UPPSALA
UNIVERSITET

Haplotype inference based on HMMs in the QTL-MAS multi-generational dataset

QTLMAS XIV Poznan, Poland 2010-05-17

Carl Nettelblad

Uppsala University, Department of Information Technology

UPPSALA UNIVERSITET

Introduction

- Highly accurate haplotype inference for multiple generations and thousands of markers
- Haplotyping followed by a basic model for QTL fitting on this year’s dataset

UPPSALA UNIVERSITET

Global Haplotype Inference

- Marker data is generally unphased
  - If markers demonstrate limited variability, tracing inheritance to founder individuals gets increasingly hard
  - Founder individuals might not be homozygotic for QTL
- Methods frequently based on approximations
  - Local windows fail for limited variability
  - Heuristics-based methods succeed when some cases are “easy”
    - Repeated inference and logical implications

UPPSALA UNIVERSITET

A General HMM Approach

- Existing code for determining genotype probabilities in 3-generation pedigrees (F2-like)
  - Every offspring individual analyzed independently
  - Total set of 4 states in each locus
    - Grandparental origin of offspring alleles (2^2)
- Extension into 4 generations (F3-like)
  - Total of 64 states (2^6)
  - Extension into 5 generations would mean 2^{14} states
- Transitions consisting of recombination, emissions represent marker data

UPPSALA UNIVERSITET

Haplotype Inference

- Consider 3 generations, separating strands (strand 1, 2) in founder generation
- Let each single marker observation have equal probability of pair being listed in 12 or 21 strand order
  - Parametrize the probability for 12 assignment (“skewness”)
  - Initialize as 0.5 in all markers but first heterozygote
  - First heterozygote marker serving as “anchor”, fixed at 0

UPPSALA UNIVERSITET

Marker Example

- 6 loci, 3 generations; Offspring, Father, Mother, grandparents
- Diploid data, strand assignment unknown
- If phasing was known, the ambiguity in marker 1 would be eliminated
  - Linkage would give stronger information in markers 3 and 4

FF (AA)	1 1	1 1	1 1	1 3	3 3	1 1
FM (BB)	2 2	2 2	1 3	1 3	4 4	2 2
MF (BB)	2 2	3 3	4 4	3 4	1 2	1 1
MM (AA)	1 1	4 4	4 4	3 4	1 2	2 2

F	1 2	1 2	1 1	3 3	3 4	1 2
M	1 2	3 4	4 4	3 4	1 2	1 2

O	1 2	1 3	1 4	3 3	1 4	2 2
	AB BA	AB	AA AB BA BB	AA AB BA BB	BA BB	BA

UPPSALA
UNIVERSITET

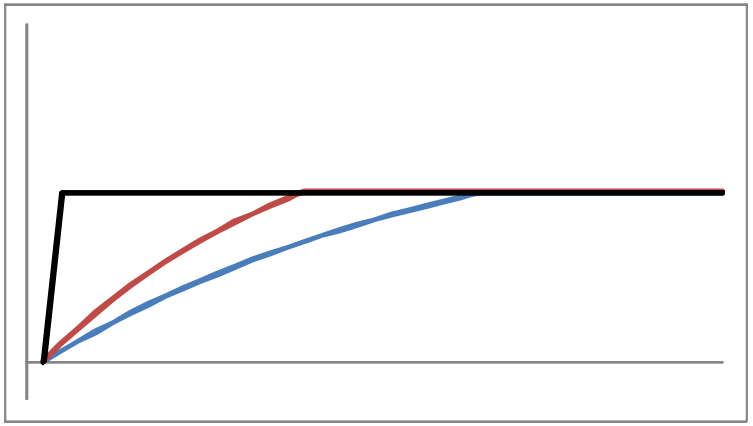
Practical Concerns

- When computing genotype probabilities in this model, we are marginalizing over possible strand assignments
  - Many will be impossible
- A HMM training algorithm can optimize the strand assignment parameters
  - Repeated analysis of local 3-generation pedigrees (grandparents, parents, offspring)
  - Baum-Welch superior to Viterbi

UPPSALA UNIVERSITET

The Training Process

- One single marker fixed in the start



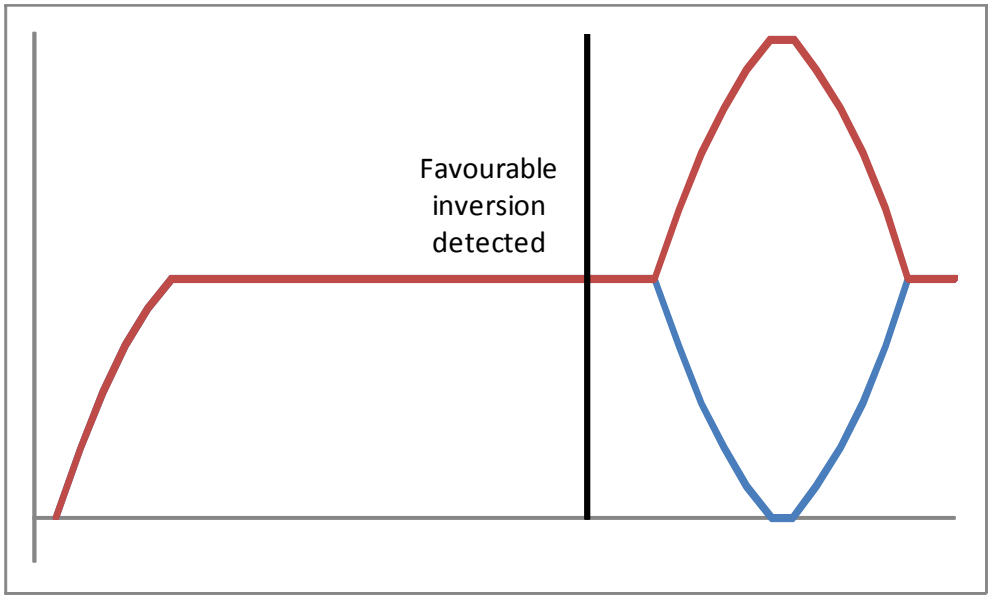
- Successive iterations move this further based on linkage detected in offspring and strand information in parents

- If mapping distances are unknown, these can be trained simultaneously

UPPSALA UNIVERSITET

Inverted Bubbles

- The information on linkage can be very limited (e.g. long homozygous regions)



- Strand assignment is arbitrary in first generation
- Specific inversion sweeps during iterations detect situations when all skewness values downstream should be swapped

UPPSALA UNIVERSITET

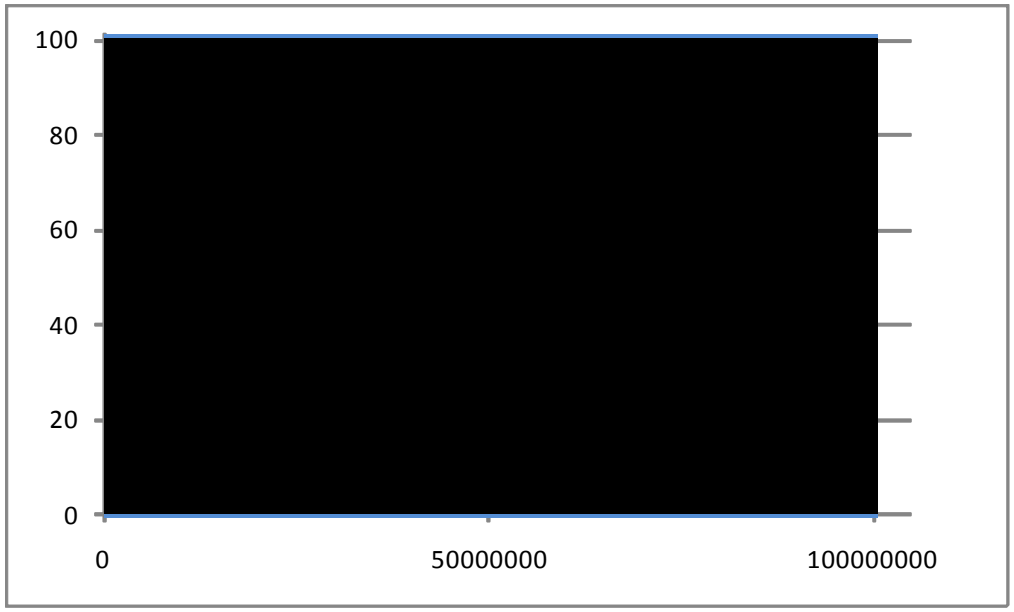
Convergence Rate

- Convergence dependent on population structure and chromosome length
  - Not number of markers, more markers help!
- After 20 iterations
  - All but 3 heterozygous pairs were phased in generation 1, chromosome 1
  - 38 were phased with some uncertainty
  - 15,205 converged
    - Precision expected to be high
    - 100 % recall for all practical purposes
  - In generation 5, only 88 out of 680,945 pairs did not converge

UPPSALA UNIVERSITET

Marker Map

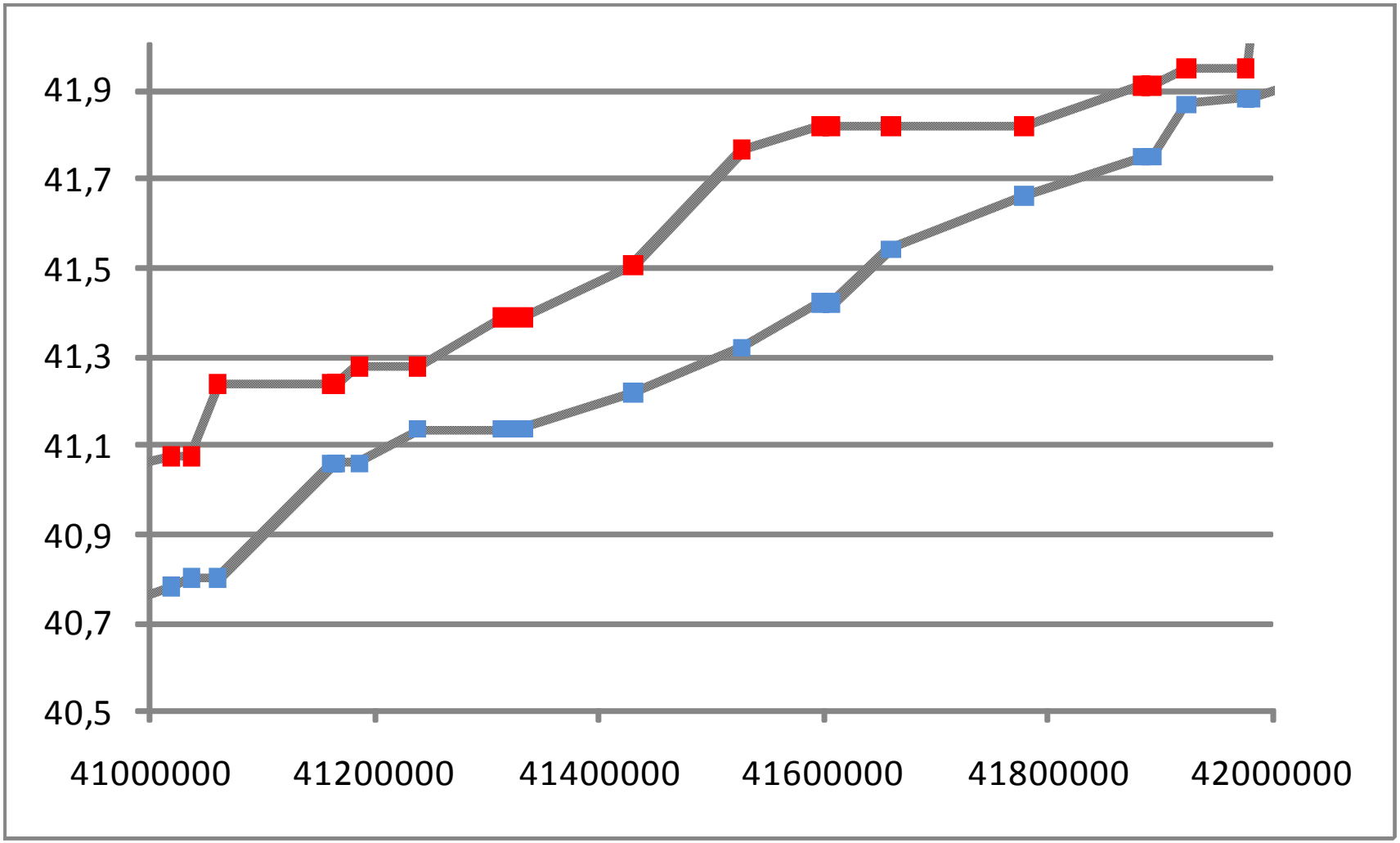
- Position in cM vs. bp



- Sex-specific marker map, no enforced recombination rate, inter-marker distances *not* initialized to sum up to 100

UPPSALA UNIVERSITET

Zooming In



UPPSALA UNIVERSITET

Analysis of Q Phenotype

- Very simple model fitting:
  - a litter effect (27.7% of variance)
  - succesively adding fixed allele effects for each of 20*2 founder alleles in a forward-selection manner
- Resulted in 5 significant QTL explaining 14.6 %
- Identical model fitted with non-haplotyped data (marker map still used)
  - Roughly identical positions for 5 first QTL, total explained variance 6.97 %
  - Permutation testing indicates that this difference is not only due to varying effective no. of degrees of freedom

UPPSALA UNIVERSITET

Discussion

- When eliminating the litter effect, explained variance for 30 fitted QTL amounted to 33.4%
  - A plateau reached after this, with only 37.6% explained for 40 QTL, seemingly matching true genetic architecture
- Litter effect and free variables for all founder alleles shadowing smaller QTL within larger ones
  - These are flaws of the very simple *QTL model* used
  - Any approach that can benefit from true allele identity-by-descent data could use efficient haplotyping
  - Adding e.g. an epigenetic model with parental sex is trivial