UCLA

Genotype Calling

Jackson Pang
Digvijay Singh

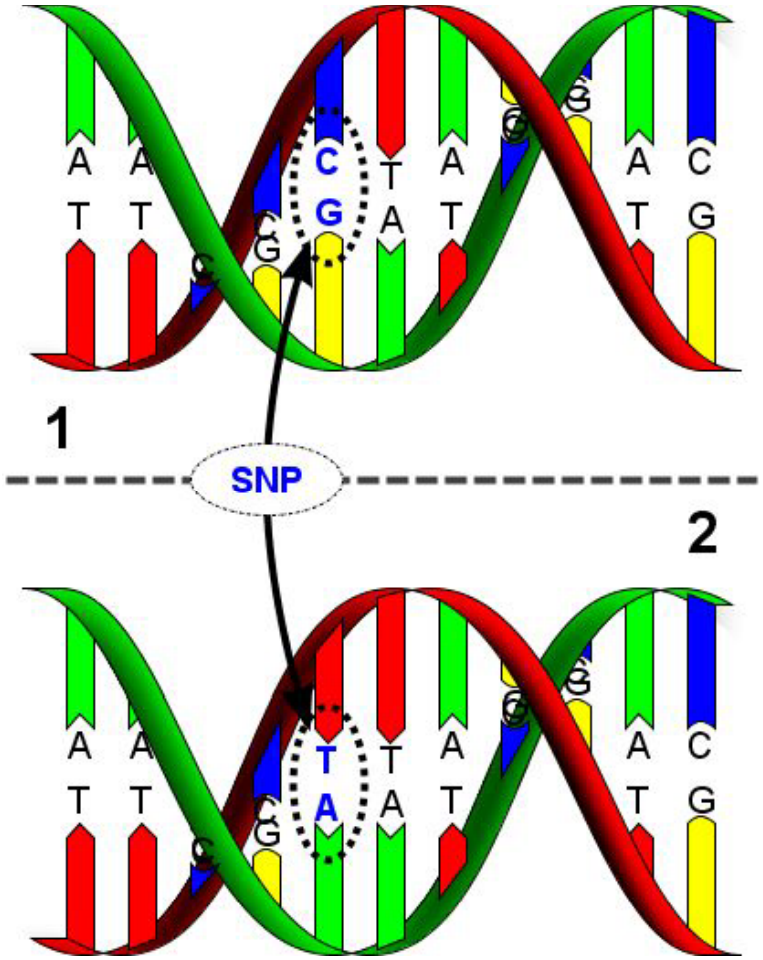
Electrical Engineering, UCLA

Outline

- **Introduction**
  - Single Nucleotide Polymorphism
  - SNP Genotyping
  - SNP Microarrays
- **The Problem**
- **Our Solutions**
- **Results**
- **Summary**

Introduction: Single Nucleotide Polymorphism (SNP)

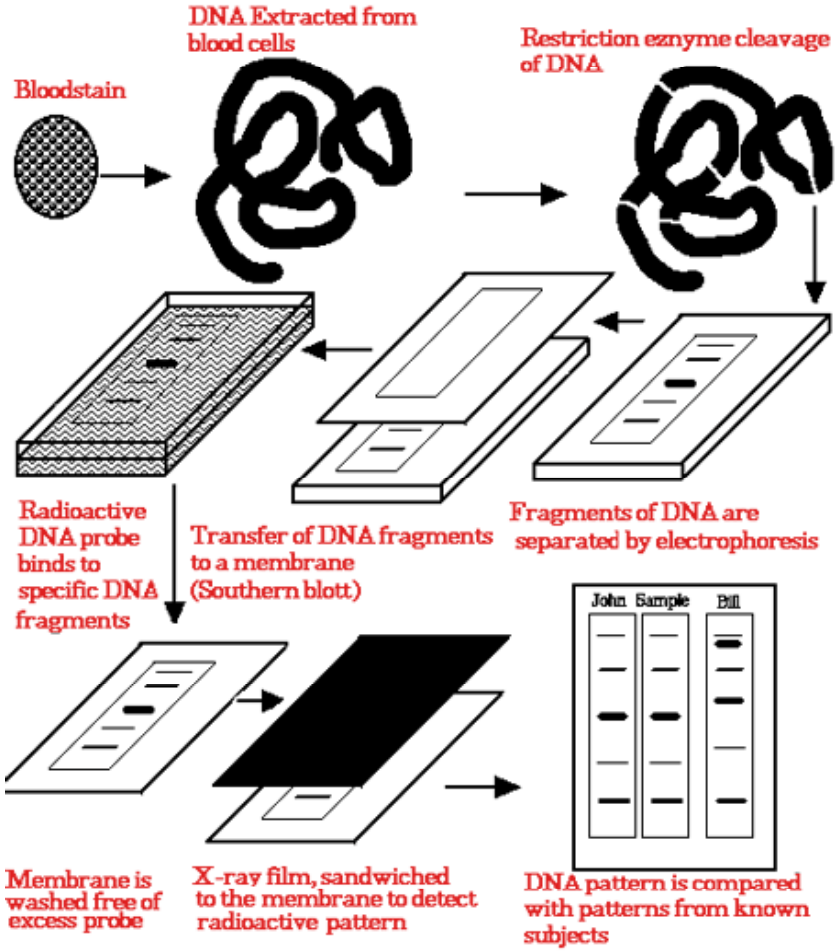
- A variation in the DNA sequence
- Mutation of a single base pair
  - Represents over 80% of human genetic variation
- Has far-reaching effects
  - Disease Risk
  - Reaction to Chemicals
  - Response to Vaccines
  - Personalized Medicine



Picture from: www.wikipedia.org

Introduction: SNP Genotyping

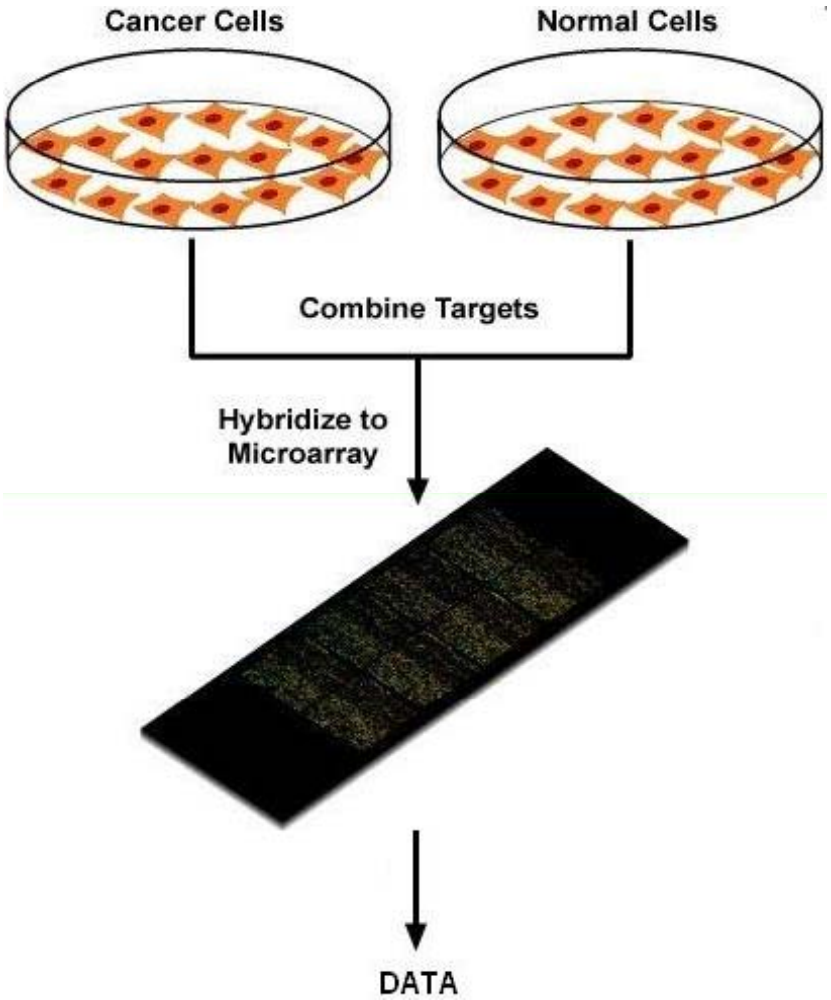
- Also termed Genotype Calling
  - Detection of SNPs
  - The SNPs are genotyped or “called”
- Many methods exist
  - Sequencing
  - Enzyme-based
    - Digest parts of genome
  - Hybridization
    - Use of probe data



Picture from: Olivier M. The Invader Assay for SNP Genotyping,2005.

Introduction: SNP Microarrays

- Hybridization-based
  - Mix control and case population DNA samples in a certain ratios
- SNP Microarrays
  - Lots of probes on a chip
  - Simultaneous inspection of thousands of SNPs
  - We use the Affymetrix 100k chip's data from the HapMap website



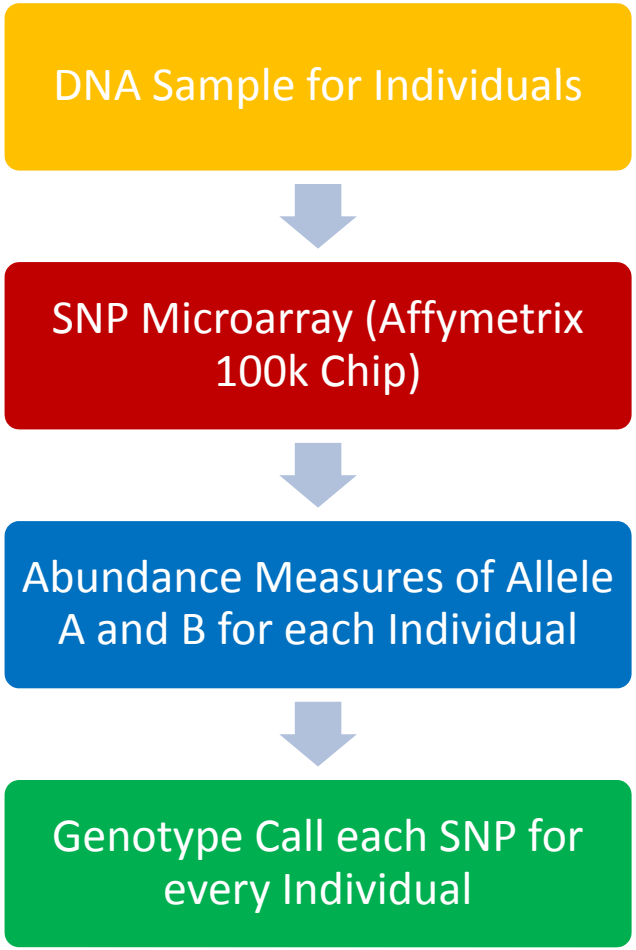
Picture from: www.wikipedia.org

The Problem

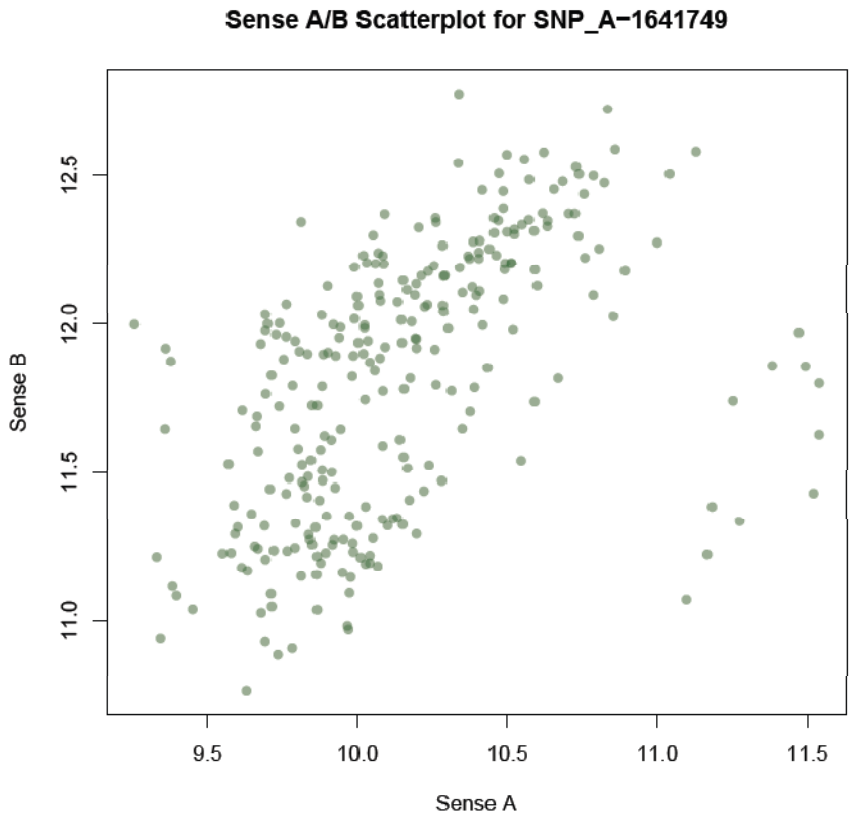
- **Introduction**
- **The Problem**
  - **Genotyping Calling from Affymetrix Data**
  - **Clustering**
- **Our Solutions**
- **Results**
- **Summary**

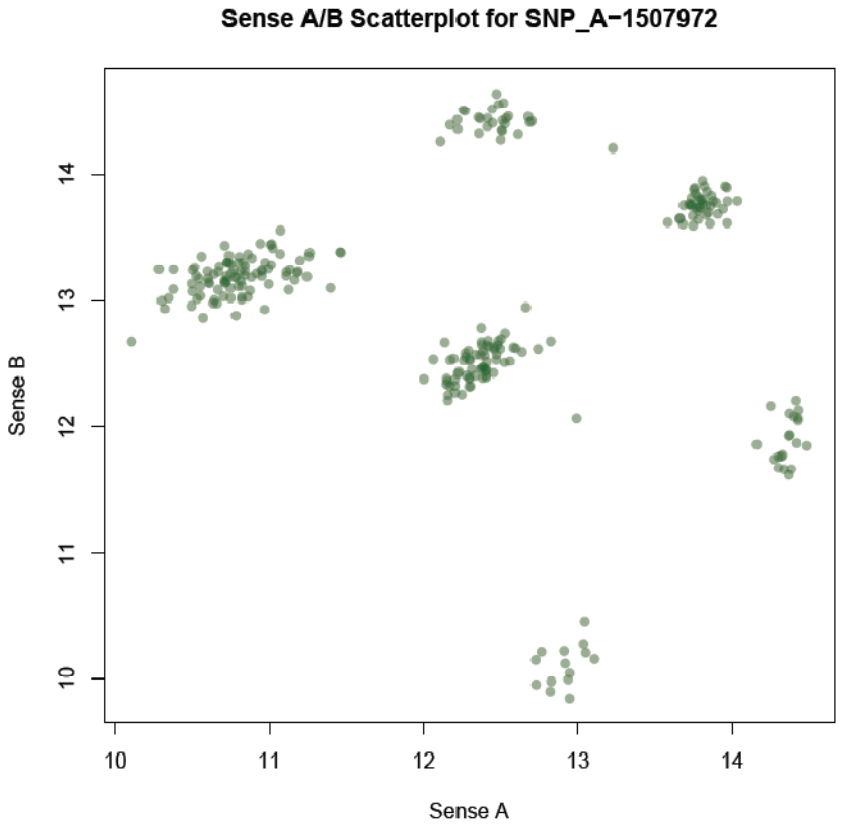
The Problem: Genotyping Calling from Affymetrix Data

- SNP has alleles A and B
- Affymetrix Microarray's probe data
  - Abundance measure for alleles A and B
- Two chromosomes
  - Each individual is "called" as AA, AB or BB
- Facilitates association
  - Calculates allele freq. in cases and controls for target SNPs



The Problem: Clustering





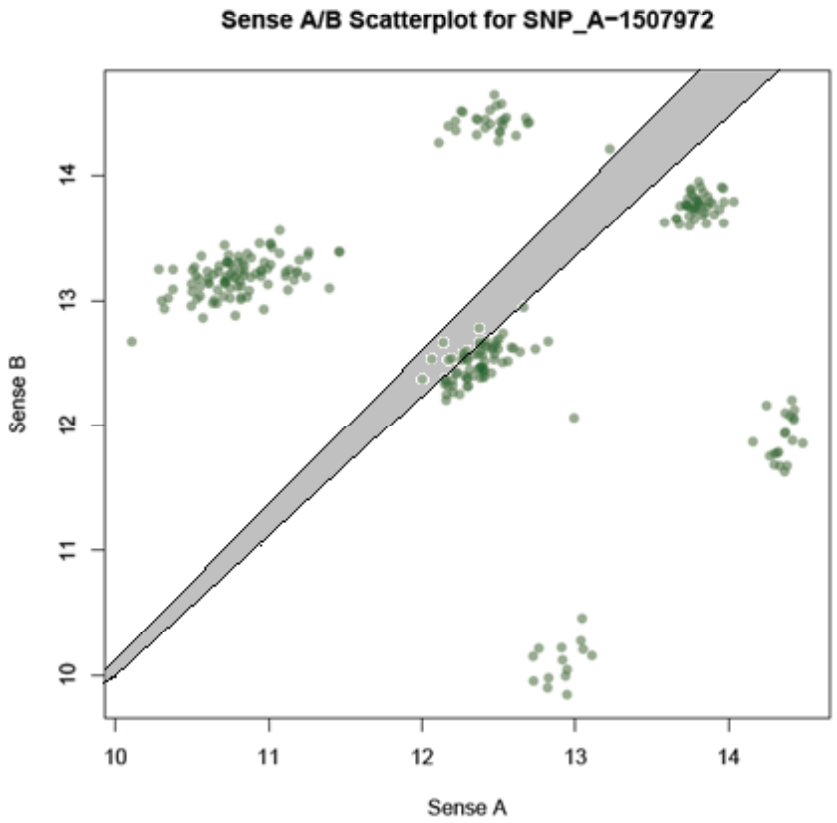
- **Processed Probe Data from Affymetrix 100k Microarray for certain SNPs**
- **The x and y axes represent abundance measures for Alleles A and B**
- **The data is for 270 individuals from a population**

Our Solutions

- **Introduction**
- **The Problem**
- **Our Solutions**
  - **Sector-sweep Clustering**
  - **K-Means Clustering**
- **Results**
- **Summary**

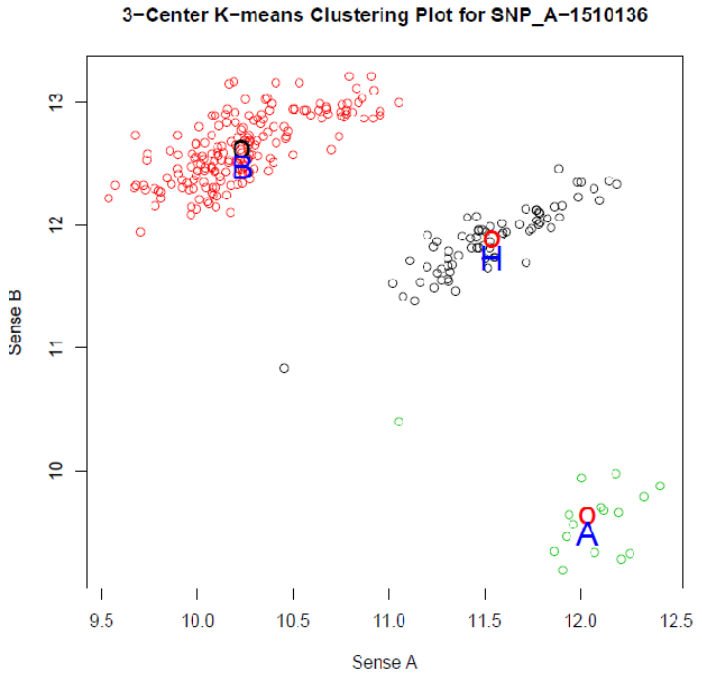
Our Solutions: Sector-Sweep Clustering

- **Basic intuition** is that all the AB genotypes will lie close to a 45 degree slope i.e. similar abundance values for A and B
- The algorithm uses a sector or slice which sweeps from the 45 degree line to 0 degrees
- An upper and lower slope threshold are determined by the sweeping sector.
- It takes 2 parameters:
  - The width of the slice (5 degrees)
  - The density drop (0.75)
- Outputs
  - Upper and lower slope thresholds
- The points above the upper threshold are classified as B
- The points below the threshold are marked as A
- All other points are marked as AB



Our Solutions: K-means Clustering

- **Basic intuition** is to partition a set of points into K groups such that the sum of squares from points to the assigned cluster centers is minimized
- Takes in 3 parameters
  - The point vector
  - # centers, # starting sets, #iterations
- Outputs
  - An array of coordinates for centers
  - Membership array
  - Withiness
- Classification of centers as genotypes
  1. The classes are relative slope ratios of the centers
     - Slope ratio of cluster 1 : center1(y) / center1 (x)
     - Highest = B; Lowest = A, Middle = H (AB)
  2. Dividing the quadrant into 3 equal sections and classifying based on the cluster's center location
     - Worse than slope ratios

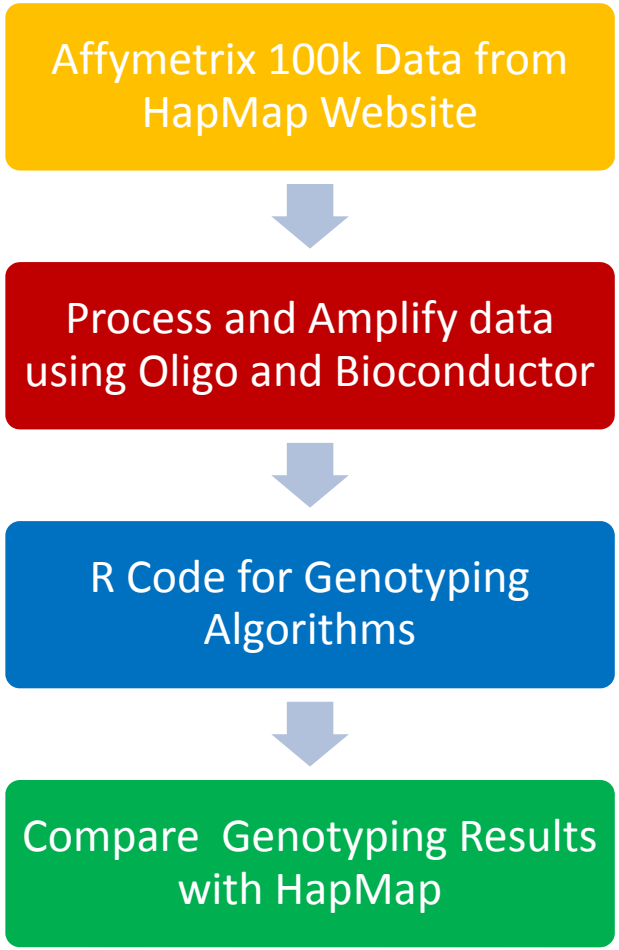


Results

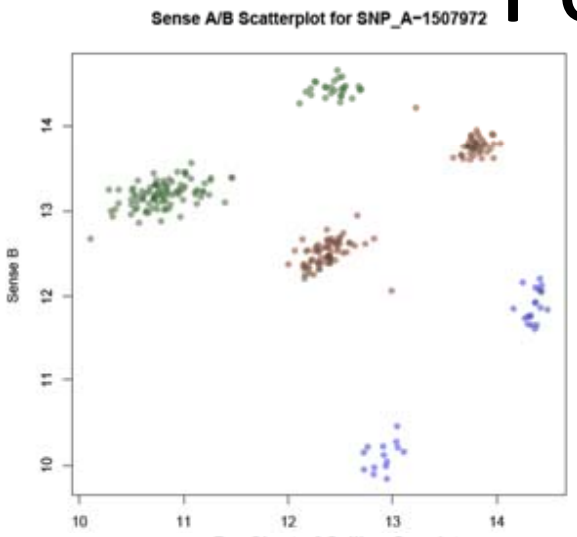
- **Introduction**
- **The Problem**
- **Our Solutions**
- **Results**
  - **Implementation Details**
  - **Sector-Sweep**
  - **K-means**
- **Summary**

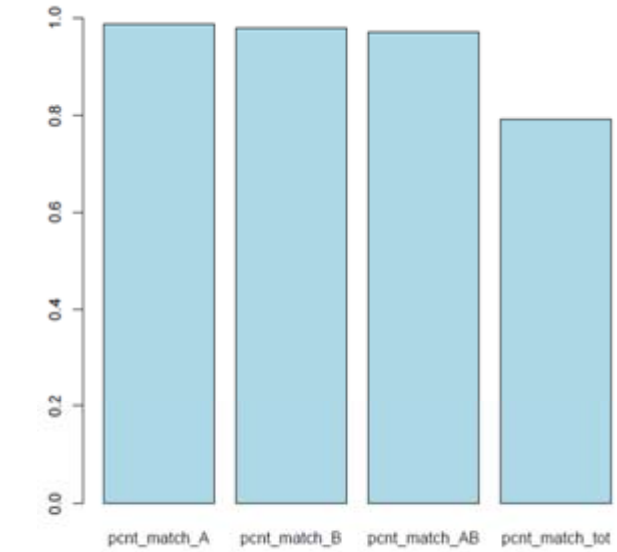
Results: Implementation Details

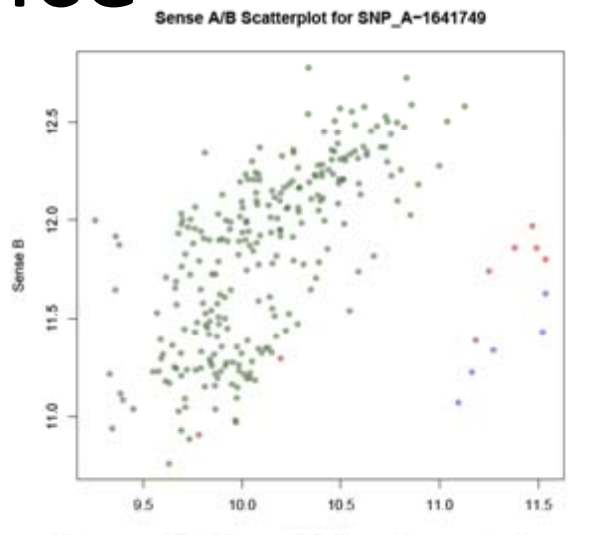
- Data
  - Affymetrix 100k Chip data used from HapMap
  - Data for 270 individuals
- Coding Paradigm
  - All coding done in R
  - Oligo and Bioconductor packages for data processing
- Goodness of Solution
  - Comparison to Hapmap Genotyped Data

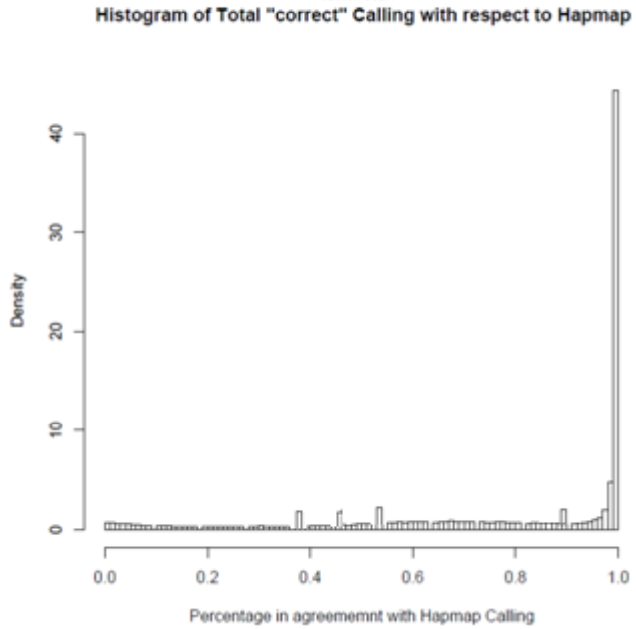


Results: Sector-Sweep Performance

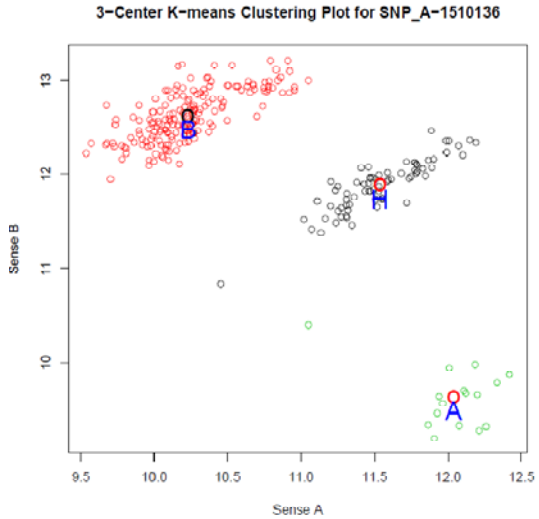


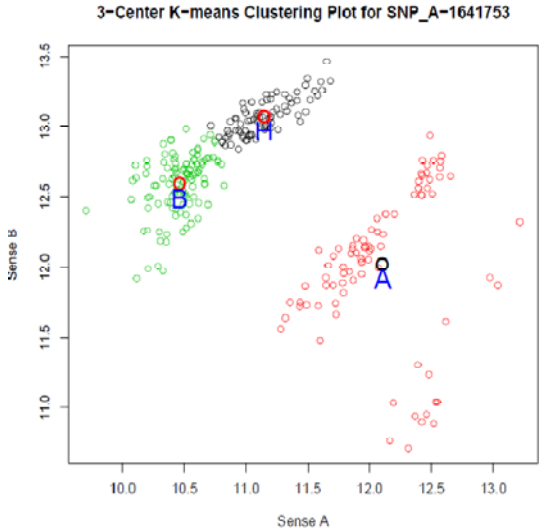


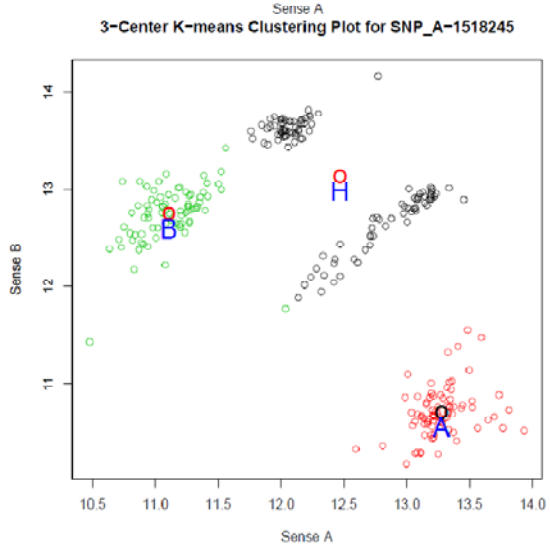


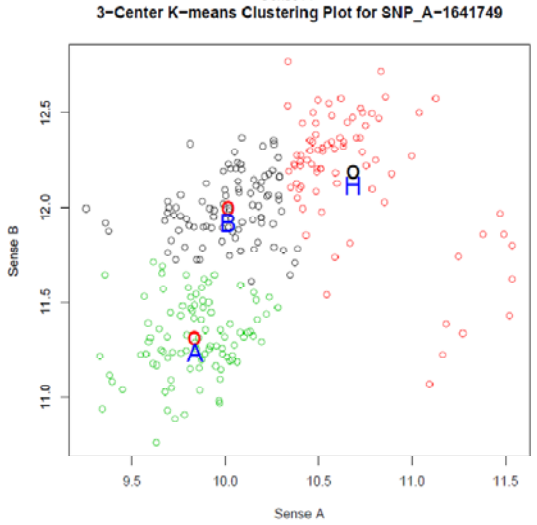


Results: Genotype Calling based on K-means with Relative Slope Classifier



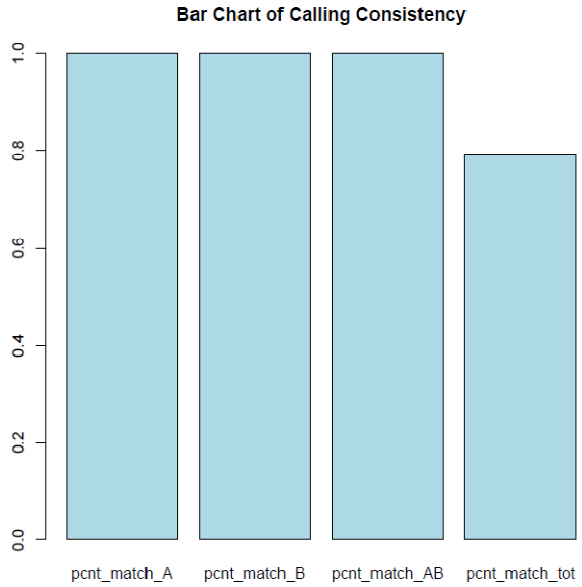


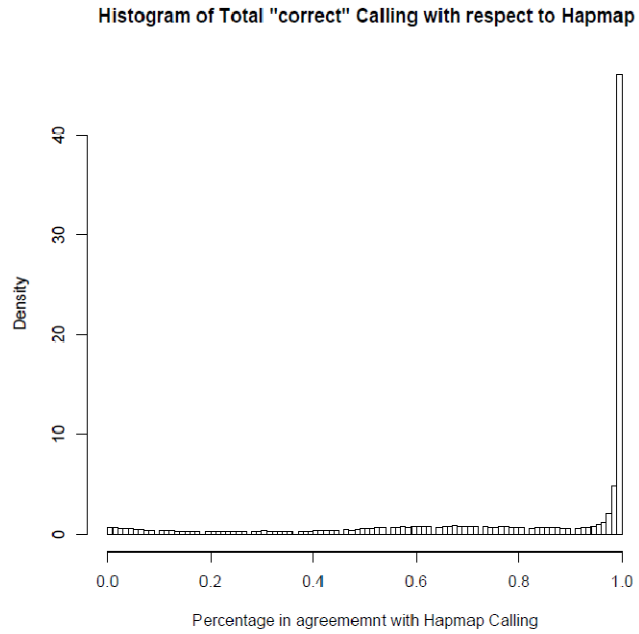




Results: K-means Performance over all HapMap SNPs

Relative Slope Classifier





Summary

- SNPs
  - Contribute to disease and reaction to drugs
  - Characterize 80% of human genetic variations
- Genotype Calling
  - Genotyping or “calling” of SNPs
  - Microarray technology being employed (Affymetrix Chip)
- A Clustering Problem
  - Need to cluster microarray data to indentify SNP alleles
- Solutions and Results
  - Sector-sweep Clustering
  - K-means Clustering
  - Comparison with HapMap Genotyped data

UCLA

Questions?

Thanks for your patience!