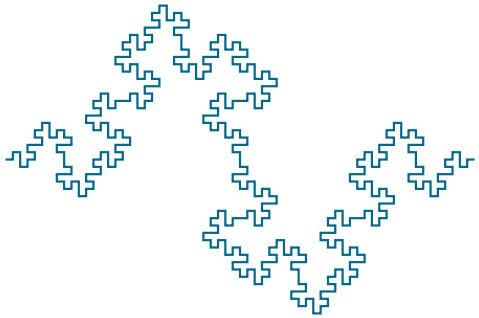
Estimating copy number polymorphisms from genotyping arrays

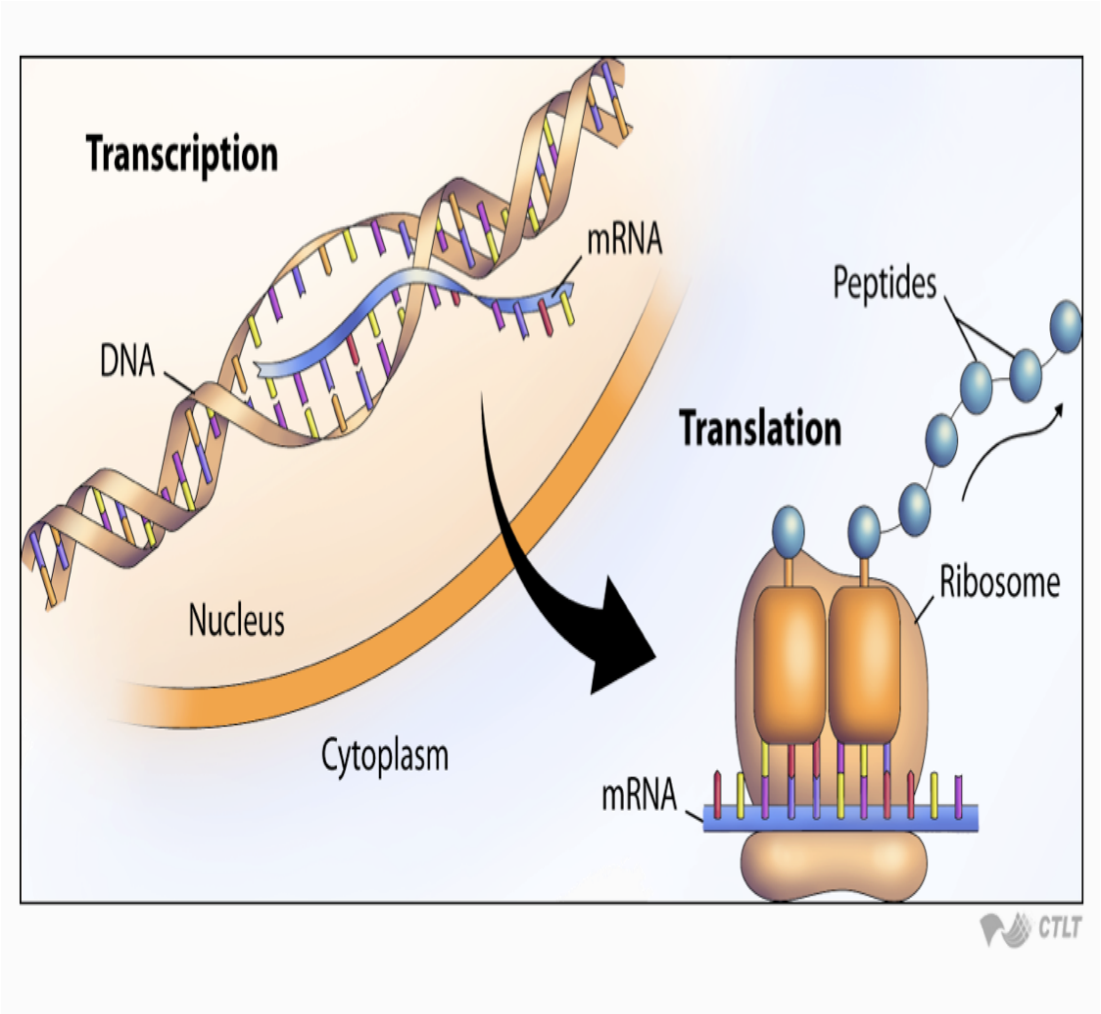
Stephen Cristiano
Johns Hopkins University



November 5, 2013

Introduction

CENTRAL DOGMA OF MOLECULAR BIOLOGY



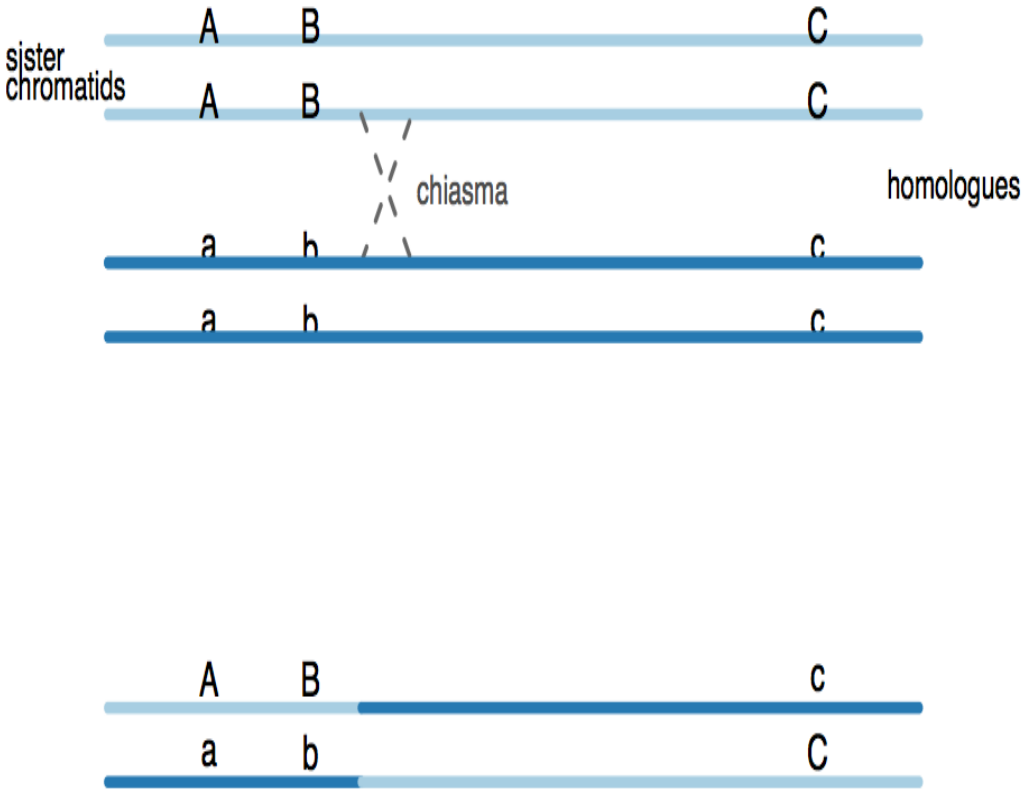
COPY NUMBER VARIATION

A loss or gain of chromosomal DNA copy number spanning hundreds to thousands of basepairs, or even entire chromosomes (aneuploidy)

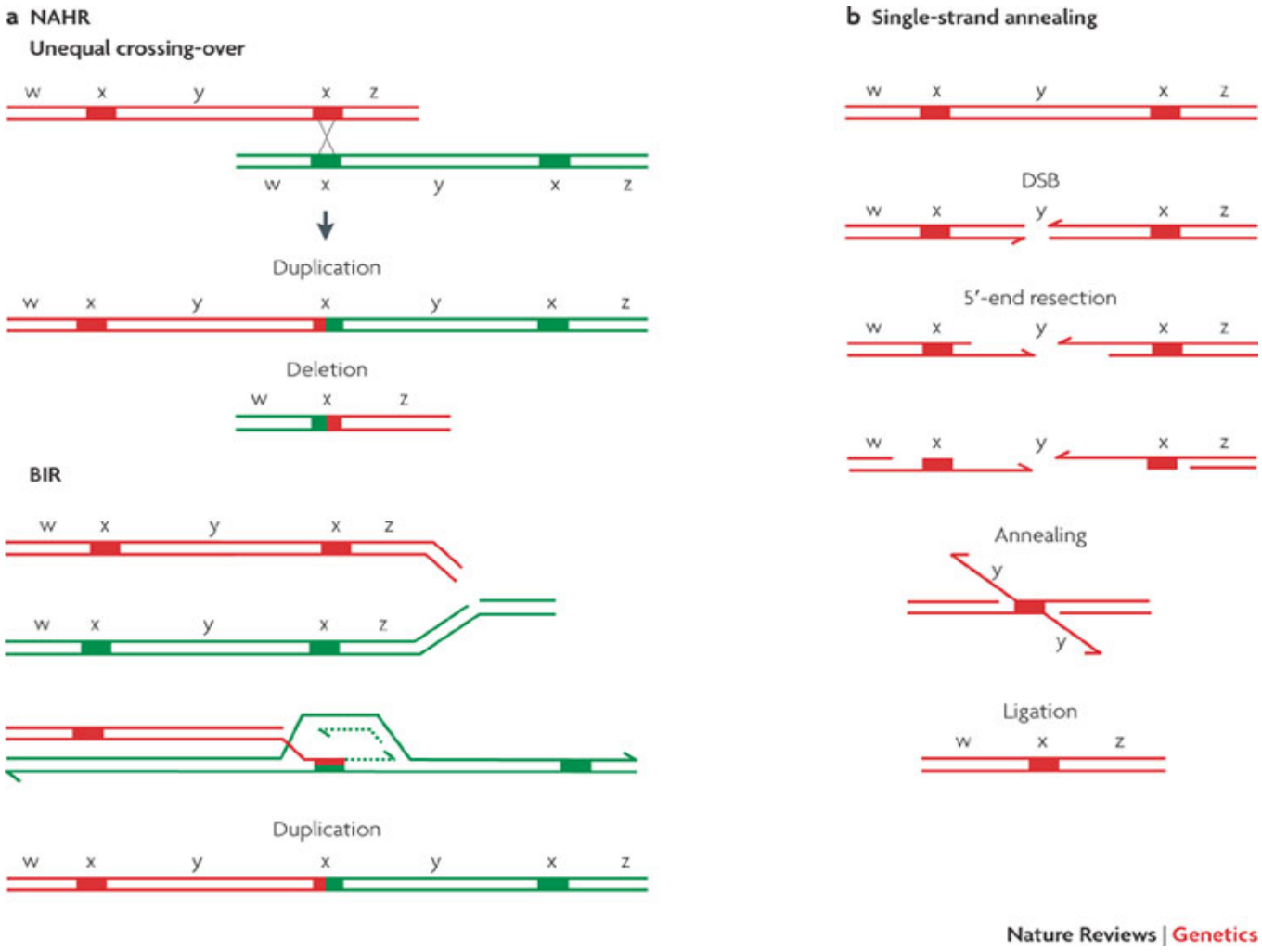
COPY NUMBER VARIATION

- Structural variation that often arises from abnormal recombination events.
- Defined as 1 kilobase or larger.
- Gain and loss of copy number indicated increase risk to common diseases such as schizophrenia and driving processes of clonal selection in tumors
- Preferentially occur in repetitive regions of the genome.
- Accounts for as much as 12% of the human genome.
- Can arise from germ line or somatic mutations. Our work is focused on germline.

NORMAL RECOMBINATION DURING MEIOSIS



CHANGE BY HOMOLOGOUS RECOMBINATION

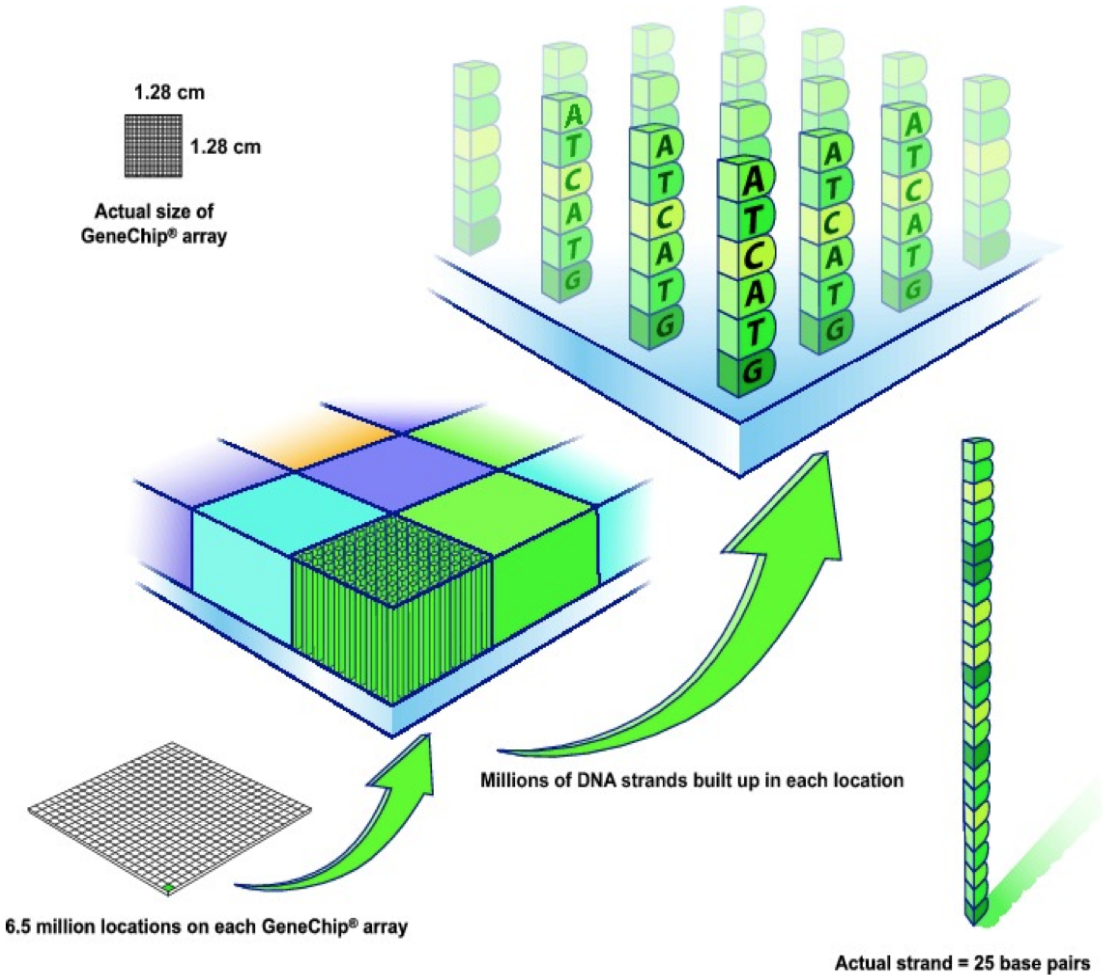


PJ Hastings, 2009: Mechanisms of change in gene copy number

NOTE

High throughput genotyping arrays can only detect low-copy repeats (0, 1, 2, 3, or 4+ copies) because of saturation of the intensities.

AFFYMETRIX PLATFORM



AFFYMETRIX PLATFORM

- Quickly scan for presence of particular genes in a biological sample.
- Each gene represented by a unique set of probe pairs (roughly 12-12 probe pairs per probe set)
- Each spot on array represents a single probe - millions of copies.
- Probes fixed to array.
- A tissue sample is prepared so its mRNA has fluorescent tags.
- mRNA samples hybridize to probes.

OTHER PLATFORMS

- Other genotyping arrays (Illumina etc).
- Comparative genomic hybridization (CGH).
- Next generation sequencing: still very challenging for surveying copy number.

CNV ESTIMATION

There are multiple modes of CNV estimation:

- By sample.
- By locus.
- Hybrid approach.

GOAL

Can we improve copy number estimates at copy number polymorphic regions?

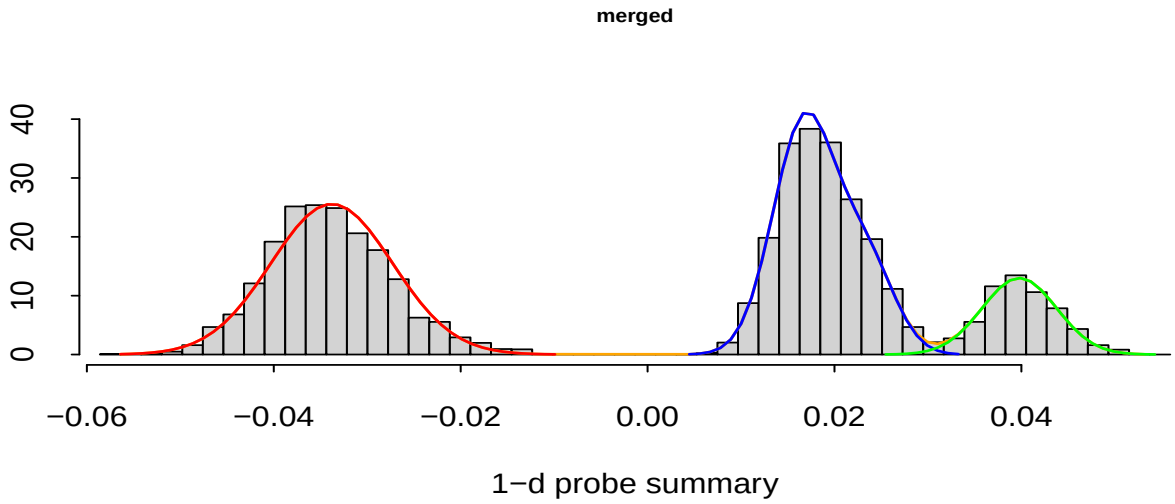
SOFTWARE

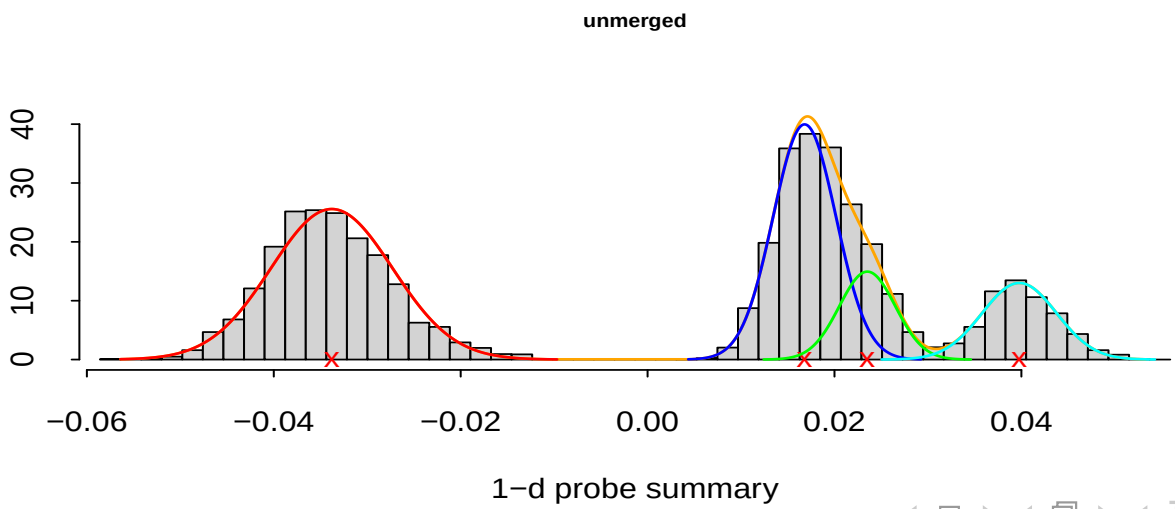
- Birdsuite (Korn, 2008).
- CNVtools (Barnes, 2008).
- cnvCall (Cardin, 2011).
- CNPbayes (Cristiano, 2013).

CARDIN (2011)

- "Bayesian hierarchical mixture modeling to assign copy number from a targeted CNV array"
- For robustness, uses a mixture of t-distributions.
- Introduces a hierarchical structure over the mean and variance across samples from different data collections.
- Uses merging algorithm to combine neighboring components with significant overlap.
- Implemented in R package cnvCall.

cnvCall


merged
0
10
20
30
40
−0.06
−0.04
−0.02
0.00
0.02
0.04
1−d probe summary


unmerged
0
10
20
30
40
−0.06
−0.04
−0.02
0.00
0.02
0.04
1−d probe summary

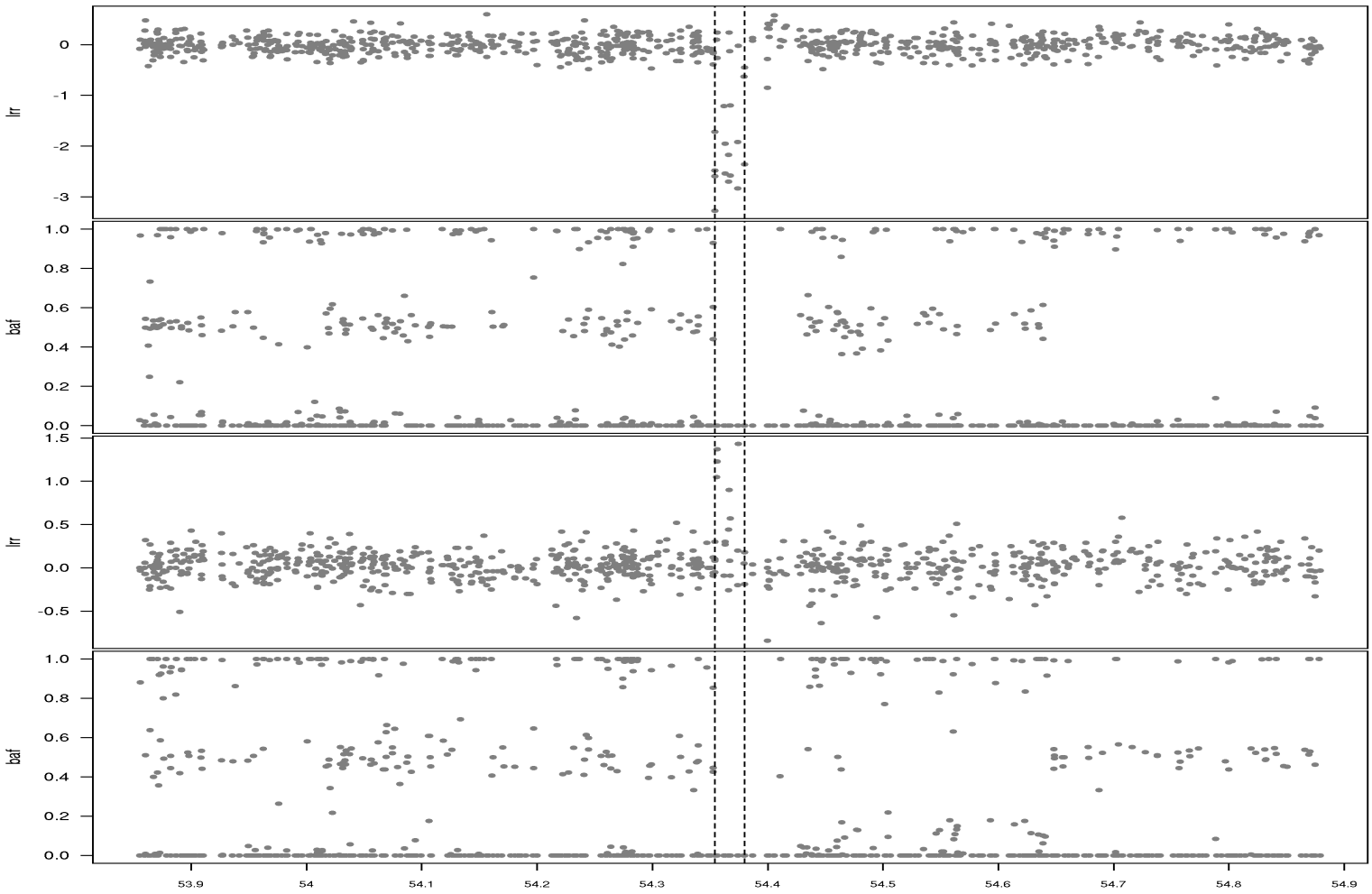
cnvCall

- Our model is most similar to CNV call.
- However, they assume copy number polymorphic regions are known.
- CNP regions will differ between populations of different ancestries, etc.
- We define CNP regions on the basis of Hidden Markov Model calls.

DATA

- 8,598 participants of European ancestry who participated in the Atherosclerosis Risk in Communities (ARIC) Study
- Genomic data: log R ratios and B allele frequencies measured from Affymetrix 6.0 arrays

LOW LEVEL SUMMARIES FOR 2 SAMPLES

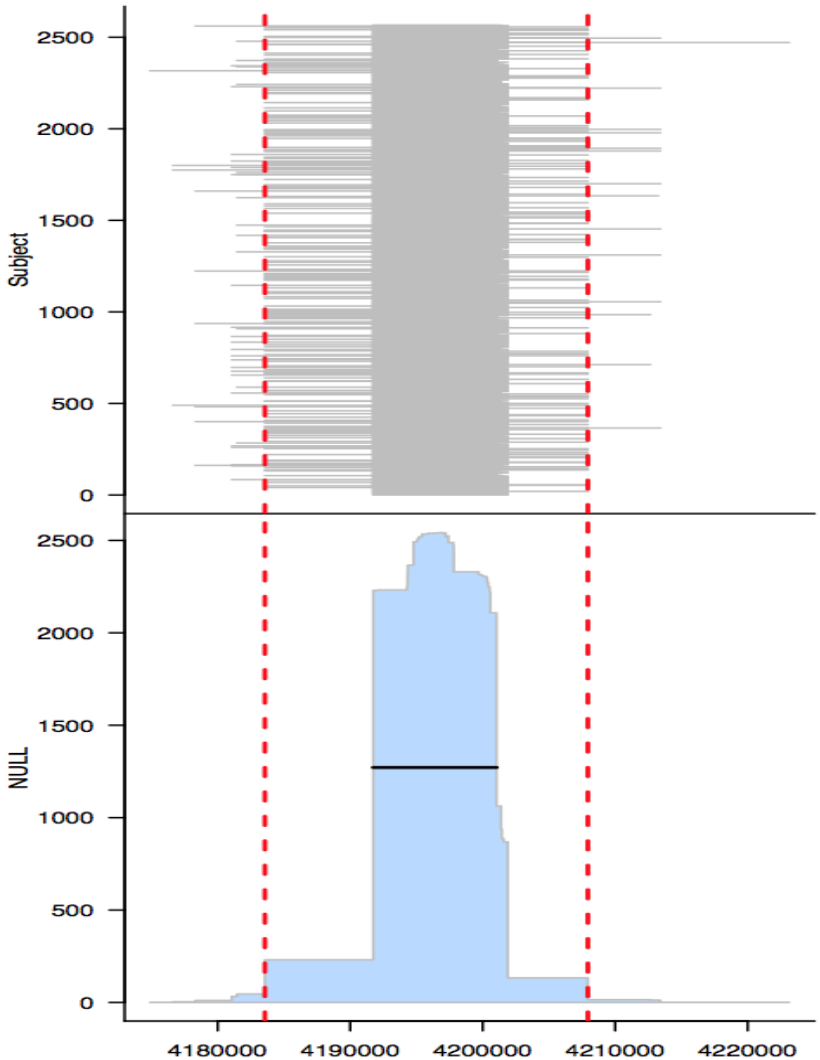


METHOD

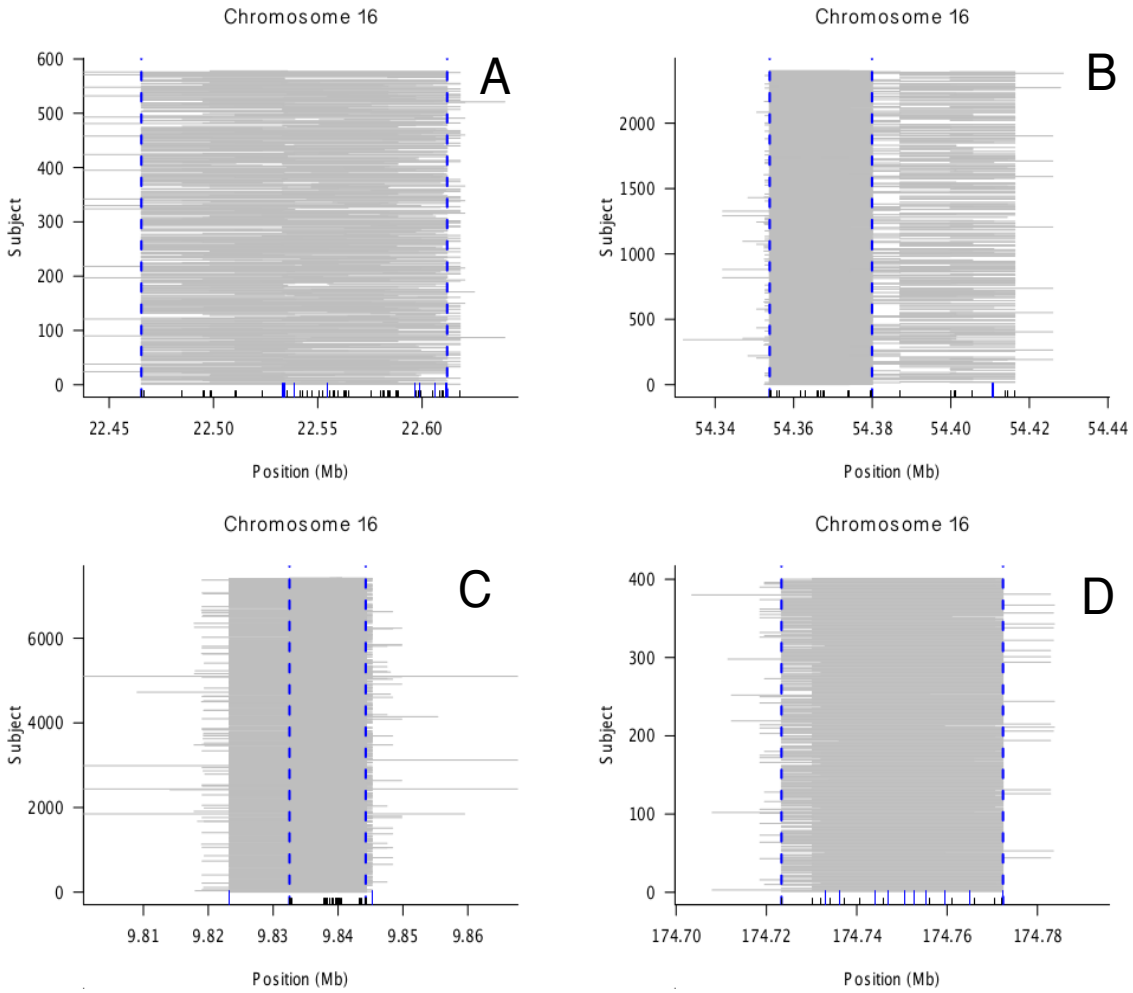
- A 6 state hidden Markov model was fit genome-wide to each subject.
- Approximately 500 regions were identified for which deletions or duplications are common in greater than 1% of subjects.
- GenomicRanges used to find copy number polymorphic loci from the HMM calls.
- A Bayesian finite Gaussian mixture model fit to the average log R ratios improves copy number estimates.

DEFINING REGIONS

- HMM gives non-perfectly overlapping sample specific regions.
- GenomicRanges used to to find copy number polymorphic loci from HMM calls.
- Regions can be complex.
- There may be large gaps in coverage of genotyping arrays.

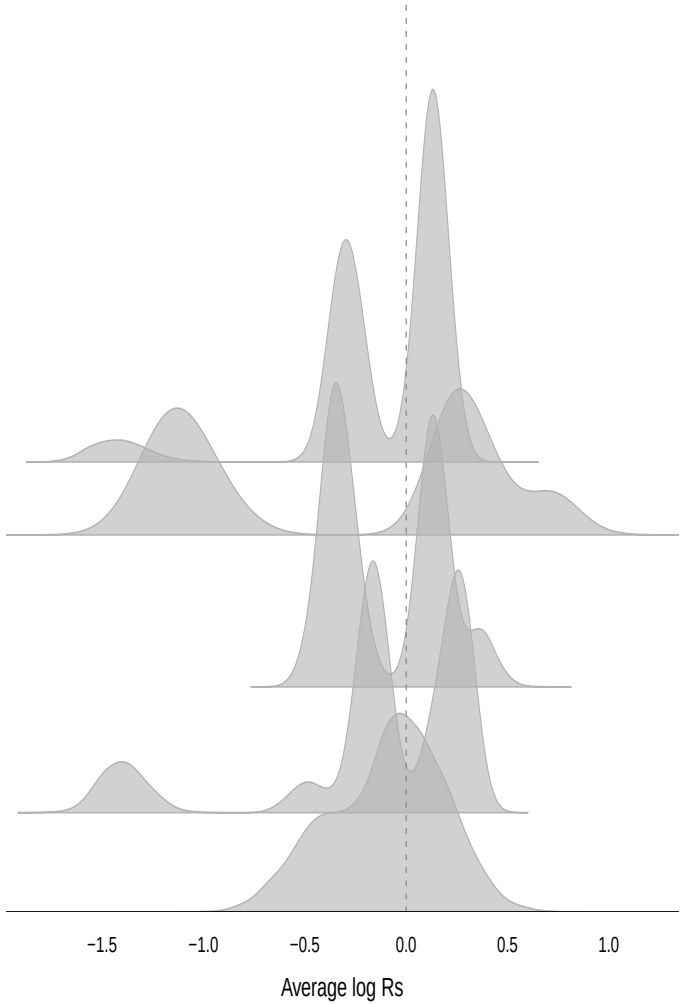


DEFINING REGIONS



EMPIRICAL ESTIMATES

- Mean and variances differ between loci .
- Expected value for diploid component is 0.
- When many deletions or duplications present, the diploid mean is biased away from 0.



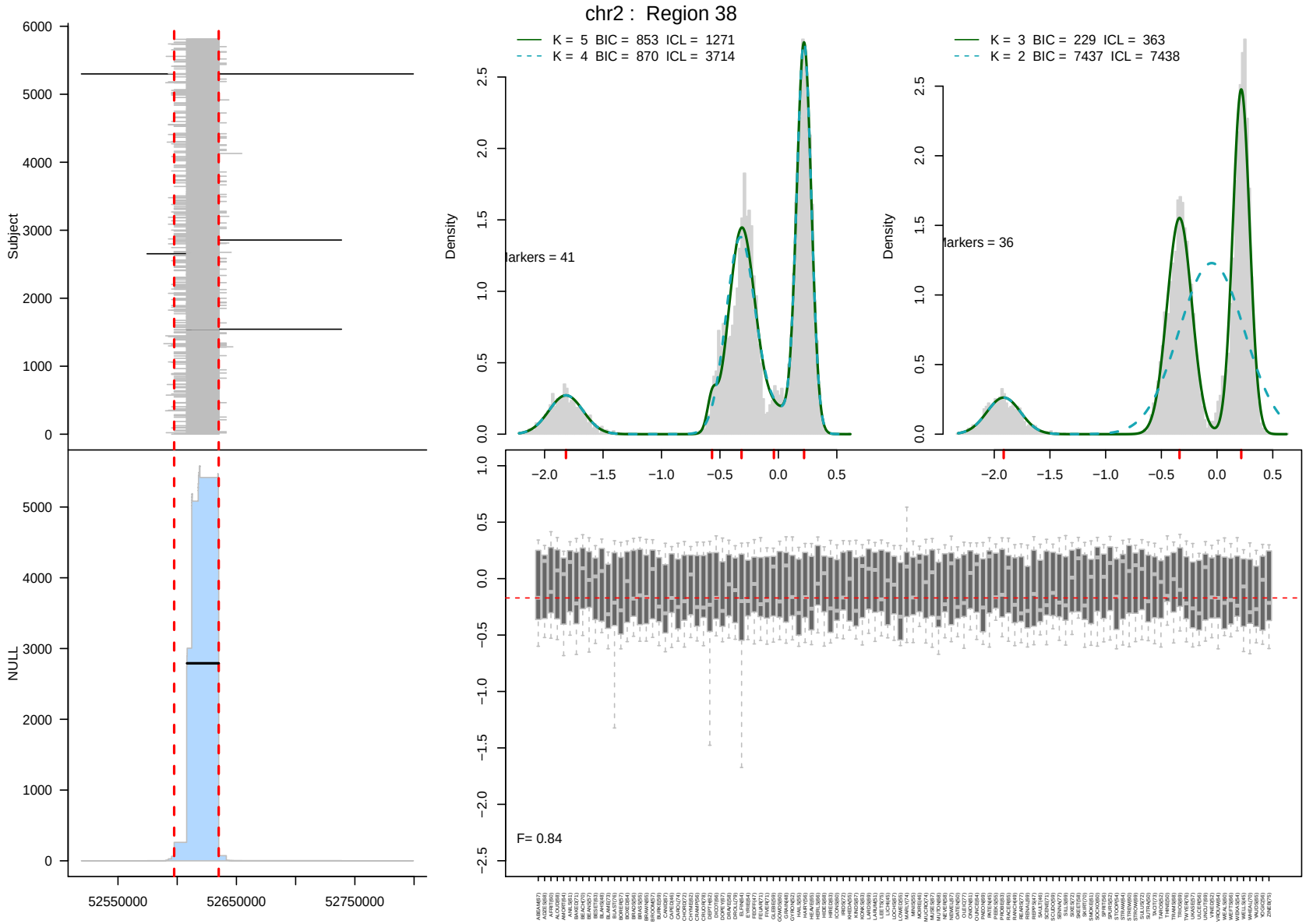
MIXTURE MODEL

- The average log R ratios follow a mixture of Gaussian distributions.
- A finite dimensional Gaussian mixture model assumes data $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbf{R}^n$ are a sample from a from a probability density function of the form

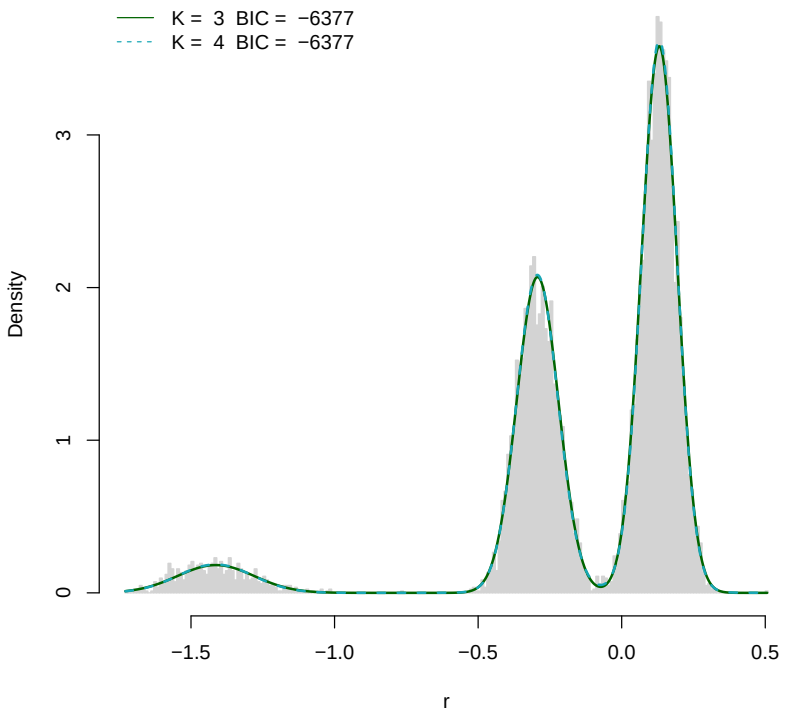
$$f(\mathbf{y}|K, \theta, \sigma^2, p) = \sum_{k=1}^{K} p_k \phi_k(\mathbf{y}|\theta_k, \sigma_k^2)$$

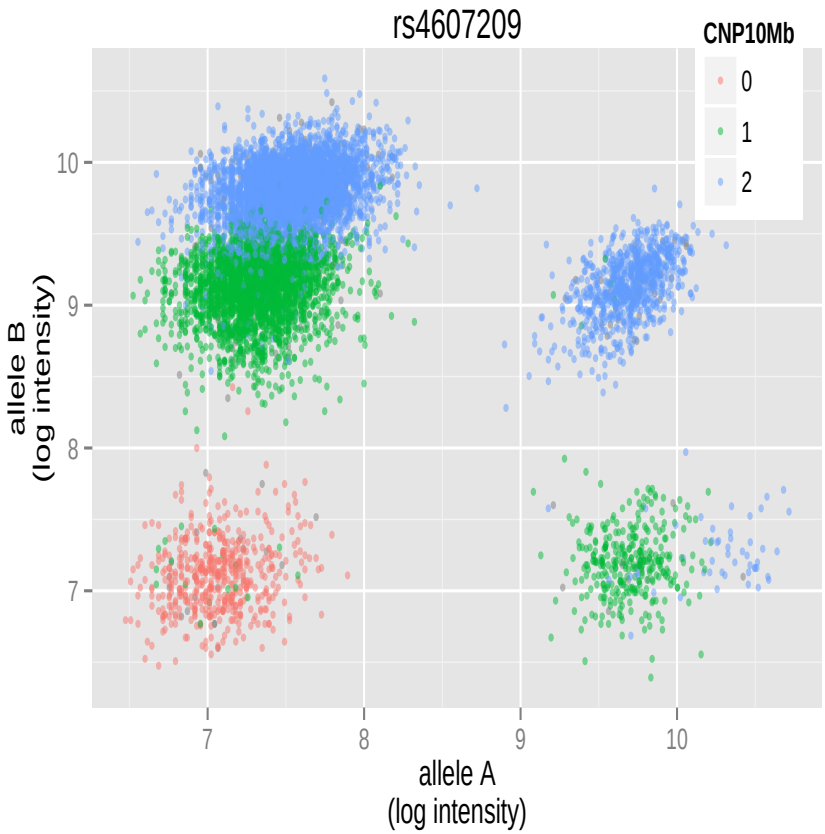
Where K represents the number of components, $\phi(\cdot\,|\theta, \sigma^2)$ is a Gaussian distribution with mean θ and variance σ^2 and $\sum_{k=1}^{K} p_k = 1$.

- Sample from a constrained full conditional on the θ's ensure identifiability and help convergence.
- Run chains of 5000 with a burn-in of 1000 for the 415 regions for each of $K = 1 \ldots 5$ and choose constraints to ensure the means have a separation of 0.2.
- The Bayesian Information Criterion (BIC) was used to assess which of the five models arising from the choices of K best fit the data.


chr2 : Region 38
K = 5 BIC = 853 ICL = 1271
K = 4 BIC = 870 ICL = 3714
K = 3 BIC = 229 ICL = 363
K = 2 BIC = 7437 ICL = 7438
Subject
NULL
Density
markers = 41
markers = 36
F= 0.84
52550000
52650000
52750000

Log-transformed intensities for the A and B allele for a SNP inside one locus on chromosome 4.

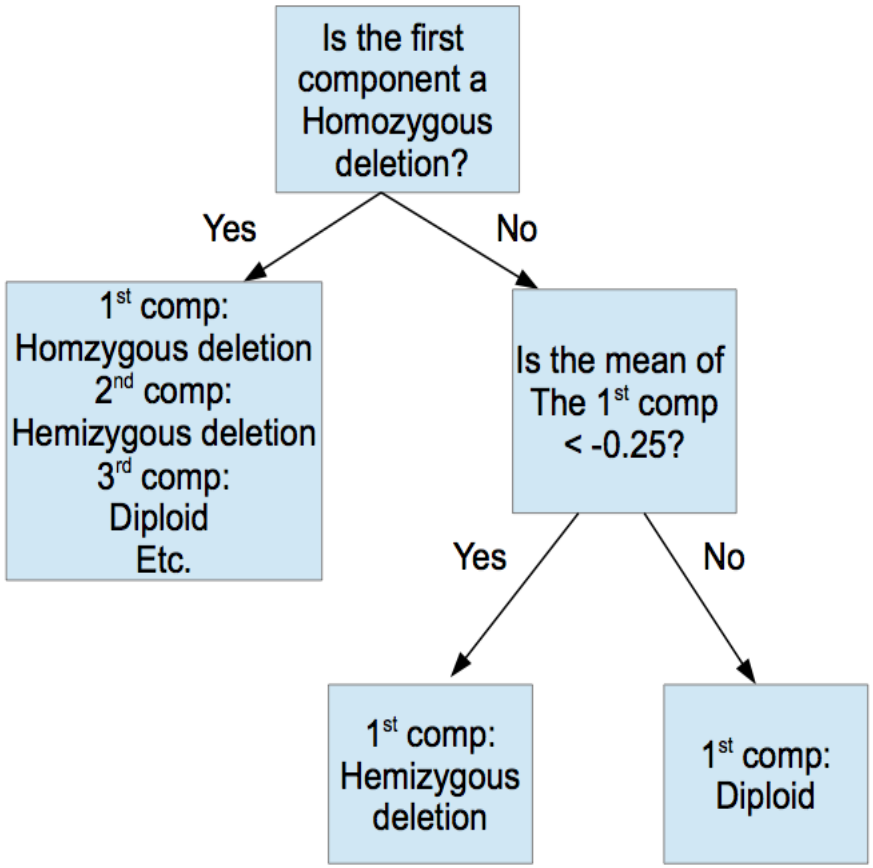




ASSIGNMENT

- Need a way to infer what copy number state each component is.
- Average log R ratios are biased in CNPs, so we can't assume the component closest to 0 is diploid.
- Many CNPs do not contain SNPs, so information about heterozygosity is often not available.
- On the log-scale, distance between homozygous deletions and hemizygous deletions is large, and homozygous deletions have a large variance relative to the other components.
- Homozygous deletions are easy to detect.

AD HOC APPROACH



DISCUSSION

- We do not necessarily need to use the maximum a priori estimates to infer copy number.
- Our model has the advantage that we can assign a probability to each copy number assignment.
- This uncertainty in copy number estimates can be propagated to association models.

COMPLICATIONS

- BIC often overestimates the number of components.
- When skew is present in one of the components, a model with an additional component to capture the skew will be preferred.
- A mixture model of skewed normal distributions may be more robust.

SKEW-NORMAL DISTRIBUTION

- A finite dimensional skew-normal mixture model assumes data $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbf{R}^n$ are a sample from a from a probability density function of the form

$$f(\mathbf{y}|K, \theta, \sigma^2, \alpha, p) = \sum_{k=1}^{K} p_k f_{SN_k}(\mathbf{y}|\theta_k, \sigma_k^2, \alpha_k)$$

  Where α a skewness parameter.

- Full conditionals are available for the proper parameter transformations and Gibbs sampling is still feasible. (Frühwirth-Schnatter, 2010)

SOFTWARE

- R package CNPbayes available on github.
- MCMC methods implemented using Rcpp for rapid computations.
- Currently being prepared for submission to Bioconductor.

WHAT NEXT

- Develop regression model for associating copy number classification with disease phenotype.
- Batch effects may be present. Consider adding a hierarchical structure to the parameters.
- Compare with other methods.

THANKS

- Rob Scharpf
- Gary Rosner
- Leonardo and Jean-Philippe