ACTA ABP BIOCHIMICA POLONICA

Vol. 62, No 1/2015
57–61
http://dx.doi.org/10.18388/abp.2014_771

Regular paper

Potential protein activity modifications of amino acid variants in the human transcriptome

Joanna Zyla[1]✉, Robert A. Bulman[2]✉, Christophe Badie[2] and Simon D. Bouffler[2]

[1]Data Mining Group, Institute of Automatic Control, Faculty of Automatic, Electronic and Computer Science, Silesian University of Technology, Gliwice, Poland; [2]Public Health England, Chilton, Didcot, United Kingdom

Background: The occurrence of widespread RNA and DNA sequence differences in the human transcriptome was reported in 2011. Similar findings were described in a second independent publication on personal omics profiling investigating the occurrence of dynamic molecular and related medical phenotypes. The suggestion that the RNA sequence variation was likely to affect disease susceptibility prompted us to investigate with a range of algorithms the amino acid variants reported to be present in the identified peptides to determine if they might be disease-causing. Results: The predictive qualities of the different algorithms were first evaluated by using nonsynonymous single-base nucleotide polymorphism (nsSNP) datasets, using independently established data on amino acid variants in several proteins as well as data obtained by mutational mapping and modelling of binding sites in the human serotonin transporter protein (hSERT). Validation of the used predictive algorithms was at a 75% level. Using the same algorithms, we found that widespread RNA and DNA sequence differences were predicted to impair the function of the peptides in over 57% of cases. Conclusions: Our findings suggest that a proportion of edited RNAs which serve as templates for protein synthesis is likely to modify protein function, possibly as an adaptive survival mechanism in response to environmental modifications.

Key words: RNA editing, amino acid variants

Received: 23 April, 2014; revised: 21 October, 2014; accepted: 12 December, 2014; available on-line: 04 February, 2015

INTRODUCTION

In a publication, which was extensively commented on, Li and coworkers (2011) reported in 2011 the occurrence of widespread RNA and DNA sequence differences (RDD) in the human transcriptome (Li *et al.,* 2011). The authors emphasized the consistent pattern of the observations and concluded that the RDDs had a biological significance and were not just "noise" (Li *et al.,* 2011). The nature of the amino acid variants in the novel RDD peptides was investigated by using mass spectrometry. The authors also suggested that the RNA sequence variation was likely to affect disease susceptibility by modifying the function of the protein. Subsequently, in 2012 Li and coworkers (2012) responded to comments which did not dispute their findings but were in disagreement on the number of RDDs. Li and coworkers (2012) also pointed out that several other research groups had reported similar phenomena. Whereas Li and coworkers (2011) had sequenced peptides from immortalized B cells in culture as well as in some primary skin cells and brain tissue, Chen and coworkers in 2012, identified other forms of RNA variants, henceforth RNA-edits, in circulating white blood cells. Despite the expected difference in gene expression between *in vitro* and *in vivo* cells of different origins, both studies reported RNA editing, which according to our results often leads to modified protein activity. Also widely discussed was the role of technical artifacts due to sequencing or sequence mapping, nevertheless they only partially explain the discovered RDDs (Pickrell *et al.*, 2012). This would suggest that it is a general, widespread editing mechanism that can affect phenotype and should be further investigated.

Until recently nsSNPs have been the major group of amino acid polymorphisms that are associated with protein stability. Worth and coworkers have estimated that up to 80% of disease-associated nsSNPs are associated with protein stabilization effects (Worth *et al.,* 2007). It is possible that the amino acid variants reported by Li and coworkers (2011) might have the potential to vary the function of the RDD peptides. By using algorithms such as Polyphen-2 (POLYmorphism PHENotyping ver. 2) (Adzhubei *et al.,* 2010) and SIFT (Sorting Intolerant From Tolerant) (Kumar *et al.,* 2009) it should be possible to determine how amino acid variants might change the function of the RDD-peptides and also other similar phenomena. The practical application of the two algorithms is now so well established that they are integrated into *Ensembl* [*www.ensembl.org/index.html*]. Other available algorithms are PANTHER (Protein Analysis Through Evolutionary Relationships) (Mi *et al.,* 2007), PhD-SNP (Predictor of human Deleterious Single Nucleotide Polymorphisms) (Capriotti *et al.,* 2006) and SNAP (Synonymous Non-synonymous Analysis Program) (Bromberg

✉e-mail: Joanna Zyla: joanna.zyla@polsl.pl; Robert A. Bulman: robert.bulman@phe.gov.uk

Abbreviations: APF, affects protein function; BLAST, the basic local alignment search tool; BNG, benign; D, disease-causing; DNA, deoxyribonucleic acid; DSLS, differential static light scattering; hSERT, human SERotonin transporter protein; MSC, median of conservation value; N, neutral; NN, non-neutral; NSP, number of sequence at position; nsSNP, nonsynonymous single-base nucleotide polymorphism; PANTHER, protein analysis through evolutionary relationships; PhD-SNP, predictor of human deleterious single nucleotide polymorphisms; Polyphen-2, polymorphism phenotyping ver. 2; PRD, probably disease-causing; PSD, possibly sisease-causing; RDD, RNA and DNA sequence differences; RI, reliability index; RNA, ribonucleic acid; Sens: Spec, sensitivity and specificity; SIFT, sorting intolerant from tolerant; SNAP, synonymous non-synonymous analysis program; subSPEC, substitution position-specific evolutionary conservation; TOL, tolerated

Table 1. Predictions by algorithms of the disease-causing potential of the nsSNPs previously selected by using differential static light scattering.
Notes [1]ΔT_{agg} = T_{agg} [Wild typeT] — T_{agg} [nsSNP Variant]. **subSPEC,** substitution position-specific evolutionary conservation; PANTHER scoring is thus: subSPEC value of -3.5 and "greater" indicates disease-causing prediction. **PRD**, probably disease-causing, **PSD** possibly disease-causing, **D**, disease-causing, **APF** affects protein function; **NN,** non-neutral; N, neutral; BNG, benign; TOL, tolerated. The use of *prd*, *bng, d* and *n* indicates that the prediction is not fully conclusive. **RI**, reliability index; **MSC**, Median of Conservation value; **NSP**, Number of Sequence at Position; **Sens : Spec**, sensitivity and specificity. INMT, indolethylamine *N*-methyltransferase. PKM2, pyruvate kinase muscle 2; SULT1A1, sulfotransferase. Results in bold are significantly deleterious. While the results in italics might appear significant they are not statistically significant.

PROTEIN & amino acid variants		PANTHER	PhD-SNP	PolyPhen-2		SIFT		SNAP
	ΔT_{agg}	subPSEC	Effect : RI	Status : Score	Sens : Spec	Effect : Score	MSC : NSP	Prediction:RI : Expected Accuracy
INMT								
F254C	–5.4	**–3.69**	*n : 1*	*prd : 1.00*	*0.0 : 1.0*	**APF : 0.01**	**3.08 : 20**	**NN: 3 : 78%**
PRMT3								
L440V	–3.2	**–3.86**	N : 6	*prd : 1.00*	*0.0 : 1.0*	**APF : 0.03**	**3.08 : 19**	N : 1 : 60%
S470C	–5.8	–6.61	*n : 1*	*bng : 0.00*	*1.0 : 0.0*	**APF : 0.00**	**3.08 : 19**	**NN : 5 : 87%**
S508N	–0.2	**–3.91**	N : 9	BNG : 0.00	1.0 : 0.0	TOL : 1.00	3.08 : 19	N : 2 : 69%
PKM2								
E28K	–6.6	**–3.51**	**D** : 6	**PRD : 0.98**	**0.72 : 0.97**	**APF : 0.01**	**3.03 : 79**	**NN : 3 : 78%**
C31F	–4.6	–4.17	**D** : 9	**PRD : 0.99**	0.68 : 0.97	**APF : 0.00**	**3.02 : 80**	NN : 2 : 70%
V71G	–4.2	–5.25	**D** : 8	**PRD : 0.99**	0.27 : 0.99	**APF : 0.00**	**3.01 : 85**	n : 0 : 53%
V292L	–5.5	–4.63	**D** : 8	*bng : 0.00*	*0.99 : 0.15*	**APF : 0.01**	**3.02 : 87**	N : 1 : 60%
Q310P	–11.1	–7.17	**D** : 7	*prd : 1.00*	*0.0 : 1.0*	**APF : 0.00**	**3.02 : 87**	**NN : 5 87%**
R339P	–4.3	–5.93	**D** : 8	BNG : 0.099	0.93 : 0.85	**APF : 0.00**	**3.02 : 87**	**NN : 3 : 78%**
S437Y	–8.5	–4.12	**D** : 4	**PSD : 0.78**	0.85 : 0.93	**APF : 0.00**	**3.02 : 87**	**NN : 3 : 78%**
V490L	–2.8	**–3.26**	N : 5	**PSD : 0.92**	**0.81 : 0.94**	TOL : 0.32	3.02 : 87	N : 7 : 94%
SULT1A1								
E151Q	–0.7	**–4.00**	N : 6	*bng : 0.001*	*0.99 : 0.15*	TOL : 0.17	3.01 : 64	*n : 0 : 53%*
E151D	0.6	–3.07	N : 7	*bng : 0.00*	*1.0 : 0.0*	TOL : 1.00	3.01 : 64	N : 2 : 69%
R213H	–0.6	–4.46	*n : 2*	BNG : 0.03	0.95 : 0.82	**APF : 0.00**	**3.00 : 63**	N : 1 : 60%
S290T	1.2	–3.38	N : 9	*bng : 0.00*	*1.0 : 0.00*	TOL : 0.67	3.01 : 64	N : 7 : 94%

et al., 2008). By using the five foregoing algorithms we have evaluated the data from Li *et al.* (2011) and Chen *et al.* (2012) as well as providing further insight into the predictive properties of the algorithms. All of the algorithms are widely known and recommended as *in silico* tools for assessing the functional effect of nsSNPs (Patnala *et al.,* 2013). We also validate the predictive properties of the algorithms by drawing upon amino acid variant data produced in different studies (Allali-Hassani *et al.,* 2009; Andersen *et al.,* 2010). In addition, the RNA editing of the GluR2 transcript GRIA2, a process which results in a 607Arg instead of Gln607 variant (Mercucci *et al.,* 2011), has also been investigated.

METHODS

The disease-causing potential of the variants was first evaluated by using algorithms Polyphen2 (Adzhubei *et al.,* 2010) and SIFT (Kumar *et al.,* 2009). Subsequently, other available algorithms, PANTHER (Mi *et al.,* 2007), PhD-SNP (Capriotti *et al.,* 2006) and SNAP (Bromberg *et al.,* 2008) were also used to evaluate the data from Li *et al.* (2011) and Chen *et al.* (2012) as well as providing further insight into the predictive properties of the algorithms. BLAST (The Basic Local Alignment Search Tool) (Altschul *et al.,* 1992), a rapid sequence similarity search tool, evaluates a submitted protein sequence to detect high degrees of amino acid sequence similarity. The algorithms then check the location of an amino acid variant throughout all available sequences. A low occurrence of similarity of an amino acid has the potential to impair a protein's function. More detail is available in reviews on the function of algorithms (Capriotti *et al.*, 2006; Mi *et al.*, 2007; Bromberg *et al.*, 2008; Kumar *et al.*, 2009; Adzhubei *et al.*, 2010).

Validation of algorithms. A validation of the predictive properties of the algorithms was sought by using the amino acid variants reported by Allali-Hassani and coworkers (2009), who noted that approximately 75% of the mutations affect the biochemical function directly. Allali-Hassani and coworkers (2009), in their extensive account, assessed the thermostability of nsSNPs and wild type proteins by differential static light scattering (DSLS) which is essentially an aggregation-based method for assessing protein stability. The authors demonstrated that 46 nsSNP amino acid variants altered the stability and activity of 16 human enzymes (Allali-Hassani *et al.,* 2009). In addition, amino acid substitution data, which were reported by Andersen *et al* in their account of mutational mapping and modelling of binding sites in the human serotonin

Table 2. Predictions by algorithms of the influence of amino acid variants on the transport of 5-hydroxytryptamine by serotonin.
subSPEC, substitution position-specific evolutionary conservation; PANTHER scoring is thus: subSPEC value of –3.5 and "greater" indicates disease-causing prediction. **PRD**, probably disease-causing, **PSD** possibly disease-causing, **D**, disease-causing, **APF** affects protein function; **NN,** non-neutral; N, neutral; BNG, benign; TOL, tolerated. The use of *prd*, *bng, d* and *n* indicates that the prediction is not fully conclusive. **RI**, reliability index; **MSC**, Median of Conservation value; **NSP,** Number of Sequence at Position; **Sens : Spec**, sensitivity and specificity. Results in bold are significantly deleterious, the italic results appear as significant but they are statistically not significant.

Variant amino acid	Transport Activity % of *WT*	PANTHER	PhD - SNP	PolyPhen-2		SIFT		SNAP
		subPSEC	Effect : RI	Status : Score	Sens : Spec	Effect : Score	MSC : NSP	Prediction : RI : Expected Accuracy
WT	100							
D98E	64±8	**–4.60**	*d : 2*	*prd* : 1.00	0.00 : 1.00	**AFP** : **0.00**	3.04 : 51	**NN** : 3 : 78%
N177S	42 ± 2	–3.91	N : 6	**PRD** : 0.98	0.76 : 0.96	TOL : 0.05	3.04 : 51	N : 2 : 68%
F341Y	57 ± 8	–4.62	**D** : 3	*prd* : 1.00	0.00 : 1.00	**AFP** : **0.02**	3.04 : 51	N : 1 : 60%
S438T	20± 4	**–3.87**	**D** : 6	**PSD** : 0.47	0.89 : 0.90	TOL : 0.05	3.04 : 51	N : 3 : 78%

transporter (Andersen *et al.,* 2010), have also been used to ascertain the performance of the algorithms.

RESULTS

Validation of algorithms

The predictive qualities of algorithms have been evaluated (Table 1) by using nsSNP data which had already been probed by the biophysical and chemical techniques described by Allali-Hassani and coworkers (2009). Data reported by Allali-Hassani and coworkers is included in Table 1, as the first two columns. Further classification of a disease-causing status is indicated in the note attached to Table 1. Examination of Table 1 indicates that over 83% of the nsSNPs in PKM2 is predicted to be disease-causing when ΔT_{agg} values range from –2.8 to –11.1. For PRMT3, 70% of the nsSNPs is classified as disease-causing when ΔT_{agg} values are –3.2 to –5.8.

We have also evaluated, the predictive properties of the algorithms by using data published by Andersen and coworkers (2010) who examined the influence of amino acid variants on the transport of (*S*)-citalopram by hSERT (Table 2). The authors noted these characteristics: (i) extension of the acidic side chain by one methylene unit (Asp98Glu) caused a 15-fold loss of potency of (*S*)-citalopram; (ii) Asn-177 contributes to formation of the surface region in the substrate binding pocket of hSERT; (iii) Phe-341 is a major determinant of the contour of the binding site of hSERT. Again, the pre-

Table 3. Evaluation of induction of phenotype variation by amino acid variants in RDD peptides.
subSPEC, substitution position-specific evolutionary conservation; PANTHER scoring is thus: subSPEC value of –3.5 and "greater" indicates disease-causing prediction. **PRD**, probably disease-causing, **PSD** possibly disease-causing, **D**, disease-causing, **APF** affects protein function; **NN,** non-neutral; N, neutral; BNG, benign; TOL, tolerated. The use of prd, *bng, d* and *n* indicates that the prediction is not fully conclusive. **RI**, reliability index; **MSC**, Median of Conservation value; **NSP,** Number of Sequence at Position; **Sens : Spec**, sensitivity and specificity. Results in bold are significantly deleterious, the italic results appear as significant but they are statistically not significant.

PROTEIN : Amino acid change	PANTHER	PhD-SNP	PolyPhen-2		SIFT		SNAP
	subPSEC	Effect : RI	Status : Score	Sens : Spec	Effect : Score	MSC : NSP	Prediction : RI : Expected Accuracy
AP2A2 : Y346D	**–8.50**	**D** : 8	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.04 : 65	**NN** : **5** : **87%**
DFNA5 : L479Q	–5.47	**D** : **3**	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.06 : 12	**NN** : **4** : **82%**
ENO1 : L224P	–4.26	**D** : **6**	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.03 : 391	**NN** : **3** : **78%**
ENO3 : V367G	–2.91	*d : 0*[1]	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.01 : 396	*nn : 0 : 58%*
FABP3 : W9R	–8.15	**D** : **8**	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.00 : 77	**NN : 6** : **93%**
FH : I52K	–1.99	*d : 1*	*bng : 0*	*1 : 0*	TOL : 0.10	3.24 : 390	N : 3 : 78%
HMGB1 : Y16N	**–5.70**	**D** : **9**	**PSD** : **0.56**	0.88 : 0.91	**APF** : **0.10**	3.02 : 174	**NN** : **6** : **93%**
NACA : D180N	–2.91	**D** : **5**	BNG : 0.021	0.95 : 0.80	**APF** : **0.00**	3.02 : 106	**NN** : **3** : **78%**
NSF : V612A	–2.55	N : 4	BNG : 0.24	0.91 : 0.88	TOL : 0.99	3.01 : 47	N : 1 : 60%
POLR2B : L444Q	–6.22	N : 8	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.04 : 218	**NN** :**1** : **63%**
RAD50 : L1018R	–4.59	*d : 2*	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.20 : 16	N : 1 : 60%
RPL12 : N103D	–0.95	**D** : **7**	BNG : 0.004	0.97 : 0.59	TOL : 0.06	3.02 : 143	N : 3 : 78%
RPL32 : A115S	–4.10	**D** : **8**	**PRD** : **0.978**	0.76 : 0.96	**APF** : **0.00**	3.01 : 137	n : 0 : 53%
SLC25A17 : E54G	–5.12	N : 6	**PSD** : **0.775**	0.85 : 0.92	**APF** : **0.01**	3.05 : 17	**NN** : **1** : **63%**
TUBA1 : E429G	–5.68	**D** : **4**	**PRD** : **0.998**	0.27 : 0.99	**APF** : **0.00**	2.97 : 288	**NN** : **2** : **70%**
TUBB2C : G269D	**–4.11**	*n : 0*	*prd : 1*	*0 : 1*	**APF** : **0.00**	3.07 : 336	**NN** : **5** : **87%**
% Disease–causing variant	68.7**%**	**D** 56.3% **D**+*d* 75%	**PSD**+**PRD** = 25% **PSD+PRD**+*prd* = 80.0%		81.3%	**–**	**NN** 62.5%; **NN**+nn = 68.8%

Table 4. Evaluation of amino acid variants identified as part of an investigation of personalized medicine.
subSPEC, substitution position-specific evolutionary conservation; PANTHER scoring is thus: subSPEC value of –3.5 and "greater" indicates disease-causing prediction. **PRD**, probably disease-causing, **PSD** possibly disease-causing, **D**, disease-causing, **APF** affects protein function; **NN,** non-neutral; N, neutral; BNG, benign; TOL, tolerated. The use of *prd*, *bng, d* and *n* indicates that the prediction is not fully conclusive. **RI**, reliability index; **MSC**, Median of Conservation value; **NSP**, Number of Sequence at Position; **Sens : Spec**, sensitivity and specificity. Results in bold are significantly deleterious, the italic results appear as significant but they are statistically not significant.

PROTEIN : Amino acid change	PANTHER	PhD – SNP	PolyPhen-2		SIFT		SNAP
	subPSEC	Effect : RI	Status : Score	Sens : Spec	Effect : Score	MSC : NSP	Prediction : RI : Expected Accuracy
IGFBP7 : K95R	–1.99	N : 9	BNG : 0.005	0.97 : 0.74	TOL : 0.04	2.99 : 27	**NN : 1 : 60%**
BLCAP : Q5R	–3.07	N : 5	*prd : 1*	*0 : 1*	*apf : 0.0*	*3.60 : 13*	**NN : 2 : 70%**
BLCAP : Y2C	**–5.04**	**D** *: 1*	*prd : 1*	*0 : 1*	*apf : 0.0*	*3.60 : 13*	**NN : 4 : 82%**
AZIN1 : S367G	–3.58	N : 6	*bng : 0*	*1 : 0*	TOL : 0.13	3.04 : 30	N : 5 : 89%
CYFIP2 : Y897S	–3.97	*d : 2*	BNG : 0.41	0.89 : 0.89	**APF : 0.02**	**3.04 : 37**	**NN : 2 : 70%**
CYFIP2 : T1067A	–1.59	N : 8	BNG : 0.32	0.90 : 0.89	**APF : 0.03**	**3.04 : 37**	N : 6 : 92%
FBXO25 : R83H	–1.98	*n : 1*	BNG : 0.01	0.96 : 0.78	TOL : 0.17	3.01 : 30	N : 5 : 89%
NDUFV1 : I439L	–1.88	N : 6	**PSD : 0.56**	0.88 : 0.91	TOL : 0.22	3.00 : 224	N : 6 : 92%
PAFAH1B3 : N86D	–0.63	*n : 1*	**PSD : 0.63**	0.87 : 0.91	**APF : 0.01**	**3.03 : 40**	**NN :1 : 60%**
APOBEC3F : P244S	–1.43	N : 4	BNG : 0.17	0.92 : 0.82	TOL : 0.51	3.04 : 29	N : 4 : 85%
% Disease–causing variant amino acids	40.0%	10%	**PSD** 20%; **PSD** + prd 40.0%		**APF** 30%; **APF**+apf 50%		50%

Table 5. Summary of the percentage of amino acid variants predicted by the algorithms to vary the phenotype.

	PANTHER	PhD-SNP	PolyPhen-2	SIFT	SNAP
RDD	69%	56%	25%	81%	56%
RNA-edits	40%	10%	20%	30%	50%

dictions of the algorithms support the independently acquired data reported by Andersen and coworkers (2010).

The variant amino acid 607-arginine in the GluR2 transcript GRIA2 (Mercucci *et al.,* 2011) abolishes the 100% impermeability to calcium. Of the five algorithms, Polyphen2 and SIFT fail to predict this established variation in the phenotype.

DISCUSSION

Our findings suggest that a proportion of edited RNAs which serve as templates for protein synthesis is likely to modify protein function. The process of editing RNA, which is still not well known, is possibly an adaptive survival mechanism in response to environmental modifications. By being able to check an amino acid sequence for a single variance, it is possible to identify a potential effect that might induce an alteration of clinical significance in the organism. The predictions by the algorithms of disease-causing states for some amino acid variants are important for clinical practice. As shown in Table 1, the value of such predictions is strengthened by the results reported by Allali-Hassani and coworkers (2009). This outcome indicates that the evaluated algorithms have a satisfactory level of prediction and as such are in agreement with the experimentally obtained data reported by Allali-Hassani and coworkers (2009). However, when the algorithms are being clinically applied it is necessary to recognize that there are methodological differences, which might lead to a variation in the results (Hicks *et al.,* 2011). We suggest that checking RNA editing by only one of the methods is not reliable. Also, it is important to recognize that the algorithms indicate only a possible effect and as such might not be wholly adequate for medical diagnosis. Nevertheless, all of them are helpful tools to identify a potentially relevant and interesting polymorphism from potential candidates and, also, could help to assess the effects of RNA editing.

This study provides a means of independently evaluating the predictive qualities of the algorithms. We have used data pertaining to RDD variants, Table 3, and amino acid variants in RNA-edit, Table 4, to evaluate the likely impact on the phenotypes of RDD and RNA-edit proteins. A comparison of Tables 3 and 4 reveals striking differences in the characteristics of the RDD and RNA-edit peptides. With the exception of the predictions by Polyphen-2, for AP2A2, DFNA5, ENO1, ENO3, FABP3 and FH, the predictions for phenotype variations in Table 3 are quite similar between algorithms. In contrast, in Table 4 the extent of the similarity in the predictions by the algorithms is much reduced. Predictions for a disease-causing state by PANTHER and SNAP are similar for a few amino acids. Only a few amino acid variants in Table 4 are predicted to be disease-causing by Polyphen-2 and SIFT. PhD-SNP does not report any of the amino acid variants to be disease-causing. In Table 5 we summarize the percentage of disease-causing amino acid variants in the RDD (Li *et al*., 2011). and RNA-edit proteins Chen and coworkers (2012).

The algorithms have been also used to examine a form of RNA editing where there is a clear variation of the phenotype. In the GluR2 transcript GRIA2 the 607-arginine amino acid variants abolishes the 100% impermeability to calcium. Of the five algorithms, Polyphen2 and SIFT fail to predict this established variation in phenotype.

The data we have reported here provides an opportunity to consider the predictive properties of algorithms that are likely to be important aids for identifying variations in function of a wide variety of proteins, for example those reported by Li and coworkers (2011). The data for PKM2, Table 1, is in quite good agreement. The pre-

dictions by Polyphen-2 of the variant amino acids in Tables 3 and 4 are distinctly different. In Table 4 there is a much lower occurrence of disease-causing amino acid variants than there is in Table 3. In Table 3, with the exception of Polyphen-2, the algorithms predict that many amino acid variants are disease-causing.

In summary, we have demonstrated: (i) in Tables 1 and 2 that the five algorithms are largely in agreement with independently reported experimental data, published by Allali-Hassani and coworkers (2009); (ii) the algorithms predict that many of the amino acid variants in the RDD peptides will vary the phenotype of the RDD species reported by Li and coworkers (2011) (Table 3); (iii) in Table 4 the extent of potential disease-causing amino acid variants reported by Chen *et al.* (2012) is lower than that reported in the RDD proteins reported by Li and coworkers (2011).

Our findings suggest that a proportion of edited RNAs which serve as templates for protein synthesis is likely to modify protein function, possibly as an adaptive survival mechanism in response to environmental modifications.

Acknowledgements

This work was partially funded by NCN grant number DEC-2013/08/M/ST6/00924 (JZ). Additionally JZ is holder of the DoktoRis scholarship — a Scholarship program for Innovative Silesia. Only the authors are responsible for the content and writing of the paper. This report is a work commissioned by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health. Calculations were carried out using the infrastructure of GeCONiI (POIG.02.03.01-24-099/13).

Authors' contributions

All the authors (J.Z., R.B., C.B., S.B.) have made substantial contributions to the conception and design of the study, drafting the article or revising it critically for important intellectual content and final approval of the version to be submitted.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicticting damaging missense mutations. *Nat Methods* **7**: 248–249.

Allali-Hassani A, Wasney GA, Chau I, Hong BS, Senisterra G, Loppnau P, Shi Z, Moult J, Edwards AM, Arrowsmith CH, Park HW, Schapira M, Vedadi M (2009) A survey of proteins encoded by non-synonymous single nucleotide polymorphisms reveals a significant fraction with altered stability and activity. *Biochem J* **424**: 15–26.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Andersen J, Olsen L, Hansen KB, Taboureau O, Jørgensen FS, Jørgensen AM, Bang-Andersen B, Egebjerg J, Strømgaard K, Kristensen AS (2010) Mutational mapping and modeling of the binding site for [S]-citalopram in the human serotonin transporter. *J Biol Chem* **285**: 2051–2063.

Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. *Bioinformatics* **24**: 2397–2398.

Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**: 2729–2734.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**: 1293–1307.

Hicks S, Wheeler DA, Plon SE, Kimmel M (2011) Prediction of missense mutation functionality depends on both algorithm and sequence alignment employed. *Hum Mutat* **36**: 661–668.

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073–1083.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**: 53–58.

Li M, Wang IX, Cheung VG (2012) Response to comments on widespread RNA and DNA sequence differences in the human transcriptome. *Science* **335**: 1302.

Marcucci R, Brindle J, Paro S, Casadio A, Hempel S, Morrice N, Bisso A, Keegan LP, Del Sal G, O'Connell MA (2011) Pin1 and WWP2 regulate GluR2 Q/R site RNA editing by ADAR2 with opposing effects. *EMBO J* **30**: 4211–422.

Mi H, Guo N, Kejariwal A, Thomas PD (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucl Acids Res* **35**: 247–252.

Patnala R, Clements J, Batra J (2013) Candidate genes association studies: a comprehensive guide to useful *in silico* tools. *BMC Genetics* **14**: 39.

Pickrell JK, Gailad Y, Pritchard JK (2012) Comment on widespread RNA and DNA sequence differences in the human transcriptome. *Science* **335**: 1302.

Worth CL, Bickerton GR, Schreyer A, Forman JR, Cheng TM, Lee S, Gong S, Burke DF, Blundell TL (2007) A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms [nsSNPs] and their relation to disease. *J Bioinform Comput Biol* **5**: 1297–1238.