Performance analysis of methods to infer missing genotypes

Christine Sinoquet

LABORATOIRE D'INFORMATIQUE DE NANTES-ATLANTIQUE

Performance analysis of methods to infer missing genotypes

Christine Sinoquet

Computer Science Institute of Nantes-Atlantic (Lina), U.M.R. C.N.R.S. 6241, University of Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex, France

— Bioinformatics —

RESEARCH REPORT

N° hal-00326741

October 2008

LINA, Université de Nantes – 2, rue de la Houssinière – BP 92208 – 44322 NANTES CEDEX 3
Tél. : 02 51 12 58 00 – Fax. : 02 51 12 58 12 – http://www.sciences.univ-nantes.fr/lina/

Christine Sinoquet

Performance analysis of methods to infer missing genotypes

16 p.

Les rapports de recherche du Laboratoire d'Informatique de Nantes-Atlantique sont disponibles aux formats PostScript® et PDF® à l'URL :
http://www.sciences.univ-nantes.fr/lina/Vie/RR/rapports.html

Research reports from the Laboratoire d'Informatique de Nantes-Atlantique are available in PostScript® and PDF® formats at the URL:
http://www.sciences.univ-nantes.fr/lina/Vie/RR/rapports.html

Performance analysis of methods to infer missing genotypes

Christine Sinoquet

christine.sinoquet@univ-nantes.fr

Abstract

Complex analyses such as genetic mapping, disease association studies, disease mapping in the context of environmental health and environmental epidemiology studies rely on high-throughput genotyping techniques. These analyses thoroughly examine genetic variations between subjects, in particular through Single Nucleotide Polymorphism (SNP). Nonetheless, though nowadays genotyping techniques impose high-quality standards, one still has to cope with the issues of missing data and genotyping errors. Typically, the percentage of missing data - or missing calls - now ranges in interval $[5\%, 10\%]$. Computational inference of missing data represents a challenging alternative to genotyping again the missing regions. This document first briefly reviews the various methods designed to infer missing SNPs. Then, it reports performances published for these inference methods. The present report carefully describes the characteristics of the different benchmarks generated by the designers (missing data percentage, correlation between SNPs). We show that most methods provide accuracies in the range $[90\%, 96\%]$. However, we also emphasize that no algorithm garantees constant high accuracies: an algorithm may perform well on some benchmarks and show in contrast relatively poor results on others.

Introduction

DNA polymorphism denotes a DNA sequence variation between members of a species (or between paired chromosomes in an individual). Single Nucleotide Polymorphism, or SNP, occurs when the possible nucleotides observed over a population, for a given locus, restrain to less than the four variants A, C, T and G. Almost all common SNPs have only two variants, also called alleles.

Though nowadays high-throughput genotyping techniques tend to produce data of increasing quality, the generation of data with missing Single Nucleotide Polymorphisms (SNPs) remains prejudicial to analyses such as association studies, for example. These analyses aim at dissecting the genetic susceptibility of complex diseases. Actually, it was shown that even low missing data percentages are likely to impact detection power (Chen *et al.*, 2004; Liu *et al.*, 2006; Dai *et al.*, 2006; Croiseau *et al.*, 2007). To avoid genotyping again the missing data - a prohibitive task both in terms of cost and time - *in silico* inference methods have been proposed.

The present document reviews performances for missing genotype inference methods. In this report, we are only interested in accuracies, not in running times. Besides, we only focus on off-line SNP inference. Indeed, some software packages dedicated to association studies and haplotype inference also allow missing data handling (Qin *et al.*, 2002; Niu *et al.*, 2002; Xie *et al.*, 2005; Dai *et al.*, 2006). In such cases, literature only provides quality indicators for the task of interest, not for the SNP imputation.

1 Description of the off-line methods selected

Tables 1A and 1B briefly describe various off-line methods designed to infer missing genotypes. All methods described here are statistical methods, except one (Roberts *et al.*, 2007).

reference	method
IPI **Su et al. (2005)**	**haplotype block partitioning based on entropy measure** A two-step iterative partition-inference algorithm is used. At first step, a dynamic programming algorithm is used to partition haplotypes into blocks. At second step, a process inspired from the expectation-maximization algorithm infers missing SNPs for each haplotype block, minimizing each block entropy. The algorithm iterates these two steps until the total block entropy is minimized. *Tuning parameters:* none
fastPHASE **Scheet and Stephens (2006)**	**hidden Markov model** To capture the fact that, over short regions, subjects in a population share only a few haplotypes and to tackle the related problem of dealing with unknown bounds of block-like patterns of linkage disequilibrium, a hidden Markov model allows cluster membership variation of haplotypes along the whole chromosome *Tuning parameters:* none
cited in **Yu and Schaid (2007)**	**k-nearest neighbour method:** Subjects with similar flanking SNPs are used to predict a missing call. *Tuning parameters:* window size, number of nearest neighbors
cited in **Yu and Schaid (2007)**	**weighted k-nearest neighbour method:** The method above is improved through contribution weighting of each neighbour. The weight is proportional to the similarity between the neighbour's flanking SNPs and those of the marker which is inferred. *Tuning parameters:* the same as above method
cited in **Yu and Schaid (2007)**	**regression tree method:** A partition tree is built through recursive binary splits. The value assigned to a missing call is the prevailing SNP observed over subjects belonging to the same subtree. *Tuning parameters:* complexity parameter to prevent unjustified splitting, minimal number of subjects assigned to a node

Table 1A Categorization of off-line missing SNP inference methods.

reference	method
Yu and Schaid (2007)	**linear regression with backward elimination:** The predictors are selected from flanking SNPSs, using a backward stepwise process. *Tuning parameters:* number of candidate SNPs used as predictors, on each side of the marker which is inferred.
Yu and Schaid (2007)	**linear regression with Least Angle Regression (**LARS**)** After picking the predictor most correlated with the response, forward stepwise selection is enhanced successively (i) bringing a new predictor into the model if it shows as much correlation with residual as the previously selected predictors, (ii) moving in a direction "equiangular" between all selected predictors (Enfron *et al.*, 2004). *Tuning parameters:* the same as above method
Yu and Schaid (2007)	**linear regression with Single Value Decomposition (**SVD**)** Single value decomposition provides Eigen vectors for SNPs. Such linear combinations of SNPs are used as covariates in linear regression. *Tuning parameters:* minimal percentage of variance explained by selected Eigen vectors, number of candidate SNPs considered on each side of the marker which is inferred.
NPUTE **Roberts** *et al.* **(2007)**	**nearest neighbor method combined with window sliding** The marker array is scanned with a sliding window. Nearest-neighbour SNP inference is performed within the frame of current window, for all missing calls in the row located at the centre of this window. The very point fundamental to this algorithm is the efficient knowledge updating of current window from previous overlapping window. *Tuning parameters:* window size
Sun et Kardia (2008)	**neural network (**NN**)** For each missing call, the χ^2 independence test is performed to identify the five most correlated SNPs.Then all 31 possible NN models are tested (1 SNP, 2 SNPs ... 5 SNPs. *Tuning parameters:* weights of the single hidden layer NN

Table 1B Categorization of off-line missing SNP inference methods.

2 Variation range for the performances of SNP imputation methods

Tables 2A, 2B and 2C compile the performances published for off-line inference methods. We dismissed the records relative to methods based on SVD linear regression, regression tree and brute applications of k-nearest neighbour approaches: such methods did not pass the 90% accuracy threshold in nearly all cases.

Our compilation would be meaningful without a detailed description of the benchmarks used. The difficulty of inference depends on data structure: local haplotype-block patterns or in contrast, mozaic patterns; density of markers; percentage of missing calls (We recall the reader that a genotype is the combination of two homologous haplotypes). Linkage disequilibrium is a main factor among various other parameters determining haplotype-block patterns, and thus local constraints on genotypes. In the following, for benchmark description, we refer to linkage disequilibrium as LD. LD describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected through the random formation of haplotypes from alleles, based on their frequencies. Among determining factors, the density of markers is unevenly described by authors.

In addition to missing data percentage, some studies generate simulated data, also constraining LD level. Reporting fastPHASE's performances for the need of their study, Yu and Schaid selected SNPs from the Human HapMap data, imposing the satisfaction of three LD constraints (**III**). The 100 first ranked SNPs that showed minor allele frequencies above 5% and p-values for the Hardy-Weinberg equilibrium test greater than 0.01 were considered as SNPs in strong LD. Then, SNPs in weak LD were selected such that the square of Pearson correlation coefficient between any two adjacent markers would be less than 0.1. The threshold for SNPs in no LD has been set to 10^{-4}, in this work.

In their study, Sun and Kardia also controled the LD parameter (**VI**). The three LD thresholds chosen are 0.2, 0.5 and 0.8. Besides, this work illustrates the use of the *ms* program to generate a coalescent model, therefore controling recombination and mutation rates (Hudson, 2002).

In Tables 2A through 2C, we observe a wide range of variations: most algorithms provide accuracies in interval $[90\%, 96\%]$. Algorithm fastPHASE shows the highest performances, possibly reaching the percentage of 97%. fastPHASE efficiency is mostly confirmed by all comparative analyses available (not reported here). Nonetheless, we learn from Table 2 that algorithms may perform well on some benchmarks and show in contrast relatively disappointing results on other datasets. This remark includes fastPHASE. For example, NPUTE provides accuracies close to 97% on the so-called 150 k dataset, whereas the range is around 93-94% for the mouse Perlegen benchmark.

Conclusion

Skimming over the published results of off-line imputation methods is enlightening enough to show that so far, the accuracy rates mostly range in interval $90 - 96\%$, depending on the data. Thus, in itself, any attempt to improve accuracy in SNP inference, if it were only for 2%, seems all the more valuable.

It could also be worth investigating whether an iterative process would improve accuracy for missing genotype inference. Confronting the imputations achieved by, say, two algorithms chosen amongst the best performing ones, one could fix as definitely resolved the missing calls whose variants are similarly inferred by these algorithms. After updating the SNP dataset with these newly imputed SNPs, another inference round would be performed. The entire process would be iterated until all remaining missing calls can not be fixed.

In the same line, for algorithms which implement the scanning of the SNP array through a sliding window, it could be worth implementing an iterative process, that time confronting the results of parsing in opposite directions. Thus, the marker array would successively be processed through rounds involving a scan from top to bottom and a scan from bottom to top.

In the field of association studies, it is expected that weak association identification will all the more benefit from SNP inference as the latter is more accurate. In particular, the identification of associations between a combination of SNPs and a disease where each SNP contributes only a few should have much to gain from more and more accurate SNP imputation.

References

Barkley, R.A., Chakravarti, A., Cooper, R.S., *et al.* 2004. Family blood pressure program: positional identification of hypertension susceptibility genes on chromosome 2. *Hypertension* 43, 477–482.

Chen, J., Peters, U., Foster, C., *et al.* 2004. A haplotype based test of association using data from cohort and nested case-control epidemiologic studies. *Hum. Hered.* 58, 18–29.

Dai, J.Y., and Ruczinski, I., and LeBlanc, M., *et al.* 2006. Imputation methods to improve inference in SNP association studies. *Genet. Epidemiol.* 30, 8, 690–702.

Daly, M.J., and Rioux, J.D., *et al.* 2001. High-resolution haplotype structure in the human genome. *Nat. genet.* 29, 229–232. doi: 10.1038/ng1001-229.

Efron, B., and Hastie, T., and Johnstone, I., *et al.* 2004. Least angle regression. *Annals of statistics* 32, 2, 407-499.

FBPP Investigators. 2002. Multi-center genetic study of hypertension: the Family Blood Pressure Program (FBPP) *Hypertension* 39, 3–9.

Frazer, K.A., Eskin, E., Kang, H.M. *et al.* 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* 448, 7157, 1050-1053.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics.* 18, 337–378.

Liu, W., Zhao, W., and Chase, G.A. 2006. The impact of missing and erroneous genotypes on tagging SNP selection and power of subsequent association tests. *Hum. Hered.* 61, 31–44.

Niu, T., Qin, Z. S., Xu, X., *et al.* 2002. Bayesian haplotype inference for multiple linked Single-Nucleotide Polymorphisms. *Am. J. Hum. Genet.* 70, 1, 157–169.

Patil, N., Berno, A.J., Hinds, D.A., *et al.* 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294, 5547, 1719–1723.

Qin, Z.S., Niu, T., and Liu, J.S. 2002. Partition ligation expectation maximization algorithm for haplotype inference with Single Nucleotide Polymorphisms. *Am. J. Hum. Genet.* 71, 5, 1242–1247.

Roberts, A., McMillan, L., Wang, W., *et al.* 2007. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* 23, 13, i401–i407. doi:10.1093/bioinformatics/btm 220.

Scheet, P., and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase, *Am. J. Hum. Genet.* 78, 4, 629–644.

Su, S.-C., Jay Kuo, C.-C., and Chen, T. 2005. Inference of missing SNPs and information quantity measurements for haplotype blocks. *Bioinformatics* 21, 9,2001–2007.

Sun, Y.V., and Kardia, S.L. 2008. Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. *Eur. J. Hum. Genet.* 16, 4, 487–95.

Wade, M., and Daly, M. J. 2005. Genetic variation in laboratory mice. *Nat. genet.* 37, 11, 1175–1180.

Xie, Q., Ratnasinghe, L.D., Hong, H., *et al.* 2005. Decision Forest Analysis of 61 Single Nucleotide Polymorphisms in a case-control study of esophageal cancer; a novel method. *BMC Bioinformatics*, 6, 2, doi:10.1186/1471-2105-6-S2-S4.

Yu, Z., and Schaid, D.J. 2007. Methods to impute missing genotypes for population data. *Hum. Genet.* 122, 5,

495–504.

	algorithm	method description	data description				
			number of SNPs	number of subjects	p_{miss} (%)	accuracy (%)	comment
I	IPI	haplotype block partitioning	103	387	1	94.92	benchmark from Daly *et al.* (2001)
		based on entropy measure	""	""	5	92.66	accross a 500 kb region (1)
		Su et al. (2005)	""	""	10	92.02	
			5200	20	1	92.31	benchmark from Patil *et al.* (2001)
			""	""	5	90.77	
			""	""	10	90.58	
II	fastPHASE	cluster membership variation	216	24	5	94.7	SeattleSNPs Variation Discovery Res., AA
		guided by hidden Markov model	""	23	""	97.6	Seattle benchmark, ED
		Scheet and Stephens (2006)	""	47	""	96.3	"", AA + ED
			41018	60	10	96.6	CEPH HapMap data, chromosome 7
			""	""	25	95.9	accross a 159 Mbp region
			15532	""	10	96.7	CEPH HapMap data, chromosome 22
			""	""	25	96.1	accross a 35 Mbp region
III	fastPHASE	in **Yu and Schaid (2007)**					extracts from a comparative analysis HapMap data, chromosome 22
			100	60	5	91.8 - 95.1 (2)	CEU, SNPs in strong LD (3)
			""	""	""	88.4 - 93.0	CEU, SNPs in weak LD, $r^2 < 0.1$ (4)
			""	""	""	61.5 - 63.5	CEU, SNPS in no LD, $r^2 < 10^{-4}$ (5)
			100	90	5	92.6 - 95.6	J/C, (3)
			""	""	""	89.3 - 94.4	J/C, (4)
			""	""	""	64.9 - 66.1	J/C, (5)
			100	60	5	87.7 - 90.8	YRI, (3)
			""	""	""	83.0 - 86.1	YRI, (4)
			""	""	""	67.9 - 69.1	YRI, (5)

Table 2A Accuracy percentages for various off-line imputation methods. p_{miss}: percentage of missing data; AA: African American population; ED: population of European descent; CEPH: Utah residents with ancestry from northern and western Europe; CEU: 60 founders from the Centre d'Etude du Polymorphisme Humain; J/C: 45 Japanese from Tokyo, Japan, and 45 Han Chinese from Beijing, China; YRI: 60 founders from the Yoruba in Idaban, Nigeria; LD: linkage disequilibrium; (1): density of markers; (2): accuracy range over various tuning parameter values; (3): the 100 top-ranked SNPs showing minor allele frequencies above 5% and p-values for the Hardy-Weinberg equilibrium test greater than 0.01 were considered as SNPs in strong LD.

	algorithm	method description	data description				
			number of SNPs	number of subjects	p_{miss} (%)	accuracy (%)	comment
IV		linear regression with					HapMap data, chromosome 22
		backward elimination	100	60	5	89.1 - 93.1(2)	CEU, (3)
		Yu and Schaid (2007)	""	""	""	85.2 - 90.6	CEU, (4)
			""	""	""	52.3 - 64.1	CEU, (5)
			100	90	5	88.4 - 93.8	J/C, (3)
			""	""	""	86.6 - 92.7	J/C, (4)
			""	""	""	61.2 - 66.7	J/C, (5)
			100	60	5	83.4 - 87.1	YRI, (3)
			""	""	""	78.9 - 83.0	YRI, (4)
			""	""	""	60.4 - 69.5	YRI, (5)
V		linear regression					HapMap data, chromosome 22
		LARS	100	60	5	89.2 - 94.2	CEU, (3)
		Yu and Schaid (2007)	""	""	""	85.3 - 91.6	CEU, (4)
			""	""	""	62.7 - 64.5	CEU, (5)
			100	90	5	88.4 - 94.3	J/C, (3)
			""	""	""	86.8 - 93.1	J/C, (4)
			""	""	""	66.2 - 66.9	J/C, (5)
			100	60	5	83.5 - 89.3	YRI, (3)
			""	""	""	79.2 - 83.6	YRI, (4)
			""	""	""	68.9 - 69.9	YRI, (5)

Table 2B Accuracy percentages for various off-line imputation methods. p_{miss}: percentage of missing data; CEU: 60 founders from the Centre d'Etude du Polymorphisme Humain; J/C: 45 Japanese from Tokyo, Japan, and 45 Han Chinese from Beijing, China; YRI: 60 founders from the Yoruba in Idaban, Nigeria; LD: linkage disequilibrium; (2): accuracy range over various tuning parameter values; (3): strong LD: the 100 top-ranked SNPs showing minor allele frequencies above 5% and p-values for the Hardy-Weinberg equilibrium test greater than 0.01 were considered as SNPs in strong LD; (4): weak LD, $r^2 < 0.1$; (5): no LD, $r^2 < 10^{-4}$.

	algorithm	method description	data description				
			number of SNPs	number of subjects	p_{miss} (%)	accuracy (%)	comment
VI		neural network					coalescent model generated by the
		Sun et Kardia (2008)					*ms* program (Hudson, 2002)
			1000	10^4	1	95.9	recombination and mutation rates
			""	""	5	94.7	both equal to 10^{-8} (*)
			""	""	10	94.7	across a 6 Mpb region
			680	10^4	1	93.1	$r^2 < 0.8$ selected from (*)
			""	""	5	92.9	""
			""	""	10	92.1	""
			552	10^4	1	92.5	$r^2 < 0.5$ selected from (*)
			""	""	5	92.7	""
			""	""	10	91.6	""
			288	10^4	1	93.2	$r^2 < 0.2$ selected from (*)
			""	""	5	92.3	""
			""	""	10	92.0	""
			1962	90	1	96.2	CEPH HapMap chromosome 22
			""	""	5	95.4	""
			""	""	10	95.1	""
			126	1458	1	86.8	chromosome 2, GENOA, FBPP Investigators (2002)
			""	""	2	86.5	data from Barkley *et al.*, 2004
			""	""	5	83.1	""
VII	NPUTE	nearest neighbour method	1024	46	5	$\sim$ 97 (6)	150 k benchmark
		combined	""	""	10	(6)	combined SNPs from the 140 k Broad/MIT mouse
		with window sliding	""	""	15	(6)	Wade and Daly (2005)
		Roberts *et al.* **(2007)**	""	""	20	(6)	and the 10 k GNF mouse dataset
			""	""	25	$\sim$ 96 (6)	
			1024	16	5	94.1	Perlegen mouse dataset
			""	""	10	94.2	Frazer *et al.* (2007)
			""	""	15	93.5	(http://mouse.perlegen.com)
			""	""	20	93.4	consecutive SNPs extracted
			""	""	25	92.8	from a high-resolution set of 8.3 million SNPs

Table 2C Accuracy percentages for various off-line imputation methods. p_{miss}: percentage of missing data; GENOA: Genetic Epidemiology Network of Arteriopathy; (6) accuracy percentages are reported from a low-resolution plot; the total accuracy decrease between 5% and 25% missing data percentages is estimated to be around 1%.

Performance analysis of methods to infer missing genotypes

Christine Sinoquet

Abstract

Complex analyses such as genetic mapping, disease association studies, disease mapping in the context of environmental health and environmental epidemiology studies rely on high-throughput genotyping techniques. These analyses thoroughly examine genetic variations between subjects, in particular through Single Nucleotide Polymorphism (SNP). Nonetheless, though nowadays genotyping techniques impose high-quality standards, one still has to cope with the issues of missing data and genotyping errors. Typically, the percentage of missing data - or missing calls - now ranges in interval $[5\%, 10\%]$. Computational inference of missing data represents a challenging alternative to genotyping again the missing regions. This document first briefly reviews the various methods designed to infer missing SNPs. Then, it reports performances published for these inference methods. The present report carefully describes the characteristics of the different benchmarks generated by the designers (missing data percentage, correlation between SNPs). We show that most methods provide accuracies in the range $[90\%, 96\%]$. However, we also emphasize that no algorithm garantees constant high accuracies: an algorithm may perform well on some benchmarks and show in contrast relatively poor results on others.

LINA, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3