LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INTER CROSSING

Haplotype-phasing in polyploids

Soumya Ranganathan

Academic Supervisor:
Dirk Metzler

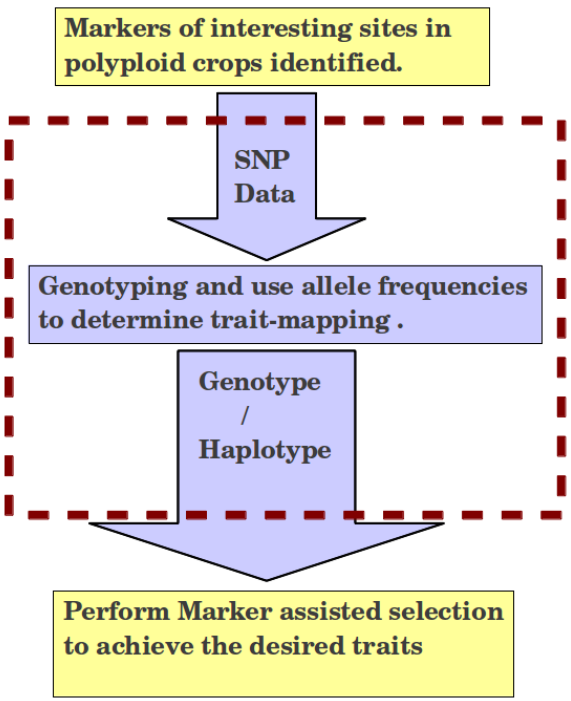
Industry Supervisor:
Andrzej Czech

Partner:
Fabian Grandke

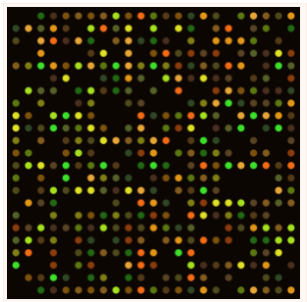
Genetwister

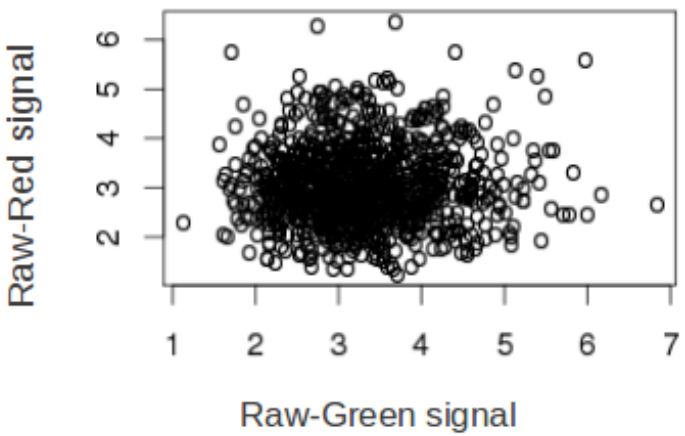
European Commission

Motivation



1.Objective of data analysis





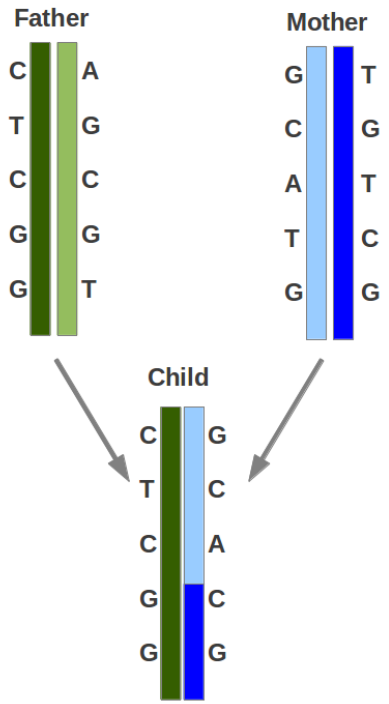
2.Micro array data

Haplotype-phasing

Bi-allelic SNPs: Two possible variants of nucleotide.

- What is Phasing?
  Separating haplotypes from different copies of chromosome.
  - Eg. Genotype of child:
    (G,C),(T,C),(A,C),(G,C),(G,G)
  - Haplotypes:
    - C-T-C-G-G
    - G-C-A-C-G
- Why phasing?

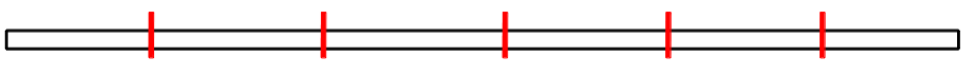
Genotype	Haplotype	Diagnosis
(G,C),(G,A)	G-A and C-G	**No Disease**
(G,C),(G,A)	G-G and C-A	**Disease**
(G,C),(G,G)	G-G and C-G	**Disease**



[Source:Lo, Christine. "Algorithms for Haplotype Phasing."]

Phasing-approaches

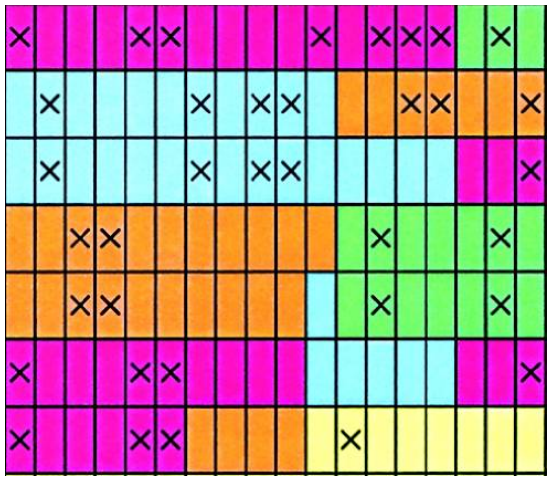
- Parsimony : Minimize total number of observed haplotypes in sample. Eg. Clarks algorithm(1990)
- Haplotype blocks
  - M. Stephens, P. Donnelly (2003) : The genome is divided into blocks of certain numbers of loci.
  - haplotype frequencies are estimated for each block and blocks are iteratively ligated with adjacent blocks into larger blocks.



Phasing model

Model fastPhase(Scheet & Stephans 2006):

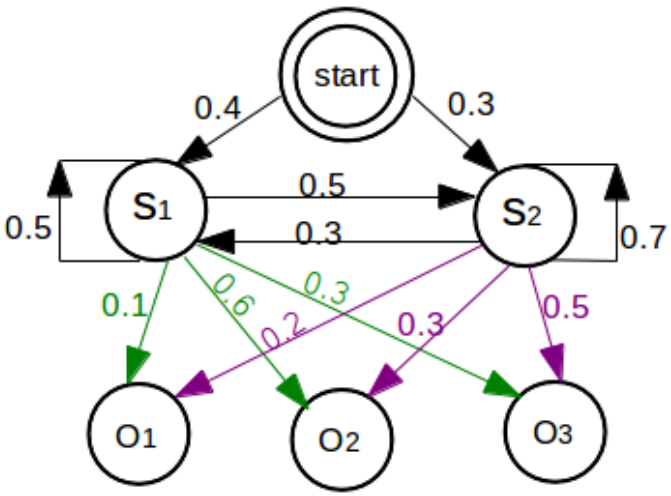
- Over short regions, haplotypes cluster into groups of similar haplotypes
- Recombination - mosaic of cluster memberships along the genome
- Model parameters
  - physical dist bet markers, d_m
  - num of ances clusters, K
  - α_{km}, freq of cluster k at marker m
  - θ_{km}, freq of allele '1' from clust k at marker m
  - r_m, recombination freq bet adj markers, m-1 and m



Each column represents a SNP, two alleles indicated by open and crossed squares. A pairs of rows represent haplotypes for an individuals. Colors represent estimated cluster memberships.

Hidden Markov model(HMM)

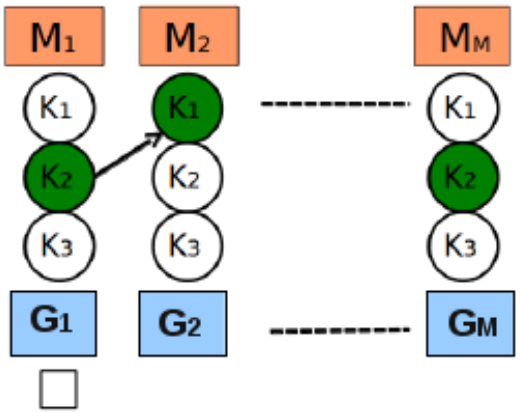
- State space: $X = \{s_1, s_2\}$
- Output variables: $Y = \{o_1, o_2, o_3\}$
- Initial state prob: $\pi = [p_{s_1}, p_{s_2}]$
- Transition prob : $A_{ij} = p(X_t = s_j | X_{(t-1)} = s_i)$
- Emission prob : $B_{ij} = P(Y_t = O_j | X_t = S_j)$



- Observed variables: $o_2 \to o_3 \to o_3 \to o_1 \to o_2 \to o_1 \ldots$
  System(steps) : $s_1 \to s_1 \to s_2 \to s_2 \to s_1 \to s_1 \ldots$

HMM in our phasing model

Haploid model, $\binom{3}{1} = 3$ choices

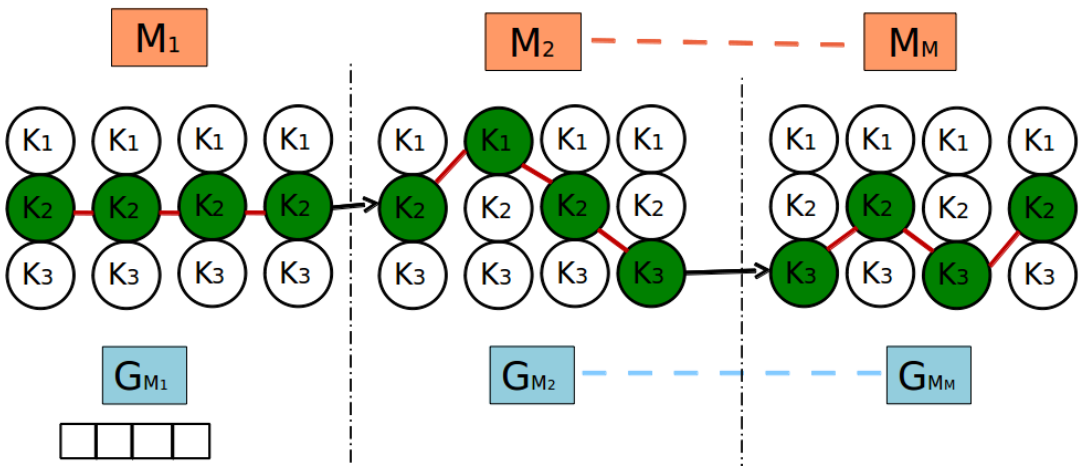


α_{km}, freq of cluster k at marker m

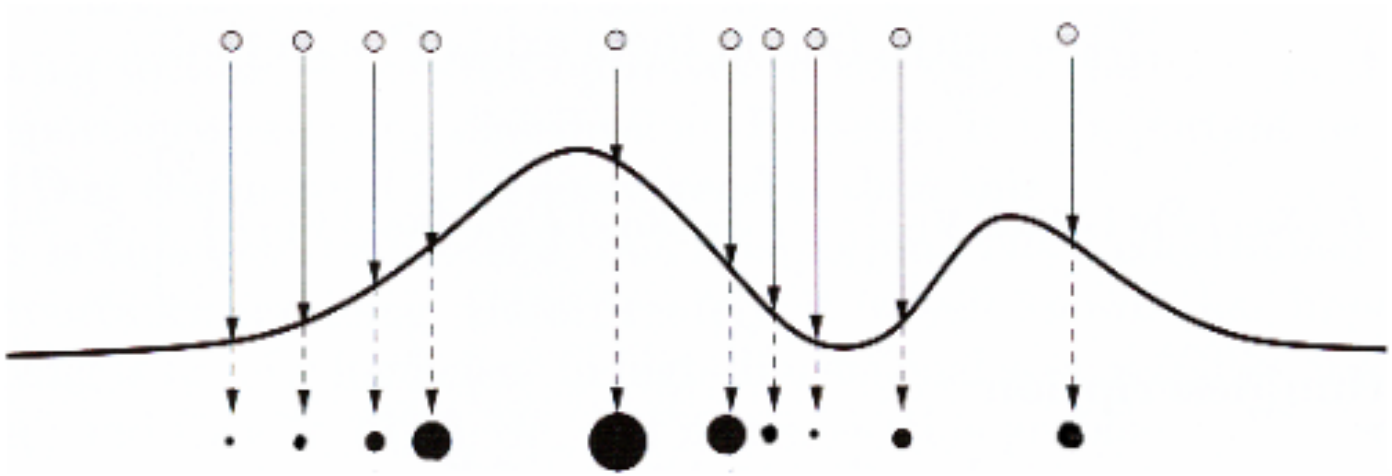
θ_{km}, freq of allele '1' from clust k at marker m

r_m, recombination freq bet adj markers, m-1 and m

Polyploid model, if K=3 and N=4, $\binom{4+3-1}{4} = 6$, but for K=8, 330 choices.



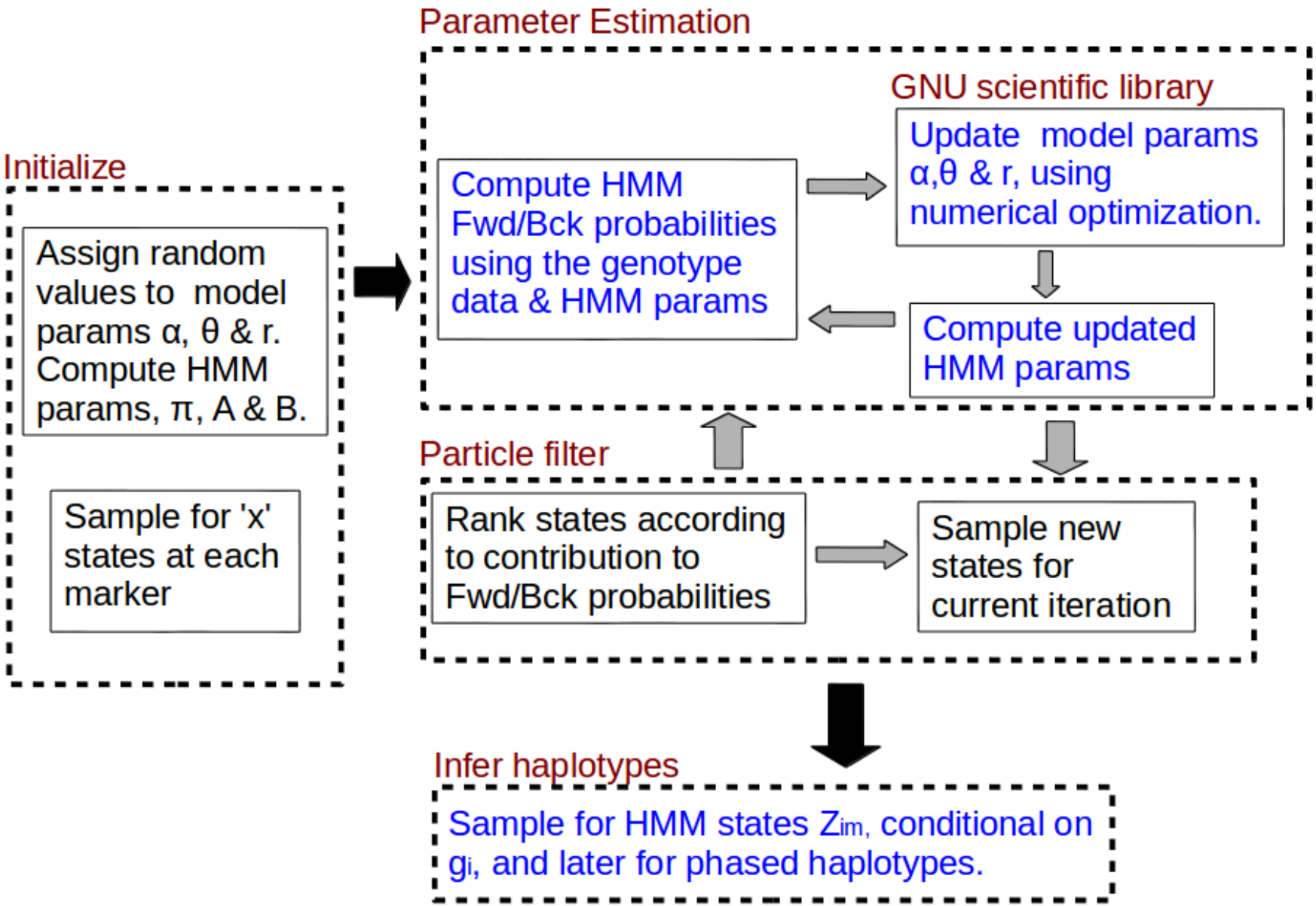
Particle filter



Source:A brief Introduction to Particle Filters by Michael Pfeiffer

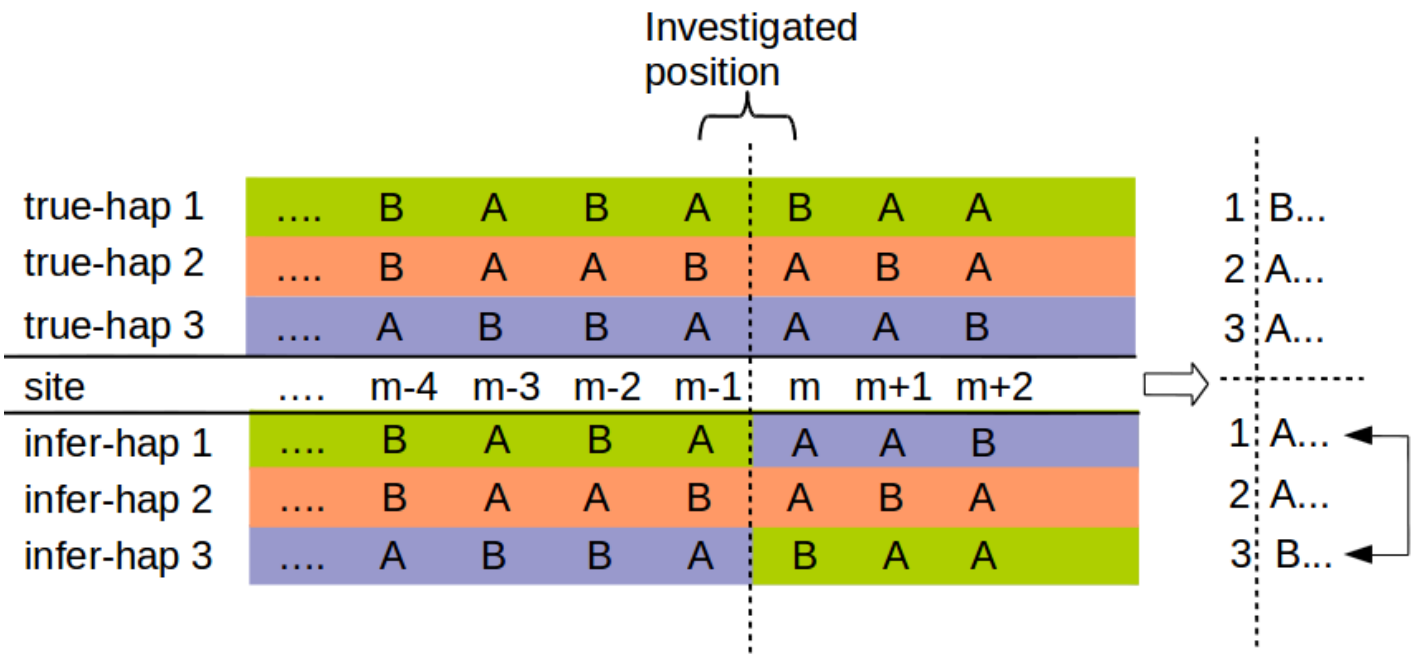
- If K is number of clusters and N is ploidy, we sample for 'X' states where, $X < \binom{N+K-1}{N}$
- Particle filter
  - Exclude states which make less contribution to the HMM inference values.
  - Include states which are similar to the ones that make significant contribution.

Phasing Flow



Phasing-Validation

Switch Error: $\frac{\text{Min Number of switches required to obtain the true haplotype phase}}{(\text{heterozygote markers in the individuals genotype }) - 1}$



Eg. Switch Error in Triploid phasing[Su. et al 2008]

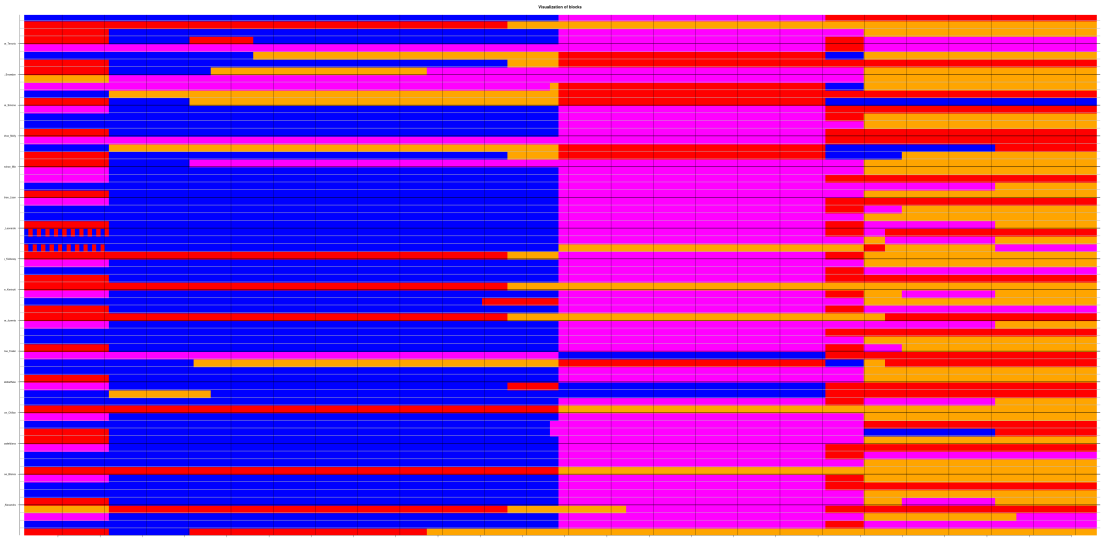
- Assign inferred haplos to corresponding true haplos
- Own method to compute Switch error using HMM

Placement- Genetwister Tech.

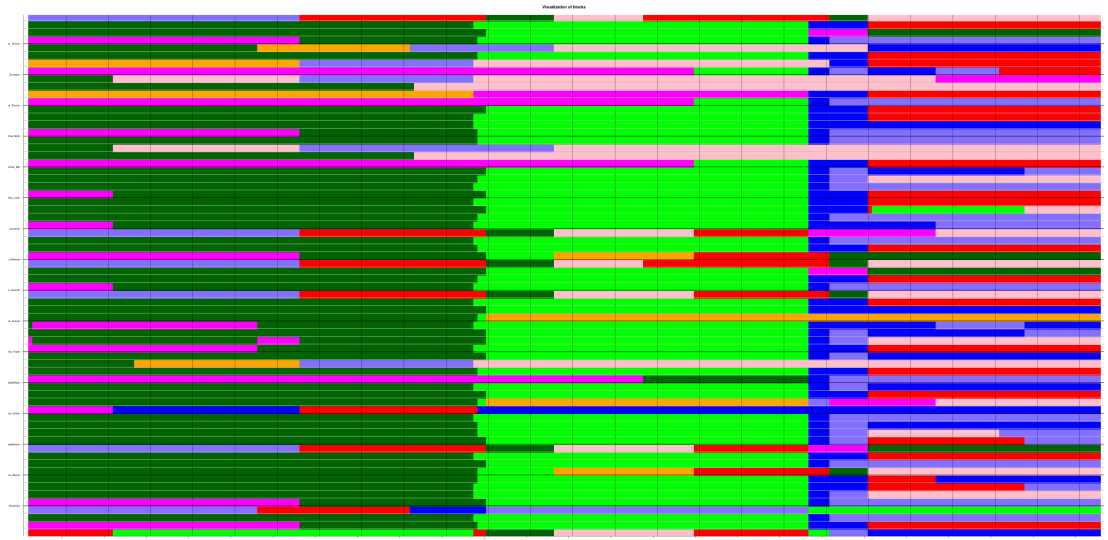
- First part: Looked into Genotyping methods, fittetra and beadarrayMSV
- Second part: Explored the real datasets
  - data formats and conversions from allele signals to genotype values
  - Analysed polyhap runs
    - Polyploid ornamental for 20 ind over 400 markers, Tetra and Hexa ploid
    - Diploid vegetable for 300 ind over 8 chromosomes
  - Conclusions
    - Selection of K(num of ances clusters)
    - Distribution of clusters is not consistent.

Results-ornamental(tetra)

K=4



K=8



Legend:

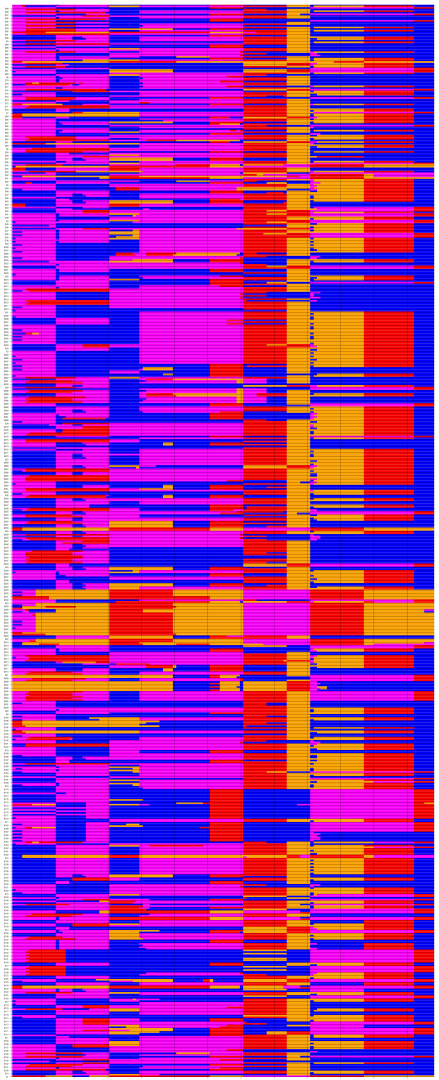
X-axis. markers

Y-axis:

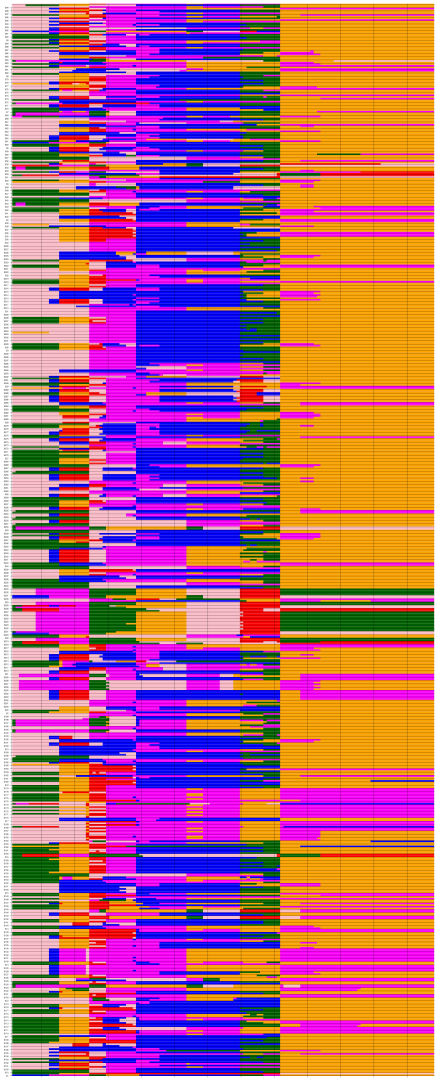
Individual cluster memberships

Results-diploid

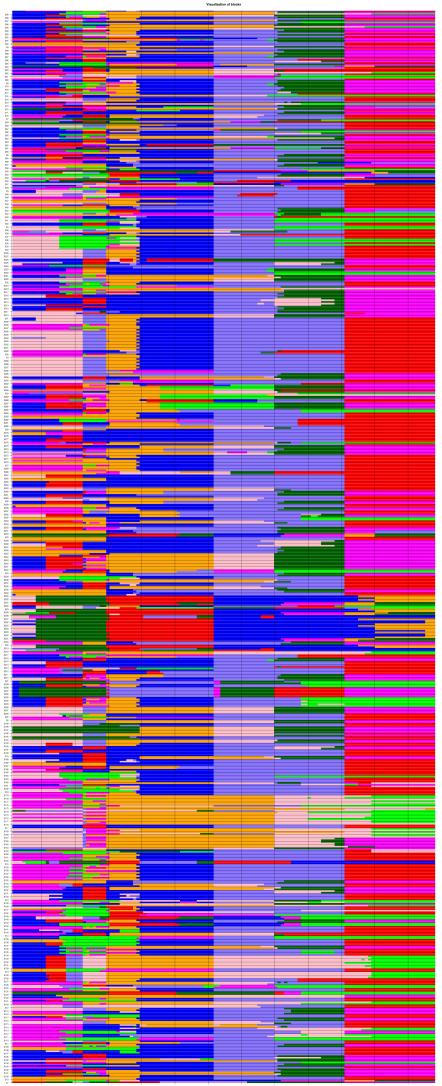
Legend:X-axis markers, Y-axis Individual cluster memberships



K=4



K=6



K=8

Questions?

Thank you!