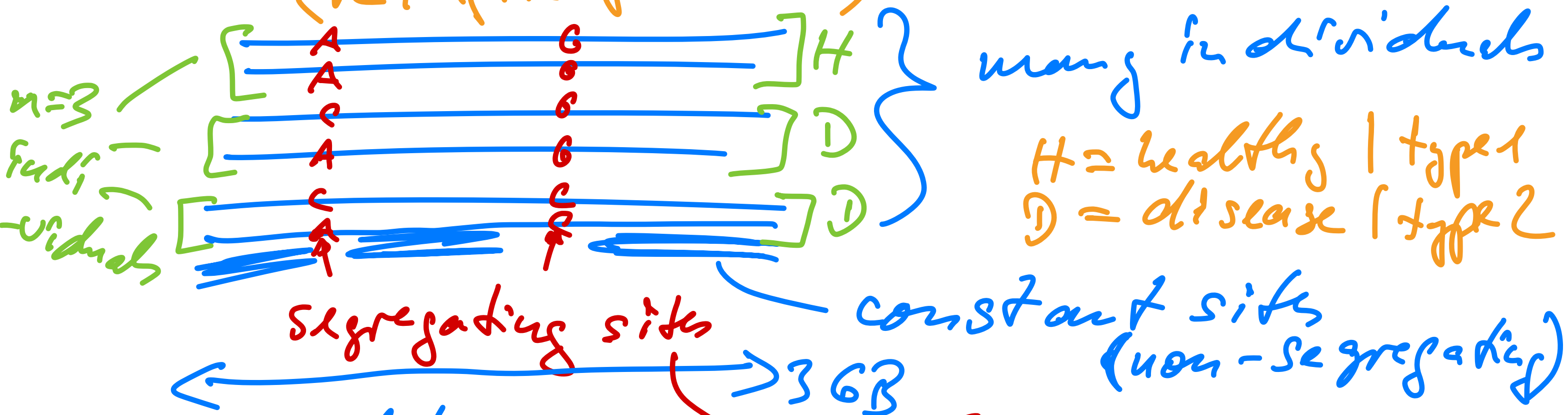
Jan 8th, 2021

SNV-disease association mapping

SNV: single nucleotide variant

(population genomics)

(VCF file format)



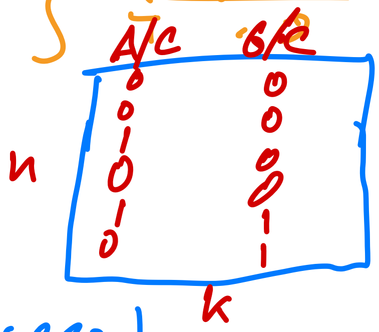
81 MB

typical data:

n = thousands of individuals

k = millions of segregating sites

1000 genomes project



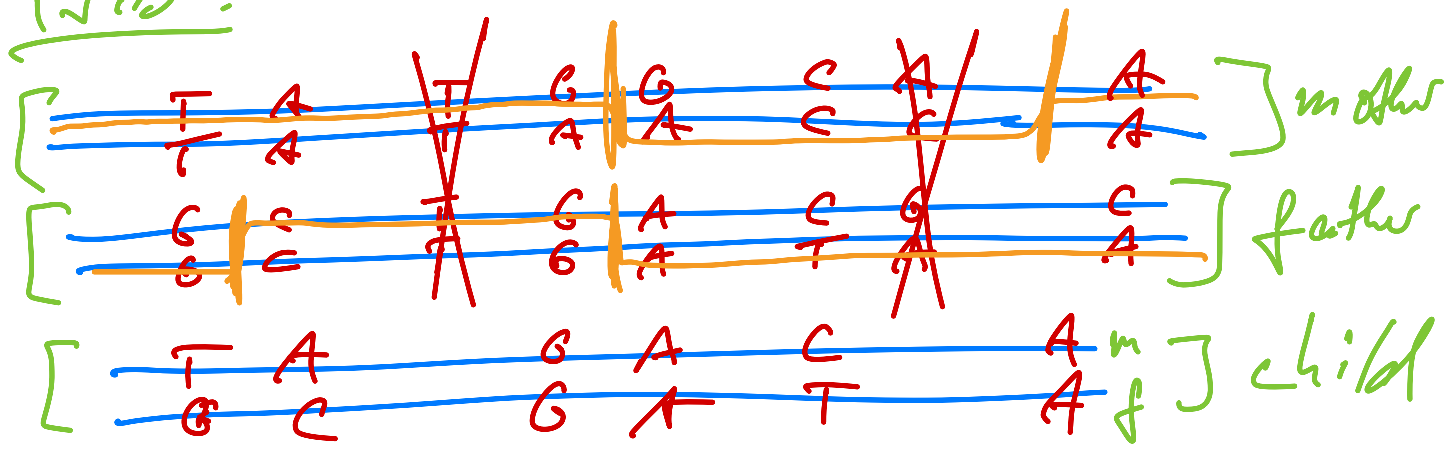
→ 4×10^{11} bits

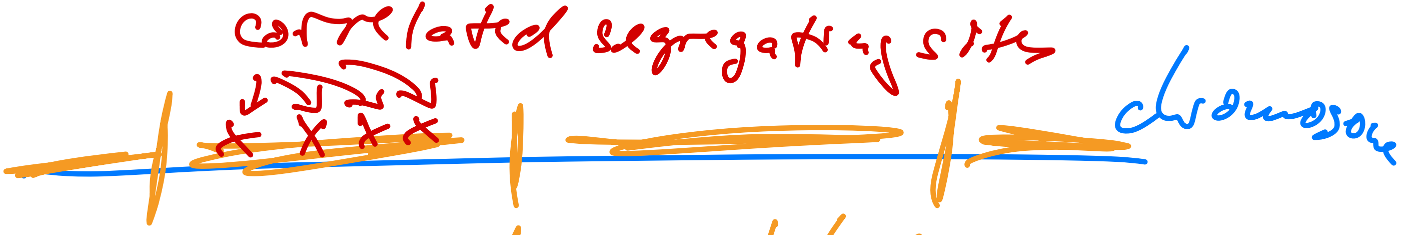
"big data"

(n = 2500 → 5000 haplotypes)

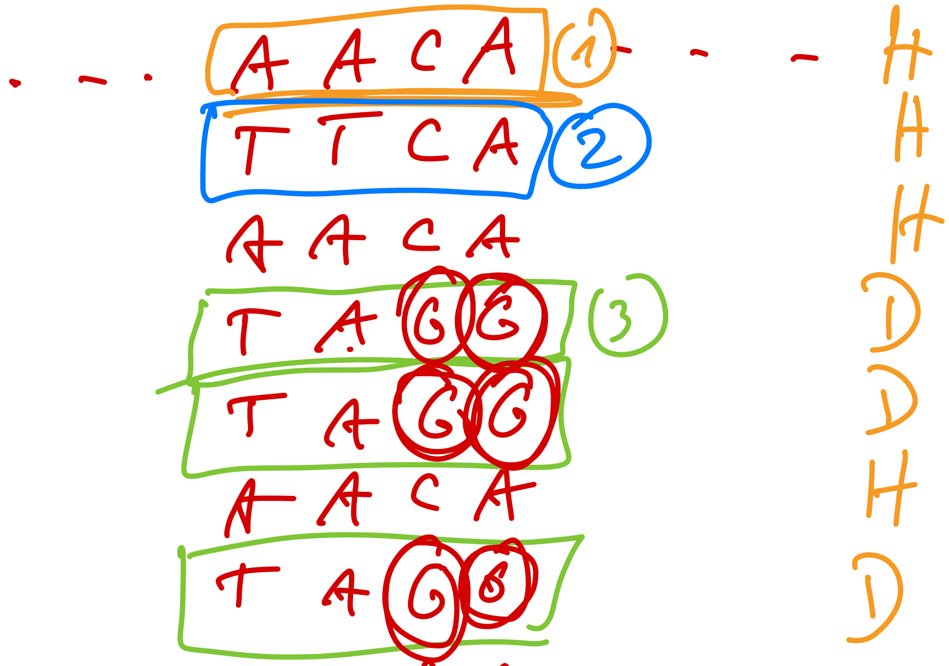
(k = 81.000.000 segregating sites)

Trio:





haplotype blocks



. . . AACA (1) - - - H

TTCA (2) H

AACA H

TAGG (3) D

TAGG D

AACA H

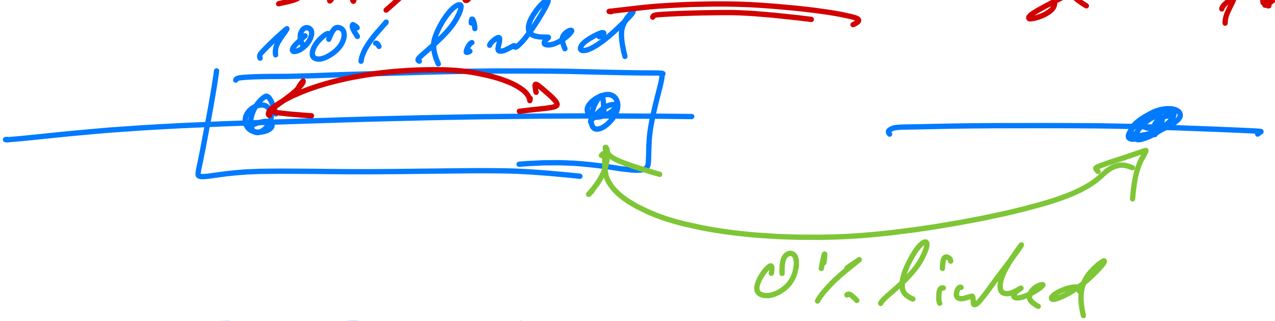
TAGG D

need: ~~- - C G~~

~~G~~ does not exist

term: linkage disequilibrium (LD)

"sites are i. absolute linkage disequilib."



"sites are in linkage equilibrium"

Problem:

Given an n x k (binary) haplotype matrix with p "cases" and q "healthy" individuals (p+q=n), find segregating sites that correlate with the case/disease pattern.

classical solutions

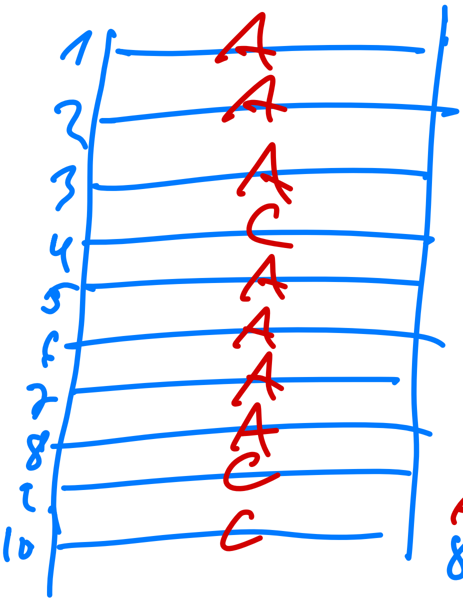
apply statistical tests, site by site.

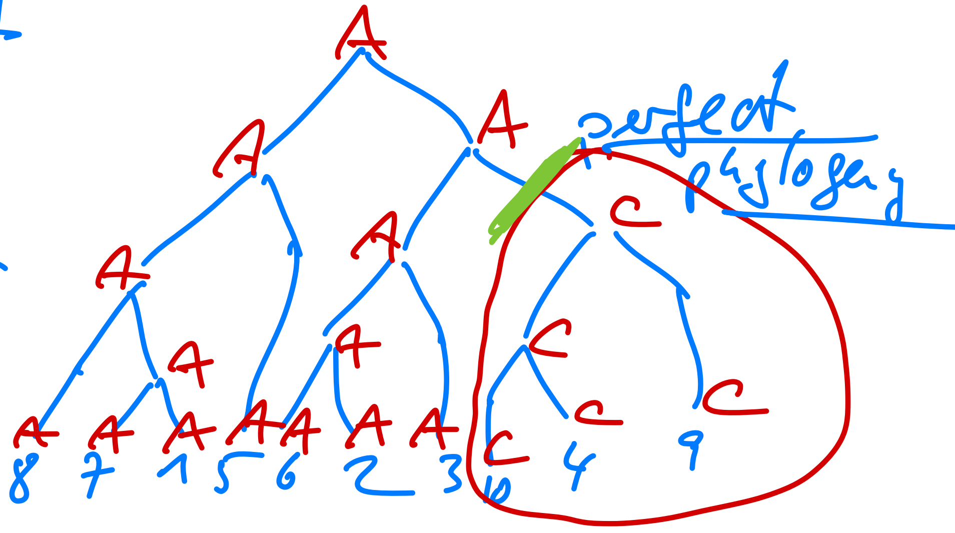
- Fisher's exact test
- chi-square test

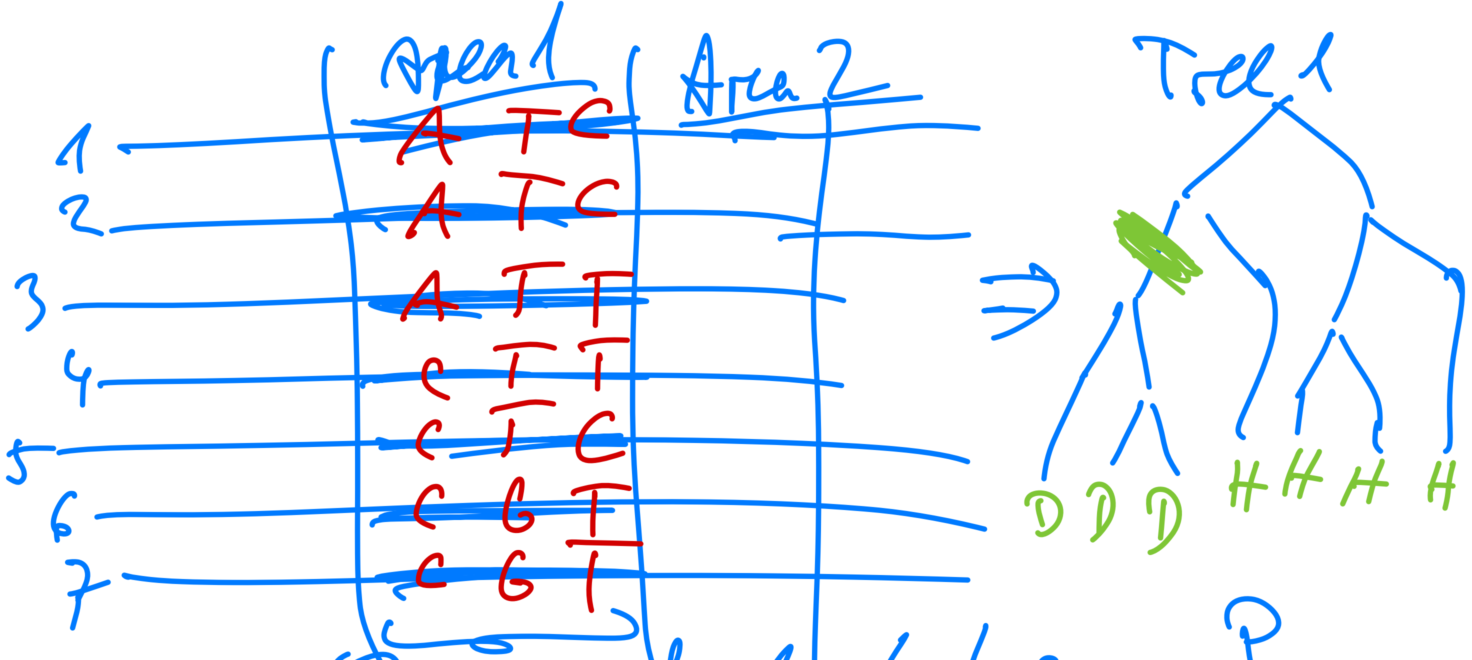
⇒ slow, especially for multiple sites

Another method (Blossoc): [infinite sites assumption]

first, focus on a single haplotype block (no crossing-over = recombination)





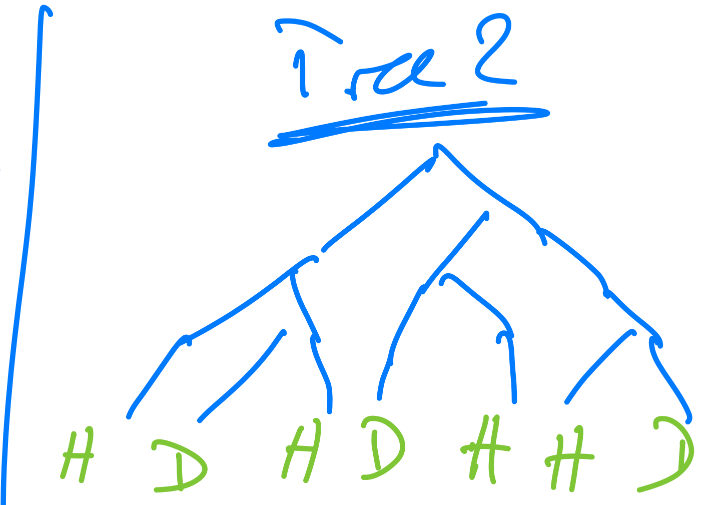


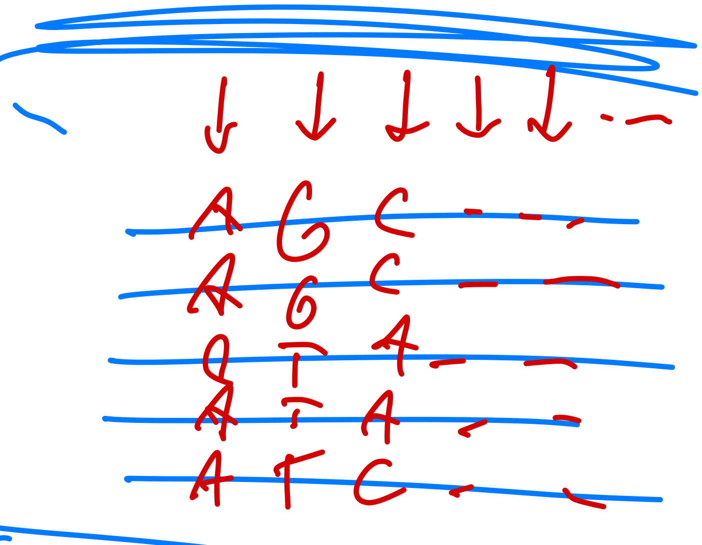
① has perfect phylogeny P

② P correlates with the genealogy of the disease

Algorithm:

1. Look at each SNV site from left to right.
2. From this, extend to the right without violating the 4-gamete test.
3. For each such region, build a perfect phylogeny (which must exist).
4. Compare all these trees with the case-disease pattern.


Tree 2
H D H D H H D


↓ ↓ ↓ ↓ ↓ . –
A G C – – –
A G C – –
? T A – –
A T A – –
A T C – –