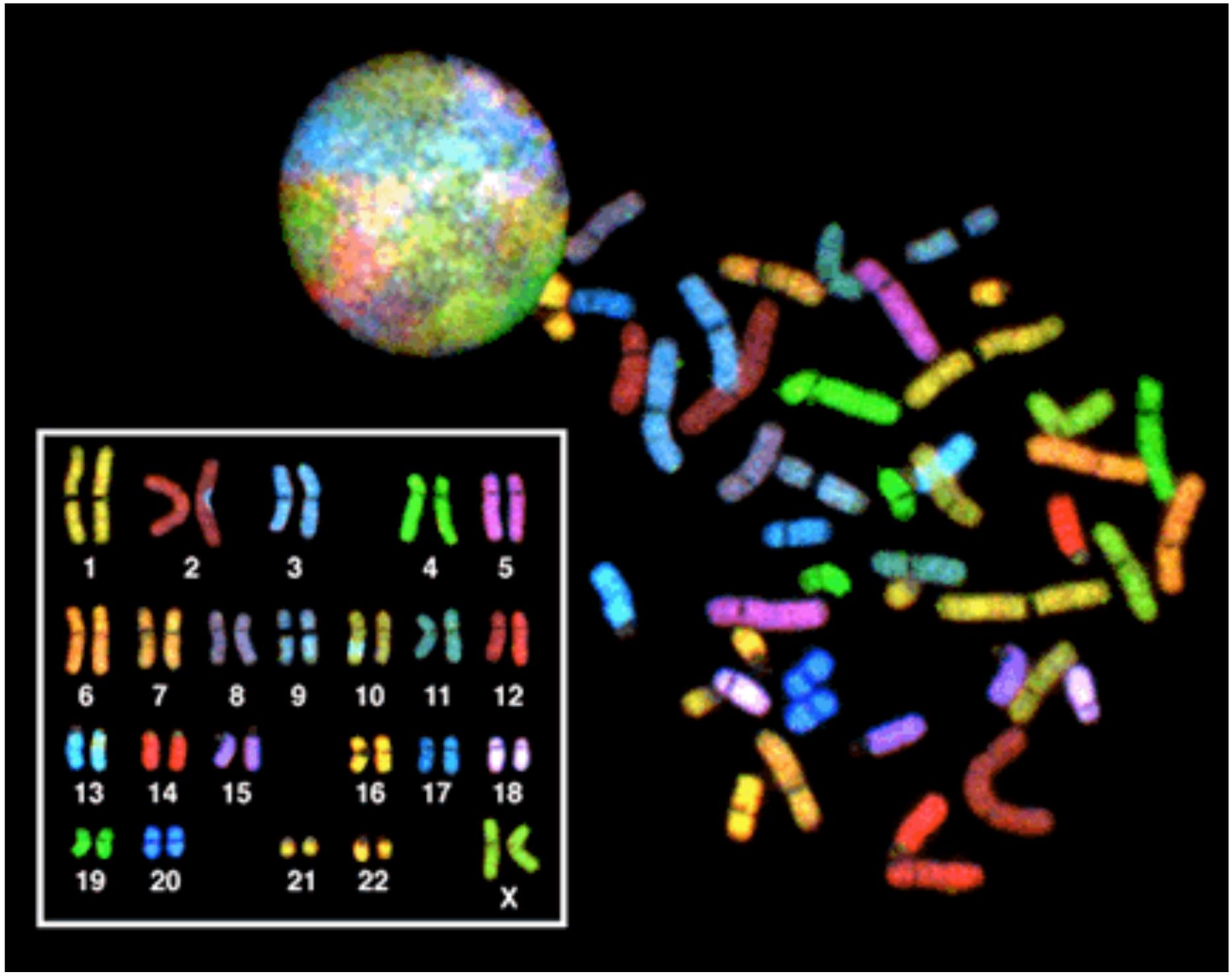
BMI/CS 776
Lecture #30
Genotypes & Haplotypes

Colin Dewey
May 8, 2008

Human genome



Genomic variation

- Single Nucleotide Polymorphisms (SNPs)
- Tandem repeat length variation (e.g. micro-satellites)
- Transposable elements (e.g., Alu elements)
- Large-scale structural variation
  - Copy number variants (CNVs)
  - Inversions
  - Transpositions

SNPs

- Position in genome that varies in ≥ 1% of the population
- *Alleles*: the variants at a specific position
  - Usually only 2 alleles: *ancestral* and *derived*
- The genomes of two random people differ at roughly 1 out of every 1,000 base pairs
  - We are >99.9% identical
  - We are ~98-99% identical to chimpanzees

dbSNP

- http://www.ncbi.nlm.nih.gov/projects/SNP/
- Central repository for all variation information
- ~6.9 SNPs in build 122, (56,789 non-synonymous)
- Other variations: insertions, deletions, etc.

HapMap

- http://www.hapmap.org
- International consortium to genotype a large number of individuals
- 269 samples, four populations
  - Yoruba (Nigeria), Northern/Western European (Utah), Han Chinese, Japanese
- 3.1 million SNPs genotyped in Phase II
- Average spacing is one SNP per 875 bp

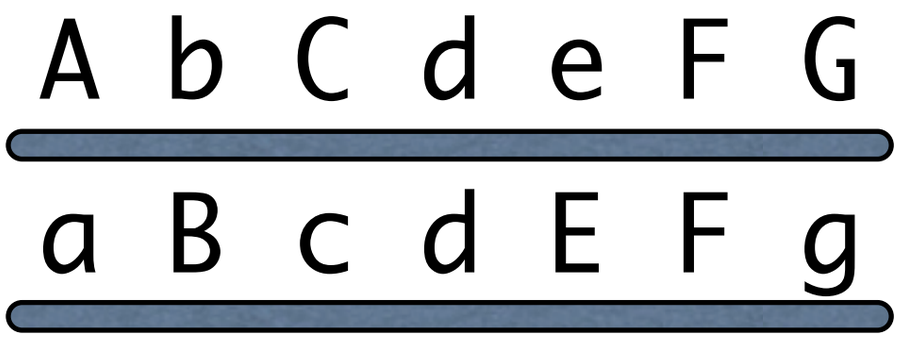
What to do with SNP data?

- Genome-wide association studies
  - Find SNPs or sets of SNPs that are associated with diseases
- Study human evolution
  - Are there any SNPs that are involved in recent adaptations?
- Genotype - Phenotype relationships
  - how does variation affect gene expression?

Genotypes and Haplotypes

- *Genotype*: The set of alleles in the genome of an individual
  - A *diploid* individual has two alleles per locus (i.e., a gene or single nucleotide)
- *Haplotype*: The set of alleles on a single chromosome

Haplotypes

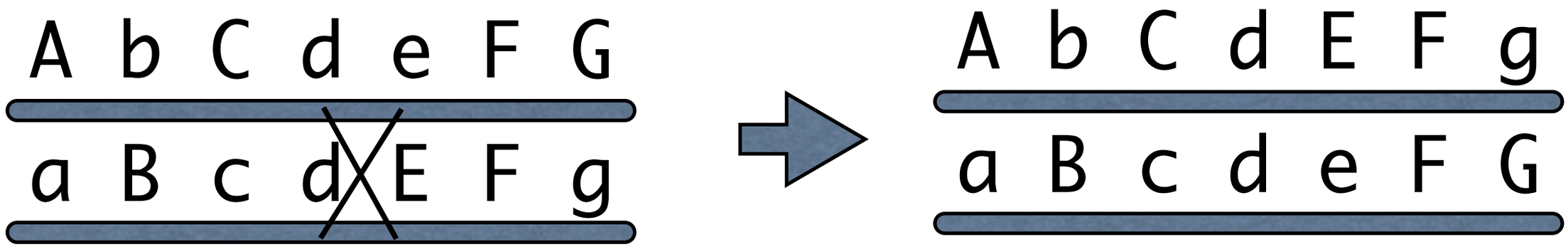


Genotype

Aa Bb Cc dd Ee FF Gg

Recombination and haplotype blocks

- *Recombination* shuffles combinations of alleles on single chromosomes



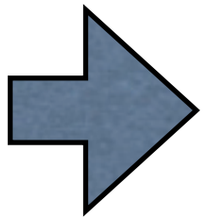
- *Haplotype blocks:* Intervals of the genome in which little recombination has occurred.
  - Combinations of alleles within these blocks tend to stay together

Linkage Disequilibrium (LD)

- When a new SNP arises it is initially strongly associated with the neighboring alleles on the chromosome
- Recombination slowly degrades these associations
- Observation of combination of alleles in haplotype more frequent than expected by chance → *linkage disequilibrium*

The Phasing Problem

- Difficult to get haplotype data directly
- Easy to get genotype
- Haplotype reconstruction (phasing) problem:
  - Given: genotypes of *n* individuals
  - Do: predict the 2*n* haplotypes

Genotypes		Haplotypes
Aa Bb Cc dd Ee FF Gg AA Bb cc DD Ee Ff GG Aa bb CC Dd ee FF Gg		?

Infinite-sites model

- Number of sites large compared to mutation rate and time period considered
- At a given SNP site:
  - mutation occurred at that site only once in human history
  - no back or multiple mutation at that site

Perfect Phylogeny

- Assuming infinite-sites model and no recombination → haplotypes described by *perfect phylogeny*
  - Each edge labeled by site(s) that mutated along that edge
  - Each polymorphic site labels *exactly one* edge

Test for perfect phylogeny

- Definition (compatibility):
  - A set C of complete cladistic characters (e.g., SNPs) are **compatible** if it allows for a perfect phylogeny
- Proposition (pairwise compatibility):
  - A set C of complete cladistic characters is compatible if and only its elements are pairwise compatible

Binary characters

- Convenient to encode characters (e.g. SNPs) as 0, 1
- Directed binary characters: 0 is the ancestral state and 1 is derived state

Theorem (Estabrook et al.) A set C of complete directed binary cladistic characters is compatible if and only if the 1-states of every pair of characters are either disjoint or one contains the other.

1-state of c: Set of all taxa for which the state of c is 1

Undirected characters

- Most often have binary undirected characters

Proposition (McMorris): Let C be a set of complete undirected binary cladistic characters. Let C' be the set of complete directed binary characters obtained from C by rooting each c in C at whichever state has the most taxa breaking ties arbitrarily. Then C is compatible if and only if C' is compatible.

Perfect phylogeny algorithm

PERFECTPHYLOGENY(*S*, *C*):

S: set of taxa

C: set of binary characters

sort characters by non-increasing cardinality of 1-states

initialize tree *T* as single node *v*, labeled *S*

for each s ∈ S **do**

lowest(s) ← v

for each c ∈ C **do**

choose some $s \in l_c$, let *v* = lowest(s)

if ∃ $t \in l_c$, s.t. *lowest*(*t*) ≠ v **then**

return characters are not compatible

else

create node *u,* labeled l_c, as child of *v*

for each $t \in l_c$ **do**

lowest(*t*) ← *u*

return *T*

(Agarwala et al.)

Perfect phylogeny → Phasing

- Haplotype reconstruction (phasing) problem:
  - Given: genotypes of *n* individuals
  - Do: predict the 2*n* haplotypes
- Ideal solution: Find 2n haplotypes that form a perfect phylogeny (assuming no recombination)