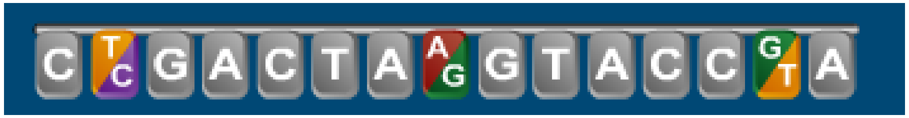
Haplotype Assembly

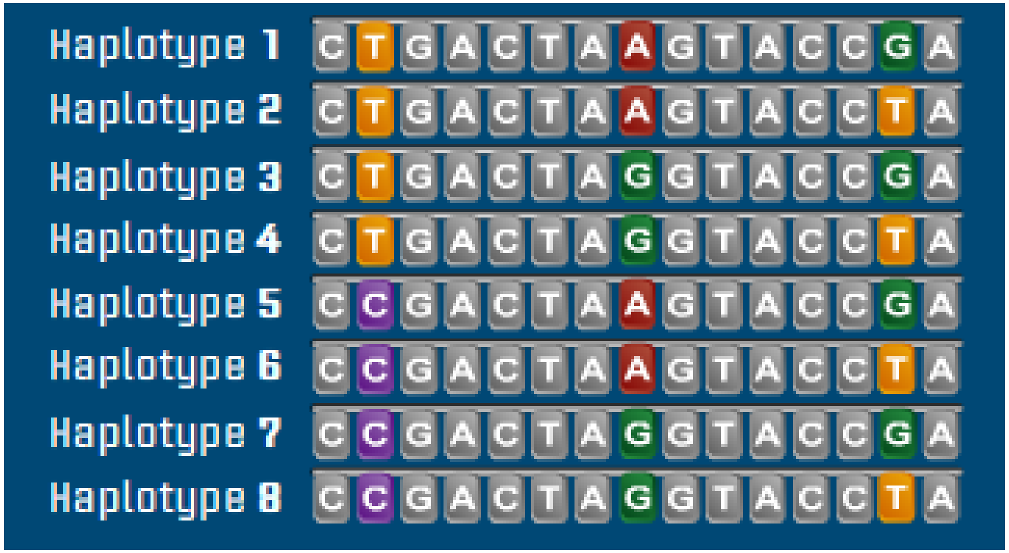
Neil Marion

Background

- 99.1% of DNA common, variations are SNPs
- SNPs - Single Nucleotide Polymorphisms



- Haplotypes - Possible Combination of SNPs

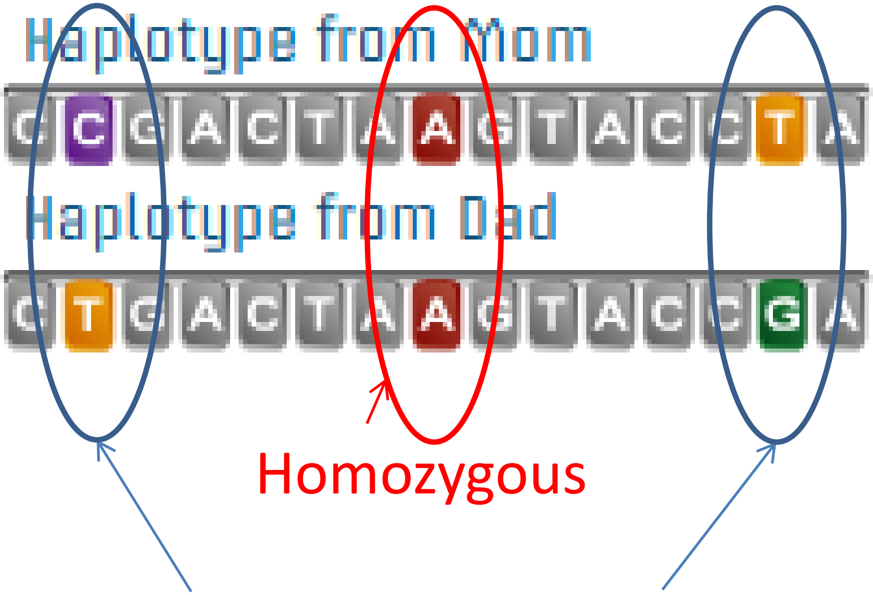


Haplotype 1	C	T	G	A	C	T	A	A	G	T	A	C	C	G	A
Haplotype 2	C	T	G	A	C	T	A	A	G	T	A	C	C	T	A
Haplotype 3	C	T	G	A	C	T	A	G	G	T	A	C	C	G	A
Haplotype 4	C	T	G	A	C	T	A	G	G	T	A	C	C	T	A
Haplotype 5	C	C	G	A	C	T	A	A	G	T	A	C	C	G	A
Haplotype 6	C	C	G	A	C	T	A	A	G	T	A	C	C	T	A
Haplotype 7	C	C	G	A	C	T	A	G	G	T	A	C	C	G	A
Haplotype 8	C	C	G	A	C	T	A	G	G	T	A	C	C	T	A

Images courtesy of:
http://learn.genetics.utah.edu/content/pharma/snips/

Background

- People have 2 Haplotypes from homologous chromosomes, one from each parent
- The Haplotype pair sites can be homozygous or heterozygous



Homozygous

Heterozygous

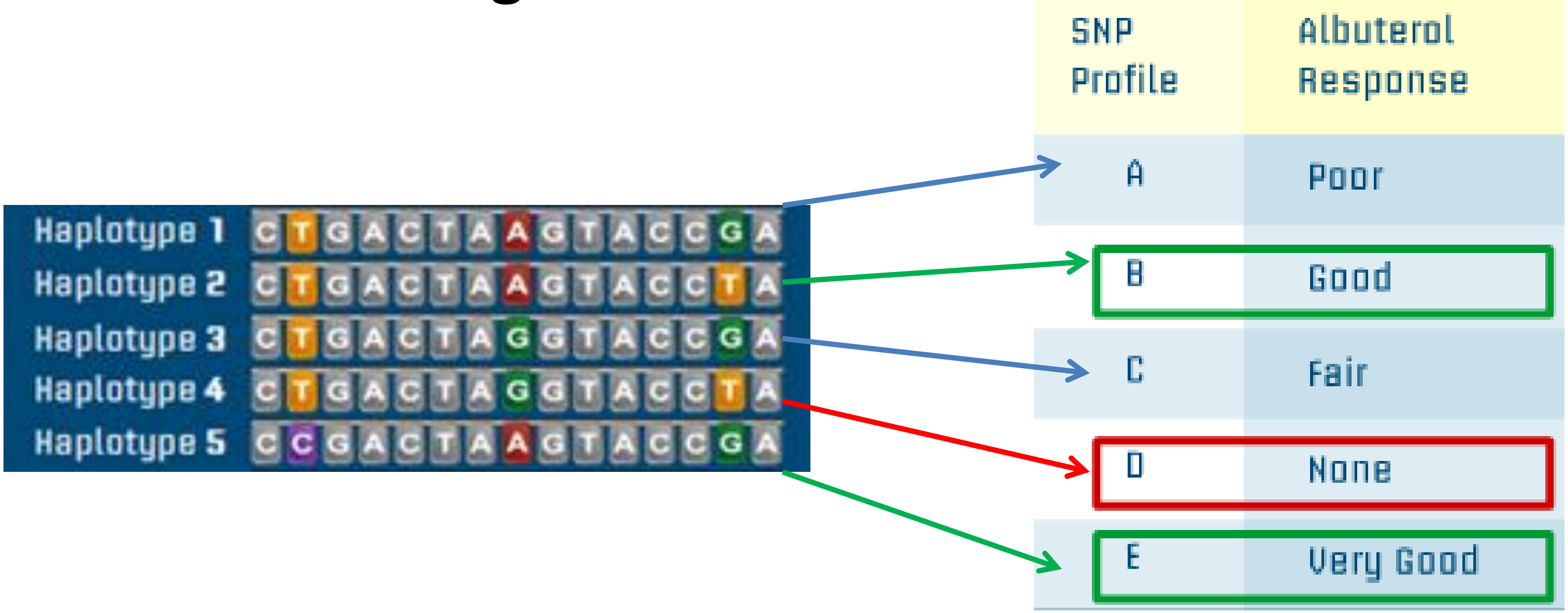
Images courtesy of:
http://learn.genetics.utah.edu/content/pharma/snips/

Biological Problem

- Gene sequencers generate reads from both haplotypes
- Reads from the same region can contain data from either haplotype, making it difficult to determine which haplotype an individual read is from
- The goal is to reassemble the reads into the separate haplotypes using a greedy algorithm based on the differences in heterozygous sites

Bio Problem - Motivation

- Why assemble the haplotypes?
- Consider the earlier haplotype SNP profiles and the drug albuterol used to treat asthma



Haplotype 1 C T G A C T A A G T A C C G A
Haplotype 2 C T G A C T A A G T A C C T A
Haplotype 3 C T G A C T A G G T A C C G A
Haplotype 4 C T G A C T A G G T A C C T A
Haplotype 5 C C G A C T A A G T A C C G A

SNP Profile	Albuterol Response
A	Poor
B	Good
C	Fair
D	None
E	Very Good

Images courtesy of: http://learn.genetics.utah.edu/content/pharma/snips/

Computational Problem

- Input: *N* x *M* Read Matrix

SNPs[M]

Reads[N]

_	_	1	1	0
_	_	_	_	0
_	_	0	0	1
_	_	_	0	1
1	1	0	_	_

- Output: Complementary haplotypes of length M

H1 =	1	1	0	0	1
H2 =	0	0	1	1	0

Benchmarks:

- Computational time to analyze the read matrix
- Accuracy of output - % of the solution haplotype that matches the actual haplotype

Simulating the Read Matrix

- Generate random positions of a certain number of heterozygous SNPs along a sequence of a certain length. Randomly assign each SNP to ‘0’ or ‘1’
  - For each read, randomly choose a start position in the sequence and read a certain number of positions (read length). Randomly decide which haplotype the read is coming from
    - Ignore homozygous SNPs and common pairs, looking only for heterozygous SNP sites
    - To assemble the read, mark unread SNP sites with ‘_’
  - Combine all the reads together to form the completed read matrix

Baseline Method

1. Given M SNPs with possible values 0 or 1, 2^M possible haplotypes exist

0	0	0	0	0
0	0	0	0	1
0	0	0	1	0
0	0	0	1	1
.	.	.	.	.

2. Compare each of the 2^M possible haplotypes against the read matrix, discarding those that conflict with both the read and the complement of the read

Benefits: If the reads overlap and cover all the SNPs, will eventually find the correct solution

Disadvantages: SLOW and inefficient. Each of the 2^M solutions may compare with all N of the reads

Greedy Method

1. Sort the read matrix by first observed SNP position

1	1	0	_	_
_	0	1	_	_
_	_	0	0	_
_	_	_	0	1
_	_	_	_	0

2. In order, proceed down the sorted matrix and compare the overlaps of the reads one by one
   – If the overlaps matches, combine the reads. If the overlaps don’t match, combine the 1st read with the complement of the 2nd read

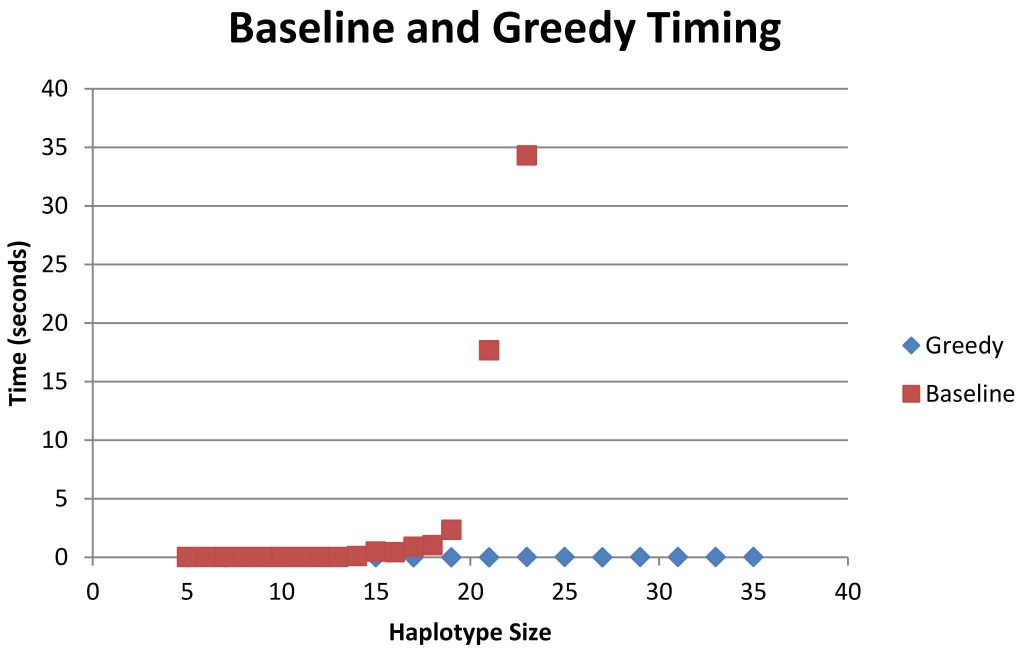
Benefits: If the reads overlap and cover all the SNPs, will find the correct/optimal solution. FASTER than the baseline method

Disadvantages: Doesn’t account for read errors (flipped value in an individual read)

Analysis - Speed

1. Baseline – Exponential
   - $2^M * 2N$ possible comparisons
   - Inefficient

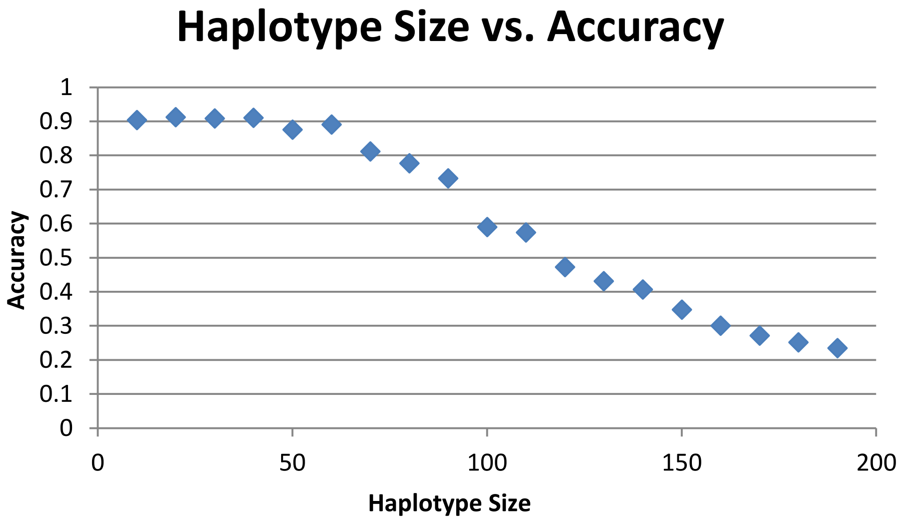
2. Greedy – Linear
   - Sorting done in single pass through matrix (more like rearranging)
   - Haplotype assembly through read comparisons also done in single pass



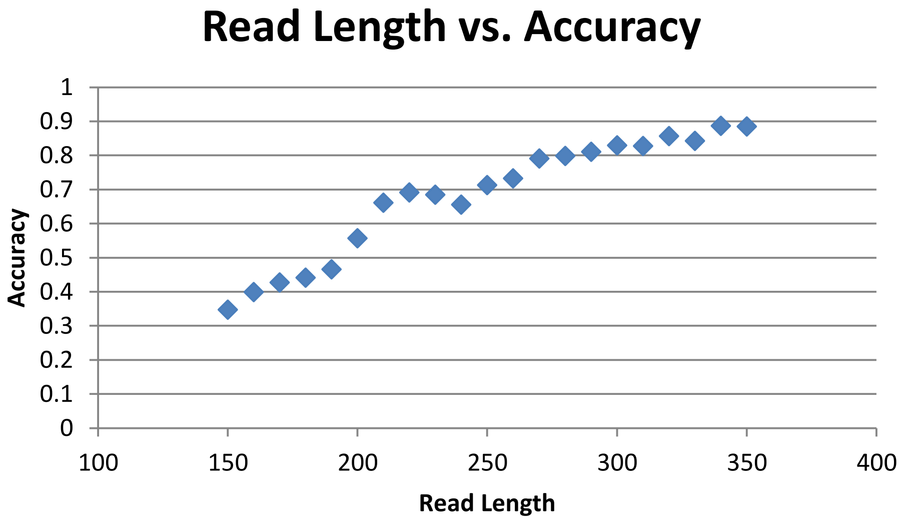
Analysis – Greedy Accuracy

1. Varying the size of the haplotype
   - Read length and number of reads kept constant
   - More reads have to overlap to cover the entire haplotype

2. Increasing the read length
   - Haplotype size and number of reads held constant
   - Longer reads means more overlaps occur



Haplotype Size vs. Accuracy



Read Length vs. Accuracy

Observations

- When the reads are extensive enough, greedy provides the correct and optimal solution
- However, when not enough overlaps are present in the reads, greedy is reduced to guessing which haplotype the read came from
- Number of previous reads from a given haplotype doesn’t change the probability that the next read is from that haplotype
  - $P(R_{10} \text{ from } H_1 \mid R_1 – R_9 \text{ from } H_1) = P(R_i \text{ from } H_1)$