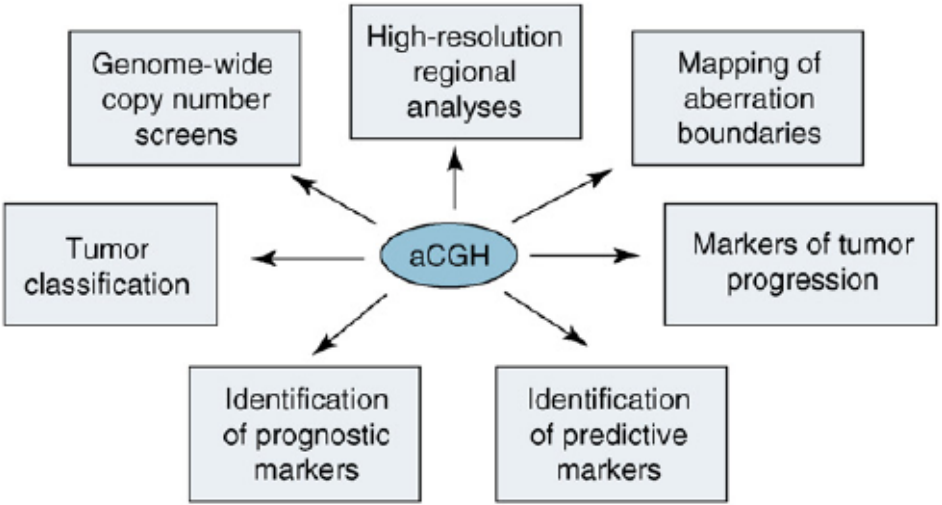
Array CGH

Motivation

- A variety of cancers (and other genetic diseases) are characterized by DNA sequence copy-number variations
- Deletions of tumor suppressors or duplication of oncogenes
- Array CGH is one method to detect copy-number variations throughout the genome
- Theoretical resolution: 1kbase

Uses of aCGH



Curren Opinion in Biotechnology

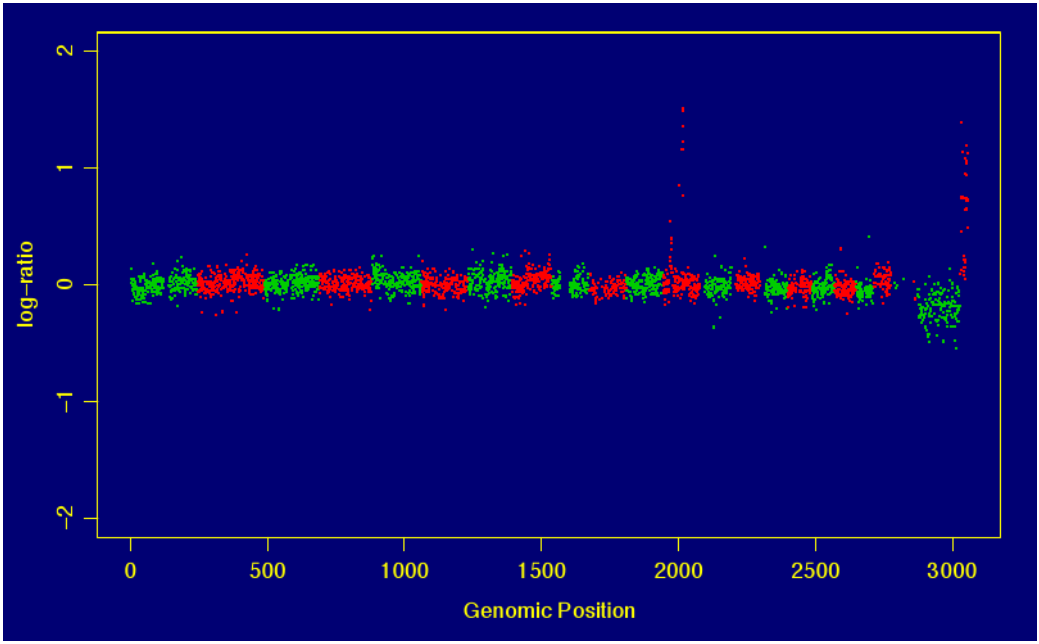
Comparative Genomic Hybridization (without arrays)

- Metaphase chromosomes from tumor and gender-matched normal DNA isolated
- labeled with Cy3 and Cy5 (for example)
- imaged separately w/ fluorescent microscope
- fluorescence quantified by imaging software
- large relative difference in fluorescence indicates DNA gain or loss

CGH vs aCGH

- instead of quantification by fluorescent microscopy, DNA hybridized to microarray probes
- **different kinds of probes:**
    - BACs, long oligos, short oligos (SNP arrays)
- tradeoff between resolution and noise
- could use one- or two-color arrays
- Result: copy number log-ratio between case and control

aCGH example results



Computational segmentation and CNV calling

- "Eyeballing" CNVs is okay...but we would like more statistically rigorous methods to distinguish real CNVs from noise
- **3 steps:**
    - **normalization** - center logFC at 0
    - **segmentation** - DNAcopy
    - **calling** - CGHcall

Computational aCGH segmentation

- Goal: identify "change-points" which will divide each chromosome into segments with equal copy number
- **Lots of approaches exist:**
    - mixture modeling
    - 3-means clustering & dynamic programming
    - Markov models
    - **circularized binary segmentation** (DNAcopy)

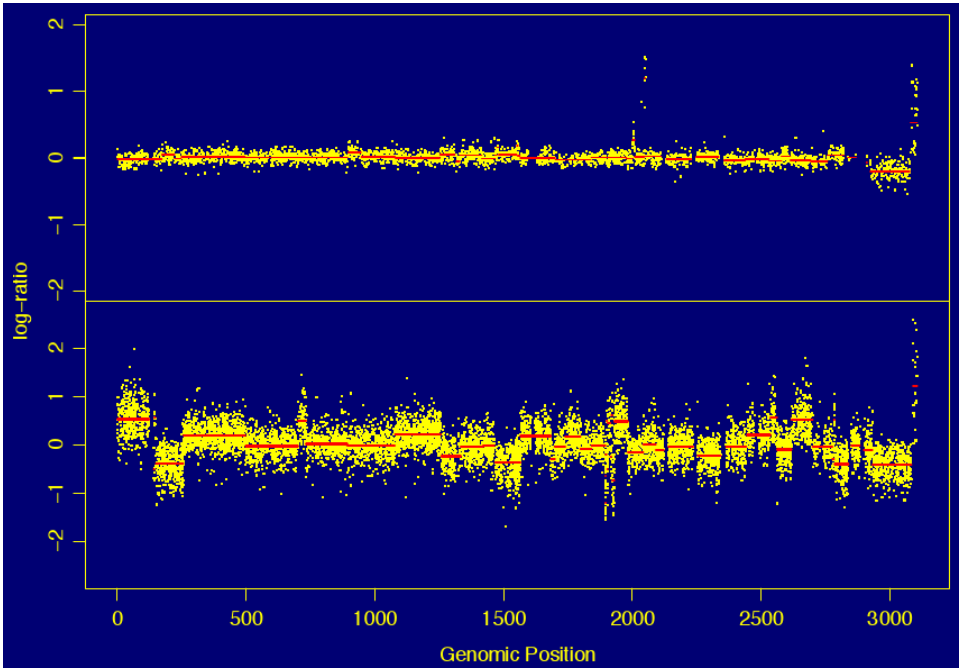
DNAcopy and the binary segmentation algorithm

- A t-like-statistic (Venkatraman & Olshen, 2006) can test the "left" and "right" sides of a potential change-point for significant difference
- **Algorithm:**
    - Begin with a large segment (chromosome).
    - Starting from the left, test all potential change-points (midpoints between probes).
    - If a significant point is found, recurse with left and right segments.

DNAcopy particulars

- This algorithm has a tunable parameter, alpha, which is the FPR. A higher FPR means more segments will probably be found.
- **Circularized** binary segmentation treats the chromosome as a circle rather than as a line, and allows detection of two change-points simultaneously. For reasons we won't digress into, this increases accuracy.
- CBS performed well (top 2 of 10) in a comparison of segmentation algorithms (Lai, 2005)

DNAcopy example segmentation



CGHcall

- Segmentation algorithms do not actually call copy number, they call any *difference from normal*.
- **Distinct copy number states of interest:**
    - Double and single deletion (0 and 1 copies)
    - Normal (2)
    - Single and double gain (3 and 4 copies)
    - Amplification (5+ copies)
- CGHcall wraps DNAcopy and provides both segmentations and calls

CGHcall algorithm

CGHcall creates a Gaussian mixture model of 6 states of interest (2 deletions, normal, 2 gain, amplification).

- Assumption: within-segment variance is uniform.

GMM parameters learned by expectation-maximization, and call is made by maximum likelihood.