coli* strains
- 39 known *E. coli* bacteriophages
- 104 known *E. coli* virulence genes



### Experimentally:

- 2 sequenced control strains (W3110 and EDL933), 3 replicates
- 2 non-sequenced strains (D1 and 3538), 3 replicates
- Bacteriophage:  $\phi$ 3538 ( $\Delta$ *stx2::cat*), 2 replicates

**Ratio problems:** some genes might be present on query strain but not on the known reference strain

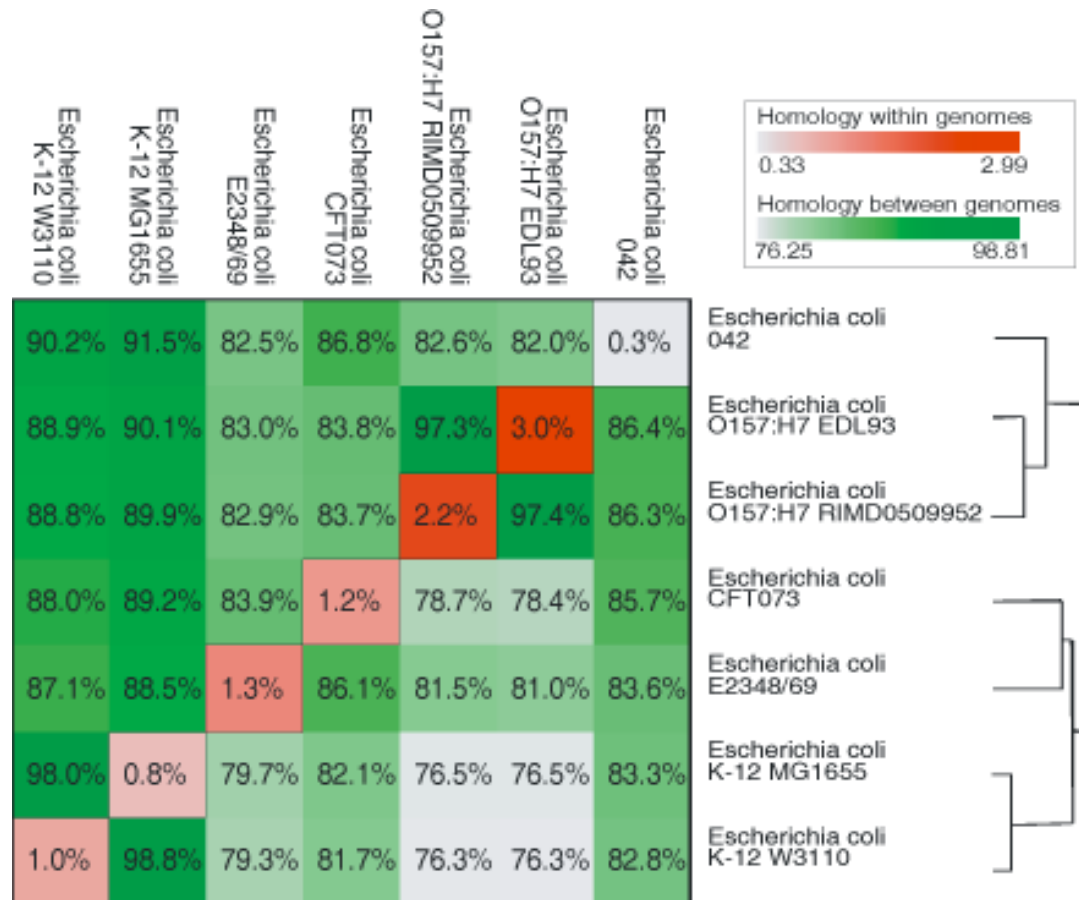
Single channel microarrays or dual channel microarrays?

- In this case, we used an Affymetrix single channel custom-made array (**NimbleExpress**)

**Partly present** genes versus **similar but distinct** genes

---

# The 7 *E. coli* strains included on the microarray

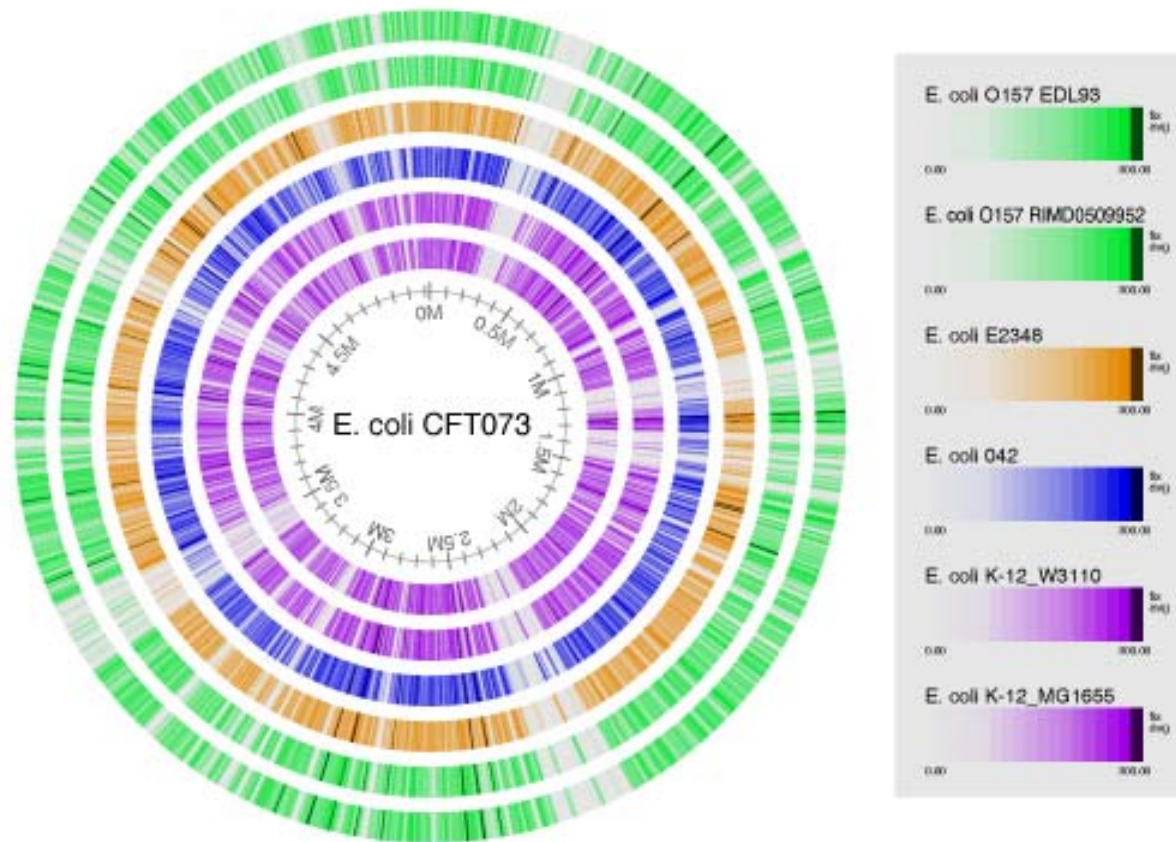


Very high similarity between the two K-12 strains and between the two O157:H7 strains.

Percentage of homologues for *E. coli* genomes in columns found in *E. coli* genomes in rows.

Willenbrock et al. *Journal of Bacteriology*. 2006 Nov;188(22):7713-21.

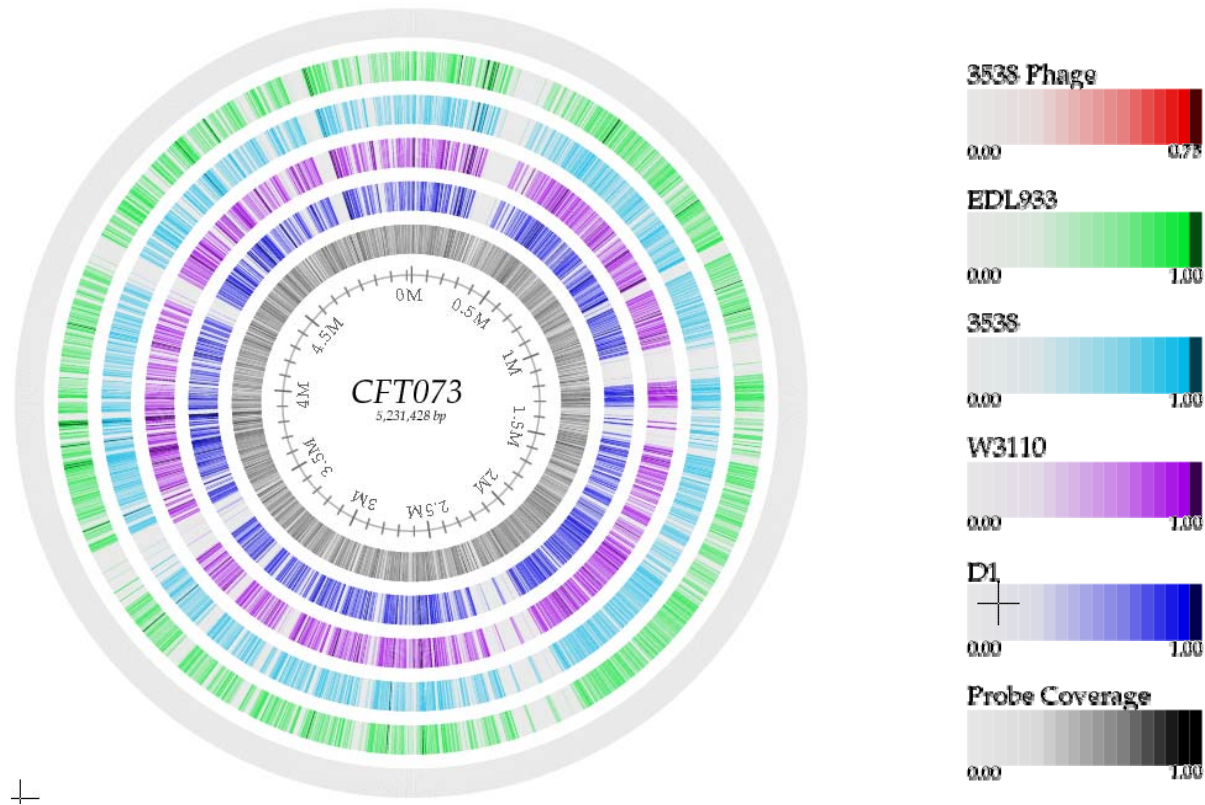
# BLAST Atlas



Willenbrock et al. *Journal of Bacteriology*. 2006 Nov;188(22):7713-21.

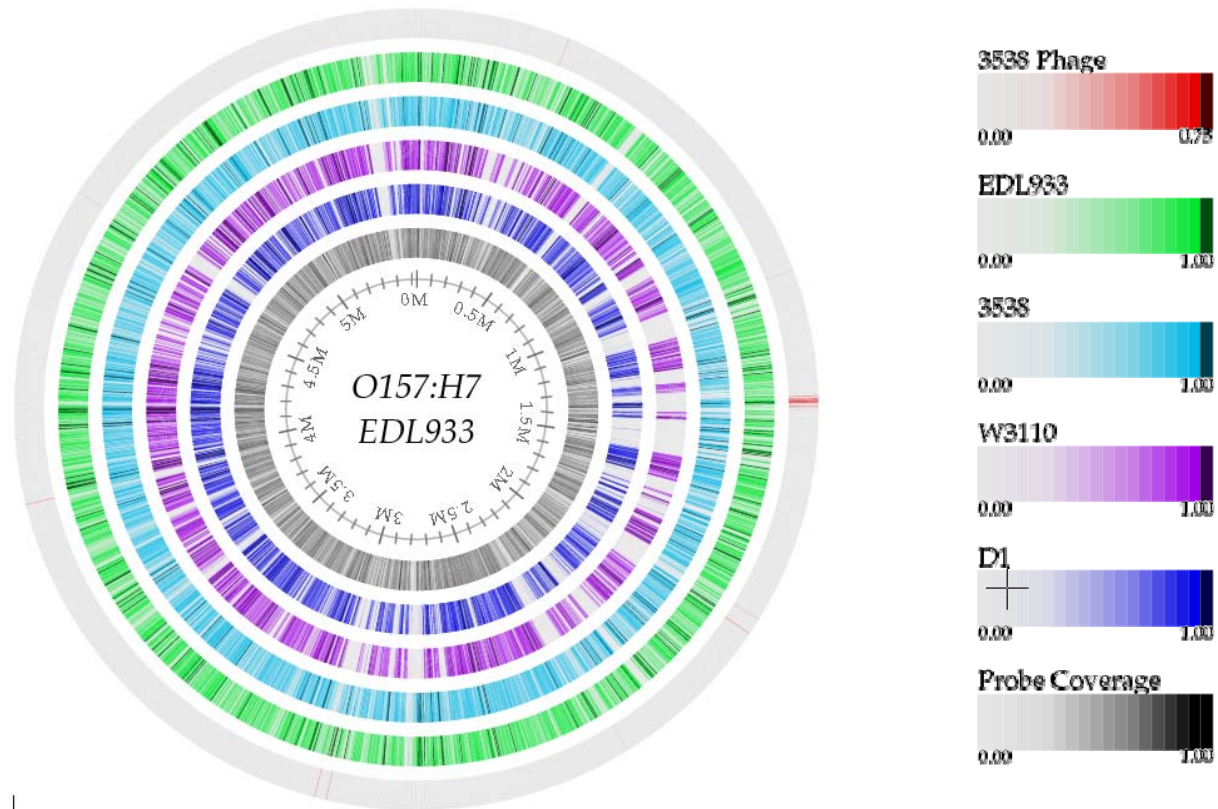
# Hybridization Atlases

Probe hybridizations for experiments (samples) result in a similar pattern as expected from the BLAST atlas



Willenbrock et al. *Journal of Bacteriology*. 2006 Nov;188(22):7713-21.

# Mapping the phage $\Phi$ 3538 ( $\Delta$ stx2::*cat*)

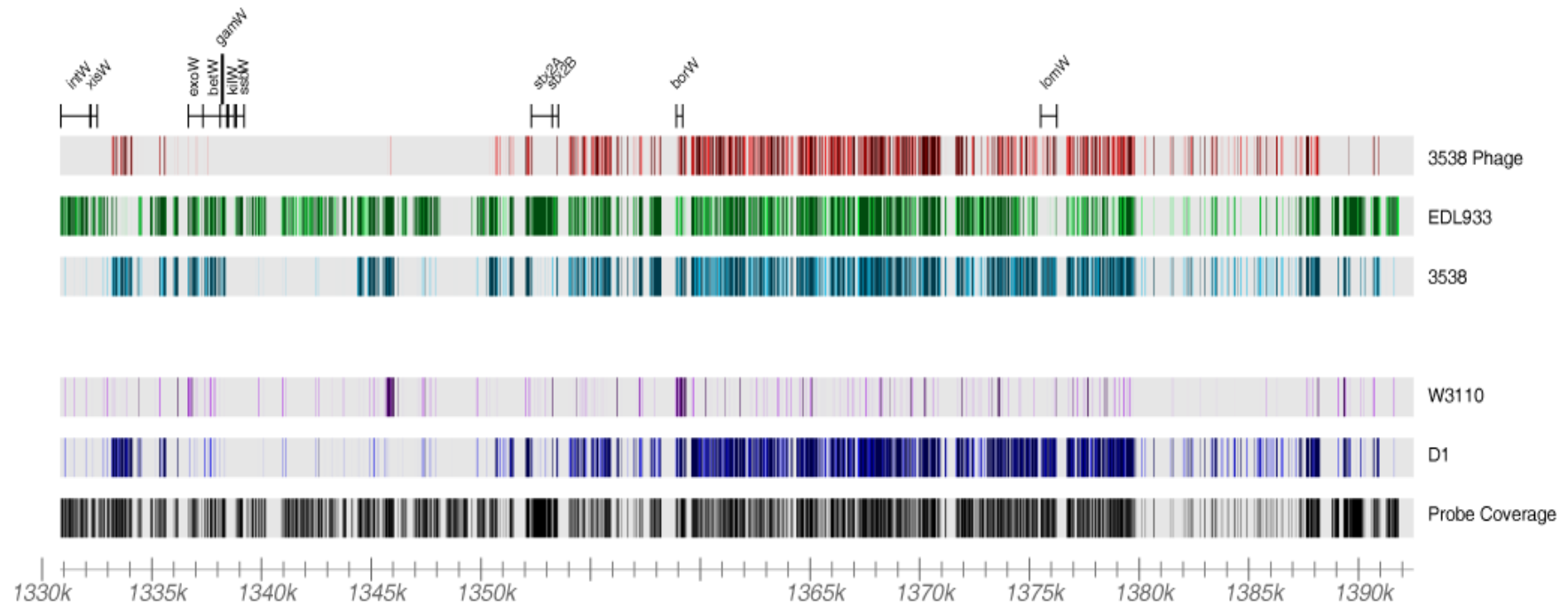


Willenbrock et al. *Journal of Bacteriology*. 2006 Nov;188(22):7713-21.

# Zoom of phage $\Phi$ 3538 ( $\Delta stx2::cat$ )

The hybridization pattern is very similar for the

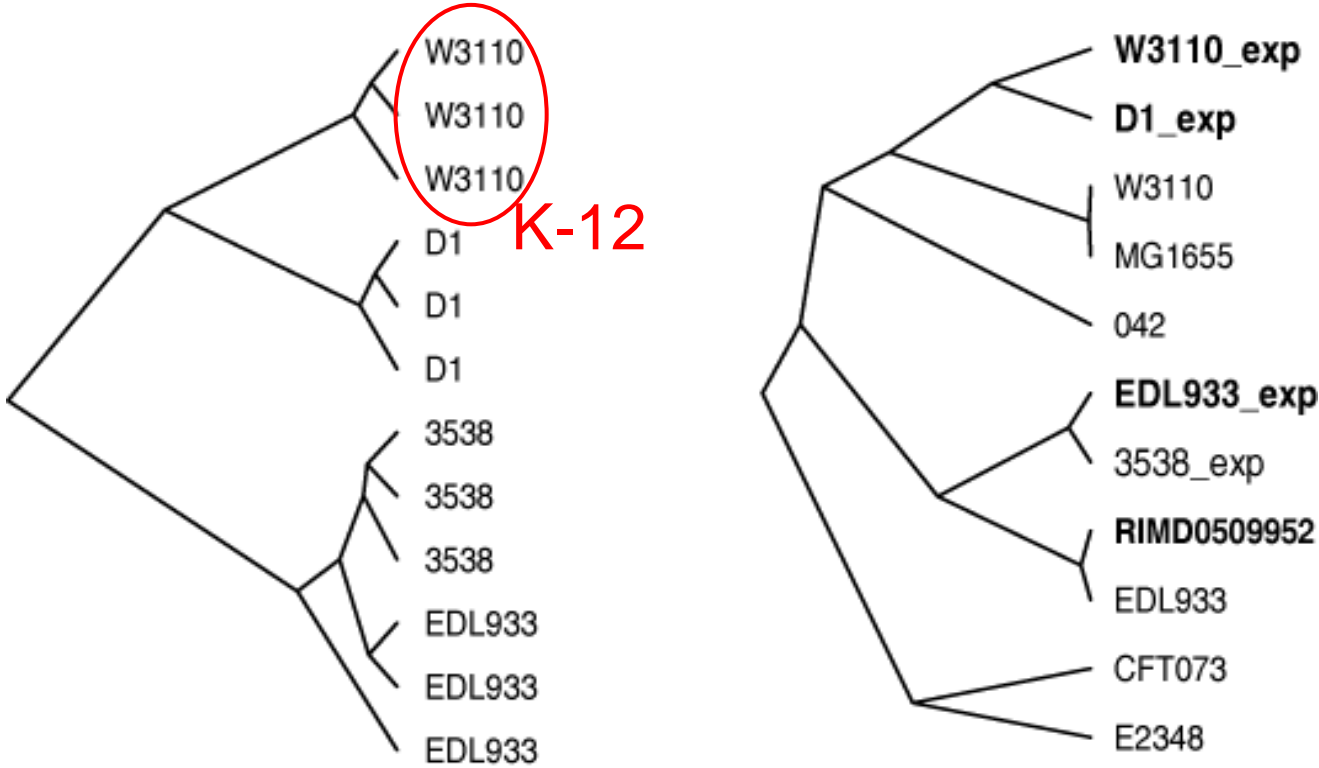
*E. coli* O157:H7 EDL933 (1.33 Mbases - 1.39 Mbases)



Willenbrock et al. *Journal of Bacteriology*. 2006 Nov;188(22):7713-21.

# Hierarchical Cluster Analysis

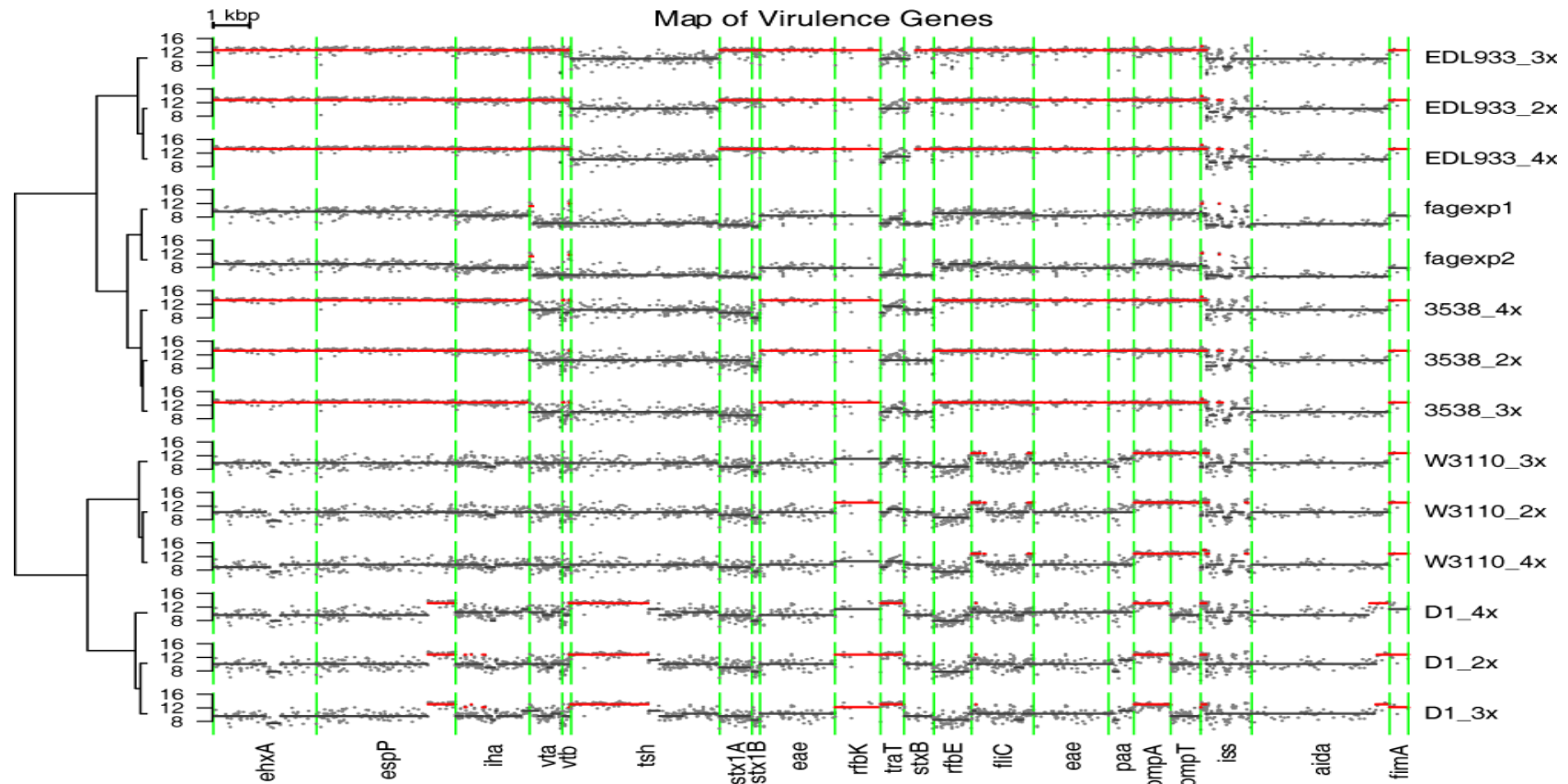
D1 is very similar to the K-12 type strains  
(W3110 + MG1655)



# E. coli virulence genes

D1 is probably still a commensal strain

- An organism participating in a symbiotic relationship from which it benefits while the other is unaffected



# Summary

---

Comparative genomic profiling of two E. coli strains

- 0175:H16 D1
- 0157:H7 3538

Identification of virulence genes and phage elements

## Conclusions:

D1 is similar to the K-12 type strains

Characterization of D1 and 3538 genes:

- Identification of a number of genes involved in DNA transfer and recombination

# Summary

---

Numerous methods have been introduced for segmentation of DNA copy number data and breakpoint identification.

- Important to benchmark against existing methods
- (however, only feasible if the software is publicly available)

Currently, CBS (DNACopy package) has the best overall performance

Merging of segmentation results improves copy number phenotype characterization

Study types:

- Study of copy number in cancer samples
  - Comparison of bacterial strains
  - Etc.
-