“Inferring parental genomes from offspring's DNA”

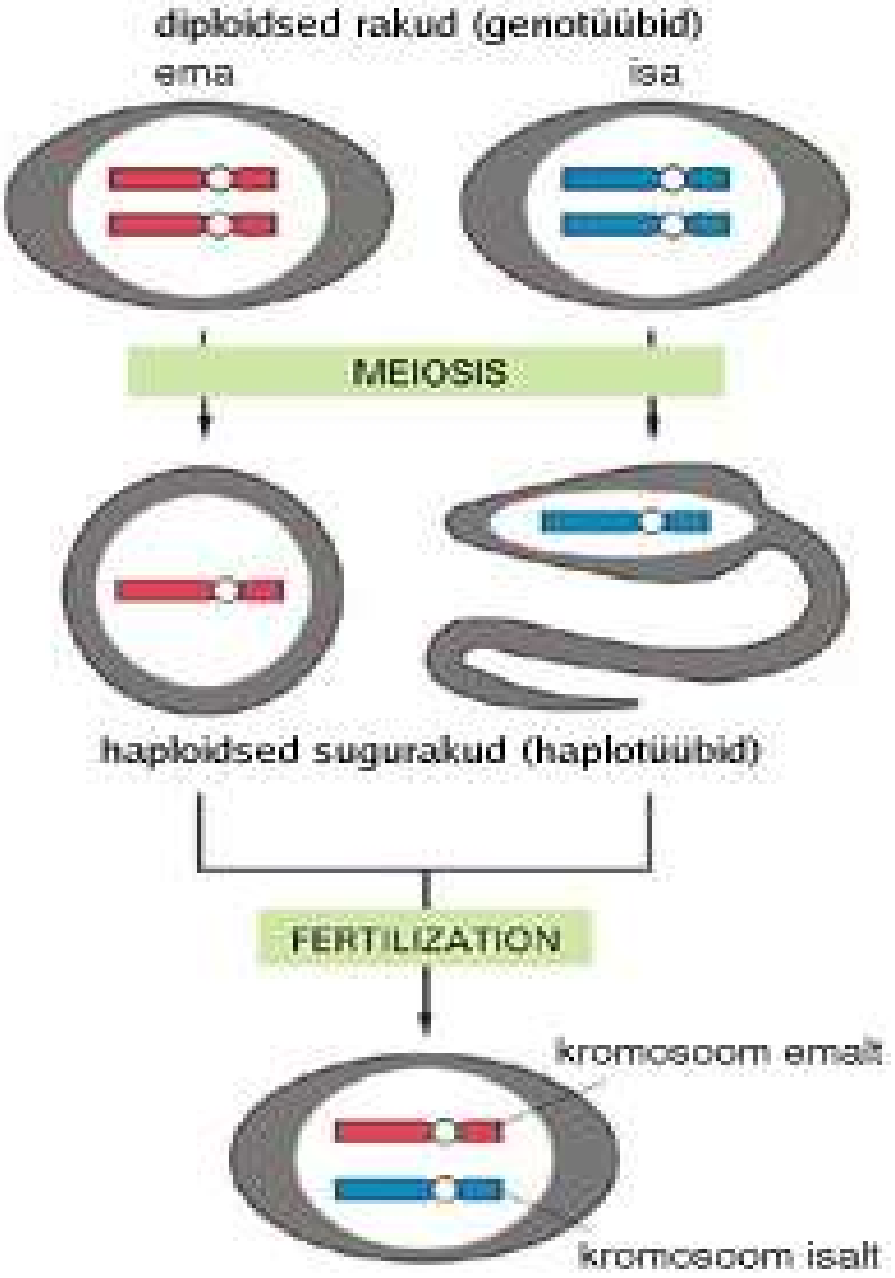
Kristo Käärmann,
t-days @ Veskisilla '04
joint work with Sven Laur

Talk outline

- What is that `haplotype` and why do we need to know about it?
- *Task 1*: Identify haploid genomes for person X
  - Family method
  - Computational methods
- *Task 2*: Reconstruct block structure of human genome
- Further ideas to look into ...

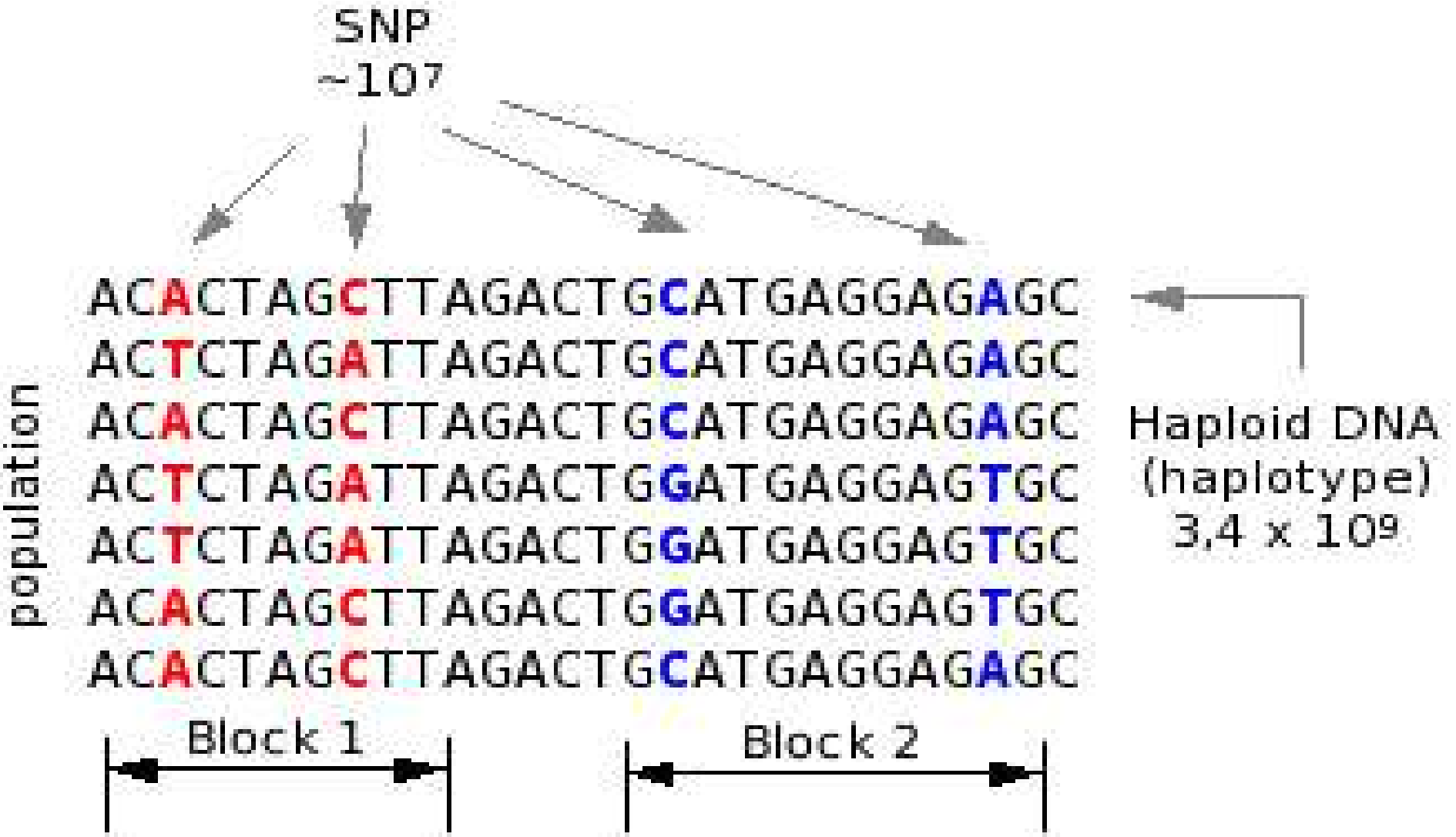
Human genome

- DNA ~ 3.4 billion base pairs
- 99.6% same for all
  - Genetic variation in 10 million SNP-s
    - SNP - Single Nucleotide Polymorphism
    - More or less equally distributed
- Genome comes in 2×23 chromosomes
  - Every cell posesses two versions of each chromosome – one from both parents
  - Recombination merges two genomes into one (takes chunks from each)


diploidsed rakud (genotüübid)
ema
isa
MEIOSIS
haploidsed sugurakud (haplotüübid)
FERTILIZATION
kromosoom emalt
kromosoom isalt

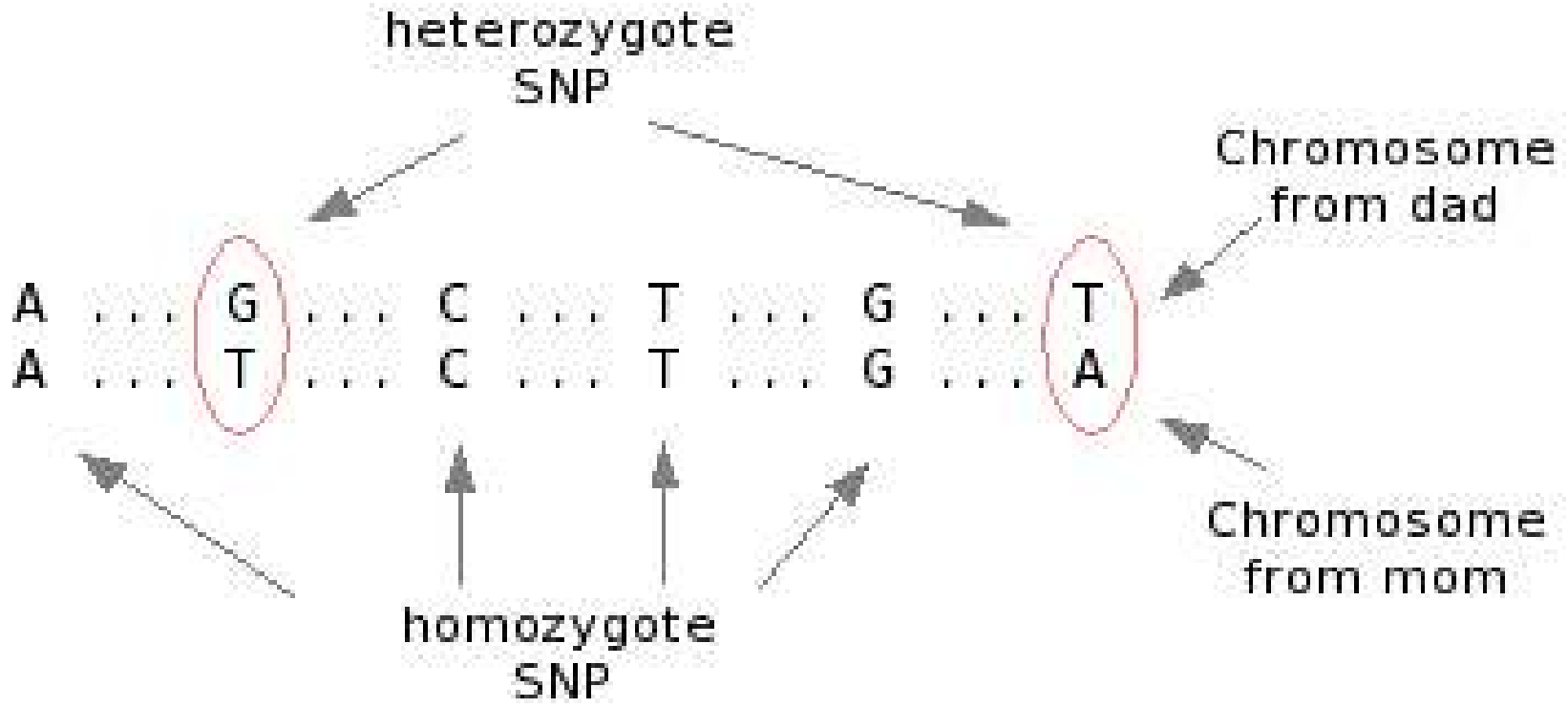
Block structure ?

- Recombination in large chunks (10 Mb)
  - Size and position of chunks varies
  - Rubik's cube effect
  - *hotspots*, conserved regions
- Haplotype block: conserved region in DNA (~20-100Kb)
  - *tag*SNPs: set of polymorphisms that uniquely specify all block alleles
- HapMap project (250 × 400 000 SNPs)
  - Map the block structure of human genome


SNP
~10⁷
ACACTAGCTTAGACTGCATGAGGAGAGC
ACTCTAGATTAGACTGCATGAGGAGAGC
ACACTAGCTTAGACTGCATGAGGAGAGC
ACTCTAGATTAGACTGGATGAGGAGTGC
ACTCTAGATTAGACTGGATGAGGAGTGC
ACACTAGCTTAGACTGGATGAGGAGTGC
ACACTAGCTTAGACTGCATGAGGAGAGC
population
Haploid DNA
(haplotype)
3,4 x 10⁹
Block 1
Block 2

Haplotype & genotype

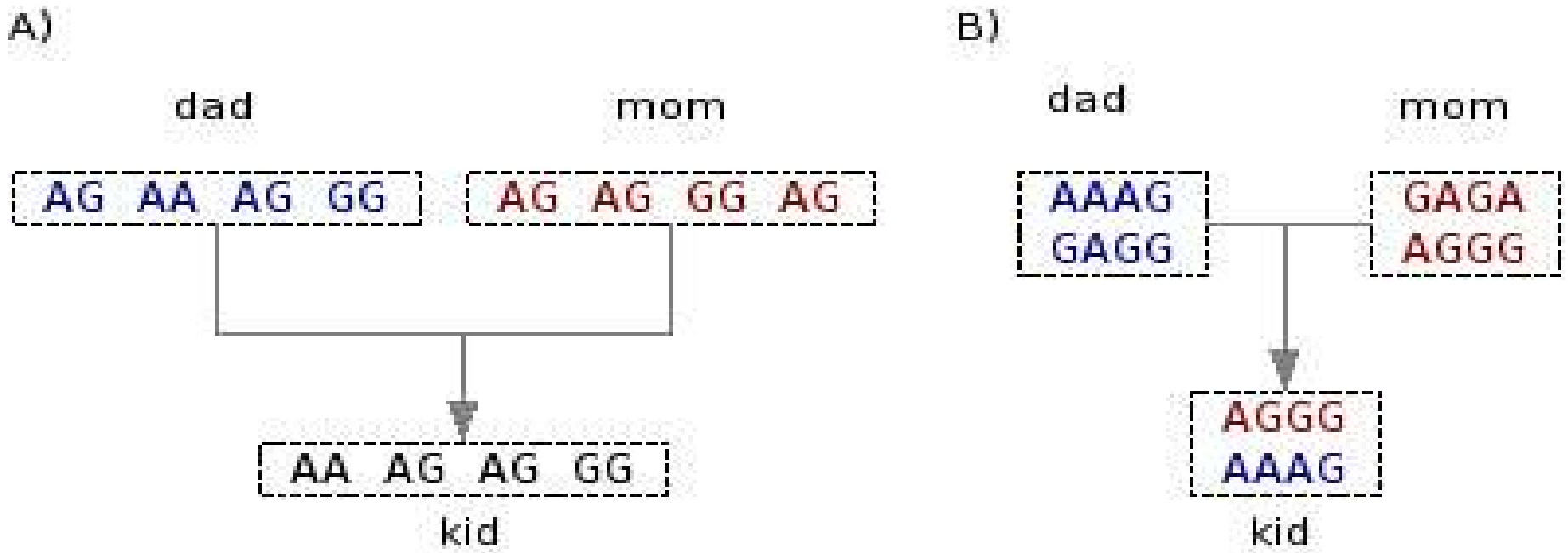
- Genotyping – identifying a base pair ($0.1 per SNP)
  - In positsion *x* is a pair {A,G}
- No reasonable lab method for reading haplotypes – that is two separate sequences {ATG .. TGA, TAG .. GCC}
- Positions that differ in chromosomes are heterozygote SNP-s
- Block structure in haplotypes!



	emalt	isalt
variant 1	GA	TT
variant 2	TA	GT
variant 3	GT	TA
variant 4	TT	GA

Haplotypes by family method

- Requires genotypes from mom-dad-child trios
- 12.5% SNP-s unsolvable by this method
- Expensive, not very accurate



Haplotypes by computation

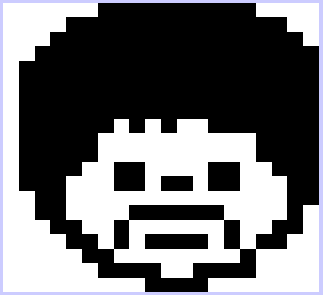
- No family data required
- Methods require a population to run on
  - Greedy algorithm (*parsimony*)
  - EM algorithm ('95)
    - Haplotype frequencies in population
  - Markov chains ('01)
    - Gibbs sampling (PHASE)
    - Haplotype inference for every single individual
    - Accurate, slow
  - Phylogeny tree
    - Reconstructing evolution

Testing results

- How to test produced haplotypes?
  - No lab method to test against!
  - Simulating populations over generations
    - 10 random genotypes x10 random mating
    - Thinning
  - Not a trivial task
- PHASE
  - Accuracy on HapMap data: ~98.8%
  - 100 SNP: 3h, 500 SNP: 48h

Sample population

Illustrative experiment



11111111100001111111111110000110
00011111111100111111110111111100
10100001010011111001101111110110 0
11100111100000011110111001111111 1
11111100100011100100100111001000 0
11001111001100010000011111111100
00100000111111111100001000000101 0
11000000010000000000000000001000
00000000000000001000000000000000
00010000000000000000001100000000
00000000011100000000000000001111
10000000000000011111111000000000
1111

ATAGCTAGACGATAGAATGCTCCCTAG
ATACTTGTTCGCTAGCTAGCTTAGATCG
ATTCGAGGATCGTAGATGCCCATGATC
GATCGAATGCATGCAGGGAGGAAATCG
ACTGACTGATGCATGCATCTTACGTACG
ATCATCACTAGTCGAGTCAGCAGCATC
GACTGACATGCTTGACTCGATGTTGACT
ACGTTGCTAGCATCCTCTAGCATCGATC
TAGCTGATCGTAGTACGACTGACTGAT
GCATTAGCATGC.....

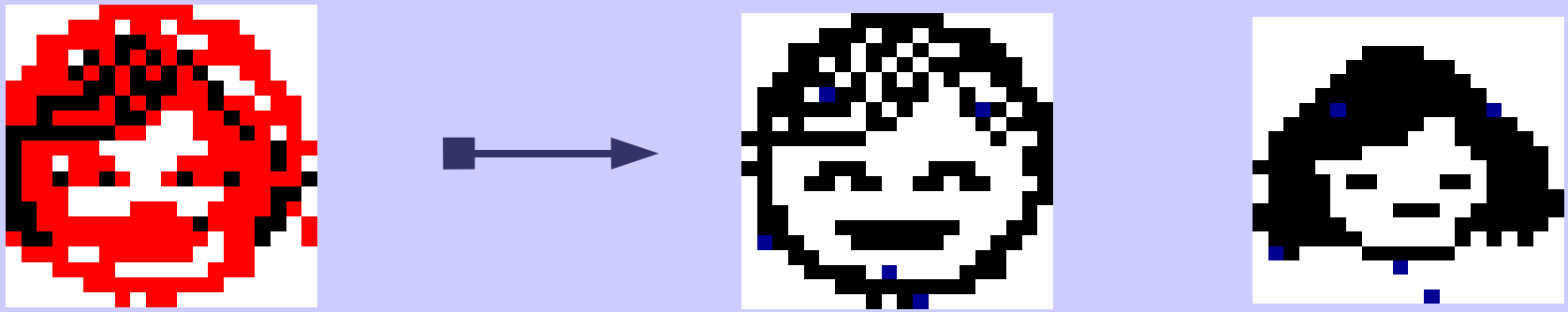
20 x 20 monochrome bitmap

400 bits

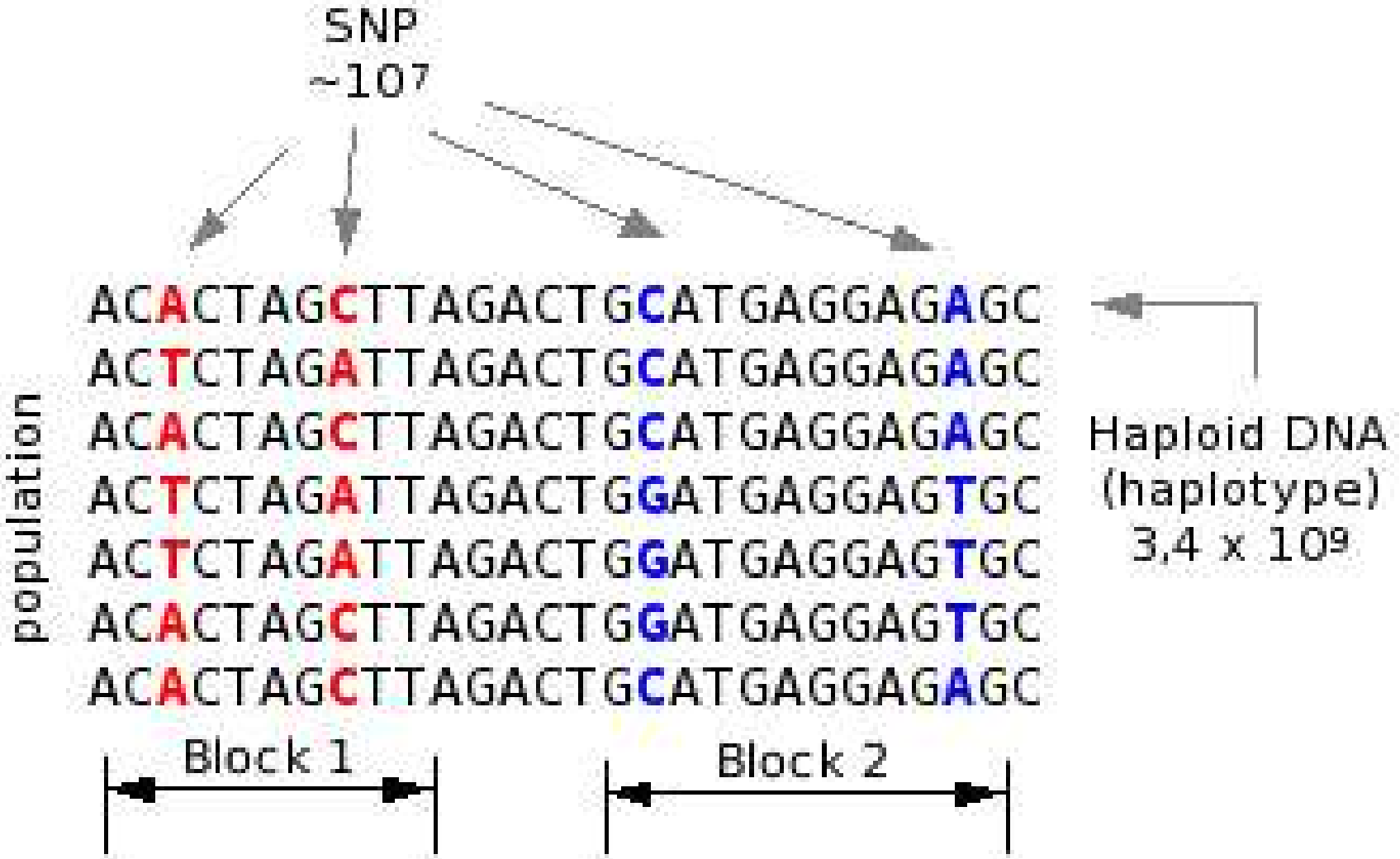
400 SNPs = 100 kB region

Offsprings from random xxx

Haplotypes by computation



Noise rate: 1.2% = 5 pixels

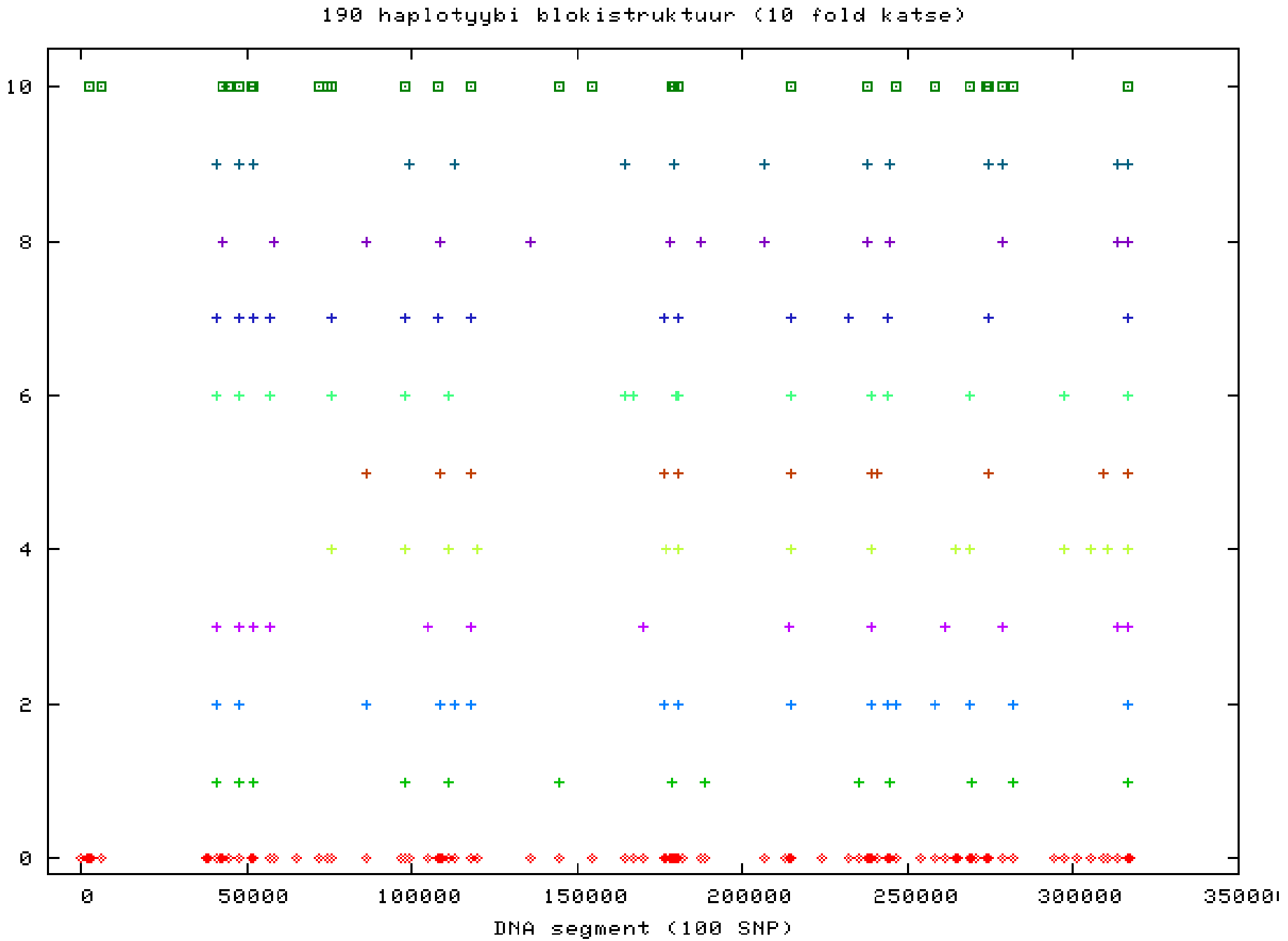

SNP
~10⁷
population
ACACTAGCTTAGACTGCATGAGGAGAGC
ACTCTAGATTAGACTGCATGAGGAGAGC
ACACTAGCTTAGACTGCATGAGGAGAGC
ACTCTAGATTAGACTGGATGAGGAGTGC
ACTCTAGATTAGACTGGATGAGGAGTGC
ACACTAGCTTAGACTGGATGAGGAGTGC
ACACTAGCTTAGACTGCATGAGGAGAGC
Haploid DNA
(haplotype)
3,4 x 10⁹
Block 1
Block 2

Mapping block structure

- Find conserved regions in DNA, claim tagSNP-s
- Methods
  - Correlation based (four-gamete test)
  - Dynamic programming
  - Mininmum description length (MDL)
- How to test the results?
  - Unsupervised learning ... (generalisation!)

Dynamic progr. (Zhang)

- $S_j = \min_{i<j}[S_i + f(i,j)]$
  - Mininmising additive cost function
  - f(i,j) = number of *tag*SNPs, which uniquely identify x% of block [i, j] alleles
  - Block is defined by *tag*SNPs
- Experimental input:
  - 190 haplotypes (95 individuals)
  - 100 SNPs over 350Kb (~0.01% of DNA)


190 haplotyybi blokistruktuur (10 fold katse)
10
8
6
4
2
0
0
50000
100000
150000
200000
250000
300000
350000
DNA segment (100 SNP)

What is to be done ...

- Haplotype inference
  - Scalability! Processing time exp to segment length
  - Partition-Ligation
    - Hard to merge, overlapping?
  - Parallel processing
- Mapping into blocks
  - Model-based approach
  - How well can we generalise?
- Blocks from genotypes?!