Allele-specific transcript isoforms in human

Victoria Nembaware[a,b], Kenneth H. Wolfe[c], Fabiana Bettoni[d], Janet Kelso[b], Cathal Seoighe[a,*]

[a] Computational Biology Group, University of Cape Town, Rondebosch 7701, Cape Town, South Africa
[b] South African National Bioinformatics Institute, University of the Western Cape, Bellville 7535, South Africa
[c] Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland
[d] Laboratory of Molecular Biology and Genomics, Ludwig Institute for Cancer Research, 01509-010, São Paulo, SP, Brazil

Received 6 August 2004; revised 4 October 2004; accepted 4 October 2004

Available online 19 October 2004

Edited by Takashi Gojobori

Abstract Estimates of the number of human genes that produce more than one transcript isoform through alternative mRNA splicing depend on the assumption that the observation of multiple transcripts from a gene can be attributed entirely to alternative splicing. It is possible, however, that a substantial proportion of cases where multiple transcripts have been observed for a gene result from differences between alleles. Many examples of genes that are spliced differently from different alleles have been reported but no systematic estimate of the proportion of alternatively spliced genes that are affected by such polymorphisms has been carried out. We find that alternative transcript isoforms are non-randomly associated with closely linked nucleotide polymorphisms, based on an integrated analysis of the dbSNP, dbEST and ASAP databases. From the observed level of association between transcript isoforms and polymorphisms, we estimate that 21% of alternatively spliced genes are affected by polymorphisms that either completely determine which form of the transcript is observed or alter the relative abundances of some of the alternative isoforms. We provide a conservative lower bound of 6% on this estimate and point out that alternative splicing cannot be confirmed absolutely unless more than one transcript is observed from the same allele.

Keywords: Alternative splicing; Polymorphism

1. Introduction

Numerous large-scale studies of alternative splicing in humans and other organisms have been carried out using expressed sequence tags (ESTs) and cDNAs [1–6] as well as, more recently, microarrays [7]. Estimates from these studies of the proportion of genes that produce alternative transcript isoforms are as high as 74% [7]. The methods that have been used to identify alternatively spliced transcripts have so far not distinguished between allele-specific transcript isoforms (i.e., different alleles produce different transcript isoforms) and true alternative splicing that affects transcripts from the same allele. With some exceptions (e.g. [8]), the contribution of polymorphism to transcript variation is frequently overlooked and remains unqualified.

Individual mutations that affect mRNA splicing have been reported, several of which have been implicated in heritable diseases or disease susceptibility (e.g. [9–15]). In many instances, an exon is skipped due to a mutation at, or close to, the donor splice site in the intron immediately downstream of it. In other cases, a transcript containing a premature stop codon mutation is degraded by the nonsense-mediated decay pathway, resulting in the accumulation of a minor alternative splice form in which the exon is skipped [16]. More surprisingly, there are several reports where aberrant splicing of a disease allele has been associated with a missense or a silent change in an exon and not with any changes in introns (reviewed in [16]). Nucleotide changes that alter splicing or that induce nonsense-mediated mRNA decay may frequently have a radical effect on the encoded protein [17] and play an important and often neglected part in genetic variation and in disease [18,19].

We used ESTs that can be mapped both to single nucleotide polymorphisms (SNPs) and to exon junctions to estimate the fraction of allele specific transcript isoforms among genes reported in the ASAP database [20] to be alternatively spliced. It is likely that many nucleotide polymorphisms that alter splicing are in introns. The SNPs that we can map to EST sequences are exonic and in many cases may not be the cause of the allele-specificity of transcript isoforms, but instead are in tight linkage disequilibrium with the causal polymorphism. Our hypothesis is that some detected exonic SNPs might be *associated* with particular transcript isoforms because they are tightly linked to other unseen intronic polymorphisms that *cause* isoform allele-specificity. For simplicity, our analysis assumes that two separate mutations, one that caused isoform allele-specificity and one that gave rise to a nearby detectable exonic SNP, occurred at different times in an unknown order. However, our approach is valid even if the observed exonic SNP is the cause of the allele-specificity.

We consider the situation in which two mutations have occurred sequentially, giving rise to two linked polymorphisms, then (in the absence of recombination in the intervening sequence) the more recent polymorphism will always be found in association with a single form of the earlier polymorphism (Fig. 1). If one polymorphism determines completely the transcript isoform observed and the other is detectable as an exonic SNP, this means that one combination of transcript isoform and SNP should never be present in EST databases.

[*] Corresponding author. Fax: +27-21-650-5321.
E-mail address: cathal@science.uct.ac.za (C. Seoighe).

Abbreviations: SNP, single nucleotide polymorphism; EST, expressed sequence tag

0014-5793/$22.00 © 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2004.10.018

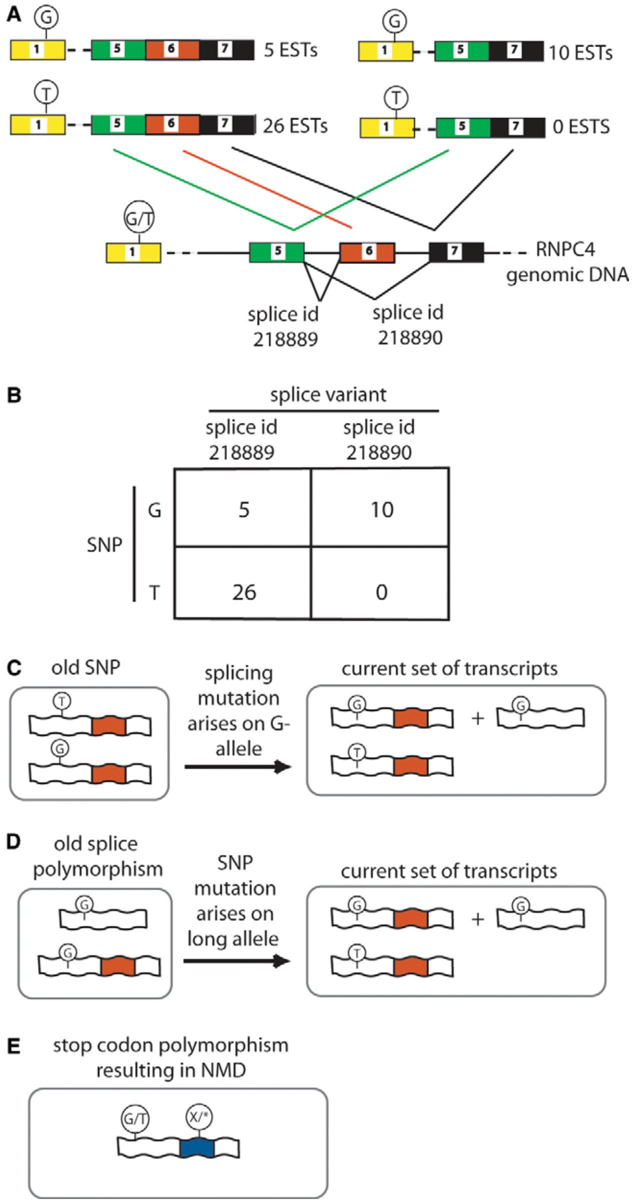

Fig. 1. Sequential mutations model. (A) A possible example of isoform allele-specificity affecting the *RNPC4* gene. ESTs consistent with the most common (longer) transcript isoform are shown on the left above the genomic sequence and ESTs consistent with the minor isoform are shown on the right. There are 10 ESTs that exclude exon 6, all of which have a G at a SNP site in exon 1. Of the ESTs from transcripts that include exon 6, five have a G at the polymorphic site and 26 have T at this site. (B) Data matrix corresponding to the *RNPC4* example (see Section 2). (C–E) Three scenarios through which sequential mutations could have given rise to the SNP and to the alternative isoforms of the gene. In (C), the SNP is more ancient and was followed by a mutation that altered mRNA splicing. The mutation that altered splicing occurred on an allele bearing a G at the SNP position and resulted in the skipping of exon 6 (red) in transcripts. In (D), a mutation occurred that altered splicing prior to the SNP mutation. The SNP mutation occurred on an allele that included exon 6 and consisted of a substitution from G to T. (E) A stop-codon polymorphism in a cassette exon (blue) of a generic alternatively spliced gene. The stop codon polymorphism is linked to an exonic SNP elsewhere on the gene, resulting in nonsense-mediated decay of the longer isoform from the allele containing the stop codon.

On the other hand, if polymorphisms do not affect the transcript isoform that is observed, the choice of transcript isoform should be independent of the SNP allele and all four isoform/SNP combinations could occur in ESTs. We count the ESTs manifesting each isoform/SNP combination in a 2 × 2 data matrix where the columns correspond to transcript isoforms and the rows correspond to SNP alleles (Fig. 1B).

Two types of allele-specific transcripts were considered, which we refer to as completely and partially allele-specific. In the case of complete allele specificity, one allele always gives rise to the major form of the transcript and another allele always gives rise to the minor form. Alternatively, if the transcript isoform is partially allele-specific, then each allele may give rise to both the major and the minor form of the transcript but with a probability that is allele-dependent. Partial allele specificity therefore involves changes in the frequency of competing transcript isoforms.

2. Data and methods

Genomic locations of ESTs and SNPs were downloaded from the UCSC Genome Browser database [21] on 24/03/04. ESTs and SNPs that mapped onto more than one genomic position were discarded. We used these genomic coordinates to calculate the positions of SNPs from release 120 of dbSNP on ESTs and retrieved the nucleotide at each SNP position from the EST sequences from release 140 of dbEST. The ASAP database of alternatively spliced genes [20] was downloaded on 02/04/04. The data included short sequence fragments from exon junctions associated with Unigene clusters. We mapped splice junctions involved in alternative splicing to EST sequences from the same Unigene cluster by scanning the ESTs for exact matches to the splice junctions.

2.1. Data matrices

We considered only the two most common alleles at each SNP position and pairs of splice junctions that are associated with different splice isoforms from the ASAP database. This resulted in 2 × 2 matrices of splice junctions and SNP alleles. In order to eliminate potential bias resulting from multiple ESTs derived from the same tissue sample, we restricted the matrices to a single EST per clone library per SNP allele. Because multiple splice junctions and SNPs from the same gene may not be independent, we considered only one SNP and one splice junction pair per Unigene cluster for the estimates of the prevalence of allele-specific isoforms. In each case, the rarest splice junction and a highly represented SNP were selected to maximize the chance of picking up polymorphic splice isoforms that are at low frequency. The restriction to a single SNP and splice junction per gene was not necessary for the detection of individual candidates for allele-specific splicing from statistical associations, but the restriction to just one EST per allele per library was maintained. The matrices as well as the mappings of SNPs and splice junctions from which they were generated are available on our website (http://www.sanbi.ac.za/snp2estmap/febs/).

2.2. Simulations

Randomized replicates of the dataset were constructed under the constraint that row and column sums for each matrix were conserved and with no association between transcript isoform and the nucleotide at the polymorphic site. The number of matrices in the replicate datasets with a zero cell was counted in order to estimate the expected number of matrices with a zero cell in the absence of an association between transcript isoform and SNP. Confidence intervals were determined that included the values obtained from 95% of the replicates. Given the number of zeros in the random datasets, we could estimate the number of additional zeros that result from non-random relationships between the rows and columns of the matrices. In the case of completely allele-specific isoforms, every matrix affected necessarily has at least one zero cell (Fig. 1). We also constructed randomized datasets with a bias such that a strong, but not exclusive, association between a row and a column of the matrix was introduced in a randomly selected proportion of the matrices and estimated the proportion of affected matrices required to reproduce the observed number of matrices with a zero cell.

Fisher's Exact Test and Bonferroni correction were used to test for statistical associations between SNPs and isoforms from individual genes. We restricted the matrices tested to only those that could achieve a maximal P-value of 0.001 based on their row and column totals to reduce the correction for multiple testing.

3. Results

We obtained 1295 matrices corresponding to ESTs that we mapped to splice junctions and SNPs as described in Section 2. Of these, 139 contained rows and columns that each summed to at least two and were thus informative using our approach based on counting the number of matrices with a zero cell (matrices with a row or column that sums to one necessarily having a zero cell). Restricting the analysis to these matrices also ensured that only transcript isoforms represented in at least two different cDNA libraries or from two different alleles within a single library were considered. Between 4 and 137, different cDNA libraries were represented in each of these matrices. Because we restricted the analysis to a single matrix per Unigene cluster, these matrices correspond to 139 different Unigene clusters. In the observed data, 85 matrices had a zero cell and were thus consistent with allele-specificity. In equivalent sets of randomized matrices, on an average 71 matrices had a zero element (Fig. 2). Only four out of 1000 replicates had as many or more matrices with a zero cell as found in the observed data. The observed number of matrices with at least one zero cell (indicated by the arrow in Fig. 2) is not likely to have resulted from random data with no relationship between SNP allele and transcript isoforms and instead indicates the presence of allele-specific isoforms in the data. The 14 additional matrices with a zero cell in the observed data correspond to 21% of the average number of matrices without a zero cell in the randomized datasets.

If we consider complete allele-specificity as the only alternative to random association (i.e., partial allele-specificity is not permitted), then a minimum 6% of the isoforms in the dataset are completely allele-specific (the proportion most consistent with the data is 21% and the maximum proportion possible is 36%; Fig. 3A). Partial allele-specificity, where different alleles have different but incomplete isoform biases, can also increase the number of matrices with a zero cell in the data but it does so less efficiently. We modeled the situation where a proportion of the matrices in the dataset showed a strong, but not exclusive, association between transcript isoform and nucleotide polymorphism and used simulation to estimate the expected number of matrices with a zero cell. The association between rows and columns for the affected matrices was such that, for an individual EST, the probability of it being isoform k given that it is SNP allele i was four times the probability of being isoform k given that it is SNP allele j. With this fourfold association, at least 14% of the isoforms in the dataset would need to be partially allele-specific in order to explain the observed number of matrices with a zero cell. The real data are likely to contain a mixture of completely and partially allele-specific isoforms. Combinations of proportions of completely and partially allele-specific forms that would be most likely to result in the number of zeros in the observed data are shown in Fig. 3B.

We tested individual data matrices for non-random associations between isoforms and SNPs. An example is shown in Fig. 1 for the gene *RNPC4*. The P-value for the association between the rows and columns of the matrix shown is 0.0002 (from a 2-tailed Fishers Exact Test). The highest statistical significance ($P < 2 \times 10^{-8}$) of any matrix in the dataset was achieved for the association between this same isoform and another SNP that was located closer to the alternatively spliced exon junction. This association remains highly significant after correction for multiple testing, suggesting that this gene-does indeed have polymorphic splice variants that are found at sufficiently high frequency in the population to be

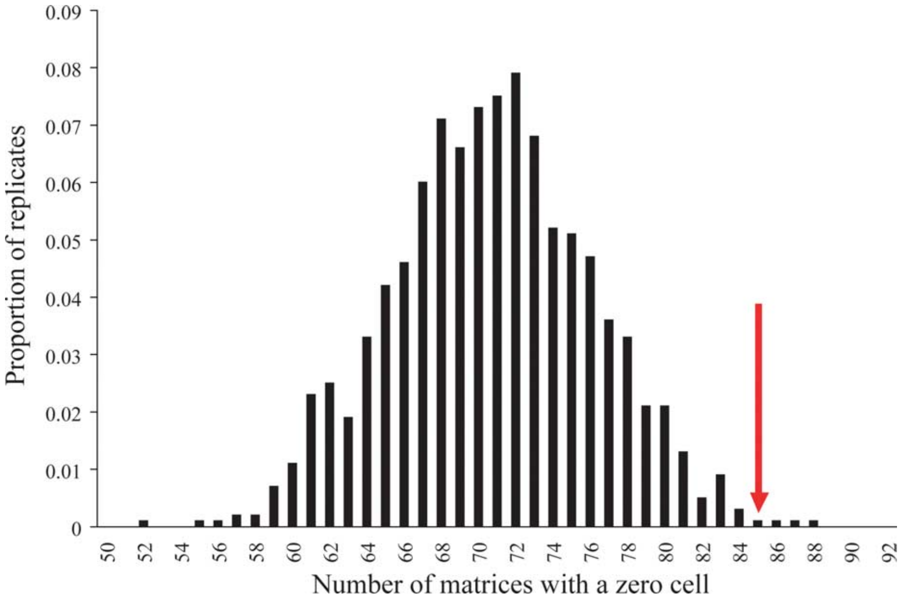

Fig. 2. The number of matrices consistent with completely allele-specific splicing in randomized data. Histogram showing the proportion of matrices with a zero cell from 1000 randomized replicates of the dataset. The arrow indicates the number of matrices with a zero cell in the observed data.

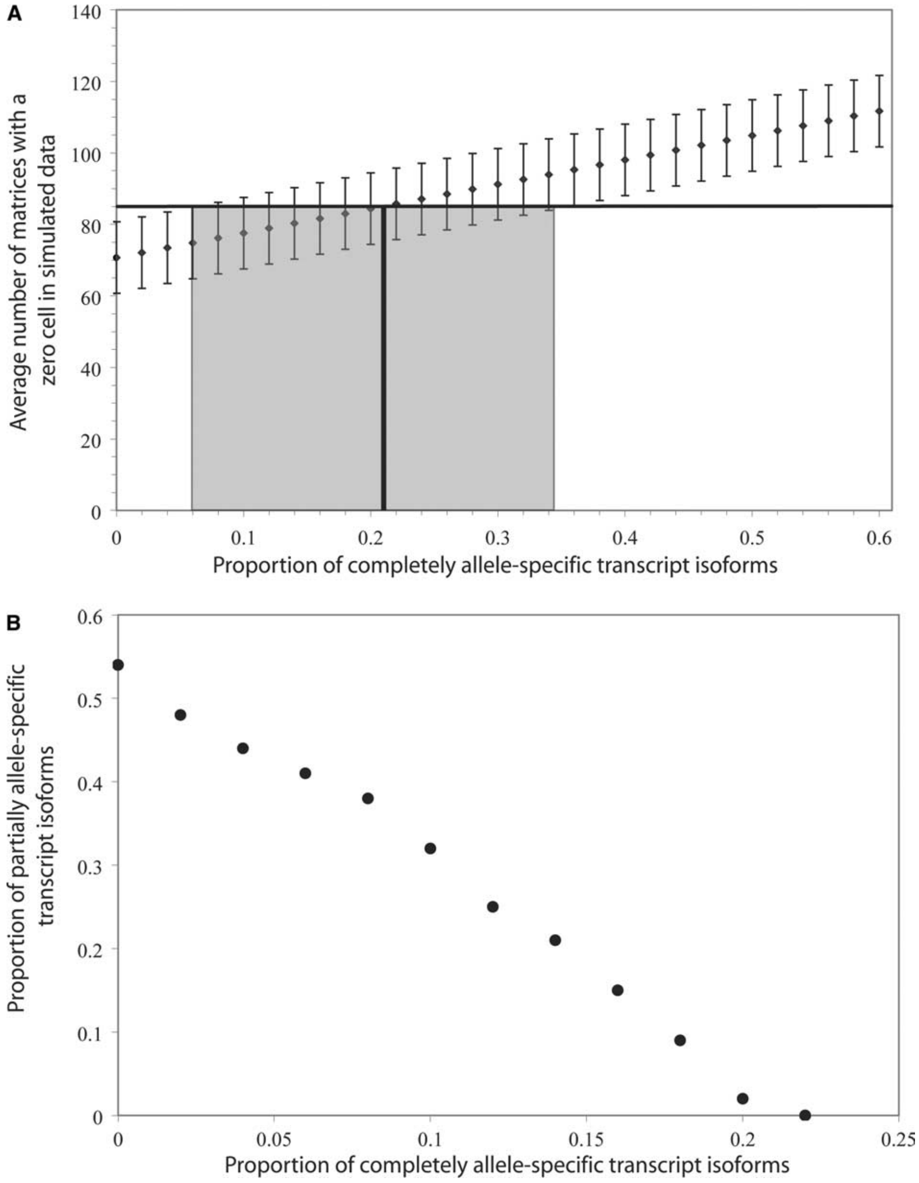

Fig. 3. Simulations of complete and partial allele-specific transcript isoforms. (A) Average numbers of matrices with a zero cell from 1000 simulated replicates of the dataset in which a proportion of the simulated matrices are derived from completely allele-specific transcript isoforms. The horizontal line shows the number of matrices with a zero cell in the observed data. The shaded area shows the confidence interval and the dark line the proportion of completely allele-specific isoforms most consistent with the observed number of matrices with a zero cell. Error bars were derived from the simulation as described in Section 2. (B) Combinations of completely and partially allele-specific transcript isoforms that could explain the observed number of matrices with a zero cell in the data. The points on the graph show the proportions of completely (y-axis) and partially (x-axis) allele-specific isoforms in the simulations that result in approximately the same number of matrices with a zero as we have observed in the data. In this example the strength of the association between rows and columns of the simulated matrices was such that, for an individual EST, the probability of its being in column k given that it was in row i was a factor of four higher than the probability of its being in column k given that it was in row j.

detectable using this method. Data matrices from 32 genes for which we have been able to detect associations between SNPs and splice junctions can be accessed at our website (http://www.sanbi.ac.za/snp2estmap/febs/). However, while all data matrices corresponding to completely allele-specific transcript isoforms are expected to have a zero cell (in the absence of recombination between the exon junctions and the SNP position), statistical associations between the matrix rows and columns occur only under specific conditions for realistic amounts of data (e.g., the second mutation occurs on a rare allele of the earlier polymorphism). For most individual matrices, there are insufficient data to establish allele-specificity with any confidence, especially when correction for multiple testing is applied.

4. Discussion

The factors that control alternative splicing in mammalian genes are imperfectly understood. Most previous studies have concentrated on variation in splicing pattern across tissues, which is probably controlled by the availability of *trans*-acting factors [22]. Our results indicate that *cis*-acting allelic variation also plays a role in determining which splice variant is observed. However, the use of public transcript data to estimate either the tissue-specificity [22] or the allele-specificity (as studied here) of transcript isoforms is complicated by the fact that multiple overlapping ESTs from the same individual are present in dbEST. If there is a significant proportion of allele-specific transcript isoforms in human populations, then any conservative test for tissue specificity of isoforms should be restricted to a single EST per allele from any one library to get a random sample of alleles. Alternatively, if the predominant variation among transcript isoforms is between tissues within a single individual, then any test for allele-specificity of isoforms should be restricted to a single EST per allele per library. In order to provide a conservative estimate of the contribution of allele-specific isoforms to transcript variation, we have restricted our analysis to one sample per allele per clone library and also to one data matrix per gene.

We propose that the allele-specificity that we report is likely to result from nucleotide polymorphisms that affect either the meaning or the strength of splice signals or enhancers. Examples of mutations in splice signal regions that affect mRNA splicing have previously been reported [9]. It is possible, however, that other causes contribute to the observed isoform allele-specificity. For instance, allele-specificity could also result from a polymorphism that determines whether nonsense-mediated decay of one isoform occurs. If the longer form of a transcript contains a nucleotide polymorphism that results in a stop codon in the translated product and if this polymorphic site is absent from the shorter form, then the longer form of the transcript would be subject to nonsense-mediated decay only in the case of the allele containing the stop codon (Fig. 1E). This could result in under-representation of the longer isoform derived from the allele containing the stop codon. In this case, both of the transcript isoforms are produced from each allele (true alternative splicing) but the likelihood of observing both isoforms would be reduced for the allele with the stop codon. While we feel that this mechanism could, in theory, contribute to allele-specificity, we are not aware that any examples of isoform allele-specificity by this mechanism have been reported. It is also possible that an apparent alternatively spliced isoform could result directly from a polymorphic deletion spanning an entire exon. An example of such a variant was reported for the human growth hormone receptor [23].

Allele-specificity of transcript isoforms may be complete or partial. A nucleotide polymorphism may alter splicing of all transcripts of the gene so that each allele produces just one form of the transcript (complete allele-specificity). Alternatively, more than one transcript isoform may be generated from a single allele (true alternative splicing) but the frequency at which each isoform is found may vary between alleles (partial allele-specificity). An example of the latter type of polymorphism was reported in the human *XPC* DNA repair gene [15], whereas in the *GNB3* gene there is almost complete association between the shorter splice variant and the 825T allele [24,25]. Complete and partial allele-specificity can both result in an association between transcript isoform and nucleotide polymorphisms but the former will cause the strongest association. As a consequence, our estimate of the proportion of completely allele-specific transcript isoforms required to explain the number of observed matrices with a zero cell should be considered as a lower bound on the proportion of the dataset that is affected by either complete or partial allele-specificity.

Recombination events, which should be rare when the distance between the splice junction and the SNP being monitored is small compared to typical human haplotype block sizes, can reduce the signal in the data and could cause underestimation of the level of isoform allele-specificity but cannot introduce false positive results. It is possible that transcript isoform choice is also affected by an individual's genotype at other unlinked loci, but dbEST-based methods are unable to detect this. Similarly, although EST sequences are known to be error-prone, random sequencing errors cannot introduce false positive associations.

There are multiple potentially independent alternatively spliced junctions for many of the genes represented in the ASAP database. Because our estimate of the prevalence of allele-specific isoforms took account of just one alternative isoform per gene, the total proportion of genes from the alternative splicing databases for which at least one allele-specific isoform exists may be far higher than the lower bound presented here.

The existence of polymorphic transcript isoforms within populations is consistent with the recent discovery that exons found only in minor splice forms are often completely absent from the orthologous gene in human/rodent comparisons [26–28], because evolutionary changes in gene structure must originate as polymorphisms within species. We provide a lower bound of 6% for the proportion of alternatively spliced genes for which either completely allele-specific or partially allele-specific transcript isoforms are present in the dataset. This lower bound was calculated based only on exon junctions from transcript isoforms that are present in at least two different EST libraries and excludes the contribution to transcript diversity from rare mutations. Interestingly, no natural upper bound on the proportion of alternatively spliced genes that are affected by polymorphism emerges from our analysis and we argue that the contribution of allele-specific transcript isoforms should be stated as a caveat in future estimates of the prevalence of alternative splicing. Our results emphasize that true alternative splicing cannot be confirmed unless more than one transcript is observed from the same allele and also caution that any inference of tissue-specific splicing must take account of allele-specificity.

Acknowledgements: We are grateful to Chris Gehring, Win Hide, Chris Muller, Yi Xing and Rüdiger Bräuning for critical comments, suggestions and assistance and to the South African National Bioinformatics Network for financial support. The manuscript has benefited from the comments of anonymous reviewers, one of whom suggested the possibility of allele-specificity through NMD.

References

[1] Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P. and Mattick, J.S. (2000) Nat. Genet. 24, 340–341.

[2] Modrek, B. and Lee, C. (2002) Nat. Genet. 30, 13–19.

[3] Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001) Nucleic Acids Res. 29, 2850–2859.

[4] International Human Genome Sequencing Consortium, 2001. Nature 409, 860-921.

[5] Hide, W.A., Babenko, V.N., van Heusden, P.A., Seoighe, C. and Kelso, J.F. (2001) Genome Res. 11, 1848–1853.

[6] Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999) Genome Res. 9, 1288–1293.

[7] Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003) Science 302, 2141–2144.

[8] Sorek, R., Shamir, R. and Ast, G. (2004) Trends Genet. 20, 68–71.

[9] Garbarz, M., Tse, W.T., Gallagher, P.O., Picat, C., Lecomte, M.C., Galibert, F., Dhermy, D. and Forget, E.G. (1991) J. Clin. Invest. 88, 76–81.

[10] Cogan, J.D., Phillips III, J.A., Schenkman, S.S., Milner, R.D. and Sakati, N. (1994) J. Clin. Endocrinol. Metab. 79, 1261–1265.

[11] Pohlenz, J., Dumitrescu, A., Aumann, U., Koch, G., Melchior, R., Prawitt, D. and Refetoff, S. (2002) J. Clin. Endocrinol. Metab. 87, 336–339.

[12] van Leusden, M.R., Pas, H.H., Gedde-Dahl Jr., T., Sonnenberg, A. and Jonkman, M.F. (2001) Lab Invest. 81, 887–894.

[13] Hellwinkel, O.J., Bull, K., Holterhus, P.M., Homburg, N., Struve, D. and Hiort, O. (1999) J. Steroid Biochem. Mol. Biol. 68, 1–9.

[14] Cox, L.A., Jett, C. and Hixson, J.E. (1998) J. Lipid Res. 39, 1319–1326.

[15] Khan, S.G., Muniz-Medina, V., Shahlavi, T., Baker, C.C., Inui, H., Ueda, T., Emmert, S., Schneider, T.D. and Kraemer, K.H. (2002) Nucleic Acids Res. 30, 3624–3631.

[16] Cartegni, L., Chew, S.L. and Krainer, A.R. (2002) Nat. Rev. Genet. 3, 285–298.

[17] Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S. and Sunyaev, S. (2003) Trends Genet. 19, 124–128.

[18] Caceres, J.F. and Kornblihtt, A.R. (2002) Trends Genet. 18, 186–193.

[19] Pagani, F. and Baralle, F.E. (2004) Nat. Rev. Genet. 5, 389–396.

[20] Lee, C., Atanelov, L., Modrek, B. and Xing, Y. (2003) Nucleic Acids Res. 31, 101–105.

[21] Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. (2003) Nucleic Acids Res. 31, 51–54.

[22] Xu, Q., Modrek, B. and Lee, C. (2002) Nucleic Acids Res. 30, 3754–3766.

[23] Pantel, J., Machinis, K., Sobrier, M.L., Duquesnoy, P., Goossens, M. and Amselem, S. (2000) J. Biol. Chem. 275, 18664–18669.

[24] Siffert, W., Rosskopf, D., Siffert, G., Busch, S., Moritz, A., Erbel, R., Sharma, A.M., Ritz, E., Wichmann, H.E., Jakobs, K.H. and Horsthemke, B. (1998) Nat. Genet. 18, 45–48.

[25] Rosskopf, D., Manthey, I. and Siffert, W. (2002) Pharmacogenetics 12, 209–220.

[26] Modrek, B. and Lee, CJ. (2003) Nat. Genet. 34, 177–180.

[27] Lareau, L.F., Green, R.E., Bhataagar, R.S. and Brenner, S.E. (2004) Curr. Opin. Struct. Biol. 14, 273–282.

[28] Thanaraj, T.A., Clark, F. and Muilu, J. (2003) Nucleic Acids Res. 31, 2544–2552.