Haplotype tagging efficiency and tagSNP sets portability in worldwide populations in NAT2 gene

Marquage d'haplotypes du gène NAT2 : efficacité dans différentes régions du monde

A. Sabbagh, P. Darlu, A. Langaney et E. S. Poloni

Référence électronique
A. Sabbagh, P. Darlu, A. Langaney et E. S. Poloni, « Haplotype tagging efficiency and tagSNP sets portability in worldwide populations in NAT2 gene », Bulletins et mémoires de la Société d'Anthropologie de Paris [En ligne], 19 (3-4) | 2007, mis en ligne le 03 février 2011, Consulté le 01 octobre 2016. URL : http://bmsap.revues.org/5223
Haplotype tagging efficiency and tagSNP sets portability in worldwide populations in NAT2 gene
Marquage d’haplotypes du gène NAT2 : efficacité dans différentes régions du monde

A. Sabbagh, P. Darlu, A. Langaney et E. S. Poloni

Introduction

1 The N-acetyltransferase 2 (NAT2) gene plays a crucial role in the metabolism of xenobiotics, including many clinically useful drugs and exogenous chemicals present in the diet, cigarette smoke and the environment (Hein 2002). Extensive polymorphism in NAT2 gives rise to a wide interindividual variation in N-acetylation capacity. In particular, a clear bimodal distribution is observed that segregates the rapid acetylator phenotype, associated with a normal acetylation capacity, from the slow acetylator one, characterized by a reduced enzyme activity. These two main metabolic phenotypes occur with varying prevalence in populations of different ethnic origin (Upton et al. 2001).

2 The clinical consequences of the acetylation polymorphism can be severe if standard drug doses are applied, exposing patients to an increased risk of adverse drug reactions or a lack of therapeutic efficacy (Meisel 2002). In addition, in the last decades, an increasing number of epidemiological studies have attempted to relate acetylation phenotype to a variety of complex human disorders, such as bladder cancer, atopic diseases, diabetes, Parkinson’s disease and many others (see Butcher et al. (2002) for a review). However, up to now, association studies in NAT2 have led to conflicting results among (and even within) human populations and most association findings have been difficult to replicate.
One reason for these inconsistencies may relate to the fact that almost all studies focused on a limited number of candidate polymorphisms, which were not necessarily the same from one study to another (Goldstein 2003; Goldstein et al. 2003). A shift toward a gene-based approach in which all common variation within a gene is considered jointly is advocated for future association studies (Neale, Sham 2004). By capturing all of the potential risk-conferring variations within NAT2, this approach should resolve much of the controversial issues of candidate-polymorphism studies.

At present, it does not seem reasonable to consider the sequencing of the entire NAT2 gene in thousands of individuals and some form of data reduction is necessary. The adoption of a marker selection strategy, such as the haplotype-tagging approach, can significantly reduce the scale and cost of genotyping. The key idea is to use linkage disequilibrium (LD) to gain maximal information from typing a selected subset of highly informative single-nucleotide polymorphism (SNP) markers, referred to as “haplotype tagging SNPs” (htSNPs) (Halldorsson et al. 2004). This approach thus requires a thorough description of LD patterns in NAT2 in the targeted populations. Moreover, in view of the marked population heterogeneity in NAT2 allele frequencies, it is especially important to characterize variability in haplotype structure within and between human populations.

Using data collected from an extensive survey of the literature, this study aimed to characterize the worldwide patterns of LD at the NAT2 gene and to evaluate haplotype tagging efficiency at this locus. We defined population-specific sets of htSNPs to be used for future association studies and examined how well continent-specific htSNPs sets, defined by grouping the population haplotypes in each continental region, perform in populations within a same continent. The performance of a “cosmopolitan” htSNPs set suitable for all human populations was also empirically evaluated.

Materials and methods

Population samples and NAT2 molecular data

We selected from published reports up to July 2006 all the population samples that were genotyped for the seven most common SNPs at NAT2 and for which genotype data was available. These seven SNPs are the main polymorphisms occurring in human populations at NAT2. The analysis of these seven variants has been shown to be highly predictive of the acetylation phenotype with a prediction rate close to 100% (Mrozikiewicz et al. 1996; Meisel et al. 1997; Gross et al. 1999; Jorge-Nebert et al. 2002; Lee et al. 2002). Among these seven SNPs, all located in the coding exon, four result in an amino acid substitution that leads to a significant decrease in acetylation capacity (G191A, T341C, G590A, G857A). The other three are either silent mutations (C282T, C481T) or non-synonymous substitutions that do not alter the phenotype (A803G).

The collected data consisted of 3,994 individuals (7,988 chromosomes) from 28 human populations representing major geographic regions (table I): Europe and North Africa (13 samples), sub-Saharan Africa (7), East Asia (5), Central/South Asia and Central America (3). Sample sizes range from 24 (Somali) to 844 (German) individuals, with an average of 140 individuals per sample. A full description of each selected sample and the corresponding references are provided elsewhere (Sabbagh et al. 2008b).
Table I—Haplotype tagging of the NAT2 gene

<table>
<thead>
<tr>
<th>Sample</th>
<th>N at each SNP</th>
<th>Haplotype tagging efficiency</th>
<th>Tagging sets portability in worldwide population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>10 SNPs</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>Study 2</td>
<td>12 SNPs</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>Study 3</td>
<td>15 SNPs</td>
<td>0.98</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Statistical analyses

7 From the unphased multi-locus genotypes provided by each study, we inferred NAT2 haplotypes using the Bayesian method implemented in PHASE v.2.1 (Stephens, Donnelly 2003), using the default parameter values in the Markov chain Monte Carlo simulations. For each data set investigated, we applied the algorithm ten times with different seeds for the random number generator, and checked for consistency of the results across the independent runs in order to verify that the algorithm did not converge to a local, rather than global, mode of the posterior distribution. We chose the results from the run displaying the best average goodness-of-fit of the estimated haplotypes to the underlying coalescent model. Haplotypes inferred by PHASE were then used as inputs in the DnaSP and tagSNPs programs used in subsequent analyses.

8 Using DnaSP (Rozas et al. 2003), the $r^2$ statistic (Hill, Robertson 1968) was computed to estimate pairwise LD between the seven genotyped SNPs, after the exclusion, in each population, of SNPs with minor allele frequency (MAF) < 0.05. Statistical significance of LD between SNP pairs was assessed using Fisher’s exact tests followed by Bonferroni corrections. Mantel tests to compare $r^2$ matrices were performed using the program CADM (Legendre, Lapointe 2004). Comparisons were made between populations within each continental group. Subsequently, $r^2$ values were recalculated for populations pooled into geographical groups and Mantel tests were again applied. Based on the inter-marker LD patterns observed, we selected, in each sample and each geographical region, the most informative subset of SNPs that retain most of the haplotype information within NAT2.
For that purpose, we used the haplotype statistic (Stram et al. 2003), implemented in the tagSNPs program (Stram et al. 2003), which identifies a minimal set of tag markers that optimizes the predictability of common haplotypes (frequency > 0.05). We required the minimum estimated haplotype value for all common haplotypes to be $\geq 0.80$. To assess the savings in genotyping offered by tagging, we defined a ‘saving index’, calculated as the total number of polymorphic markers (with MAF > 0.0) genotyped in a sample divided by the number of htSNPs selected. We also evaluated the tagging efficiency of each htSNPs set by dividing the mean coefficient of determination of all common haplotypes in a sample (mean) by the number of htSNPs in the set.

Results

A total of 3,994 individuals from 28 worldwide samples were analyzed for their genotype at the seven common SNPs of the NAT2 gene. The genotypic distributions at each SNP in the different populations were not significantly different from Hardy-Weinberg proportions (all $p$-values were smaller than 5% after Bonferroni correction for multiple testing). The LD structure of NAT2 was investigated by computing the $r^2$ measure of LD between SNP pairs. We first tested whether the amount of LD differed between human populations. All population samples displayed similar levels of LD within each geographic area, except Somali who showed higher LD at NAT2 (average $r^2$ value = 0.589) than other sub-Saharan African samples and were, in that respect, more similar to Europeans (Table I). The mean pairwise $r^2$ value between the seven SNPs in the European samples (0.567 ± 0.075, including Moroccans) was significantly higher (Wilcoxon’s test, $P \leq 0.0002$) than in both East Asians (0.276 ± 0.023) and Africans (0.243 ± 0.050, without Somali). No difference in the level of LD was found between East Asian and African populations. However, the proportion of SNP pairs with $r^2 \geq 0.5$ was far smaller in sub-Saharan Africans (6.7%) than in both Europeans (40%) and East Asians (33.3%). Ashkenazi Jews exhibited the highest level of LD (average $r^2$ value = 0.763); such an excess of LD is often observed in founder populations that recently grew from relatively small sizes (Shifman, Darvasi 2001). We then tested whether the structure of LD was similar among populations by computing the correlation between $r^2$ matrices of LD. Mantel’s tests gave highly significant correlation values both between population pairs within geographic areas and for pairs of continental regions ($P < 0.0001$ with 10,000 permutations). Thus, although Europeans exhibited higher levels of LD at NAT2, the pattern of LD in this gene was similar across human populations.

The number of haplotypes inferred from the unphased multi-locus genotype data in each sample is displayed in Table I, as well as the haplotype diversity. Sub-Saharan African populations displayed greater haplotype diversity than either Europeans or Asians. A larger number of haplotypes of similar frequencies occur indeed in these populations, generating a huge number of distinct genotypes. By contrast, in populations of Asian origin, only a few major haplotypes were found, namely NAT2*4, NAT2*6A and NAT2*7B. The mean haplotype diversity was estimated to be 0.79 ± 0.06 in Africans and 0.59 ± 0.11 in East Asians. Europeans displayed an intermediate value of 0.69 ± 0.04.

For each of the 28 worldwide samples, we defined sets of htSNPs that captured most of the haplotype diversity of the NAT2 coding region (Table I). We used the haplotype criterion (Stram et al. 2003), which reflects the degree to which a given htSNPs set explains the variability in the haplotypes it is chosen to tag. Aof 0.80 was selected as a
threshold for tag selection and performance measurement. In all population samples, except sub-Saharan Africans, only a few common haplotypes explained the vast majority of all chromosomes (~90% and more), and only two to three htSNPs were sufficient to adequately characterize the common variation: the mean coefficient of determination (mean) of all common haplotypes was near or above 0.90 in each sample. The best tagging efficiency was observed in East Asians who exhibited the lowest haplotype diversity at NAT2. In contrast, in sub-Saharan Africans, where greater haplotype diversity and lower levels of LD were observed, no reduction in the genotyping requirement was possible as SNPs were poor markers of each other in these samples (except in Somali where the number of SNPs could be reduced from seven to four). We next designed continent-specific sets of htSNPs, by grouping the population haplotypes in each continent, as well as a worldwide set of htSNPs, by pooling all 28 samples into a single population, and using three different thresholds for tag selection (0.80, 0.85, and 0.90) (Table II). We evaluated the performance of these htSNPs sets in individual samples by examining whether the minimum estimated haplotype was ≥ 0.80 or not in the tested sample with the considered htSNPs set. When a threshold of 0.80 was used for htSNPs selection, the continent-specific and worldwide htSNPs sets did not perform well in ~40% of the samples, thus demonstrating low transferability. In Europe/North Africa and East Asia, a more stringent threshold of 0.90 was needed to identify htSNPs sets that consistently perform well in all samples within each geographic area. In sub-Saharan Africans, no reduction in the number of SNPs was possible whatever the threshold used. A four-htSNP set (C282T, C481T, G590A, A803G), selected using either a 0.85 or a 0.90 threshold from the pooled worldwide samples, performed successfully in all the 21 non-African samples but not at all in sub-Saharan Africans.

Table II—Transferability of continent-specific and worldwide htSNPs sets among human populations.

Table II - Transférabilité des htSNP dans les populations humaines, par continent et pour l’ensemble des continents.

<table>
<thead>
<tr>
<th>Continent-specific htSNPs sets</th>
<th>Mean haplotype</th>
<th>0.80 ≤ ( R^2 )</th>
<th>Mean haplotype</th>
<th>0.85 ≤ ( R^2 )</th>
<th>Mean haplotype</th>
<th>0.90 ≤ ( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Saharan Africa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-African</td>
<td>6</td>
<td>CHL, C481T, TMY, G590A, A803G, C481T</td>
<td>1</td>
<td>all F</td>
<td>1</td>
<td>all F</td>
</tr>
<tr>
<td>African</td>
<td>3</td>
<td>C282T, A803G</td>
<td>5% (1/20)</td>
<td>C282T, C481T, AMN</td>
<td>5% (1/20)</td>
<td>C282T, C481T, AMN</td>
</tr>
<tr>
<td>Worldwide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Saharan African</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For more details, please refer to the article.
Discussion

This study provided a thorough description of the haplotype diversity and LD structure of *NAT2* at a worldwide level. It enabled the definition of population-specific htSNPs sets which should be extremely useful for future association studies.

Our analyses revealed very similar patterns of LD among human populations at the *NAT2* locus, although disparities in the level of LD appeared among the main continental groups. Even within a small gene like *NAT2*, the SNP markers appeared to be poorly correlated in sub-Saharan Africans and the whole set of SNPs initially genotyped in these populations was needed to accurately represent the common variation at this locus (table I, II). There were therefore almost no gains from haplotype tagging in these populations. An exception is the Somali sample which displayed a higher level of LD at *NAT2* compared to the other sub-Saharan African populations examined and for which a greater reduction in the number of SNPs required to accurately represent the *NAT2* common variation was observed. Since the full *NAT2* gene diversity has not yet been investigated in other East African populations, it is not possible to conclude whether this population is peculiar with respect to this genetic system, whether it resembles other East African populations, or whether these observations result from the small size of the Somali sample (*N* = 24) which may not be representative of the *NAT2* genetic variation in the Somali population. In contrast, only 2 to 3 SNPs were sufficient to capture most of the variation in the European and Asian samples. If a reduction from 7 to 2-3 SNPs to be typed may seem negligible with the advent of high-throughput technologies for SNP genotyping, the savings in cost and time are no more trivial if such a 2- to 3-fold saving is achieved in each of the hundreds or thousands of genes to be typed in large-scale association studies, that adopt either a whole-genome or candidate-gene approach.

htSNPs sets appeared to be portable among populations from a same continent, provided that the continent-specific htSNPs sets were selected with a stringent criterion (minimum haplotype ≥ 0.90). A “cosmopolitan” htSNPs set suitable for all human populations could not be identified for the *NAT2* gene, unless considering all the seven common SNPs of this gene. However, it was possible to identify a single four-htSNP set that worked adequately in multiple human populations of non-African origin. This set was of the same size or just a little bit larger than the European and East-Asian continent-specific htSNPs sets (each composed of four and three SNPs, respectively). Additional studies are needed to determine whether this four-htSNP set still performs well in larger numbers of non-African populations. The use of a single htSNPs set in multiple populations would ensure comparability for replication efforts within complex traits and would also allow direct comparison of the role of the same variants as risk factors for different conditions (Need, Goldstein 2006). However, if the number of SNPs to be genotyped has to be significantly reduced, one should better apply SNP selection procedures separately in each individual population.

A recent study has empirically evaluated the ability of tagging markers selected in *NAT2* to predict the individual acetylator status (Sabbagh et al. 2008a): the functional variation was shown to be adequately represented by the selected tagging markers, these latter providing a classification accuracy for the individual acetylator status close to the maximal 100% value observed with the entire set of common polymorphisms in this gene. Therefore, the tagging approach appears as a valuable tool for pharmacogenetic studies.
testing NAT2 genetic variation for association with a given clinical trait. Even if tagging will probably be replaced later on by high-throughput sequencing which enable comprehensive studies of human genetic variation, indirect association studies via haplotype tagging appear as an efficient and economical approach at present and should remain so for the near future, especially when screening a large number of candidate genes.

Acknowledgements

We gratefully acknowledge the French “Fondation pour la Recherche Médicale” for supporting this work.

BIBLIOGRAPHIE

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


Gross (M.), Kruisselbrink (T.), Anderson (K.), Lang (M.), McGovern (P.), Delongchamp (R.), Kadlubar (F.) 1999 Distribution and concordance of N-acetyltransferase genotype and phenotype in an American population, Cancer Epidemiology Biomarkers and Prevention 8: 683-692.


Lee (S.Y.) and (K.A.), Ki (C.S.), Kwon (O.J.), Kim (H.J.), Chung (M.P.), Suh (G.Y.), Kim (J.W.) 2002, Complete sequencing of a genetic polymorphism in NAT2 in the Korean population, Clinical Chemistry 48: 775-777.


Meisel (P.), Schroeder (C.), Wulff (K.), Siegmund (W.) 1997, Relationship between human genotype and phenotype of N-acetyltransferase (NAT2) as estimated by discriminant analysis and multiple linear regression:1, Genotype and N-acetylation in vivo, Pharmacogenetics 7: 241-246.

Mrozikiewicz (P.M.), Cascorbi (I.), Brockmoller (J.), Roots (I.) 1996, Determination and allelic allocation of seven nucleotide transitions within the arylamine N-acetyltransferase gene in the Polish population, Clinical Pharmacology and Therapeutics 59: 376-382.


Nei (M.) 1987, Molecular Evolutionary Genetics, Columbia University Press, New York.

Rozas (J.), Sánchez-DelBarrio (J.C.), Messeguer (X.), Rozas (R.) 2003, DnaSP, DNA polymorphism analyses by the coalescent and other methods, Bioinformatics 19: 2496-2497.


Sabbagh (A.), Langaney (A.), Darlu (P.), Gérard (N.), Krishnamoorthy (R.), Poloni (E.S.) 2008b, Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history, BMC Genetics 9: 21.


Stram (D.O.), Haiman (C.A.), Hirschhorn (J.N.), Altshuler (D.), Kolonel (L.N.), Henderson (B.E.), Pike (M.C.) 2003, Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study, Human Heredity 55: 27-36.

Upton (A.), Johnson (N.), Sandy (J.), Sim (E.) 2001, Arylamine N-acetyltransferases -of mice, men and microorganisms, Trends in Pharmacological Sciences 22: 140-146.

RÉSUMÉS

Le polymorphisme génétique du gène NAT2 est responsable de fortes différences interindividuelles de l’activité d’acétylation de l’enzyme N-acétyltransférase 2 (NAT2). Cette enzyme joue un rôle fondamental dans le métabolisme de nombreux xénobiotiques et médicaments utilisés en clinique. Jusqu’à présent, la plupart des études d’association qui ont tenté de relier les phénotypes d’acétylation du gène NAT2 à plusieurs maladies complexes chez l’homme ont conduit à des résultats contradictoires, essentiellement en raison des faibles connaissances sur la variation géographique des profils de déséquilibre de liaison au locus NAT2. C’est pourquoi nous avons entrepris une étude exhaustive de la littérature relatant la diversité mondiale haplotypique du gène NAT2 et décrivant les déséquilibres de liaison entre différents SNP (Single Nucleotide Polymorphism) de ce gène. Notre échantillon comprend 28 populations (incluant 3994 individus) réparties sur quatre continents (Afrique, Europe, Asie, Amérique). Pour chacune d’elles, nous avons estimé l’efficacité du marquage haplotypique (tagging) et défini le
sous-ensemble de SNP (htSNP) qui s’avère juste nécessaire pour réaliser efficacement des études d’association. Ce marquage haplotypique permet de réduire d’un facteur 2 à 3 fois le nombre de SNP à étudier, du moins pour les populations d’Europe et d’Asie de l’Est. En revanche, la diversité haplotypique en Afrique est telle qu’aucune réduction n’est possible pour ce continent. Les marqueurs htSNP sélectionnés dans une population restent performants pour le continent dans lequel elle se situe, à la condition que les htSNP aient été sélectionnés avec un critère suffisamment «stringent». Pour ce gène NAT2, il n’est pas possible d’identifier un unique ensemble de htSNP qui serait « universel » et efficace pour toutes les populations ; mais un ensemble réduit à seulement quatre htSNP reste parfaitement efficace dans le cas des populations non africaines.

Genetic polymorphism in the NAT2 gene is responsible for pronounced interindividual differences in the acetylation activity of the N-acetyltransferase 2 (NAT2) enzyme, which plays a crucial role in the metabolism of many clinically useful drugs and exogenous chemicals. Up to now, most association studies that have attempted to relate the acetylation phenotype in NAT2 to a variety of complex human disorders have led to contradictory results in different populations. Some of these inconsistencies may result from a poor knowledge of linkage patterns at NAT2 and their geographic variation. Using data from an extensive survey of the literature, we investigated the worldwide haplotype diversity and linkage disequilibrium structure of NAT2. For 28 population samples (including 3,994 individuals) from four continental regions (Africa, Europe, Asia, America), we evaluated haplotype tagging efficiency at NAT2 and defined population-specific sets of haplotype tagging SNPs (htSNPs) to be used for future association studies. Tagging common haplotypes yielded 2- to 3-fold savings in European and East Asian samples, while no gains from tagging were observed in samples of African ancestry, which displayed high haplotype diversity at NAT2. htSNPs sets appeared to be portable among populations from a same continent, provided that the continent-specific htSNPs sets were selected with a stringent criterion. A “cosmopolitan” htSNPs set suitable for all human populations could not be identified for the NAT2 gene, but a single four-htSNP set proved to perform successfully in all the non-African populations investigated.

INDEX

Mots-clés : déséquilibre de liaison, études d’association, génétique des populations, marquage haplotypique, NAT2
Keywords: association studies, haplotype tagging, linkage disequilibrium, population genetics

AUTEURS

A. SABBAGH

INSERM U535, Génétique épidémiologique et structure des populations humaines, Hôpital Paul Brousse, BP 1000, 94817 Villejuif CEDEX, France, e-mail : audrey.sabbagh@bjn.aphp.fr; INSERM UMR745, Université René Descartes, Faculté des Sciences Pharmaceutiques et Biologiques, 4 avenue de l’Observatoire, 75270 Paris Cedex 06, France et Université Paris-Sud, UMR-S535, 94817 Villejuif, France.
P. DARLU
INSERM U535, Génétique épidémiologique et structure des populations humaines, Hôpital Paul Brousse, BP 1000, 94817 Villejuif CEDEX, France et Université Paris-Sud, UMR-S535, 94817 Villejuif, France.

A. LANGANÉY
Laboratoire de Génétique et Biométrie, Département d’Anthropologie et d’Écologie, Université de Genève, 12 rue Gustave-Revilliod, 1211 Genève 4, Suisse et Muséum national d’histoire naturelle, Direction générale, CP135, 43 rue Buffon, 75281 Paris, Cedex 05, France.

E. S. POLONI
Laboratoire de Génétique et Biométrie, Département d’Anthropologie et d’Écologie, Université de Genève, 12 rue Gustave-Revilliod, 1211 Genève 4, Suisse.