Inferring haplotypes from genotypes

David Balding
Department of Epidemiology and Public Health
Imperial College London

d.balding@imperial.ac.uk

Collaborators: Laurent Excoffier & Guillaume Laval,
Zoological Institute, University of Bern, Switzerland

Paper to appear *Human Genomics*, Oct 2003.

Background: at any genomic locus (except for Y and mtDNA) you have two alleles, which may be the same: say AA, a homozygote. At two other loci, you may be heterozygous with genotypes Bb and Cc.

Problem: classify the 6 alleles into two groups of 3 alleles having the same parent-of-origin:

A B C and A b c ?

or

A b C and A B c ?

Vocab: The unordered allele pairs form a (multi-locus) *genotype*, which is to be decomposed into its two *haplotypes*. The missing information is the *phase*.

Why do we care?

In an ideal world: rarely need to solve this problem,

- haplotypes = intermediate step, not endpoint;
- better to work directly from genotype data.

But: haplotypes, not genotypes, form a unit of inheritance; they form the natural basis for statistical genetics models, and have interpretability advantages.

In the actual world: huge interest in this problem,

- genotypes now readily generated;
- "gold rush" to link haplotypes with disease.

Expensive solutions

- **laboratory methods** are available to identify haplotypes: slow and expensive but can be accurate.

- **pedigree data**, e.g. family trios:

Locus	1	2	3	4
Father	00	01	00	11
Mother	01	11	01	01
Child	00	01	01	01

  can infer that child received 0 0 0 1 from the father and 0 1 1 0 from the mother. **But:** haplotypes cannot be deduced when all three individuals are heterozygous (01) at a locus, and not everyone has their parents' genotypes available.

Statistical methods based on population samples

Shared inheritance $\Rightarrow$ few of the possible haplotypes realised in a population. Equivalently: strong statistical dependence between neighbouring loci on a haplotype.

(Humans well-mixed so effect is universal, but stronger in isolated subpopulations)

- Clark (1990)
- EM algorithm (Excoffier & Slatkin, 1995; Hawley & Kidd, 1995; Long et al., 1995)
- PHASE 1 (Stephens *et al.* 2001)
- HAPLOTYPER (Niu *et al.* 2002)
- ELB (Excoffier, Laval & Balding 2003)
- PHASE 2 (Stephens & Donnelly 2003)

Clark's algorithm

Based on a "parsimony" principle: tries to minimise the number of distinct reconstructed haplotypes.

1. Search for resolved individuals, and record all recovered haplotypes.

2. Consider an unresolved individual. If the genotype can be decomposed into a haplotype pair that includes an existing recovered haplotype: (i) accept this assignment, (ii) add complementary haplotype to list of resolved haplotypes.

3. Repeat until all individuals are resolved or no more haplotypes can be recovered.

Problems:

- No starting point for algorithm.
- Multiple solutions.
- May leave many unresolved individuals.
- How to deal with missing data?
- Does not allow for recombinant haplotypes.

Likelihood-based algorithms: EM, PHASE, H'TYPER

Product-multinomial likelihood:

$$P(\mathbf{G}|\mathbf{h}) = \prod_j \sum_i f\left(G_j|H_i\right) \pi\left(H_{i1}|\mathbf{h}\right) \pi\left(H_{i2}|\mathbf{h}\right)$$

G_j	phase-unknown genotype of individual j
H_i	haplotype pair $= \{H_{i1}, H_{i2}\}$
$\mathbf{h}$	population haplotype frequency vector
f	indicates G_j compatible with H_i
π	multinomial probability mass function

Assumes haplotype pairs independent (Hardy-Weinberg Equilibrium); NB may not hold for case-control data.

EM algorithm obtains MLE of $\mathbf{h}$ via an arbitrary initial assignment $\mathbf{h}^{(0)}$ which is iteratively updated until convergence using **Expectation** and **Maximisation** steps. Severely limited in no. of loci because frequencies of all possible haplotypes are recorded. The H_i obtained via MLE given $\mathbf{h}$.

Both PHASE and HAPLOTYPER are **pseudo-Gibbs samplers**. Haplotypes initially assigned arbitrarily then each H_i updated iteratively conditional on $\mathbf{H}_{-i}$, the other haplotype assignments. Heuristic updating algorithm, not the conditional distribution $f\left(H_i \,|\, \mathbf{G}, \mathbf{H}_{-i}^{(t)}\right)$ under an explicit statistical model.

PHASE and HAPLOTYPER converge to a stationary distribution, and samples of output phases, suitably thinned, give an approximation to this distribution.

Because there is no explicit likelihood, it is not as easy to interpret the stationary distribution as in standard Bayesian settings.

Pseudo-posterior probabilities can be calibrated in simulation studies. Has been done for PHASE with good results.

HAPLOTYPER employs the simple **beta/Dirichlet prior** distribution. It uses a partition-ligation ("divide-and-conquer") algorithm that is very fast, handles large datasets (but only SNPs).

Not only counts of a haplotype are informative about its population frequency, counts of *similar* (e.g. differ by 1 mutation step) haplotypes are also (less) informative. PHASE uses a more complicated **prior based on coalescent theory** to incorporate this effect. It handles SNP and STR (= microsatellite) haplotypes.

Statistically, PHASE beats HAPLOTYPER beats EM beats Clark.

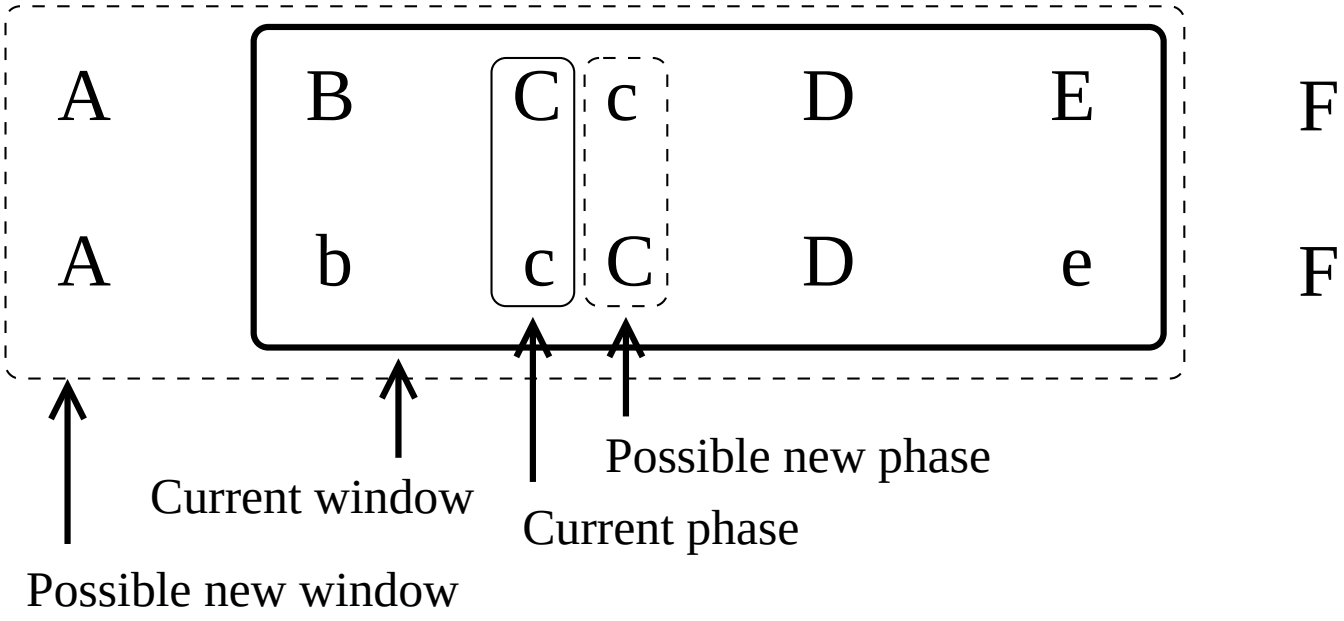
PHASE 1 is very slow. PHASE 2 borrows computational tricks from HAPLOTYPER and is much faster.

Recombination

None of the existing methods is designed to accommodate recombinations. Haplotypes can only be globally well-inferred if LD is very high throughout the region $\Rightarrow$ recombinations are rare. But ...

- recombination hot-spot can occur; whole haplotype accuracy is then infeasible, but we may wish to
  - diagnose the hot-spot,
  - infer partial haplotypes on either side.
- for large genomic regions we may seek haplotypes that are locally accurate, if not globally.
- simple haplotyping algorithm can later be built into e.g. fine mapping algorithm.

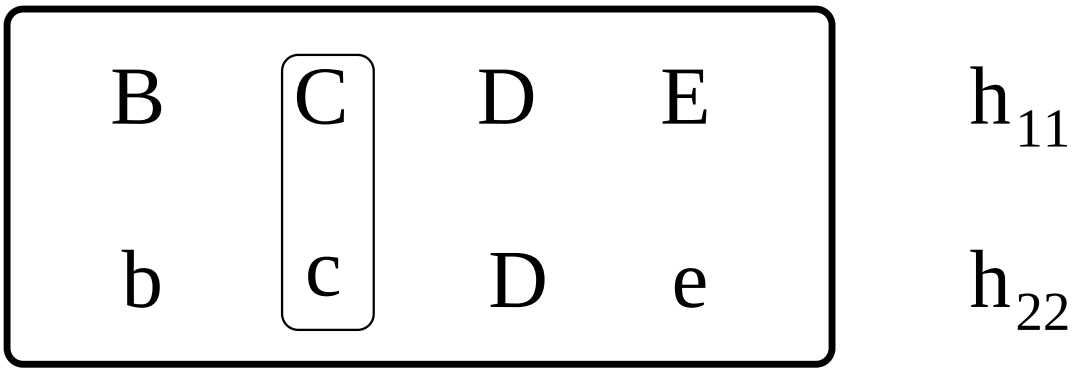
ELB algorithm uses **adaptive windows** to accommodate recombination effects. Associated with the current query locus (here C), there is a window including C and neighbouring loci (here B through E).



ELB iteratively updates (i) the window to maximise information content; (ii) the phase based on current window (here: counts of h'types BXDE, bXDe, X$\in${c,C}).

Haplotype notation for current window:

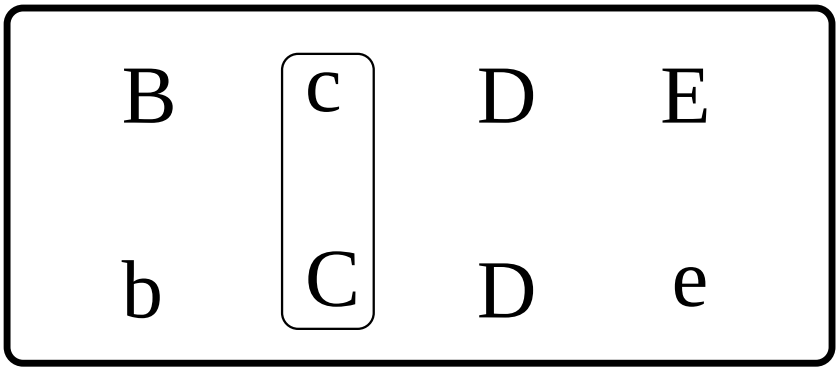
Current phase



h_{11}

h_{22}

Alternate phase



h_{12}

h_{21}

Phase updates

Choose an individual, and update each ambiguous phase in random order.

At each phase call we wish to choose between the current haplotype pair (h_{11}/h_{22}) in the window, and the alternative pair (h_{12}/h_{21}), with probabilities proportional to their joint population proportions. Assuming HWE, these are $p_{11}p_{22}$ and $p_{12}p_{21}$, where p_{ij} denotes the population proportion of h_{ij}.

The p_{ij} are not known, but can be estimated from the n_{ij}, the haplotype counts (within the current window) among the other $n-1$ individuals in the sample.

MLE of p_{ij} not useful here: it would imply that $h_{ij}/h_{i'j'}$ will never be chosen if either $n_{ij} = 0$ or $n_{i'j'} = 0$.

We choose a Bayesian posterior mean under a symmetric Dirichlet prior, and assign h_{11}/h_{22} instead of h_{12}/h_{21} with probability:

$$\frac{(n_{11}+\alpha)(n_{22}+\alpha)}{(n_{11}+\alpha)(n_{22}+\alpha)+(n_{12}+\alpha)(n_{21}+\alpha)} \qquad (1)$$

Increasing α allows more flexibility to "explore" phase assignments not currently observed – but all unobserved haplotypes are treated the same.

We fixed $\alpha = 0{\cdot}01$ in the results below, on the basis of small simulation studies that also considered $\alpha = 0$ and $\alpha = 0{\cdot}1$.

Allowing for mutation

Mutation $\Rightarrow$ rare haplotypes similar to a more common haplotype. Wish to give weight to haplotypes "close" to frequently observed haplotypes.

PHASE uses a coalescent approximation. We adopt simpler, ad-hoc solution: replace each n_{ij} in (1) with

$$n_{ij} + \epsilon x_{11}$$

where x_{ij} = sample count of haplotypes within the current window that differ from h_{ij} by one mutation.

We chose $\epsilon = 0{\cdot}01$ for SNP data, and $\epsilon = 0{\cdot}2$ for STR.

Window updates

Roughly speaking, we want windows to be as large as possible subject to:

1. few recombinations/high LD within the window

2. haplotype counts not too small.

To achieve 1, we wish to choose the window that maximizes $R = \max\{r, 1/r\}$, where $r = p_{11}p_{22}/p_{12}p_{21}$ and so is naturally estimated by

$$\frac{(n_{11}+\epsilon x_{11}+\alpha)(n_{22}+\epsilon x_{22}+\alpha)}{(n_{12}+\epsilon x_{12}+\alpha)(n_{21}+\epsilon x_{21}+\alpha)}$$

Problem: bias towards larger windows, because

$$\begin{aligned} \text{large windows} &\Rightarrow \text{small counts} \\ &\Rightarrow \text{more extreme estimates.} \end{aligned}$$

So we also considered the smoother estimator:

$$\widehat{r} = \frac{(n_{11}+\epsilon x_{11}+\alpha)(n_{22}+\epsilon x_{22}+\alpha)+\gamma}{(n_{12}+\epsilon x_{12}+\alpha)(n_{21}+\epsilon x_{21}+\alpha)+\gamma} \qquad (2)$$

and found that it conveyed no advantage for STR data, but $\gamma = 0{\cdot}5$ gave a significant advantage for SNP data.

Prior to each phase call, we consider two successive alterations to the current window for that individual at that locus: extension by one locus at one end, then reduction by one locus at the other end. Each proposed alteration is accepted/rejected with probability

$$\frac{\hat{R}_1}{\hat{R}_1 + \hat{R}_2},$$

where $\hat{R}_k = \max\{\hat{r}_k, 1/\hat{r}_k\}$ for window k, $k = 1, 2$, and $\hat{r}_k$ is given at (2).

If both proposals are rejected, the previous window is retained.

Global and local accuracy

Global accuracy: the proportion of ambiguous individuals whose entire haplotype pair is correctly recovered: not very useful for large genomic regions.

Local accuracy: several measures available. We measure the proportion of pairs of successive ambiguous loci for which the phase is correctly recovered. NB allele switch at a locus

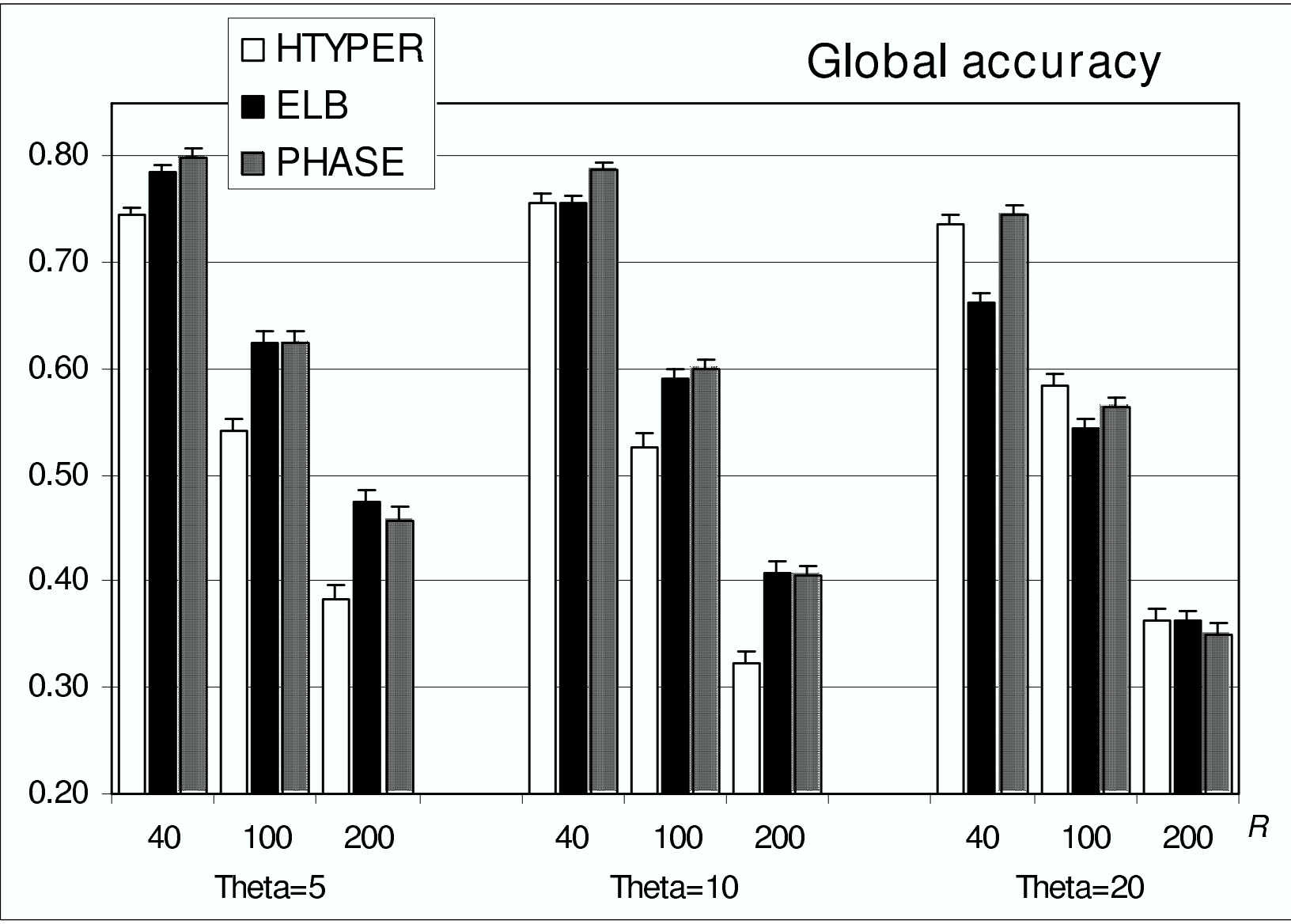
True						Recovered					
A	B	C	D	E	F	A	B	C	d	E	F
a	b	c	d	e	f	a	b	c	D	e	f

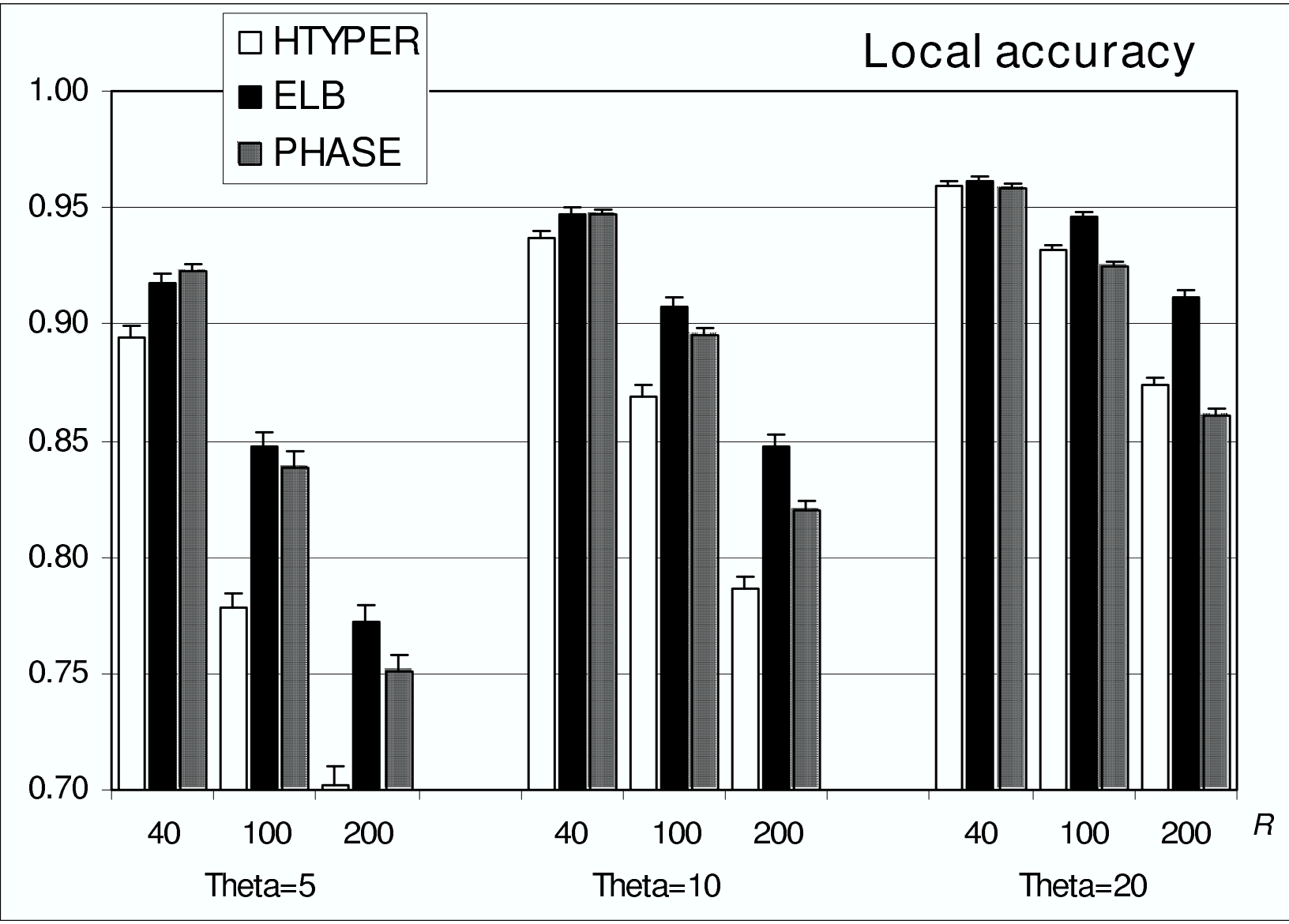
counts as two phase errors.

Simulation datasets 1: SNP

SNP = Single Nucleotide Polymorphism

	Mutat. param.	Recomb. param.	# var. sites	[range]	Pairwise discord.
1	5	40	25	[14-39]	4.8
2	5	100	25	[13-44]	4.8
3	5	200	25	[10-38]	4.9
4	10	40	49	[33-69]	9.9
5	10	100	48	[31-70]	9.6
6	10	200	48	[30-61]	9.6
7	20	40	90	[65-127]	18.3
8	20	100	90	[65-109]	18.7
9	20	200	89	[65-119]	18.5


Global accuracy
HTYPER
ELB
PHASE
0.80
0.70
0.60
0.50
0.40
0.30
0.20
40
100
200
40
100
200
40
100
200
R
Theta=5
Theta=10
Theta=20

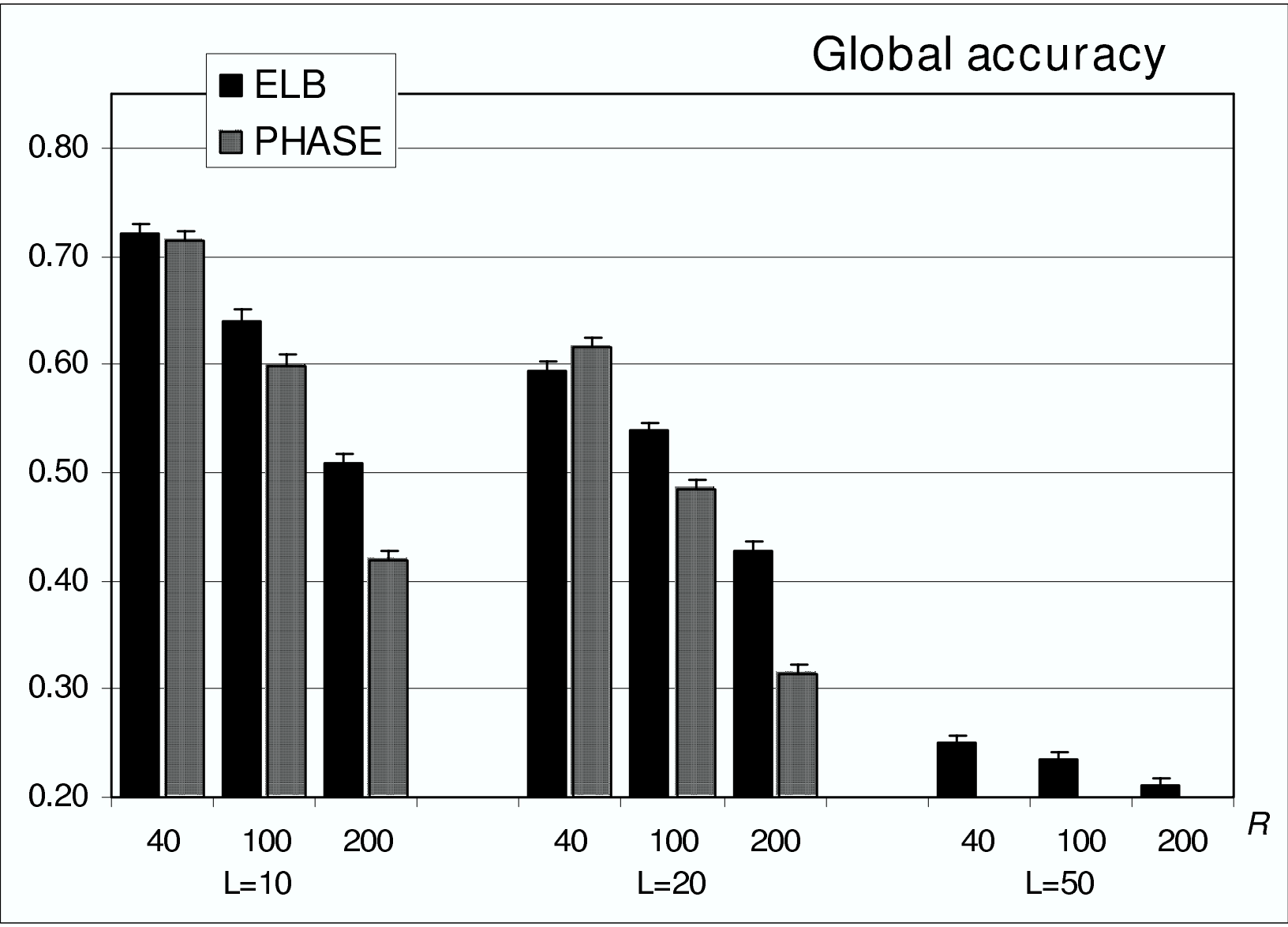

HTYPER
ELB
PHASE
Local accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
40
100
200
40
100
200
40
100
200
R
Theta=5
Theta=10
Theta=20

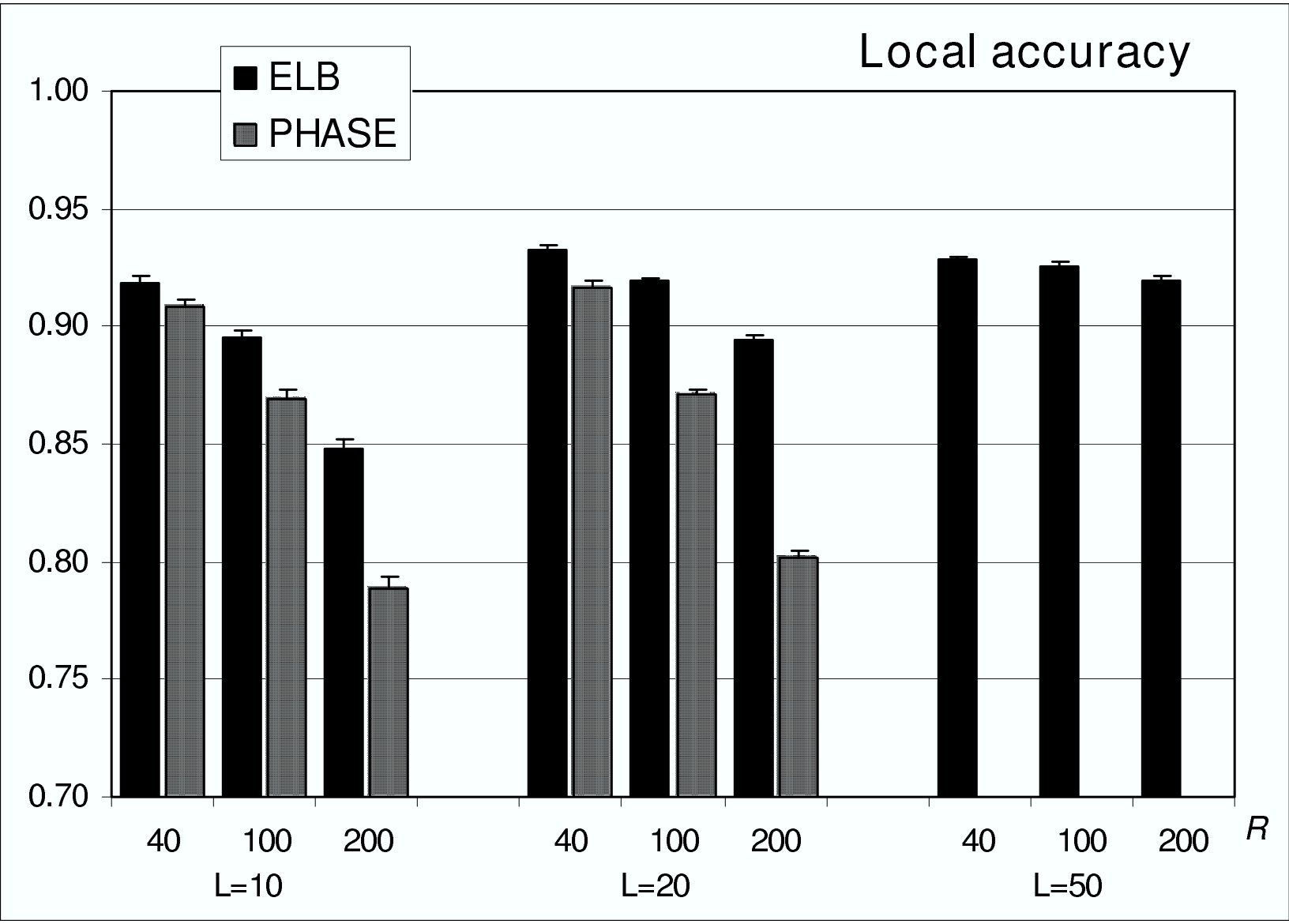
Simulation datasets 2: STR

STR = Short Tandem Repeat (or microsatellite)

	Mutat. param.†	Recomb. param.	# loci	Pairwise discord.
10	10	40	10	7.8
11	10	100	10	7.9
12	10	200	10	7.8
13	10	40	20	15.7
14	10	100	20	15.6
15	10	200	20	15.6
16	10	40	50	39.1
17	10	100	50	39.1
18	10	200	50	39.1

† per locus.


Global accuracy
ELB
PHASE
0.80
0.70
0.60
0.50
0.40
0.30
0.20
40
100
200
40
100
200
40
100
200
R
L=10
L=20
L=50

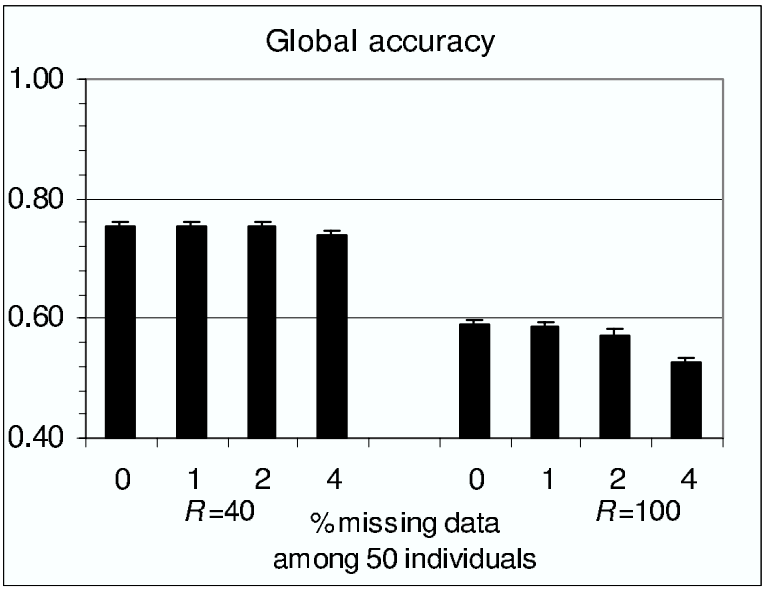

Local accuracy
ELB
PHASE
1.00
0.95
0.90
0.85
0.80
0.75
0.70
40
100
200
40
100
200
40
100
200
R
L=10
L=20
L=50

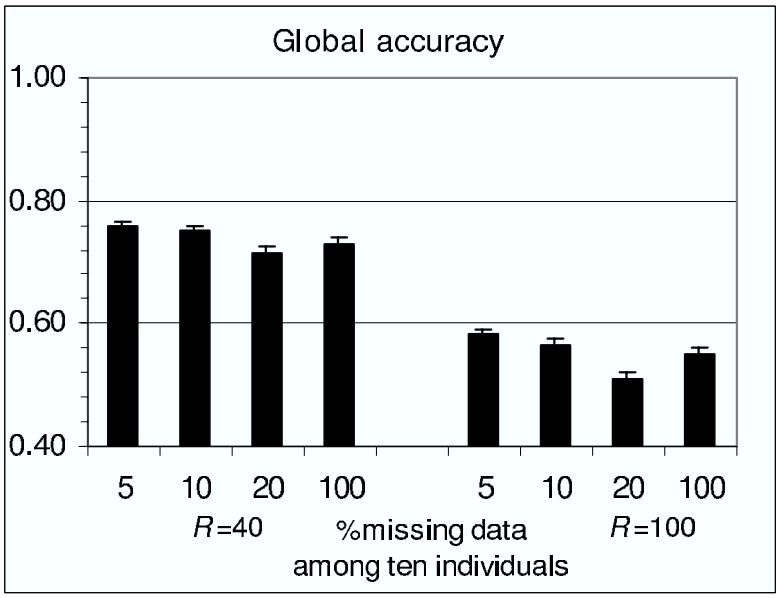
Simulation datasets 3: SNP with missing data

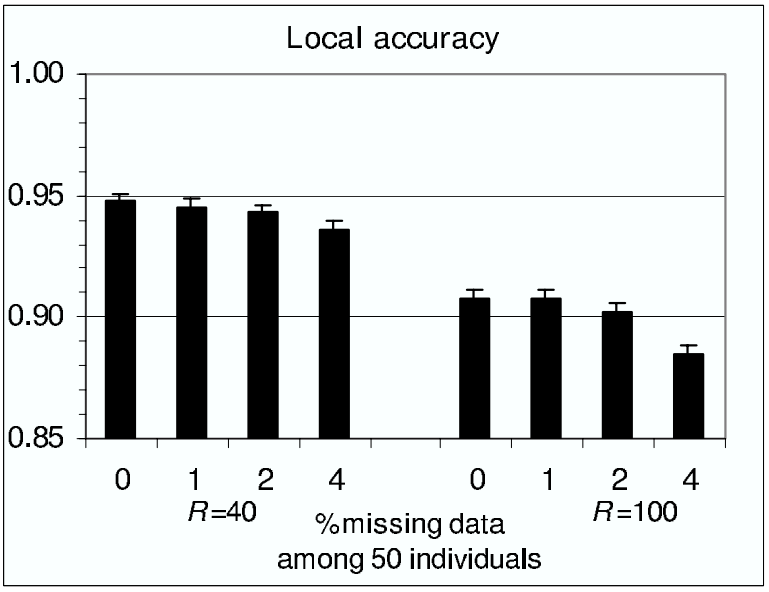
We investigated two scenarios for missing SNP data:

- A low proportion (1%, 2%, and 4%) of missing data is distributed uniformly across all individuals.

- 40 individuals were error-free, the remaining 10 individuals had 5%, 10%, 20%, and 100% missing data.


Global accuracy
1.00
0.80
0.60
0.40
0 1 2 4 0 1 2 4
R=40
%missing data
among 50 individuals
R=100


Global accuracy
1.00
0.80
0.60
0.40
5 10 20 100 5 10 20 100
R=40
%missing data
among ten individuals
R=100


Local accuracy
1.00
0.95
0.90
0.85
0 1 2 4 0 1 2 4
R=40
%missing data
among 50 individuals
R=100

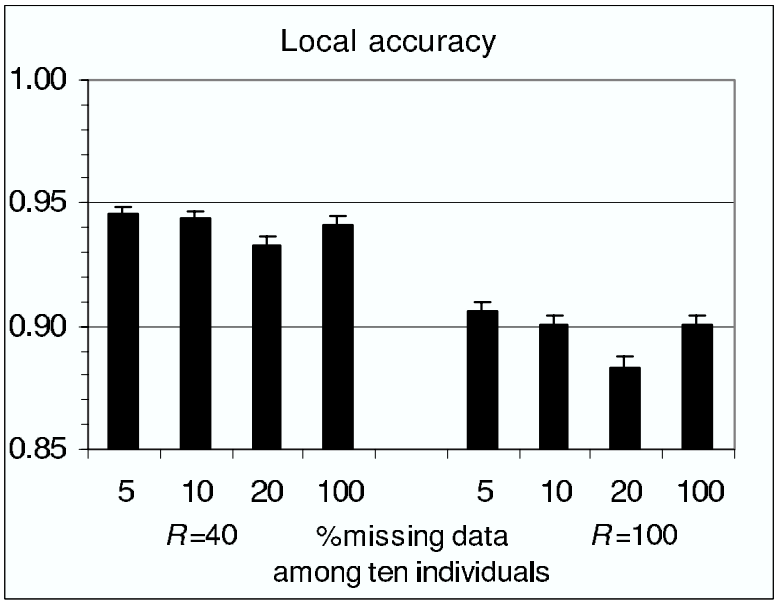

Local accuracy
1.00
0.95
0.90
0.85
5 10 20 100 5 10 20 100
R=40
%missing data
among ten individuals
R=100

Figure 3

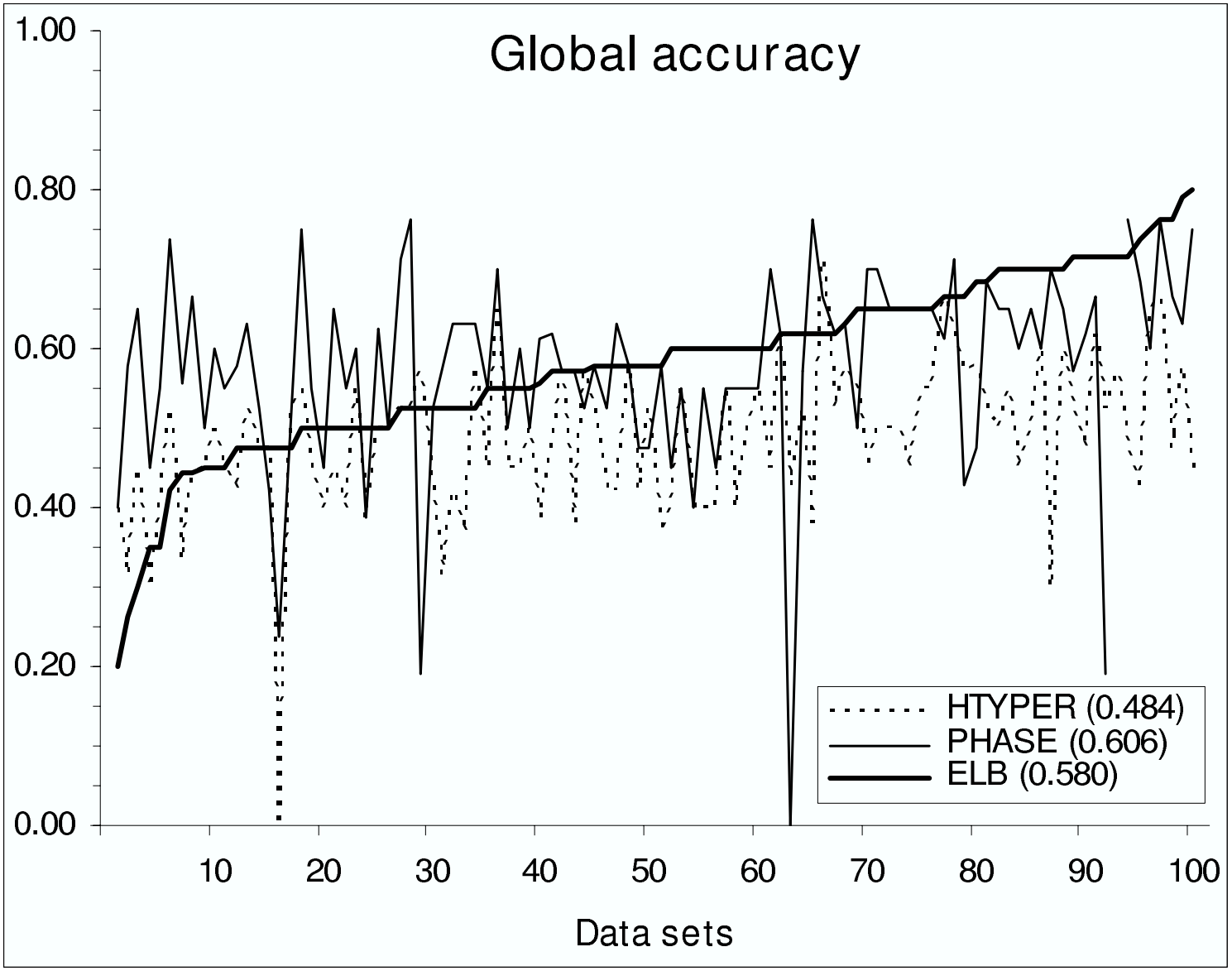
"Real" data

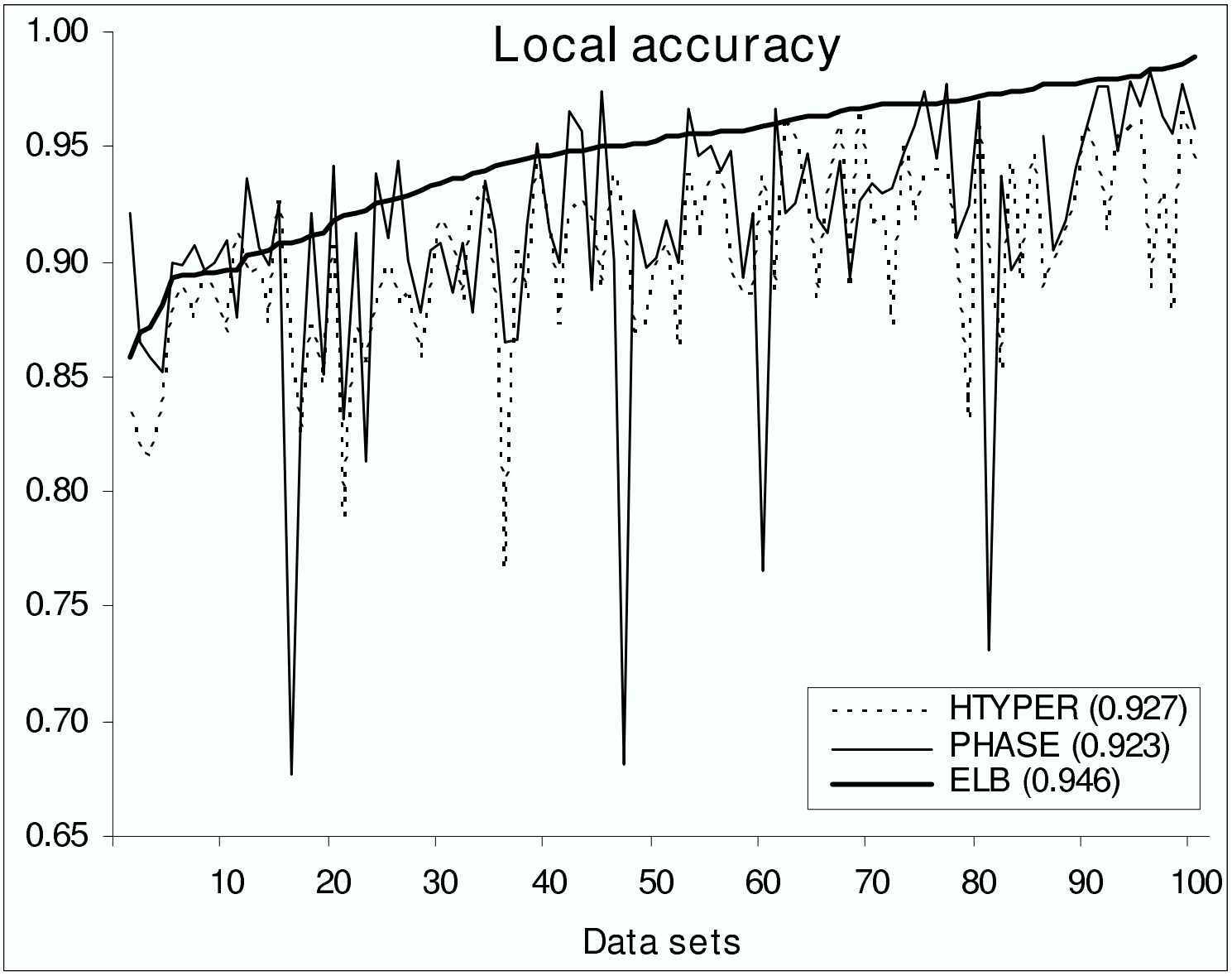
100 datasets generated by randomly pairing 42 human male X chromosomes, typed in a number of short fragments over a 193-Kb low-recombination region. There were 97 sites at which at least one chromosome differed from a reference chromosome; on average, chromosome pairs differed at 31 sites.

The chromosomes were drawn from:

23	Afrikaner men
9	Ashkenazim
3	British
3	Swedes
3	Greeks

and the reference individual was Italian (Prof. Francisco Gianelli, from Guy's Hospital London).


Global accuracy
1.00
0.80
0.60
0.40
0.20
0.00
10
20
30
40
50
60
70
80
90
100
Data sets
HTYPER (0.484)
PHASE (0.606)
ELB (0.580)


Local accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
10
20
30
40
50
60
70
80
90
100
Data sets
HTYPER (0.927)
PHASE (0.923)
ELB (0.946)

Conclusions

- ELB is based on simple heuristic ideas, contains ad-hockeries and several "fudge factors".

- best suited to large genomic regions where recombination is non-negligible.

- superior in terms of local accuracy, near optimal for global accuracy and very fast.

- substantial room for further development and optimisation.