Estimation of N-acetyltransferase 2 haplotypes

Klaus Golka[1], Mirabutaleb Samimi[1], Meinolf Blaszkewicz[1], Doris Dannappel[1], Hermann M. Bolt[1], Silvia Selinski[2]

[1]Institute for Occupational Physiology at the University of Dortmund (IfADo),

Ardeystr. 67, 44139 Dortmund, Germany

[2]Department of Statistics, University of Dortmund, Vogelpothsweg 87,

44221 Dortmund, Germany

Fax:
+49 231 1084 308 (IfADo)
+49 231 755 5303 (Department of Statistics)

E-Mail:
golka@ifado.de
selinski@statistik.uni-dortmund.de

Abstract

N-Acetyltransferase 2 (NAT2) genotyping may result in a considerable percentage in several ambiguous allele combinations. PHASE 2.1 is a statistical program which is designed to estimate the probability of different allele combinations. We have investigated haplotypes of 2088 subjects genotyped for NAT2 according to standard PCR/RFLP methods. In 856 out of 2088 cases the genotype was clearly defined by PCR/RFLP only. In many of the remaining cases the program clearly defined the most probable allele combination: In the case of *5A/*6C, *5B/*6A the probability for *5B/*6A is 99% whereas the alternative allele combination *5A/*6C can be neglected. Other combinations cannot be allocated with a comparable high probability. For example the allele combination *5A/*5C, *5B/*5D provides for *5A/*5C a probability of 69% whereas the estimation for *5B/*5D allele is only 31%. In the two most often observed constellations in our data [(*12A/*5B, *12C/*5C); (*12A/*6A, *12B/*6B, *4/*6C)] the probability of allele combination was ascertained as follows: *12A/*5B, 98%; *12C/*5C, 1.4% and *12A/*6A, 82%; *4/*6C, 17%; *12B/*6B, 0%. The estimation of the NAT2 haplotype is important because the assignment of the NAT2 alleles *12A, *12B or *13 as a rapid or slow genotype has been discussed controversially. Otherwise the classification of alleles in subjects which are not showing a clearly allocation can result in a rapid or slow acetylation state. This assignment has an important role in survey of bladder cancer cases in the scope of occupational exposure with aromatic amines.

Keywords PHASE 2.1, NAT2 genotyping, single nucleotide polymorphism

Introduction

Haplotype analysis has been used for more precise localisation of genes investigating different processes in different populations (Xu et al. 2002). It may be also used for suggesting the most probable allele combination in cases in which data do not suggest a single allele combination. For example, N-acetyltransferase 2 (NAT2) genotyping may result in a considerable percentage in ambiguous allele combinations.

One approach to overcome the problems of different possible allele combination is to combine genotyping and phenotyping. In a considerable number of cases the result of genotyping can be made more robust. Nevertheless, additional phenotyping is expensive, time consuming and in a number of cases not feasible due to practical problems.

Another approach to get more robust allele combinations is to apply a haplotype reconstruction algorithm which aims to determine the most likely allele combination by using the haplotype of a gene. The haplotype is defined as the group of single nuclear polymorphisms (SNPs) which define a genotype e.g. of a polymorphic enzyme. Several computer programs which aim to determine the most probable allele combination of a gene are available (e.g., Grinberg and Mano 2004; Sobel et al. 1995; Wang and Xu 2003; Vijayasatya and Mukherjee 2005; for review see Salem et al. 2005).

In this study, the program PHASE 2.1, designed by Stephens and co-workers (Stephens et al. 2001; Stephens and Donnelly 2003) was applied. This program, written in DOS and thus asking for data input in the DOS level, uses a Baysian approach for the reconstruction of SNPs. In this study, SNPs were derived indirectly from genotyping using RFLP and PCR. Calculation is performed by Markov-Chain Monte Carlo algorithms. The program requires information on the relative distance of the investigated gene loci (in base pair units). The loci should be located on one chromosome. More than two alleles per gene locus as well as missing values are allowed. Use of background information on the recombination rate is possible. We have investigated haplotypes of 2088 subjects genotyped for NAT2 according to standard PCR and RFLP based methods.

Subjects and Methods

A total of 2088 subjects who had visited the Central Unit Clinical Occupational Medicine in our institute for different purposes or who were members of study cohorts in different hospitals, mostly, cases of urological cancer, colorectal cancer or controls were investigated. Each subject donated 10 ml EDTA blood. DNA was extracted using standard methods. Genotyping for NAT2 was performed using PCR and RFLP based standard methods (Bolt et al. 2005). A total of 7 single nucleotide polymorphisms (SNPs) which are adequate to genotype Caucasians for NAT2 were investigated (Figure 1).

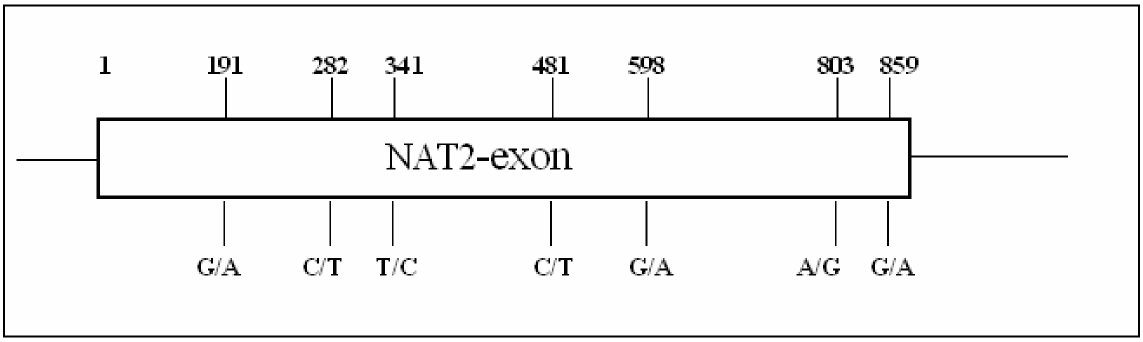

Fig. 1: Seven single nucleotide polymorphisms (SNPs) of the human NAT2 exon relevant in Caucasians.

Haplotype analysis was performed using the program PHASE 2.1, developed by Stephens and co-workers (Stephens et al. 2001; Stephens and Donnelly 2003) and provided via world wide web (http://www.stat.washington.edu/stephens/). After registration, the licence can be required for free for non-profit use. This program is running on the DOS level only. Thus data presented by spreadsheet programs like Excel must be presented using the DOS editor.

Statistical analysis

The estimation of individual haplotype pairs, their probability as well as the sample haplotype frequencies was performed using the software PHASE, version 2.1. The approach underlying PHASE is a Bayesian haplotype reconstruction method (see Stephens et al. 2001, Stephens and Donnelly 2003, for details).

Suppose we have a sample of n diploid individuals from a population. Let $G = (G_1, \ldots, G_n)$ be the observed genotypes and $H = (H_1, \ldots, H_n)$ denote the unknown corresponding haplotype pairs. Let $F = (F_1, \ldots, F_M)$ denote the set of unknown population haplotype frequencies and $f = (f_1, \ldots, f_M)$ be the set of unknown sample haplotype frequencies where $1, \ldots, M$ is an arbitrary labelling of the M possible haplotypes. Considering the unknown haplotypes as random quantities the prior information, i.e. the assumptions on genetic and demographic models, on the population haplotype frequency F etc, and the likelihood, i.e. the information given by the observed data leading, for instance, to estimates of the sample haplotype frequencies f, are combined to calculate the posterior distribution $\Pr(H|G)$ of the unobserved haplotypes H given the observed genotypes G. This is done by the use of Gibbs sampling, a Markov-Chain Monte Carlo algorithm (Gilks et al. 1996).

Hence, the individual haplotypes can be estimated from this posterior distribution by choosing, for instance, the most likely haplotype reconstruction for each individual.

Results

In 856 out of 2088 cases the genotype was clearly defined by PCR and RFLP based standard methods only. A re-analyses of the clearly defined 856 cases confirmed the results. Only in 19 of these cases an additional genotype was offered (Table 1, 2).

Table 1: Rate and absolute number of allele combinations of NAT2 in 2088 subjects.

Alleles	Rate (%)	Number (n)	Assigned Phenotype
*4/*4	4.74	99	rapid
*4/*5A	1.34	28	intermediary
*4/*6B	0.04	1	intermediary
*4/*7A	0.04	1	intermediary
*4/*7B	0.04	1	intermediary
*5A/*5A	0.09	2	slow
*5A/*5B	1.81	38	slow
*5A/*5C	0.04	1	slow
*5A/*5C, *5B/*5D	0.19	4	slow
*5A/*6A	1.62	34	slow
*5A/*6B	0.09	2	slow
*5A/*6C , *5B/*6A	24.23	506	slow
*5A/*7B	0.19	4	slow
*5B/*5B	16.52	345	slow
*5B/*5C	1.48	31	slow
*5B/*6B	0.23	5	slow
*5B/*6C	0.04	1	slow
*5B/*7A	0.09	2	slow
*5B/*7B	2.05	43	slow
*5C/*6A	0.14	3	slow
*5C7*7B	0.04	1	slow
*6A7*6A	7.80	163	slow
*6A/*6B	0.14	3	slow
*6A/*7B	0.95	20	slow
*6B/*13 , *4/*6A	2.20	46	intermediary-rapid
*6B/*6B	0.04	1	slow
*12A/*4	0.14	3	rapid
*12A/*5A , *12C/*5D , *4/*5B	17.24	360	slow-rapid
*12A/*5B , *12C/*5C	0.43	9	intermediary?
*12A/*5C	0.04	1	intermediary?
*12A/*5D , *4/*5C	0.76	16	slow-rapid
*12A/*6A , *12B/*6B , *4/*6C	0.23	5	intermediary
*12A/*6B	0.04	1	intermediary?
*12A/*7B , *12B/*7A	0.04	1	intermediary?
*12B/*5A , *13/*5B	0.19	4	intermediary
*12B/*5E , *5C/*6A , *5D/*6C	0.71	15	slow
*12B/*6A	0.09	2	intermediary
*12C/*4	0.04	1	rapid
*12C/*12C	0.04	1	rapid
*12C/*13	0.04	1	rapid
*12C/*5A	0.28	6	preliminary
*13/*4	0.14	3	rapid
*13/*13	0.04	1	rapid
*13/*6B , *4/*6A	11.78	246	slow-rapid
*13/*7A , *4,*7B	1.14	24	slow-rapid

Table 2: Frequencies of allele combinations of the NAT2 genotype in 2088 subjects clearly defined by PCR and RFLP based standard methods only.

Allele combination	**Number (n)**	**Frequency**	
*5A/*5B	17	17 x 100 %	
*5A/*6A	34	26 x 100 %,	8 x 99%
*5B/*5B	345	344 x 100%	1 x not identified
*5B/*5C	31	31 x 100%	
*5B/*7B	43	43 x 100%	
*6A/*6A	163	162 x 100%	1 x not identified
*6A/*7B	20	20 x 100%	

In many of the remaining cases with two or three possible alternative allele combinations the program clearly defined the most probable allele combination: In the case of *5A/*6C, *5B/*6A the probability for *5B/*6A is 99% and the alternative allele combination *5A/*6C may be neglected (Table 1, 3).

Table 3: Frequencies of alternative allele combinations of the NAT2 genotype in 2088 subjects.

Allele combinations	**Number (n)**	**Frequency of combinations**
*4/*5A	28	28 x *4/*5A,100 %
*5A/*6C , *5B/*6A	506	506 x *5A/*6C, 99 %
*12A/*5A , *12C/*5D , *4/*5B	360	360 x *4/*5B, 99 %
*12A/*5D , *4/*5C	16	16 x *4/*5B, 99 %
*12A/*6A , *12B/*6B , *4/*6C	3	3 x *4/*6C, 17.2%; *12°/*6A, 82 %
*12B/*5E , *5C/*6A , *5D/*6C	15	15 x *5C/*6A, 99%
*13/*6B , *4/*6A	247	247 x *4/*6A
*13/*7A , *4/*7B	24	24 x *4/*7B, 99%

Other combinations cannot be allocated with a comparable high probability. For example the allele combination *5A/*5C, *5B/*5D provides for *5A/*5C a probability of 69% whereas the estimation for *5B/*5D allele is only 31%. In the two most often observed constellations in our data [(*12A/*5B, *12C/*5C); (*12A/*6A, *12B/*6B, *4/*6C)] the probability of allele combination was as follows: *12A/*5B, 98%; *12C/*5C, 1.4% and *12A/*6A, 82%; *4/*6C, 17%; *12B/*6B, 0% (Table 1, 4).

Table 4: Present assignment of human NAT2 alleles (Butcher et al. 2002; Arylamine N-Acetyltransferase Nomenclature Committee 2003).

Present assignment of human *NAT2* alleles		
Allele	**Phenotype**	**Nucleotide change(s)**
*NAT2*4*	Rapid	none
*NAT2*5A*	Slow	341T>C, 481C>T
*NAT2*5B*	Slow	341T>C, 481C>T, 803A>G
*NAT2*5C*	Slow	341T>C, 803A>G
*NAT2*5D*	Slow	341T>C
*NAT2*5E*	Slow	341T>C, 590G>A
*NAT2*5F*	Slow	341T>C, 481C>T, 759C>T, 803A>G
*NAT2*6A*	Slow	282C>T, 590G>A
*NAT2*6B*	Slow	590G>A
*NAT2*6C*	Slow	282C>T, 590G>A, 803A>G
*NAT2*6D*	Slow	111T>C, 282C>T, 590G>A
*NAT2*7A*	Slow	857G>A
*NAT2*7B*	Slow	282C>T, 857G>A
*NAT2*10*	Unknown	499G>A
*NAT2*11*	Unknown	481C>T
*NAT2*12A*	Rapid	803A>G
*NAT2*12B*	Rapid	282C>T, 803A>G
*NAT2*12C*	Rapid	481C>T, 803A>G
*NAT2*13*	Rapid	282C>T
*NAT2*14A*	Slow	191G>A
*NAT2*14B*	Slow	191G>A, 282C>T
*NAT2*14C*	Slow	191G>A, 341T>C, 481C>T, 803A>G
*NAT2*14D*	Slow	191G>A, 282C>T, 590G>A
*NAT2*14E*	Slow	191G>A, 803A>G
*NAT2*14F*	Slow	191G>A, 341T>C, 803A>G
*NAT2*14G*	Slow	191G>A, 282C>T, 803A>G
*NAT2*17*	Slow	434A>C
*NAT2*18*	Unknown	845A>C
*NAT2*19*	Slow	190C>T

Discussion and Conclusions

The estimation of the NAT2 haplotype is important because the assignment of the NAT2 alleles *12A, *12B or *13 as a rapid or slow genotype is being discussed controversially (see Bolt et al. 2005). The classification of alleles in subjects not showing a clear allocation can result in a rapid or slow acetylation state (see figure 2 for illustration). This assignment has an important role in surveys of bladder cancer cases elicited by occupational exposures with aromatic amines. However, molecular haplotyping methods are labour-intensive and expensive. Thus, the use of haplotype reconstruction algorithms such as PHASE provides an effective alternative yielding good results (Sabbagh and Darlu 2005).

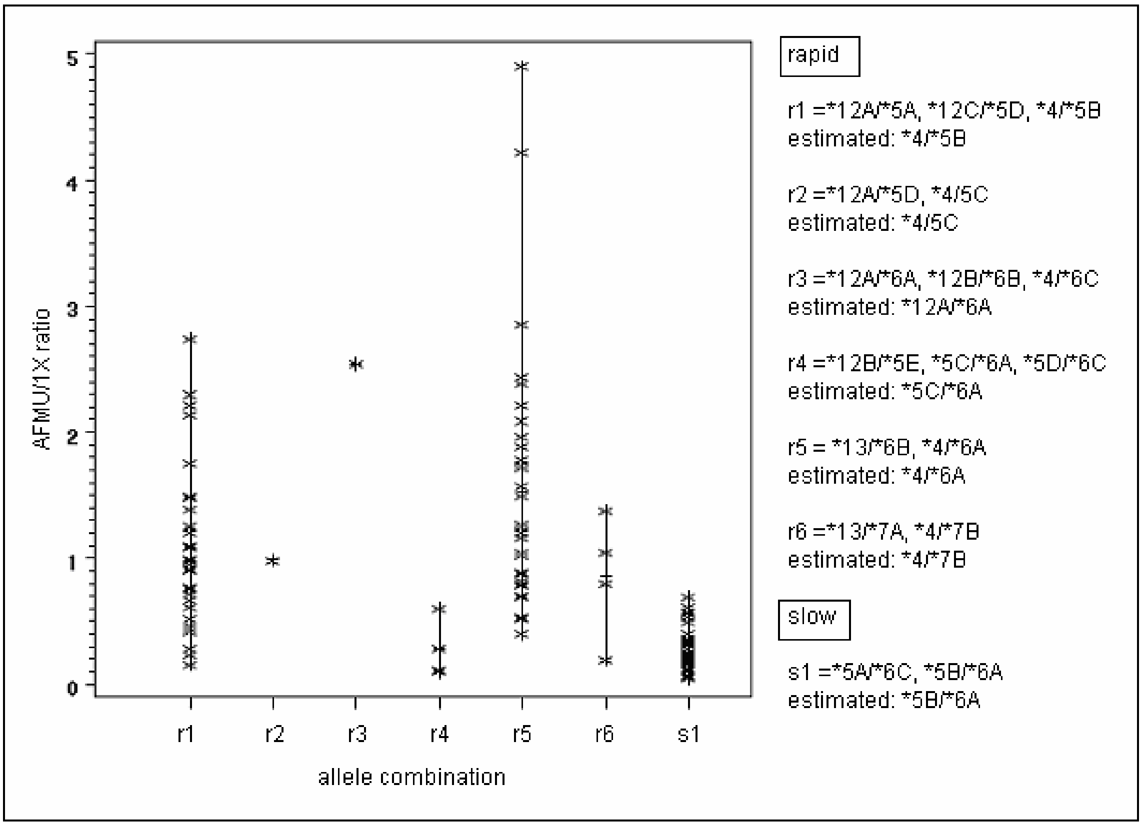

Fig. 2: Individual NAT2 phenotyping data (ratios of urinary AFMU/1X), allocated to genotypic groups of ambiguous *NAT2* alleles (extracted from Bolt et al. 2005)

A mayor drawback of this program is that it is running on the DOS level only. Thus data presented by spreadsheet programs like Excel must be presented using the DOS editor, which is a problem at least for all those not familiar with data input on the DOS level.

Acknowledgements: The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, “Reduction of complexity in multivariat data structures”) is gratefully acknowledged. Many thanks to Heinz Dieter Giller who introduced us in the basics of the DOS editor.

References

Arylamine N-Acetyltransferase Nomenclature Committee, latest update 2003 http://www.louisville.edu/medschool/pharmacology/NAT.html

Bolt H.M., Selinski S., Dannappel D., Blaszkewicz M., Golka K. (2005) Re-investigation of the concordance of human NAT2 phenotypes and genotypes. Arch. Toxicol. 79:196-200

Butcher N.J., Boukouvala S., Sim E., Minchin R.F. (2002) Pharmacogenetics of the arylamine N-acetyltransferases. Pharmacogenomics J. 2:30-42

Gilks W.R., Richardson S., Spiegelhalter D.J. (eds.) (1996) Markov-Chain Monte Carlo in practice. Chapman & Hall, London

Golka K., Prior V., Blaszkewicz M., Bolt H.M. (2002) The enhanced bladder cancer susceptibility of NAT2 slow acetylators towards aromatic amines: a review considering ethnic differences. Toxicol. Lett. 128:229-241

Grinberg S. and Mano A under guidance of Geiger D. and Fishelzom M. (2004) Haplotyping by MPE query http://bioinfo.cs.technion.ac.il/projects/Grinberg-Mano/

Sabbagh A. and Darly P. (2005) Inferring haplotypes at the NAT2 locus: the computational approach. BMC Genetic 6, http://www.biomedcentral.com/1471-2156/6/30

Salem R.M., Wessel J., Schork N.J. (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals Hum. Genomics 2:39-66

Stephens M., Donnelly P. (2003) A comparison of Bayesian methods for haplotype reconstruction. Am. J. Hum. Genet. 73:1162-1169

Stephens M., Smith N.J., Donnelly P. (2001) A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68: 978-989

Vijayasatya R., Mukherjee A. (2005) An efficient algorithm for perfect phylogeny haplotyping. Proc IEEE Comput. Syst. Bioinform. Conf. 2005:103-110

Wang L., and Xu Y. (2003) Haplotype inference by maximum parsimony. Bioinformatics 19:1773–1780

Weeks D.E., Sobel E., O'Connell J.R. ,Lange K. (1995) Computer programs for multilocus haplotyping of general pedigrees. Am. J. Hum. Genet. 56:1506-1507

Xu C.F., Lewis K., Cantone K.L., Khan P., Donally C., White N. et al. (2002) Effectiveness of computational methods in haplotype prediction. Hum. Genet. 110: 148-156