Reconstruction of Genome Ancestry Blocks in Complex Plant Populations

Diploid Multi-Parental Populations (MPP) & TetraPloid Families (TPF)

Chaozhi Zheng & Marco Bink
Acknowledgments

Co-authors

- Wageningen University & Research (NL)
  - Chaozhi Zheng,
  - Martin Boer, Roeland Voorrips
  - Fred van Eeuwijk, Hans Jansen
- Biomathematics and Statistics Scotland (UK)
  - Christine A. Hackett
- Forage Genetics International (USA)
  - Julie Ho
Outline

- Objective
- Hidden Markov Model frameworks
- Empirical Test Results - Simulated Data
- Applications - Real Data
- Concluding Remarks
Objective

- Utilize genomic marker data as available in **complex plant populations**
  - Diploid multi-parental populations
  - Increase QTL segregation probability
  - Increase QTL mapping resolution
  - Tetraploid full sib families
  - Genetic Understanding & Breeding Target

- Hidden Markov Model frameworks to reconstruct genome ancestry blocks (~IBD probabilities)
  - [MPP] RABBIT (Reconstruction of Ancestral Blocks BIT by bit)
  - [TPF] TetraOrigin
MPP Multi-parental populations

Increase QTL segregation probability
- 2-way recombinant inbred lines (RIL)
- Nested association mapping
- Arabidopsis multiparental RIL
- Mouse collaborative cross (MCC)

Increase QTL mapping resolution
- No inbreeding stage: advanced intercross lines, heterogeneous stocks, diversity outbred
- Arabidopsis multiparental advanced generation intercross (MAGIC)
MPP

Haplotype reconstruction

Four-way RIL by selfing

Homozygous founders \( F_0 \)

100011_1_0_001111001
1111000111100111_101
10011010101101010101
1010101100010_111010

Sampled genotypes (#allele1) \( F_3 \)

20002002_22102121111

Genetic predictors for QTL mapping

10001001111101011010
10001001111100111010101
Haplotype reconstruction

Previous approaches

HAPPY (Mott et al 2000. PNAS)
- Completely outbred: two independent haplotypes
- Fully inbred: two homozygous haplotypes
- General but not incorporating pedigree info

GAIN (Liu et al 2010. Bioinformatics)
- Partially inbred: Joint modelling two haplotypes
- Incorporate pedigree information
- Only applicable for MCC

Features of RABBIT approach

Applicable to
- full spectrum of inbred (complete, partial, or no)
- various mapping populations in any generation, no matter if pedigrees or breeding designs are known
- to autosomes and sex chromosomes

Handles missing data and allelic typing errors in both founders and sampled lines
### MPP  Statistical framework

#### Hidden Markov model

- Hidden state $o_i = (p,m)$ at SNP$_i$, where $p,m = \text{red, green, purple, blue}$.

- Sample individual’s genotype $y_i = 00, 01, 11$, or NN (= missing data), conditional independent.

#### The prior ancestral origin process

- Modelling $p\left[o_1 = \left(\begin{array}{c} a \\ b \end{array}\right) \rightarrow o_2 = \left(\begin{array}{c} c \\ d \end{array}\right)\right]$

- **indepModel**  Independent transition
  - $p(a \rightarrow c) p(b \rightarrow d)$

- **depModel**  Dependent transition
  - $p(b \rightarrow d \mid a \rightarrow c) = 1$

- **jointModel**  Joint transition
  - $p(a \rightarrow c) p(b \rightarrow d \mid a \rightarrow c)$
  - if completely outbred (no inbreeding)
  - if completely inbred

#### The data model

- Prior true gtp homozygous founder $p(00) = p(11) = 0.5$
- Prior true gtp sampled individual $p(00) = p(01) = p(10) = p(11) = 0.25$
- Allelic errors occur independently
- Likelihood is conditional on missing patterns
MPP Empirical Test – simulated data

**Inbreeding!**

**Mouse Collaborative Cross**

(B) CC funnel

---

**statistical Evidence (Model fit)**

**estimation Error Rate**

*jointModel* outperforms *indepModel* and *depModel* in intermediate (partially inbred) generations
MPP Comparison with GAIN & HAPPY

Mouse Collaborative Cross

- Data no errors, different prior ancestral origin process
- Estimation error rate: $\text{RABBIT} \approx \text{GAIN} << \text{HAPPY}$
- Inconsistency: $\text{GAIN}(0) \approx \text{RABBIT}(0.05\%) < \text{HAPPY}(2\%)$

- Pedigree contains little information, compared to dense mark data
MPP Individual’s Marginal Posterior Probability

Real data (Kover et al. 2009)
Arabidopsis MAGIC
19 founders

Grey level = posterior probability

HAPPY are more noisy than RABBIT
**Maximize Flexibility:**

- No strict distinction between allopolyploidy & autopolyploidy
- *A priori* allow preferential bivalent and also quadrivalent pairing (⇒ double reduction)

**Haplotype Reconstruction in a tetraploid full-sib family**

- Multi-locus approach (*genetic map required*)
- Preferential pairing, polysomic inheritance
- Outcrossing parents(!): inference on parental linkage phases

To calculate posterior probability e.g. 

\[
Pr(\text{parental origin } X_{SNP_1}^{SIB_1} = [1 \ 3 \ 7 \ 7] | D)
\]
conditional on all the dosage data \(D\).
TPF Modelling overview

Parental Haplotypes

Possible gametes by pairing in parent P1

Rectangles: observed dosage scores
Circles: random variables
Solid lines: probabilistic relationships

Directed Acyclic Graph of network model, which becomes a **Hidden Markov Model** conditional on chromosome pairings
TPF Parental origin processes

- Diploid gamete:
  - Quadrivalent formation \([1234]\) occurs with prob \(p_{\text{quad}}\)
  - Given bivalent pairing: preferential pairing occurs with prob \(p_{\text{pref}}\) Possible pairs: \([12][34], [13][24], [14][23]\)
- Two parents: 16 combinations of chromosome pairings
  \([12][34], [13][24], [14][23], [1234] \times [56][78], [57][68], [58][67], [5678]\)
- Parental origins along
  - Haploid gamete follow a Poisson process
  - Diploid gamete follow two independent Poisson processes
- The two (diploid) gametes constitute a zygote resulting from independent meiosis processes
Given the parental dosages

- the parental haplotype $H_t$ at a particular locus is *a priori* equally probable
  - E.g., $D_t^{P1} = 2$ & $D_t^{P2} = 1$
  - $H_t$ takes one of 24 combinations of $(1100,1010,1001,0110,0101,0011) \times (1000,0100,0010,0001)$
- If missing dosage: $D_t = 0,1,2,3,4$ equally probable
- Allow for errors: With prob $p_{error}$ observed dosage may be one of other possible dosages
TPF Two-stage haplotype reconstruction

I. Maximum a posteriori estimation of parental phases (haplotypes)

II. Given parent phases, independently for each sibling:
   1. Calculate posterior probabilities for each of bivalent or quadrivalent combinations.
   2. Integrate these combinations to obtain the marginal posterior probabilities at all SNPs.
**TPF Simulation study: 1 full sib family**

**Number of inconsistencies between inferred and true parental haplotypes**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Preferential pairing $p_{pref}$</th>
<th>Quadrivalent pairing $p_{quad}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DStd-M</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DPref-M</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>DQuad-M</td>
<td>0</td>
<td>2/3</td>
</tr>
<tr>
<td>DPrefQuad-M</td>
<td>1/2</td>
<td>1/2</td>
</tr>
</tbody>
</table>

$N_{SNP}=1200, N_{SIB}=200$ PedigreeSim Software (Voorrips & Maliepaard BMC Bioinf 2012)

Perfect inference of parental haplotypes, even for low numbers of sibs & SNPs!!
TPF: Individual’s Marginal Posterior Probability

Application to real potato data

36 ancestral states, i.e., 1256, 1257, ..., 3467, 3478

Grey level = posterior probability
TPF Individual’s Marginal Posterior Probability

Application to real potato data

36 ancestral states, i.e., 1256, 1257, ... , 3467, 3478

Grey level = posterior probability

(Hackett et al. 2013)
TPF Individual’s Marginal Posterior Probability

Application to real potato data

36 ancestral states, i.e., 1256, 1257, ..., 3467, 3478

Grey level = posterior probability

RABBIT – full model

ancestral state 0 lumps all quadrivalent states (1156, 1157, ..., 2488, 3488)
Concluding remarks

Two Powerful Approaches

✓ Hidden Markov Models
✓ Probabilistic Inference

✓ Multi-parental populations (MPP)
✓ TetraPloid (full sib) families (TPF)
  ✓ Extension to higher ploidy

Robust to missing & erroneous data
✓ Currently based on SNP dosages
✓ Pending extensions to GBS data
References


