Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca
Milan, Italy

A fast and practical approach to genotype phasing and imputation on a pedigree with erroneous and incomplete information

ICCABS 2012

Yuri Pirola, Gianluca Della Vedova, Stefano Biffani,
Alessandra Stella, and Paola Bonizzoni

pirola@disco.unimib.it

Outline

HI on Pedigrees with Recombinations, Errors, and Missing Data:

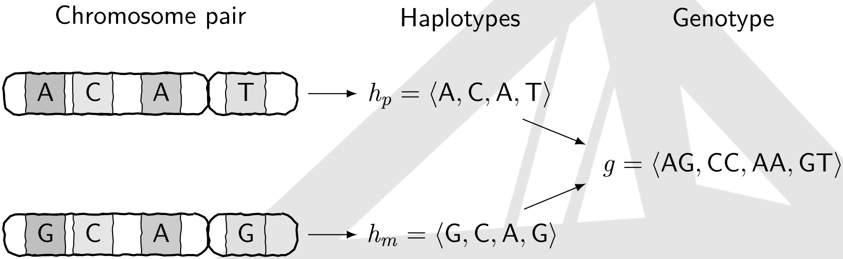
- Introduction and Background
- *Minimum Recombinant Haplotype Configuration with Bounded Errors* problem (MRHCE):
  - Exact algorithm
  - Experimental evaluation and comparison
- Conclusions

Our Contribution

Original Contributions:

- *Generalization* of an existing model for HI to a more *realistic* setting:
  - *missing genotypes* and *genotyping errors*
- *Practical* and *exact* algorithm:
  - for the *new* and the *old* formulations (*MRHCE*, *MRHC*)
  - can detect hard-to-discover genotyping errors

The two main "characters"



Haplotypes

$h_p = \langle \text{A}, \text{C}, \text{A}, \text{T} \rangle$

$h_m = \langle \text{G}, \text{C}, \text{A}, \text{G} \rangle$

Genotype

$g = \langle \text{AG}, \text{CC}, \text{AA}, \text{GT} \rangle$

Haplotypes: **useful** (e.g., genetic mapping, association studies, . . .)

Genotypes: easy to collect

Haplotype Inference (or Genotype Phasing) problem

Problem (Haplotype Inference)

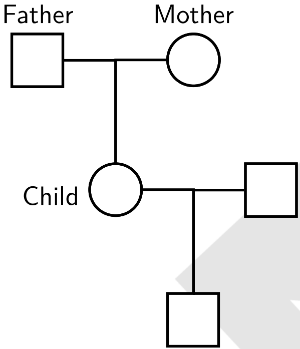
Given the genotypes of a ***population****, recover (=infer) the pairs of haplotypes of each individual.*

Different *kinds of populations* and *genetic models*

$\Rightarrow$

Different *computational problems*

HI on Pedigrees



Parental relationships

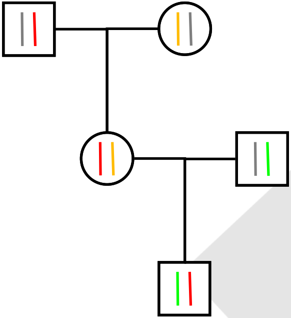
$\Longleftrightarrow$

Mendelian laws of inheritance

$\Longleftrightarrow$

Easier/More accurate HI

HI on Pedigrees



Genotyped Pedigree:
pedigree + genotypes

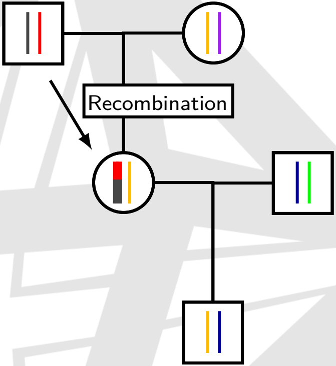
Haplotype Configuration:
assignment of haplotypes consistent with genotypes

Minimum Recombinant Haplotype Configuration (MRHC) (Qian and Beckmann, AJHG, '02)

Recombinations are (quite) common!

Main assumption:
the most likely solution is the one with the **minimum number of recombinations**

Computational problem:
MRHC
Compute the haplotype configuration with the minimum number of recombinations



Limitations of MRHC

MRHC model generally assumes:

- *complete* genotypes (some methods do not require them)
- *perfect* genotypes

$\Rightarrow$ **unrealistic!**

Our *new* computational problem:

MRHCE = **MRHC** + **Missing Genotypes**
+ **Errors**

Minimum Recombinant Haplotype Configuration with Bounded Errors (and missing data)

Minimum Recombinant Haplotype Configuration with Bounded Errors (MRHCE) problem

Given an incompletely genotyped pedigree and a bound e, compute a haplotype configuration that induces the *minimum number of recombinations and at most e genotyping errors*.

Computational Complexity: MRHCE $\in$ NP_hard

(since MRHC $\in$ NP_hard, Liu *et al.*, TCS, '07)

Algorithm (overview)

- **One main "subroutine":**

  $\texttt{solve\_reHC}(P_G, r, e)$

  Computes, if exists, a haplotype configuration for P_G with at most r recombinations and e errors.

Algorithm solve_MRHCE(P_G, e)

```
r_lb ← r_ub ← 0

while solve_reHC($P_G$, r_ub, $e$) $\neq$ NIL do
    r_lb ← r_ub
    r_ub ← $\max(1, 2$ r_ub$)$

while r_lb + 1 < r_ub do
    r ← $\lfloor$(r_lb + r_ub)$/2\rfloor$
    solve_reHC($P_G$, r, $e$)
    if solution found then r_ub ← r
    else r_lb ← r

return last computed solution
```

Algorithm `solve_reHC`(P_G, r, e)

solve_reHC(P_G, r, e):

- NP_hard problem
- Compute a solution by reduction to SAT
- SAT $\in$ NP_c but solvers are fast in practice
  (MiniSat, CryptoMiniSat, clasp, . . .)

Algorithm idea:

- Encode the instance in a logic formula
- Use a SAT solver to find a truth assignment (if exists)
- Reconstruct the haplotype configuration

Algorithm `solve_reHC(`P_G, r, e`)` - SAT formulation

Four parts:

1 - *Mendelian laws of inheritance:*

- one allele from the father and one from the mother according to the phase
- 6 clauses per individual per locus

2 - *Genotype consistency (errors):*

- the computed haplotypes are consistent with the observed genotypes otherwise $e_i[l]$ is *true*
- at most 3 clauses per individual per locus

Algorithm `solve_reHC(`P_G, r, e`)` - SAT formulation

Four parts:

3 - *Recombinations:*

- if phase changes between adjacent loci, then $r_{p,i}[l]$ is *true*
- 8 clauses per individual per locus

4 - *Cardinality constraints:*

$$\sum_{\substack{\text{individual } i\\ \text{locus } l}} e_i[l] \leq e \qquad \sum_{\substack{\text{individual } i\\ \text{parent } p \text{ of } i\\ \text{locus } l}} r_{p,i}[l] \leq r$$

- encoded via *Cardinality Networks* (Asin *et al.*, Constr., '11)
- $O(nm \log^2 \max\{r, e\})$ clauses

Implementation

`reHCstar`

`https://github.com/yp/reHCstar/`

- Open-source: GPLv3 license
- Includes: CryptoMiniSat 2.9.1
  MiniSat 2.2.0
  (can be used with other solvers as well)

Preliminary experimental evaluation

Does it work?

Preliminary experimental evaluation:

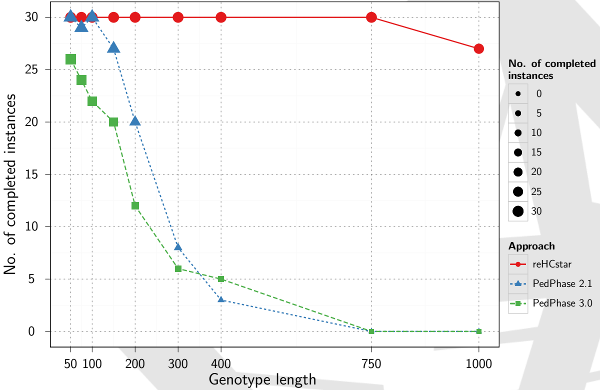
1. Comparison with PedPhase 2.1 and 3.0
2. Analysis of a real and complex cattle pedigree

Comparison with PedPhase 2.1 and 3.0

“Contenders”	
PedPhase 2.1 (Li and Jiang, JCB, ‘05)	Errors: *no* $(e = 0)$ Missing genotypes: *yes* Exact approach: ILP-based
PedPhase 3.0 (Li and Li, JBCB, ‘09)	Errors: *no* $(e = 0)$ Missing genotypes: *yes* Heuristic: concatenation of zero-recombinant blocks

Test instances: different pedigree “topology”, pedigree size, genotype length, recombination and missing rate.

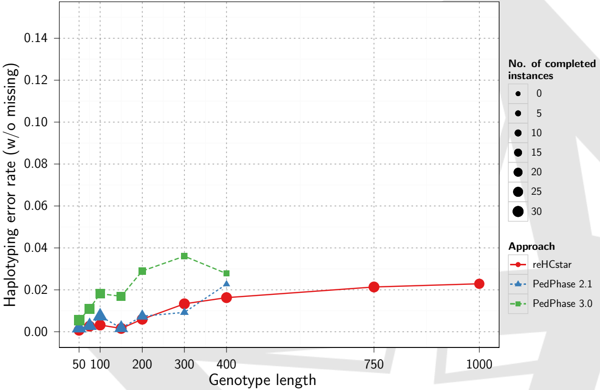
Comparison with PedPhase 2.1 and 3.0



Only `reHCstar` solved almost all the instances!

(one-hour of time-limit for each instance)

Comparison with PedPhase 2.1 and 3.0

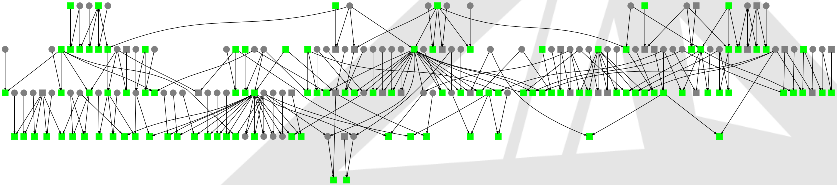


As accurate as PedPhase 2.1 (but scales better)

PedPhase 3.0 was faster but is not as accurate (and does not scale well)

Analysis of a real cattle pedigree

Pedigree: 207 individuals (105 genotyped) on 50 loci



`reHCstar` found a (likely) non-Mendelian genotyping error

Conclusions

Conclusions:

- **MRHCE**: new "realistic" formulation of HI
- `reHCstar`:
  - *Exact* and *scales well* on large/complex pedigrees
  - As *accurate* as existing approaches

Work in progress:

- Integrating NGS data

A fast and practical approach to genotype phasing and imputation on a pedigree with erroneous and incomplete information

Yuri Pirola, Gianluca Della Vedova, Stefano Biffani,
Alessandra Stella, and Paola Bonizzoni

`pirola@disco.unimib.it`

Thank you for your attention!