

Bayesian inference for SNP Arrays

Eric Brugel

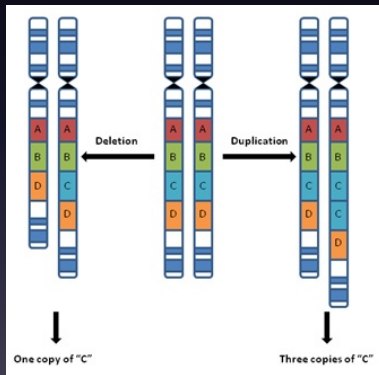
Dimacs REU
Schliep Labs

Mentor: John Wiedenhoeft

June 5, 2014

Copy Number Variation

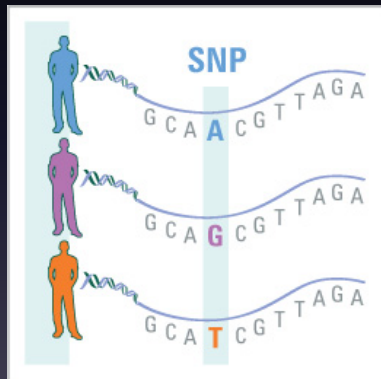
- Copy Number Variation - Variation of the number of copies of a region in DNA caused by deletion or duplication



(a) Copy Number Variation

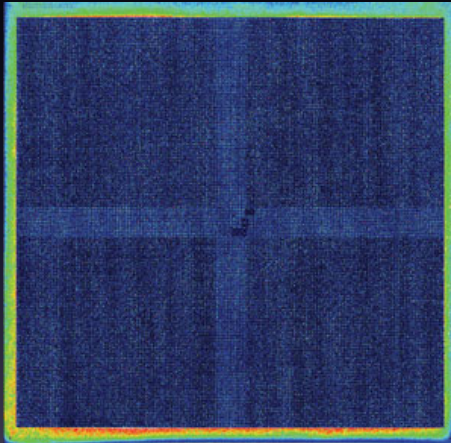
Single Nucleotide Polymorphism

- SNPs are single variations of genes in a population
- SNP arrays are a way to store the SNPs of a genome



(b) SNP

Single Nucleotide Polymorphism



(c) SNP array

Genotype

- Alleles are alternate forms of the same gene
- The genotype is the combination of alleles for a gene

Current Workflow

- Preprocess the SNP arrays
- Get point estimates of the genotype and copy number variations
- Use neighboring region's statistics to to determine regions of deletion, amplification, or loss of heterozygosity

Problems with the current method

- Statistical errors in the pre-processing
- Point estimates
 - finds the local maximum
 - loss of information
 - over-fitting data
- Could not be used with parent SNP arrays

Proposed Solution

- We will combine both of the steps and create a single model using Bayesian inference
- Bayesian inference is used to find a distribution rather than value for hidden variables given observed variables
- Our approach could be used with parents SNPs arrays to obtain a more accurate model

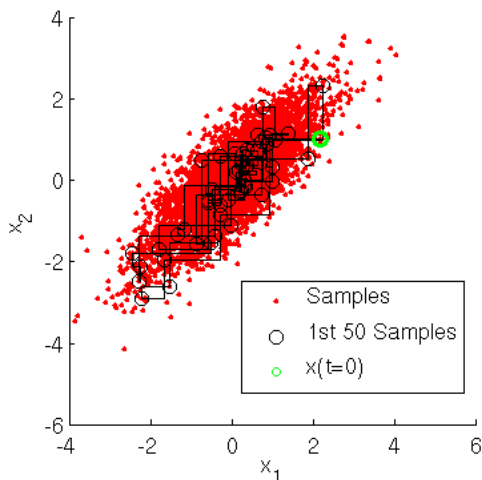
Difficulties in using Bayesian statistics

- Finding the posterior distribution involves integrating out the parameters which is analytically intractable

Gibbs Sampling

- A Gibbs sampler is a Markov Chain Monte Carlo algorithm used to approximate multivariate distributions.
- Let $p(x_1, x_2, \dots, x_n)$ be the joint distribution of $(x_1, x_2, \dots, x_n)$
- Algorithm:
Input: k the number of samples
Begin with some initial $(x_1^0, \dots, x_n^0)$
for $i \in \{1, \dots, k\}$
for each j sample x_j^i from $p(x_j | x_1^i, \dots, x_{j-1}^i, x_{j+1}^{i-1}, \dots, x_n^{i-1})$

Gibbs Sampling



(d) Gibbs Sampler

Parallelizing the Gibbs Sampler

- We create a graph representation of our model and do vertex coloring
- Nodes of different color are conditionally independently of each other allowing parallelization.

Thank You