Haplotypes, linkage disequilibrium, and the HapMap

Jeffrey Barrett

Boulder, 2009

wellcome trust sanger institute

Outline

1. Haplotypes

2. Linkage disequilibrium

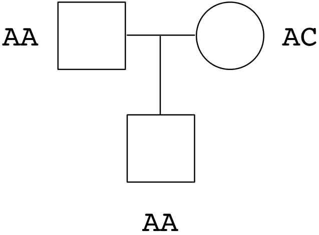
3. HapMap

4. Tag SNPs

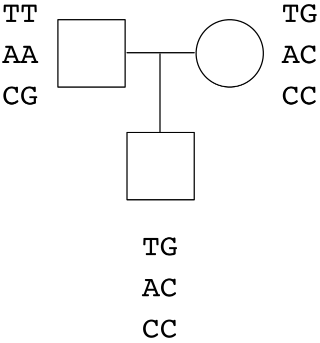
Outline

1. Haplotypes
2. Linkage disequilibrium
3. HapMap
4. Tag SNPs

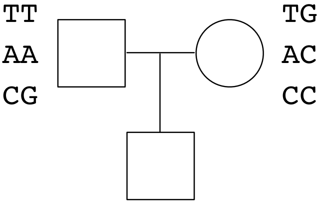
The problem: humans are diploid



The problem: humans are diploid



The problem: humans are diploid



T | G
A | C
C | C

or

T | G
C | A
C | C

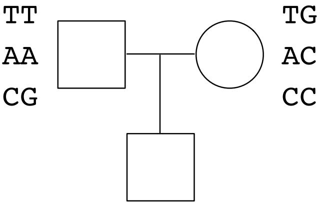
How can we resolve phase?

- Using experimental methods. (complicated and expensive)

How can we resolve phase?

- Using experimental methods. (complicated and expensive)
- Using pedigree information.

How can we resolve phase?

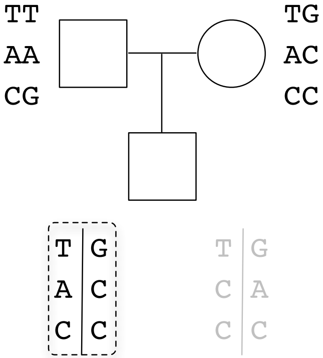


T | G
A | C
C | C

or

T | G
C | A
C | C

How can we resolve phase?



How can we resolve phase?

- Using experimental methods. (complicated and expensive)
- Using pedigree information. (requires family data)

How can we resolve phase?

- Using experimental methods. (complicated and expensive)
- Using pedigree information. (requires family data)
- Using a statistical algorithm applied to data from a population sample.

The EM algorithm

Consider a segment of the genome, with N haplotypes in the population, with frequencies $\theta_1 \ldots \theta_N$. If we have a sample of M genotypes, g, we can represent the probability of those genotypes as a sum of the probabilities of each possible pair of haplotypes which could give rise to the observed genotypes:

$$2\theta_j\theta_k \qquad j \neq k$$

$$\theta_j{}^2 \qquad j = k$$

The probability of all the genotype data is then:

$$P(g|\theta) = P(g_1)P(g_2)\ldots P(g_M)$$

The EM algorithm

Of course, we don't actually know the haplotype frequencies (θ), so we need to use a numerical algorithm to find a maximum likelihood estimate of those estimates, given the genotype data.

We do this in an iterative approach with two steps:

1. Expectation, where we compute the expected number of haplotypes each individual carries, given the haplotype frequencies.
2. Maximization, where we re-estimate the haplotype frequencies given the haplotype counts from the 'E' step.

Estimate for θ tends to converge relatively quickly. We can now make guesses about each individual's haplotypes based on these frequencies, but they're not guaranteed to be accurate!

Aspects of EM vs. alternatives

- Fast, and easy to implement.
- Works best when markers in question are relatively strongly correlated (which in practice is only in short genomic segments).
- Better for estimating population frequencies than individual haplotypes.
- Doesn't handle uncertainty in estimates very well.
- More complicated models, including that in the PHASE program, address some of these issues by incorporating knowledge of how haplotypes segregate in populations (the 'coalescent model')

Outline

1 Haplotypes

2 Linkage disequilibrium

3 HapMap

4 Tag SNPs

What creates genetic variation?

The two processes which increase genetic diversity in a population are mutation, which introduces novel variants into the population, and recombination, which re-shuffles the existing patterns of variation (haplotypes).

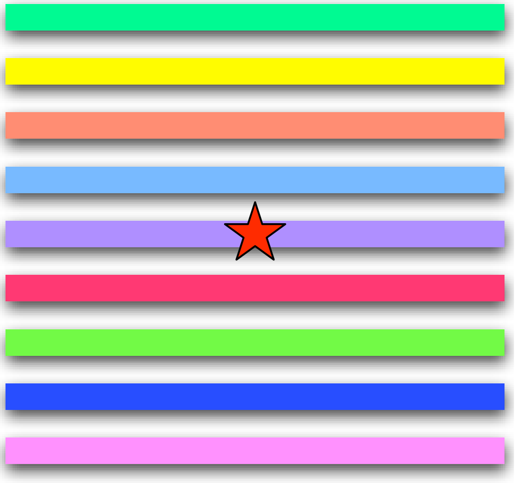
What creates genetic variation?

The two processes which increase genetic diversity in a population are mutation, which introduces novel variants into the population, and recombination, which re-shuffles the existing patterns of variation (haplotypes).

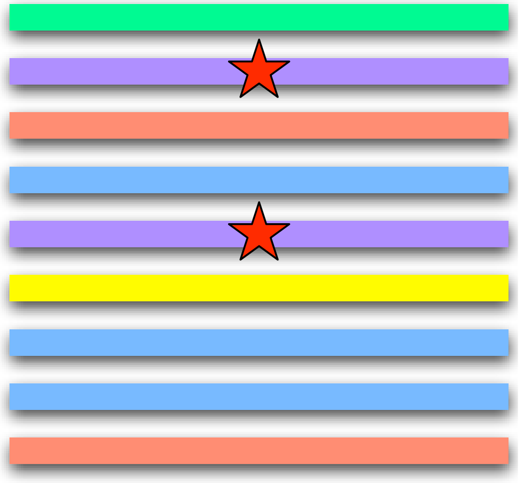
The fate of new mutations is also affected by drift, selection, and population history. What we really care about is what patterns are left behind in genetic variation because of these forces, and how they affect disease studies.

Mutation and recombination in a population

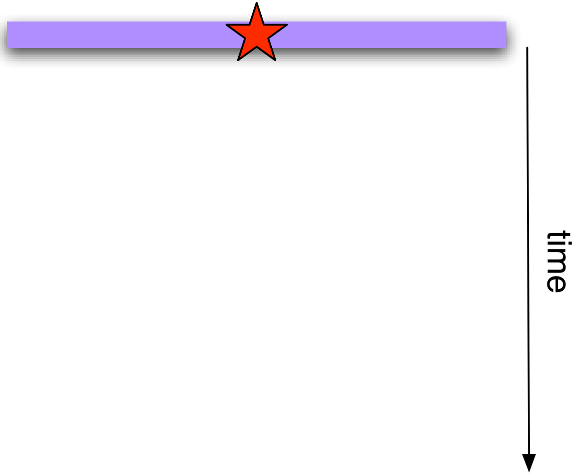
Mutation and recombination in a population



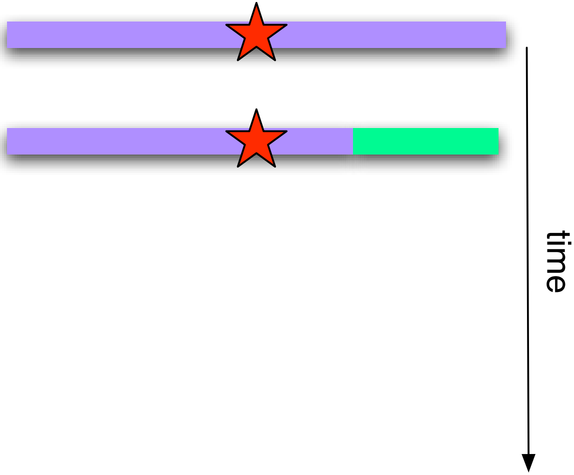
Mutation and recombination in a population



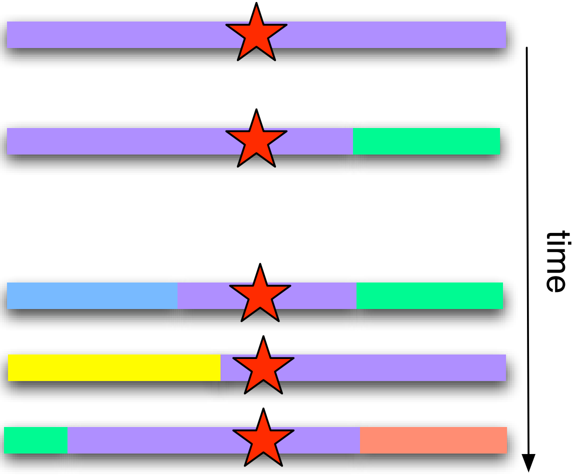
Mutation and recombination in a population



Mutation and recombination in a population



Mutation and recombination in a population



Consequences of mutation and recombination

- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.

Consequences of mutation and recombination

- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.
- In the absence of recombination this correlation would never be broken down and would extend a great distance along chromosomes.

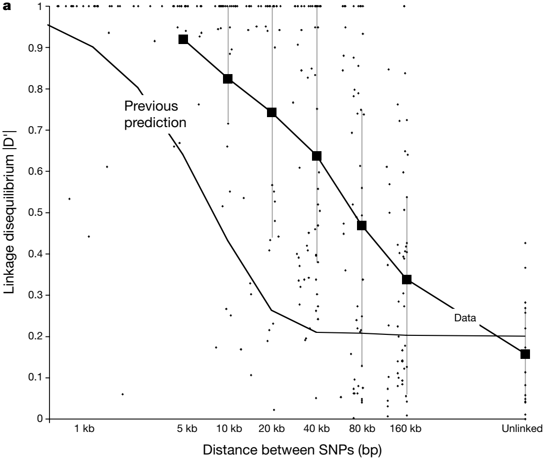
Consequences of mutation and recombination

- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.
- In the absence of recombination this correlation would never be broken down and would extend a great distance along chromosomes.
- Recombination breaks down this correlation over many successive generations, leaving a narrower and narrower window of correlation.

Consequences of mutation and recombination

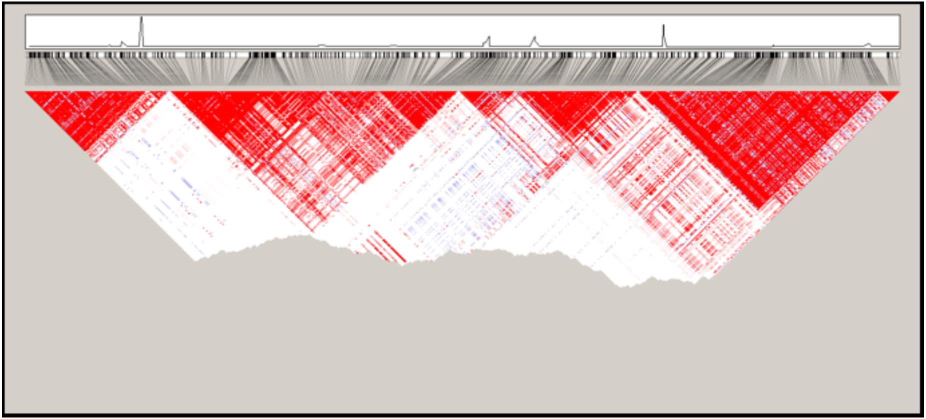
- Genetic variants are correlated because they occur on a particular haplotype background, and segregate in populations on that background.
- In the absence of recombination this correlation would never be broken down and would extend a great distance along chromosomes.
- Recombination breaks down this correlation over many successive generations, leaving a narrower and narrower window of correlation.
- Under certain assumptions (neutral evolution, random mating, homogenous recombination), we can model exactly how far this correlation should extend.

Theoretical vs. empirical patterns of LD



Reich et al, *Nature*, 2001.

Heterogeneous recombination drives observed LD patterns



Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	pq	q(1-p)
	1-q	p(1-q)	(1-p)(1-q)

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

$$D = \pi_{11} - pq$$

Quantifying LD

		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

$$D = \pi_{11} - pq$$

$$D' = D/D_{\text{max}}$$

Quantifying LD

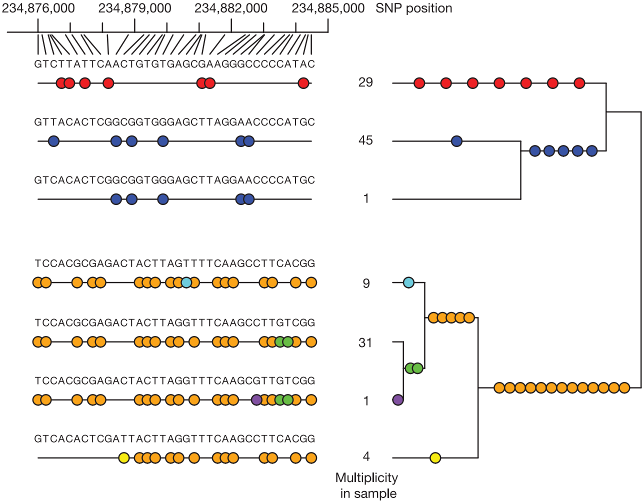
		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

$$D = \pi_{11} - pq$$

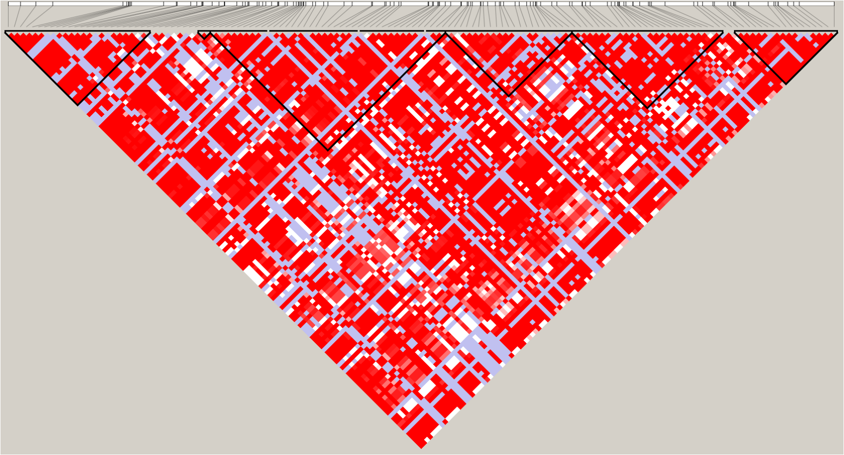
$$D' = D/D_{\text{max}}$$

$$r^2 = D/p(1-p)q(1-q)$$

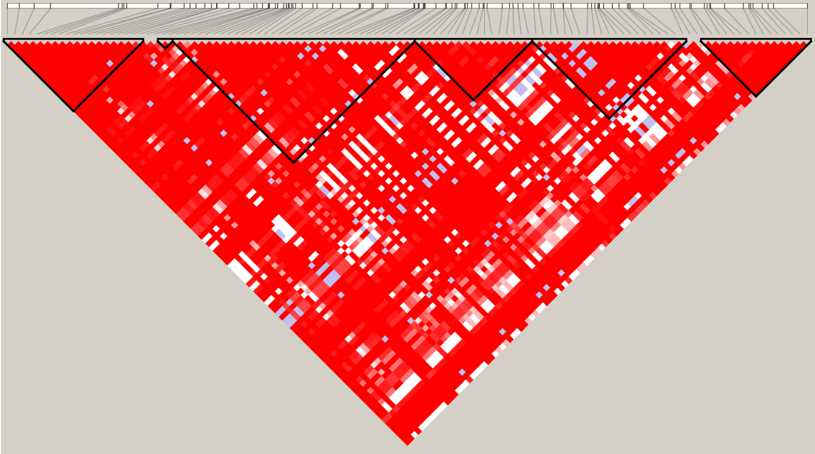
D' and r^2


234,876,000
234,879,000
234,882,000
234,885,000
SNP position
GTCTTATTCAACTGTGTGAGCGAAGGGCCCCCATAC
29
GTTACACTCGGCGGTGGGAGCTTAGGAACCCCATGC
45
GTCACACTCGGCGGTGGGAGCTTAGGAACCCCATGC
1
TCCACGCGAGACTACTTAGTTTTCAAGCCTTCACGG
9
TCCACGCGAGACTACTTAGGTTTCAAGCCTTGTCGG
31
TCCACGCGAGACTACTTAGGTTTCAAGCGTTGTCGG
1
GTCACACTCGATTACTTAGGTTTCAAGCCTTCACGG
4
Multiplicity
in sample

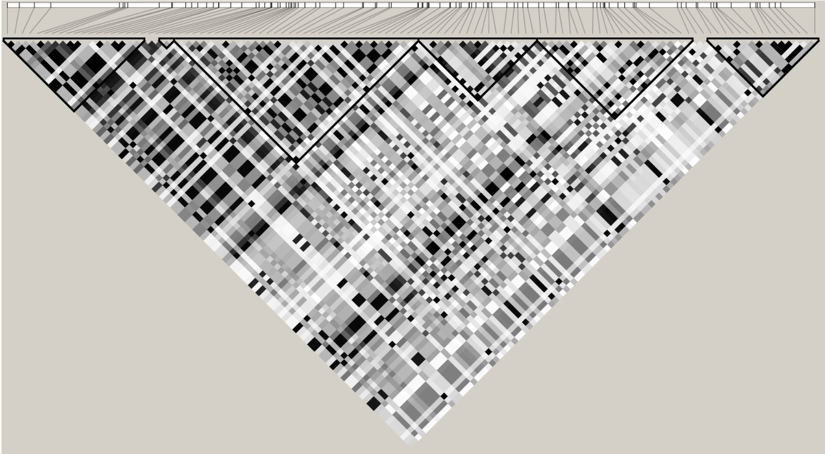
D' in a region of 100kb



D' for common SNPs in a region of 100kb



r^2 for common SNPs in a region of 100kb



Outline

1. Haplotypes

2. Linkage disequilibrium

3. HapMap

4. Tag SNPs

A haplotype map of the human genome

27 October 2005 | www.nature.com/nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE HAPMAP PROJECT
Chapter and verse on human genetic variation

Project details (Phase I/II)

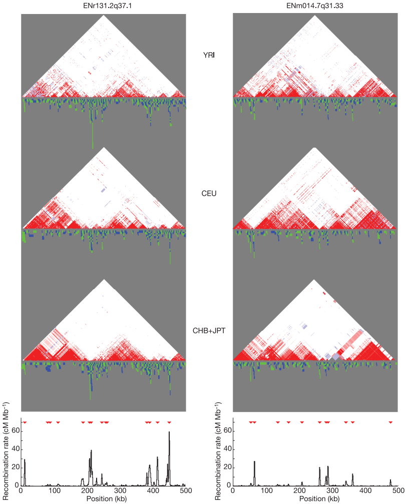
Samples:

- 90 Yoruba (30 parent-parent-offspring trios) from Ibadan, Nigeria (YRI)
- 90 CEPH samples (30 trios) of European descent from Utah (CEU)
- 45 Han Chinese from Beijing (CHB)
- 45 Japanese from Tokyo (JPT)

SNPs: Original goal was 1 SNP every 5kb, but as genotyping costs dropped, eventual catalogue included approximately 4 million polymorphic SNPs scattered across the genome.

Panel	% $r^2 > 0.8$	mean max r^2
YRI	81	0.90
CEU	94	0.97
CHB+JPT	94	0.97

Why multiple populations?

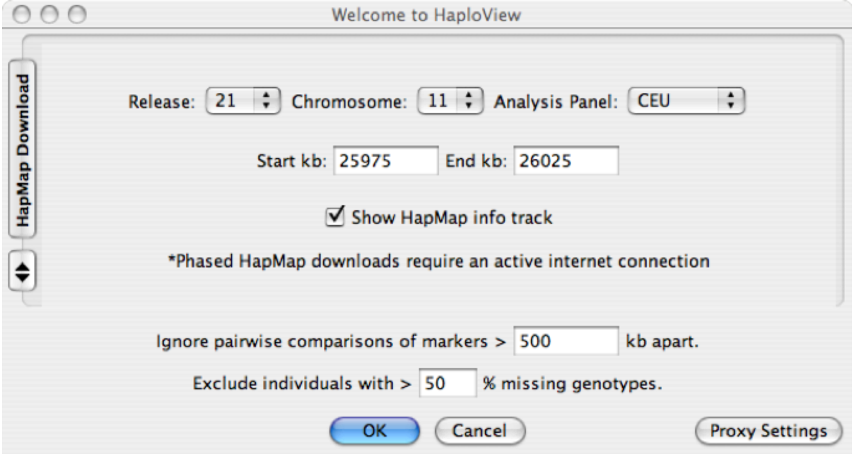


Project details (Phase III)

- African ancestry in Southwest USA (ASW)
- Chinese in Denver, CO (CHD)
- Gujarati Indians in Houston, TX (GIH)
- Luhya in Webuye, Kenya (LWK)
- Mexican ancestry in Los Angeles, CA (MEX)
- Maasai in Kinyawa, Kenya (MKK)
- Toscans in Italy (TSI)
- (additional samples from CEU, YRI, JPT, CHB)

$\sim$ 1.5 million SNPs

Accessing HapMap data with Haploview



Release: 21 Chromosome: 11 Analysis Panel: CEU

Start kb: 25975 End kb: 26025

☑ Show HapMap info track

*Phased HapMap downloads require an active internet connection

Ignore pairwise comparisons of markers > 500 kb apart.

Exclude individuals with > 50 % missing genotypes.

Outline

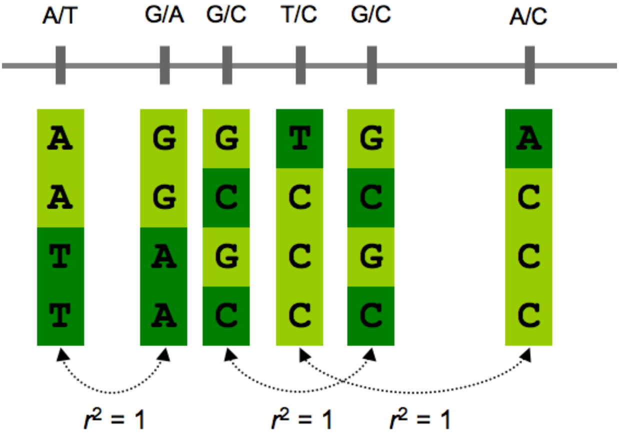
1. Haplotypes

2. Linkage disequilibrium

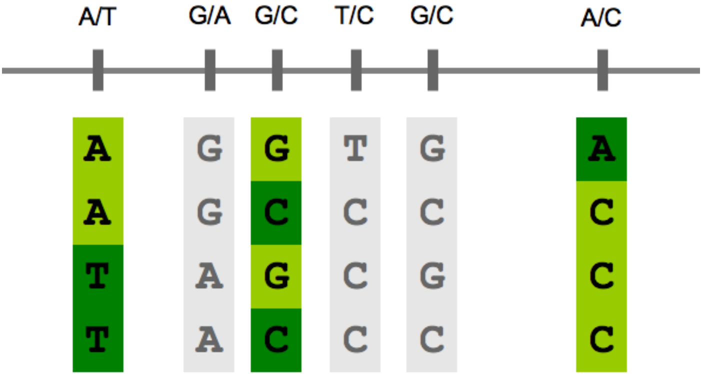
3. HapMap

4. Tag SNPs

How can we use HapMap knowledge for disease studies?

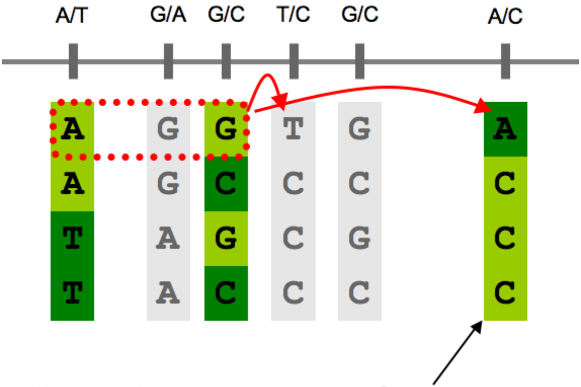


Gain efficiency by removing redundant SNPs



A/T	G/A	G/C	T/C	G/C	A/C
A	G	G	T	G	A
A	G	C	C	C	C
T	A	G	C	G	C
T	A	C	C	C	C

Haplotypes can yield additional gains in efficiency



A/T	G/A	G/C	T/C	G/C	A/C
A	G	G	T	G	A
A	G	C	C	C	C
T	A	G	C	G	C
T	A	C	C	C	C

No need to genotype this SNP