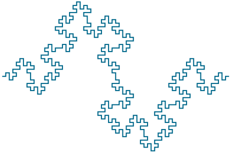
Estimating copy number polymorphisms from genotyping arrays

Stephen Cristiano
Johns Hopkins University

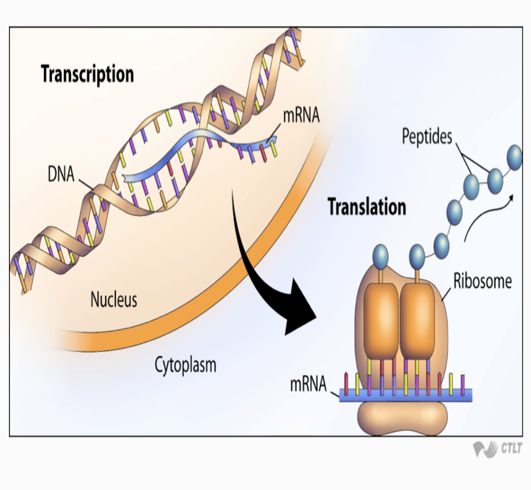


November 5, 2013

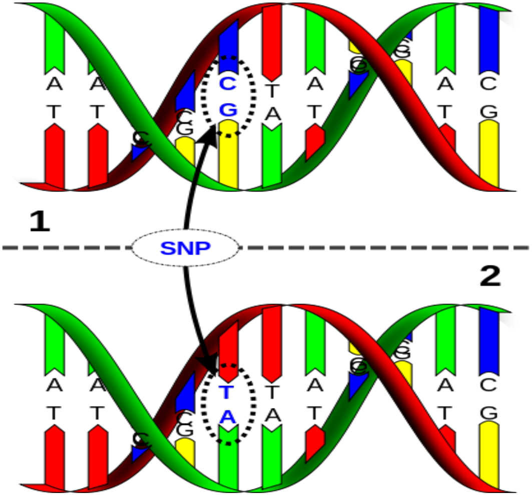
INTRODUCTION

- INTRODUCTION
- BIOLOGY
  - Central Dogma
  - Single nucleotide polymorphisms
  - Copy number variation
- PLATFORM
  - Affymetrix
  - CNV estimation
- METHODS
  - Data
  - Model
  - Bayesian Mixture Model
- DISCUSSION
  - Other models
  - Software
  - Future considerations

CENTRAL DOGMA OF MOLECULAR BIOLOGY



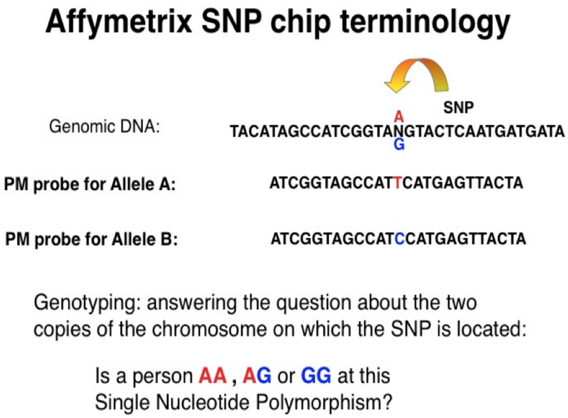
SINGLE NUCLEOTIDE POLYMORPHISM



SNPs

- Single Nucleotide Polymorphism are DNA sequence variations that differs at a single base among members of a population.
- Two common alleles at most SNPs (>1%)
- More rare can not be interrogated by high-throughput platforms. What is rare depends on the population.

AFFYMETRIX SNP CHIP TERMINOLOGY



Affymetrix SNP chip terminology

SNP

A

Genomic DNA: TACATAGCCATCGGTANGTACTCAATGATGATA

G

PM probe for Allele A: **ATCGGTAGCCATTCATGAGTTACTA**

PM probe for Allele B: **ATCGGTAGCCATCCATGAGTTACTA**

Genotyping: answering the question about the two copies of the chromosome on which the SNP is located:

Is a person **AA** , **AG** or **GG** at this Single Nucleotide Polymorphism?

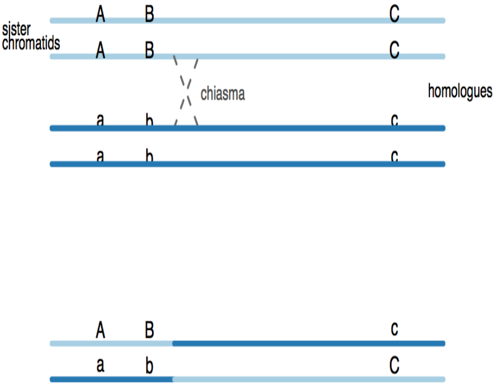
COPY NUMBER VARIATION

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

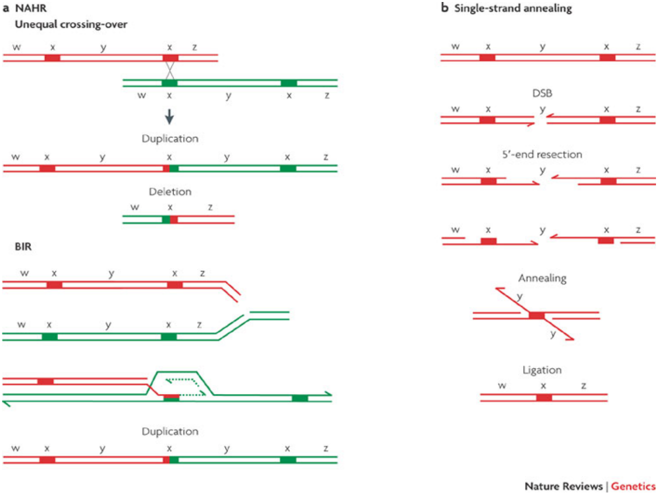
COPY NUMBER VARIATION

- 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.

Normal recombination during meiosis



CHANGE BY HOMOLOGOUS RECOMBINATION



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

Germline vs Somatic CNV

- DNA is collected from blood or tissue.
- The isolated DNA is typically amplified by PCR.
- Copy number changes during meiosis are present in all cells in an individual.
- In diseases such as cancer, recombination can occur during mitosis resulting in cells with different DNA copy numbers.
- Implication: for germline diseases, we expect the DNA copy number to be an integer. For cancer, noninteger DNA copy number is plausible due to heterogeneity of the cells within a tissue.

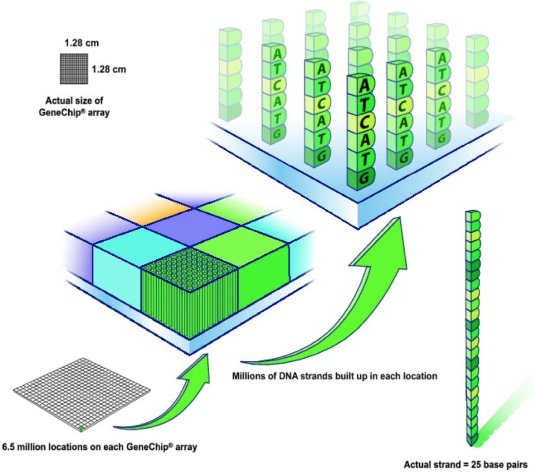
OTHER SOURCES OF DNA VARIATION

High throughput genotyping arrays can only detect low-copy repeats (0-5 copies).

Forms of DNA variation that we can not detect:

- Short or highly repetitive sequences such as LINEs and SINEs
- insertions
- inversions
- translocations

AFFYMETRIX PLATFORM


1.28 cm
1.28 cm
Actual size of GeneChip® array
A
T
C
A
T
G
Millions of DNA strands built up in each location
6.5 million locations on each GeneChip® array
Actual strand = 25 base pairs

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

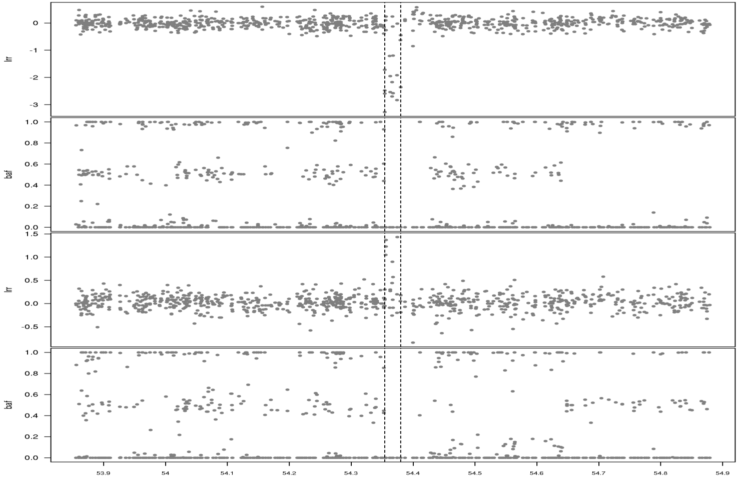
There are multiple modes of CNV estimation:

- By sample: segmentation of noisy marker-level estimates of copy number in individual genomes to infer the latent copy number.
- By locus: marker-level estimates directly in association models followed by smoothing the test statistics.
- Hybrid approach.

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

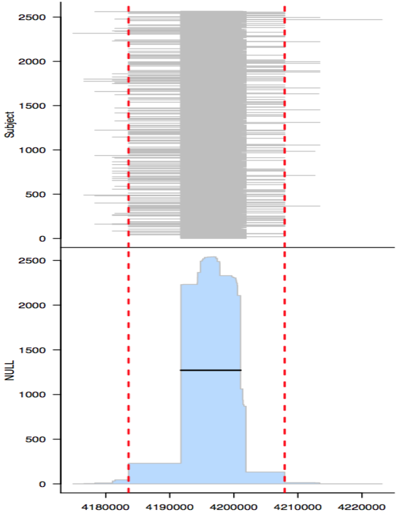


METHOD

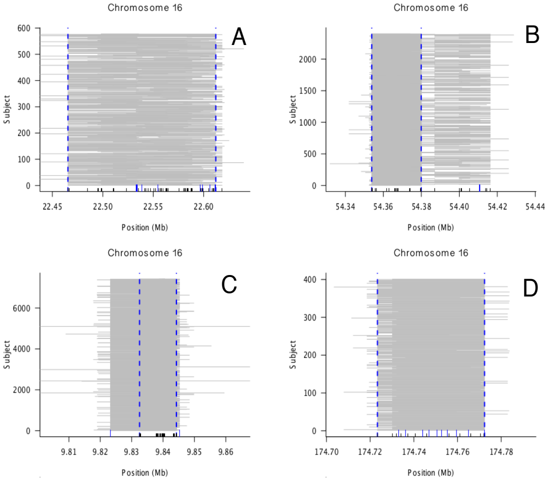
- A 6 state hidden Markov model was fit genome-wide to each subject.
- Approximately 500 regions were identified for which deletions or duplications are common in greater than 1% of subjects.
- GenomicRanges used to find copy number polymorphic loci from the HMM calls.
- A Bayesian finite Gaussian mixture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

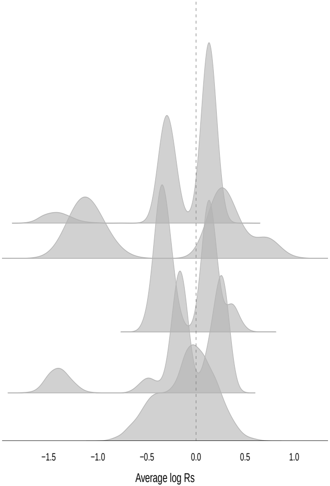


DEFINING REGIONS



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

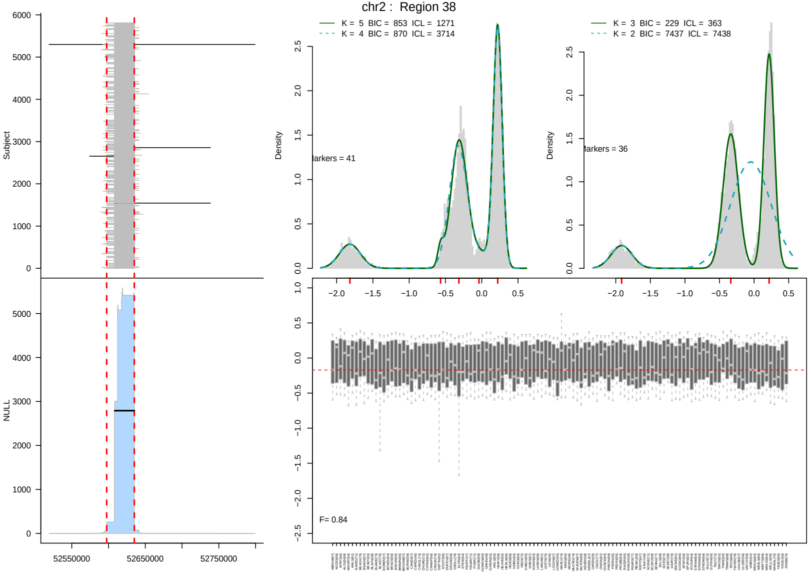
$$f(\mathbf{y}|K, \theta, \sigma^2, p) = \sum_{k=1}^{K} p_k \phi_k(\mathbf{y}|\theta_k, \sigma_k^2)$$

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

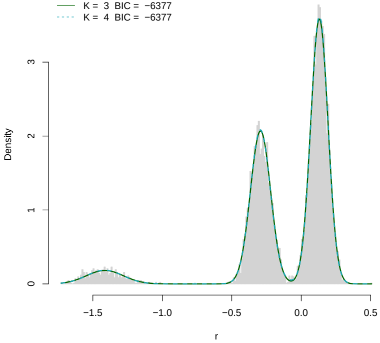
- Sample from a constrained full conditional on the θ's ensure identifiability and help convergence.

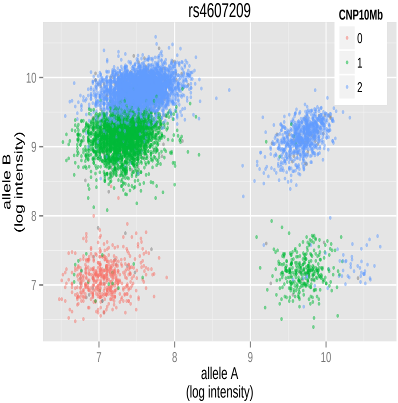
- Run chains of 5000 with a burn-in of 1000 for the 415 regions for each of $K = 1 \dots 5$ and choose constraints to ensure the means have a separation of 0.2.

- 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
Markers = 41
Markers = 36
Density
Subject
NULL
F= 0.84
52550000
52650000
52750000
6000
5000
4000
3000
2000
1000
0
2.5
2.0
1.5
1.0
0.5
0.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
−2.5

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





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

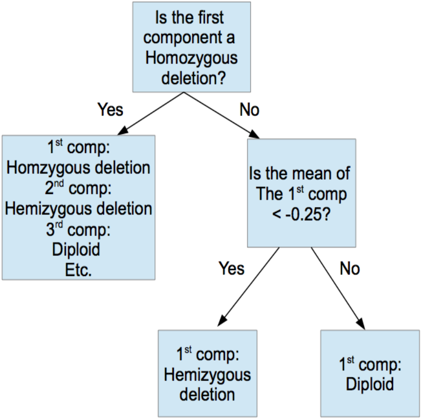
  $$f(\mathbf{y}|K, \theta, \sigma^2, \alpha, p) = \sum_{k=1}^{K} p_k f_{SN_k}(\mathbf{y}|\theta_k, \sigma_k^2, \alpha_k)$$

  Where α a skewness parameter.

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

ASSIGNMENT

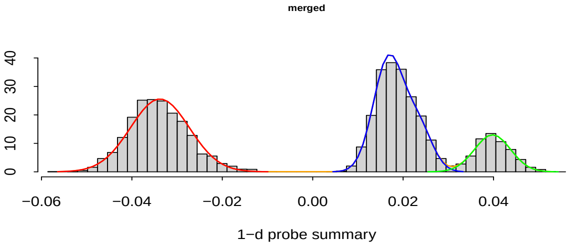
Need way to assign individuals to copy number classes.

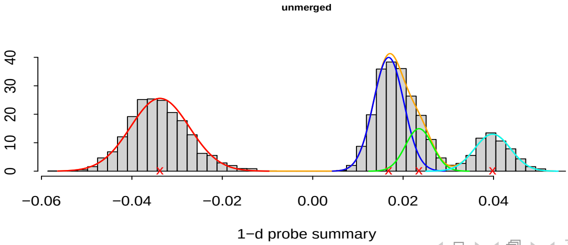

Is the first component a Homozygous deletion?
Yes
No
1st comp: Homzygous deletion 2nd comp: Hemizygous deletion 3rd comp: Diploid Etc.
Is the mean of The 1st comp < -0.25?
Yes
No
1st comp: Hemizygous deletion
1st comp: Diploid

Cardin (2011)

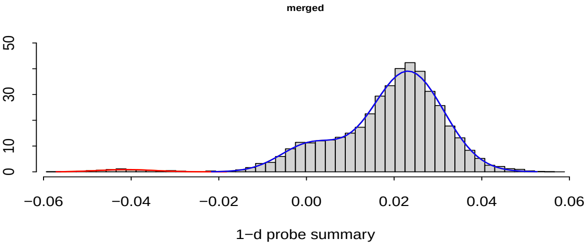
- "Bayesian hierarchical mixture modeling to assign copy number from a targeted CNV array"
- For robustness, uses a mixture of t-distributions.
- Introduces a hierarchical structure over the mean and variance across samples from different data collections.
- Uses merging algorithm to combine neighboring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

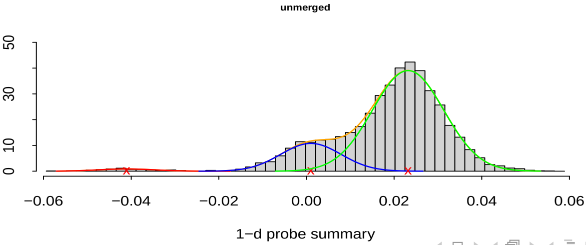
CNVCall


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


merged
50
30
10
0
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
1−d probe summary


unmerged
50
30
10
0
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
1−d probe summary

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