Efficient whole-genome association mapping using local phylogenies for unphased genotype data

Zhihong Ding, Thomas Mailund, Yun S. Song

Report by Shuwei Li
11/18/08

BLOSSOC (BLOck aSSOCiation)

- Constructs local tree-like genealogies along the genome
- Scores those genealogies according to how the cases and controls are clustered
- Relies on a deterministic, efficient algorithm to build a single tree for each locus (infinite-sites model), achieve efficient computational time.

Phase-known BLOSSOC

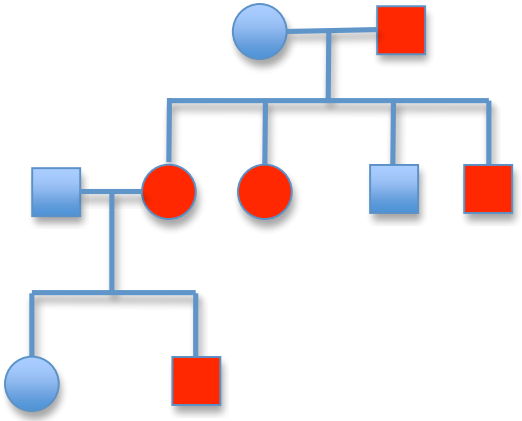
However, a major limitation is:
reliance on having phased haplotype data

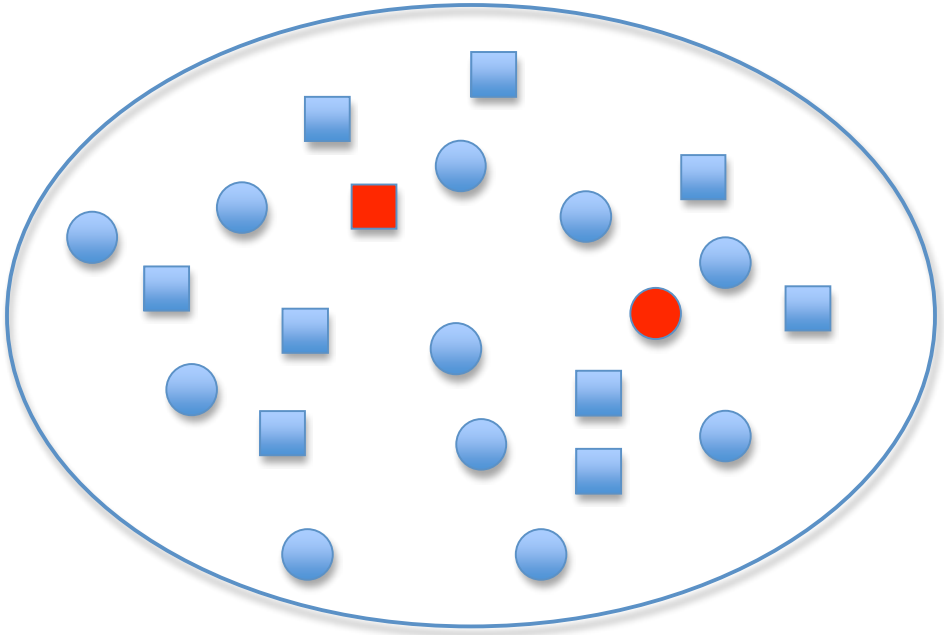
Solution: combines a linear-time algorithm for phasing genotypes on trees with the original tree-based method for association mapping

Phase-unknown BLOSSOC

ALGORITHM: Motivation and Idea

Basic idea: test for a significant clustering of affected individuals in local trees to find a location in the genome that harbors a disease-predisposing mutation.





ALGORITHM: Building local phyogenies

Input: Set of genotypes G and their disease status, user specified number m
Output: Likelihood scores for each marker

PPH: Perfect Phylogeny Haplotyping

LPPH: Linear time Perfect Phylogeny Haplotyping

If the size of I is at least m, then
  build a local tree T for genotypes in I using the LPPH algorithm
If the size of I is less than m
  Add neighboring markers to I until its size equals m
  Use the entropy minimization algorithm to infer the phase of genotypes in I
  Build a local tree T for the haplotypes in I
Score T and output the score as the score for marker i

ALGORITHM: Building local phyogenies

```
Input: Set of genotypes G and their disease
status, user specified number m
Output: Likelihood scores for each marker

For each marker i
   Find the largest interval I around
      marker i such that genotypes in I have
      a PPH solution
   If the size of I is at least m, then
      build a local tree T for genotypes in
      I using the LPPH algorithm
   If the size of I is less than m
      its size equals m
      to infer the phase of genotypes in I
      Build a local tree T for the
      haplotypes in I
   Score T and output the score as the
      score for marker i
```

Initialize X to be the set containing only x. Then alternate two steps until neither is possible:
(1)If X and the next marker immediately to the left together admit a PPH solution, then add that marker to X.
(2) If X and the next marker immediately to the right together admit a PPH solution, then add that marker to X.

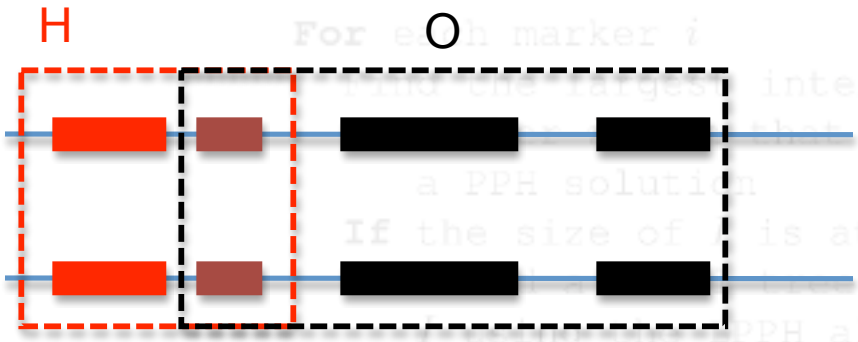
ALGORITHM: Building local phyogenies

```
Input: Set of genotypes G and their disease
status, user specified number m
Output: Likelihood scores for each marker

For each marker i
  Find the largest interval I around
     marker i such that genotypes in I have
     a PPH solution
  If the size of I is at least m, then
     build a local tree T for genotypes in
     I using the LPPH algorithm
  If the size of I is less than m
     Add neighboring markers to I until
       its size equals m
     Use the entropy minimization algorithm
       to infer the phase of genotypes in I
     Build a local tree T for the
       haplotypes in I
  Score T and output the score as the
     score for marker i
```

ALGORITHM: Building local phyogenies

Given a hapotype h and a phasing solution ϕ for a set of n genotypes in G



	H	O
H	1 X 2	1
O	1	2 X 2

$$COV(h,\phi) = n_{(h,other)} + 2n_{(h,h)}$$

$$H(\phi) = \sum_{h:COV(h,\phi)\neq 0} -\frac{COV(h,\phi)}{2n}\log\frac{COV(h,\phi)}{2n}$$

ALGORITHM: Building local phyogenies

Entropy minimization algorithm

(1) Generate a random phasing solution ϕ for genotypes G.

(2) Repeat the following:

(a) Find the pair $(g, (h_1, h_2))$ such that $H(\phi')$ is minimized, where ϕ' is obtained from ϕ by re-explaining $g \in G$ with (h_1, h_2).

(b) If $H(\phi') < H(\phi)$, then let $\phi = \phi'$, else exit the loop.

(3) Output phasing solution ϕ.

ALGORITHM: Scoring local phylogenies

Consider the tree with a hierarchical clustering structure: each partition of the tree into subtrees defines a clustering.

Given a clustering $\mathcal{C}=\{c_1,\ldots,c_n\}$ and corresponding disease risks $\Theta=\{\theta_1,\ldots,\theta_n\}$, the likelihood of the observed disease status is given by

$$L(\mathcal{C},\Theta)=\prod_{i=1}^{n}\theta_i^{A_i}(1-\theta_i)^{U_i},$$

Bayesian approach with independent β-priors $\pi(\theta)=1$.

$$L(\mathcal{C})=\prod_{i=1}^{n}\int_0^1\theta^{A_i}(1-\theta)^{U_i}\pi(\theta)\mathrm{d}\theta=\prod_{i=1}^{n}B(A_i+1,U_i+1),$$

where B is the β function. To integrate out $\mathcal{C}$, we again choose a uniform prior on clusters, obtaining the following final score:

$$L=\frac{1}{|\mathcal{C}|}\sum_{\mathcal{C}}L(\mathcal{C}),$$

BLOSSOC

Evaluation: Ranking Experiments

Table 1. Ranking experiment results showing the average number of top-10 ranked markers within a given distance of the true disease-predisposing locus

ρ	WR	GRR	1 kb			10 kb			100 kb		
			χ^2	P	U	χ^2	P	U	χ^2	P	U
100	5%	1.2	0.08	0.11	**0.12**	0.63	0.76	**0.86**	4.88	**4.89**	4.58
100	5%	1.4	0.06	**0.09**	0.08	0.66	0.75	**0.82**	4.97	5.36	**5.47**
100	5%	1.6	0.13	0.13	0.13	0.97	**1.43**	1.24	5.34	**6.14**	5.99
100	5%	1.8	0.10	0.21	**0.22**	1.14	1.75	**1.78**	6.04	**6.51**	6.40
100	5%	2.0	0.20	**0.21**	0.19	1.26	**1.59**	1.56	5.82	6.11	**6.20**
100	10%	1.2	**0.12**	0.11	0.09	0.70	**0.71**	0.69	4.42	4.50	**4.72**
100	10%	1.4	**0.09**	0.05	0.06	0.85	0.81	**0.90**	**5.37**	5.29	5.20
100	10%	1.6	0.12	0.16	**0.18**	1.25	1.83	**1.84**	6.02	**6.90**	6.84
100	10%	1.8	0.09	**0.14**	0.12	1.03	**1.64**	1.62	5.50	**6.58**	6.52
100	10%	2.0	0.07	0.16	**0.17**	1.06	**1.80**	1.70	5.82	**6.62**	6.54
400	5%	1.2	0.01	0.01	**0.03**	0.37	0.40	**0.43**	2.25	2.46	**2.67**
400	5%	1.4	0.04	**0.05**	0.03	0.40	**0.47**	0.36	2.45	2.78	**2.81**
400	5%	1.6	0.01	**0.03**	0.02	0.55	**0.73**	0.69	2.78	**3.54**	3.26
400	5%	1.8	0.03	0.05	**0.06**	0.50	**0.80**	0.63	2.89	**4.02**	3.74
400	5%	2.0	0.07	**0.09**	**0.09**	0.64	**0.95**	0.86	3.24	**4.61**	4.27
400	10%	1.2	**0.07**	0.03	0.03	0.24	**0.30**	0.27	1.77	2.23	**2.27**
400	10%	1.4	0.13	0.15	**0.16**	0.44	**0.66**	0.61	2.75	**3.46**	3.29
400	10%	1.6	0.07	**0.12**	0.11	0.68	0.82	**0.83**	2.89	3.26	**3.42**
400	10%	1.8	**0.10**	0.08	0.07	0.60	**0.72**	0.61	3.32	**4.18**	4.10
400	10%	2.0	0.10	**0.14**	**0.14**	0.81	1.19	**1.20**	3.58	**4.84**	4.70

Based on 100 simulated datasets each with 1000 case and 1000 control individuals. Columns denoted χ^2 correspond to single-marker χ^2-test results, while columns denoted 'P' ('U', respectively) correspond to BLOSSOC results using $m=5$ for phased (unphased, respectively) data. GRR, 'Genetic Relative Risk'; WR,'Wildtype Risk'.

BLOSSOC

Evaluation: Ranking Experiments

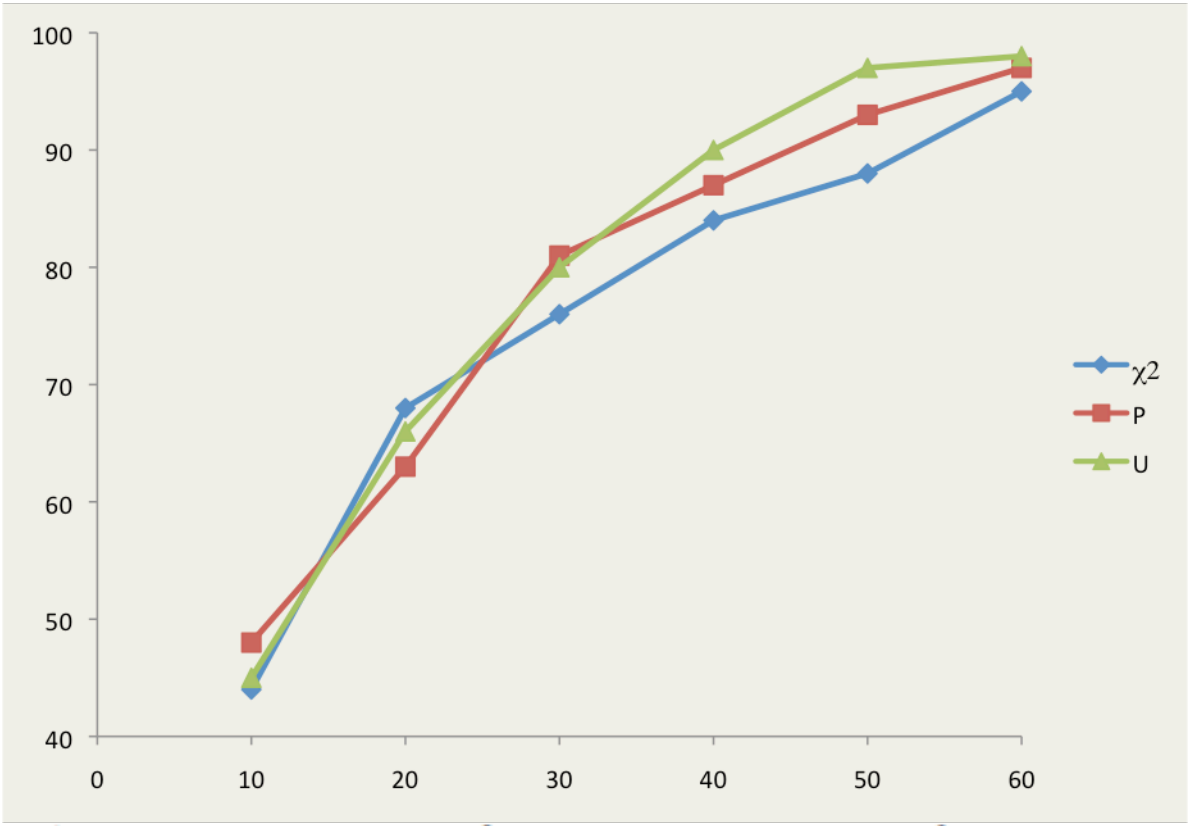
Table 2. Ranking experiment results showing the fraction of data sets with at least one top-10 marker within a given distance of the true disease-predisposing locus

ρ	WR	GRR	1 kb			10 kb			100 kb		
			χ^2	P	U	χ^2	P	U	χ^2	P	U
100	5%	1.2	**0.13**	**0.13**	0.12	0.42	**0.46**	0.39	**0.98**	0.90	0.86
100	5%	1.4	0.15	**0.22**	0.20	0.71	**0.74**	0.70	**0.99**	0.95	0.98
100	5%	1.6	0.27	**0.29**	0.28	0.73	**0.82**	0.80	0.99	0.99	0.99
100	5%	1.8	0.19	**0.30**	0.28	0.79	**0.83**	0.80	**1.00**	0.98	0.98
100	5%	2.0	0.19	**0.30**	0.28	0.81	**0.85**	0.84	1.00	1.00	1.00
100	10%	1.2	0.11	0.11	**0.14**	**0.59**	0.52	0.51	**0.98**	0.94	0.96
100	10%	1.4	0.24	**0.27**	**0.27**	**0.73**	0.66	0.63	**0.98**	0.96	0.95
100	10%	1.6	0.16	**0.26**	0.24	0.74	**0.82**	0.73	**1.00**	0.98	0.99
100	10%	1.8	0.23	**0.30**	**0.30**	**0.85**	**0.85**	0.83	1.00	1.00	1.00
100	10%	2.0	0.28	0.42	**0.43**	0.85	0.86	**0.91**	1.00	1.00	1.00
400	5%	1.2	**0.05**	0.04	0.04	**0.39**	0.27	0.28	**0.91**	0.75	0.80
400	5%	1.4	0.08	0.07	**0.09**	0.46	**0.51**	0.46	**0.95**	0.88	0.86
400	5%	1.6	0.08	**0.13**	0.11	0.45	**0.55**	0.53	**0.96**	0.88	0.94
400	5%	1.8	0.12	**0.16**	0.15	0.70	**0.78**	0.77	**0.99**	**0.99**	0.98
400	5%	2.0	0.16	0.19	**0.20**	0.69	**0.73**	**0.73**	**1.00**	0.99	**1.00**
400	10%	1.2	**0.04**	0.03	**0.04**	0.27	**0.32**	0.23	**0.89**	0.75	0.73
400	10%	1.4	0.03	**0.06**	0.05	0.47	**0.51**	0.48	**0.93**	0.88	0.91
400	10%	1.6	0.12	**0.14**	0.13	0.68	0.73	**0.74**	**0.98**	0.96	0.97
400	10%	1.8	0.14	0.16	**0.17**	0.69	0.76	**0.77**	0.99	0.99	0.99
400	10%	2.0	**0.13**	**0.13**	0.12	0.69	**0.75**	0.74	**1.00**	0.98	0.98

Based on 100 simulated datasets each with 1000 case and 1000 control individuals. Columns denoted χ^2 correspond to single-marker χ^2-test results, while columns denoted 'P' ('U', respectively) correspond to BLOSSOC results using $m=5$ for phased (unphased, respectively) data. GRR, 'Genetic Relative Risk'; WR, 'Wildtype Risk'.

Evaluation: Localization Experiments

Table 3. Percentage of datasets with the highest scoring marker within distance ϵ (in kb) from the disease-predisposing SNP, which is untyped



50	88	93	97		94	96	91		91	94	92		96	98	99
60	95	97	98		97	99	96		95	98	97		97	99	100

Based on 100 simulated datasets for each setting, with 5% WR and $\rho = 40$. Columns denoted χ^2 correspond to single-marker χ^2-test results, while columns denoted 'P' ('U', respectively) correspond to BLOSSOC results using $m = 5$ for phased (unphased, respectively) data.

Evaluation: Parkinson's disease genome-wide dataset

Table 5. Top-10 highest scoring SNPs in the Parkinson disease dataset (Fung *et al.*, 2006) analyzed using BLOSSOC

Chromosome	dbSNP ID	Location	BLOSSOC score
10p12	rs792456	22214547	29.9567
10p12	rs792455	22233428	29.4246
10p12	rs2666781	22245682	29.4223
10p12	rs2807982	22255866	25.9754
10p12	rs2666750	22259562	25.9754
7p15	rs7793103	21920080	16.1051
7p15	rs7798144	21920802	16.1051
7p15	rs11760455	21921256	16.1051
7p15	rs3829757	21921944	16.1051
8p22	rs7824519	14267167	13.1250

Locations correspond to that of NCBI Build 36.1.

Evaluation: Running Time

Table 6. Running times of BLOSSOC

Number of Individuals	m	Number of SNPs	Running time (h)
500	5	100,000	0.81
500	5	200,000	1.67
1000	5	100,000	9.17
1000	5	200,000	19.15
2000	5	100,000	50.83
4000	5	50,000	114.17
500	10	100,000	2.93
500	10	200,000	5.58
1000	10	100,000	33.83
1000	10	200,000	64.04
2000	10	100,000	149.17

See the main text for the computer spec.

Discussion: Future work

- Explore ways of finding the optimal setting for *m* for different datasets

- Develop more efficient algorithms to deal with incompatible regions.

- As a consequence of building local phylogenies, population structure within a sample can result in false-positive associations.

THANKS !

Question