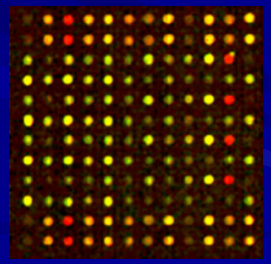
snapCGH: An R package for segmentation, normalization and processing of array CGH data

Mike Smith, John Marioni, Natalie Thorne, Simon Tavaré

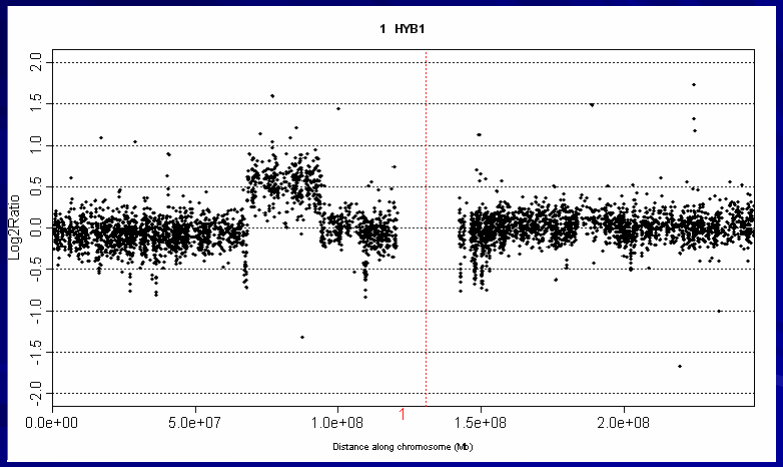
Department of Oncology
University of Cambridge

What is array CGH

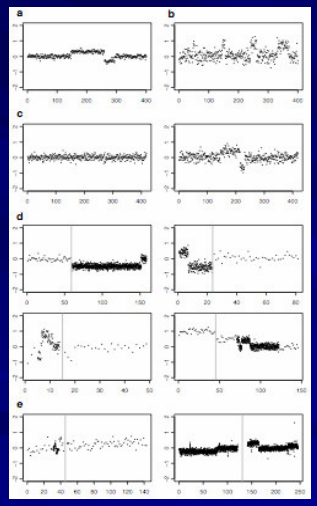
- Stands for array comparative genomic hybridization
- test and reference DNA (e.g. DNA from a tumor and DNA from a normal genome) are differentially labelled with fluorochromes (e.g. Cy3 and Cy5) and hybridized to the microarray.
- Fluorescence ratio of the test vs. reference intensities at each spot on the array indicate the relative copy number

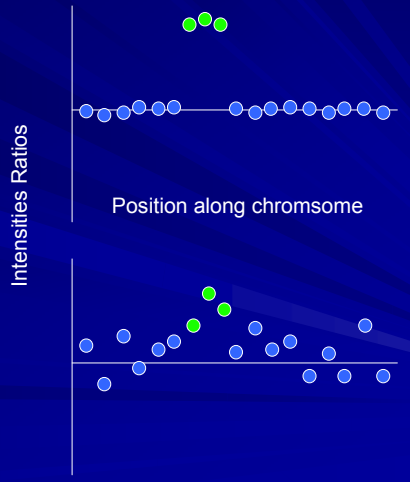


Example of aCGH data



Segmentation of aCGH data





Why segment the data?

- Noise reduction
- Detection of lost and gained regions of a chromosome
- Breakpoint analysis

Recurrent aberrations across samples may indicate:
- an oncogene or
- a tumor suppressor gene

Diversity of segmentation algorithms

- Currently many different segmentation algorithms available in a variety of platforms (R, Java, C)
- Even within R there is disparity in the format of input data for different methods
- This is partly responsible for minimal research comparisons between algorithms

snapCGH

- Designed to be compatible with the popular R package *limma* (S3)
- This allows smooth continuation between pre-processing, normalization and the segmentation step
- Potentially reduces learning curve for anyone already familiar with *limma* e.g. biologists

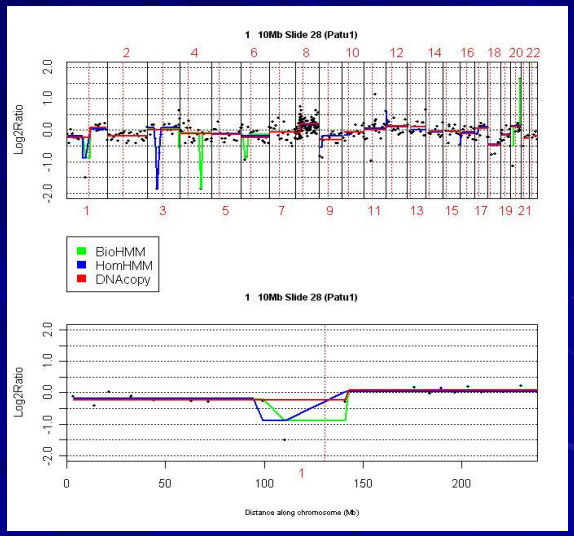
Software Wrappers

- snapCGH incorporates software wrappers for several segmentation algorithms available via the Bioconductor project.
- These include: aCGH, DNACopy, GLAD, tilingArray
- Provides a common class for output, the *SegList*, allowing easy comparison between methods

Comparison of methods

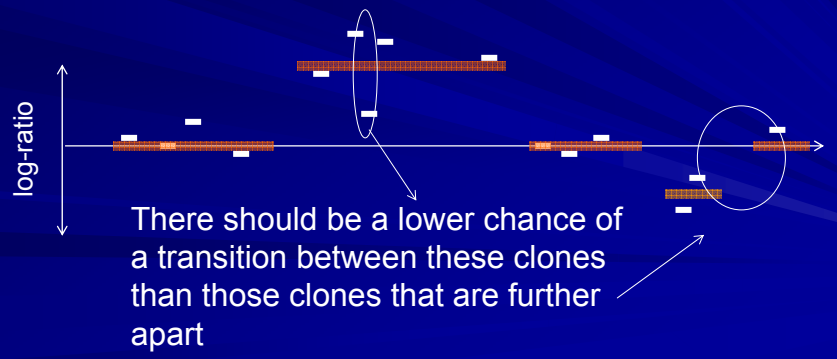
Example of three segmentation methods performed on the same Sample

Interactive allowing the user to zoom in on a particular chromosome



BioHMM

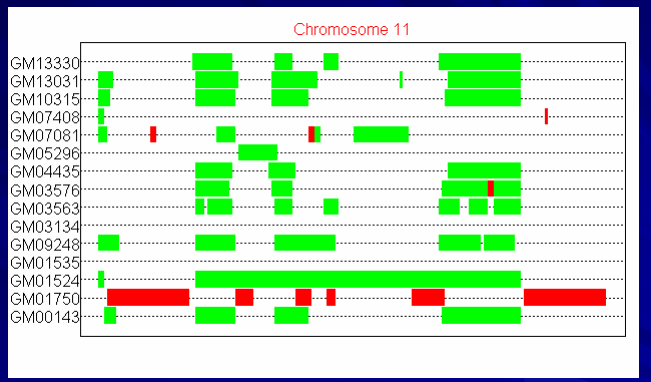
- BioHMM is a heterogeneous Hidden Markov model
- It incorporates the distance between clones



Simulation

- Simulated data is often used to assess the effectiveness of segmentation schemes
- None currently take into account the spatial nature of aCGH data
- Generates the physical location of clones along a chromosome before generating the corresponding intensity ratios
- In particular, differentiate between tiled and non-tiled regions, and different technologies

Cross Sample Analysis



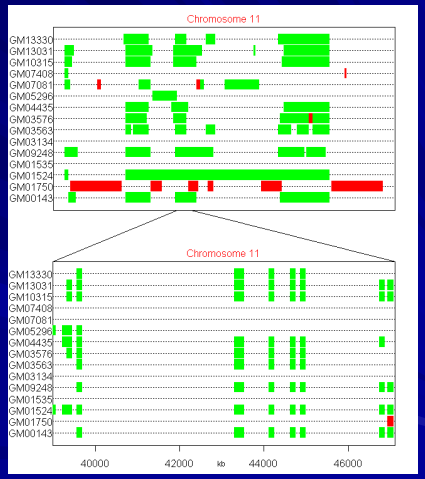
Displaying regions of gain or loss across multiple samples

Hope to find :

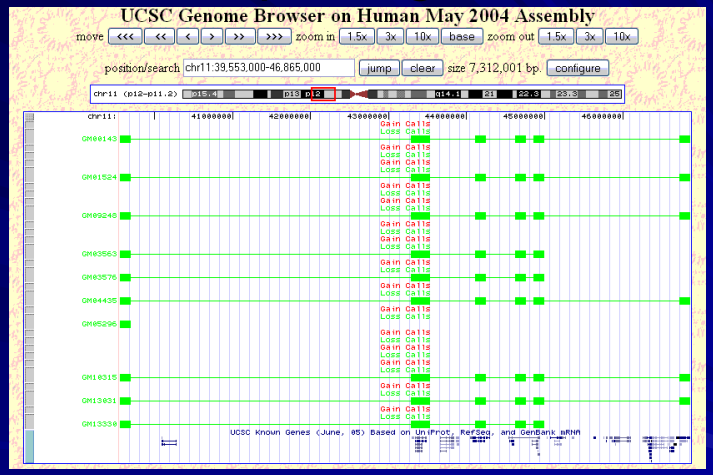
- Common breakpoints
- Minimal regions (regions of common gain or loss between samples)

Functionality

- Originally made by colouring cells in Excel manually!
- Interactive allowing user to zoom in on regions of interest



Exporting Data



- Can output information as a Gene Feature Format (GFF) file.
- Can then be imported into online databases such as the UCSC genome browser

Future Plans

- Develop methods to import data from external sources e.g. connect to the UCSC Genome Browser database
- Incorporate algorithms for finding minimal regions, rather than the interactive graphical approach currently available in snapCGH

Acknowledgements

Department of Oncology, University of Cambridge:

Paul Edwards
Jessica Pole

Carlo Caldas
Maria Garcia

The Welcome Trust Sanger Institute:
Nigel Carter
Heike Fiegler

University of California at San Francisco:
Jane Fridlyand