European Journal of Human Genetics (2005) 13, 86–95
© 2005 Nature Publishing Group All rights reserved 1018-4813/05 $30.00
www.nature.com/ejhg

ARTICLE

Single-nucleotide polymorphisms in genes relating to homocysteine metabolism: how applicable are public SNP databases to a typical European population?

Bohumila Janošíková[1], Petra Zavadáková[1] and Viktor Kožich*,[1]

[1]*Institute of Inherited Metabolic Disorders, Charles University, 1st Faculty of Medicine, Prague, Czech Republic*

To facilitate the association studies in complex diseases characterized by hyperhomocysteinemia, we collected structural and frequency data on single-nucleotide polymorphism (SNPs) in 24 genes relating to homocysteine metabolism. Firstly, we scanned $\sim$1.2 Mbp of sequence in the NCBI SNP database (dbSNP) build 110 and we detected 1353 putative SNPs with an average *in silico* genic density of 1:683. Out of 112 putative SNPs in coding regions (cSNPs), we selected a subset of 42 cSNPs and we assessed the applicability of the NCBI dbSNP to the Czech population – a typical representative of European Caucasians – by determining the frequency of the putative cSNPs experimentally by PCR-RFLP or ARMS-PCR in at least 110 control Czech chromosomes. As only 25 of the 42 analyzed cSNPs met the criterion of $\geq$1% frequency, the positive predictive value of the NCBI data set for our population reached 60%, which is similar to other studies. The correlation of SNP frequency between Czechs and other Caucasians – obtained from NCBI and/or literature – was stronger ($r^2 = 0.90$ for 20 cSNPs) than between Czechs and general NCBI database entries ($r^2 = 0.73$ for 27 cSNPs). Moreover, frequencies of all 20 putative cSNPs, for which data in Caucasians were available, were congruently below or above the 1% frequency criterion both in Czechs and in other Caucasians. In summary, our study shows that the NCBI dbSNP is a useful tool for selecting cSNPs for genetic studies of hyperhomocysteinemia in European populations, although experimental validation of SNPs should be performed, especially if the cSNP entry lacks any frequency data in Caucasians.

European Journal of Human Genetics (2005) **13,** 86–95. doi:10.1038/sj.ejhg.5201282
Published online 20 October 2004

Keywords: SNP; gene; database; frequency; predictive value; ethnicity

Introduction

Homocysteine is a thiol-containing amino acid, which occupies a key position in the metabolism of one-carbon units and of sulfur compounds. Many clinical studies revealed an association of elevated plasma levels with an increased risk of cardiovascular disease[1,2] or of other conditions.[3] These studies, however, do not prove causality as they merely demonstrate an epidemiological correlation. Homocysteine metabolism is in part determined by genetic variants, which are fixed at conception and which do not typically change throughout life. Assuming Mendelian randomization, any observed association of these genetic factors with disease would suggest that the respective allelic variants are etiologically related to disease. Although association studies require that suitable genetic markers exist, a comprehensive list of such genetic variants in the field of homocysteine research is not available.

Polymorphism is defined as a heritable DNA change occurring in at least 1% of alleles; variants with frequency

*Correspondence: Dr V Kožich, Institute of Inherited Metabolic Disorders, 1st Faculty of Medicine-Charles University, Ke Karlovu 2, 128 08 Praha 2, Czech Republic. Tel: +420 224967679; Fax: +420 224919392; E-mail: Viktor.Kozich@LF1.CUNI.CZ.

Received 3 March 2004; revised 24 June 2004; accepted 5 August 2004

higher than 10% are considered common polymorphisms. Single-nucleotide polymorphisms (SNPs) represent the most frequent type of polymorphisms in human population and may be useful in association studies, as they may actually be functionally relevant, and/or might be in linkage disequilibrium with other such variants, which may have any effect. The number of discovered SNPs has increased tremendously over the past few years. The SNPs are present in different parts of human genome; variations in coding region together with changes in regulatory regions are believed to have the highest impact on phenotype.

Public SNP databases (dbSNPs) are a highly valuable resource of information about polymorphisms in the candidate genes. At present, several dbSNPs exist in public domain, their SNP content significantly overlaps but also complements.[4] The dbSNP of National Center for Biotechnology Information is one of the central repositories for newly discovered genomic and cDNA sequence variations, both single base changes and short deletions and insertions.[5] In this dbSNP, almost six million unique SNPs had been deposited as of November 2003 (dbSNP build 117). The quality of database entries was evaluated in several studies employing positive predictive value (ie the probability that a putative SNP entry in a database is indeed a true polymorphism for a given population, with frequency of the rare allele higher than 1%) and sensitivity (ie the probability that all existing SNPs are deposited in the database).[6–9] The above studies analyzed samples of mixed ethnic origin,[6–9] and to our knowledge, the role of ethnicity on predictive value of SNPs databases has been evaluated in only a few studies.[10–12] It is also important to note that the above-mentioned reports examined genes that were otherwise not a subject of intense research in clinical samples, which may have caused a rather low sensitivity of dbSNP in one of these studies.[6]

The aims of our study were (a) to collect all available information on SNPs in 24 genes relating to homocysteine metabolism (either directly in the methionine cycle or indirectly in metabolism of vitamins) and (b) to assess the applicability of database entries to a typical Caucasian population from Central Europe. The applicability of database was evaluated for a subset of 42 putative SNPs in seven genes of folate and homocysteine metabolism by calculating the positive predictive value after determining the population frequency by PCR-RFLP or ARMS-PCR in at least 100 control Czech chromosomes.

Methods

SNP data mining from database

The SNPs in 24 genes relating to homocysteine metabolism were searched at the NCBI web page as of January 2003 (build 110); detailed information on the analyzed genes is given in Table 1. The *in silico* search was based on gene name or symbol, the candidate SNPs were manually localized to 5′UTR, introns, exons and 3′UTR of the particular gene using the GenBank reference sequence and recommended numbering starting with adenosine in the first ATG. The use of this numbering system led to discrepancies to some previously published SNPs (eg c.677C>T, c.1298A>C and c.1305C>T in the MTHFR gene). To collect the recent data on individual SNPs for Table 2, we updated the frequency using build 117 (November 2003) of the NCBI dbSNP.

Literature searches

To collect recent data on SNPs and their frequencies, we also explored the literature, using Medline searches with specific gene names to identify the relevant studies published as of November 2003. In addition, data from conference proceedings were used for completing the list of known polymorphisms.

Genotyping and determination of frequency in the Czech population

To evaluate the positive predictive value of dbSNP, we selected all 42 SNPs available in the build 110 of dbSNP (as of January 2003), which were localized in the coding regions of seven genes relating directly to homocysteine metabolism. The frequency of additional SNPs rising from dbSNP build 117 (as of November 2003) was not determined. The frequency of selected 42 cSNPs was estimated experimentally in the Czech population using PCR-RFLP or ARMS-PCR with allele-specific primer pairs (see Table I in web supplement). Quality control of each batch of samples was ensured by (i) the presence of an additional internal restriction site, (ii) complete cleavage of wild-type PCR product for SNP that destroys a naturally occurring restriction site, (iii) including samples with known genotype or (iv) using a different PCR product containing restriction site as an external control (for details see Table I in web supplement). Samples of genomic DNA from healthy controls aged between 18 and 65 years from a homogenous Caucasian population in the Czech Republic have been employed;[28] at least 110 alleles (range 110–1194 alleles, median 300 alleles) were examined for the presence of each variant. Frequency of SNP was determined by counting the number of chromosomes carrying and lacking the variant.

Positive predictive value of database subset for Czech population

Positive predictive value was calculated in a subset of 42 SNPs as a ratio of the number of true polymorphisms (with frequency of the rare allele higher than 1%) to the total number of the putative SNPs that were found by *in silico* searches. Correlation was calculated using Prophet 5.0 software (BBN Systems and Technologies).

Table 1 Genes included in this study

					Coding sequence		Genomic sequence	
Gene	*OMIM*	*Symbol*	*EC number*	*Localization*	*GenBank #*	*Length (bp)*	*GenBank #*	*Length (bp)*
S-adenosylhomocysteine hydrolase	1800960	*AHCY*	3.3.1.1	20cen-q13.1	NM_000687	1299	NT_028392	23 116
Betaine-homocysteine methyltransferase	602888	*BHMT*	2.1.1.5	5q13.1–15	NM_001713	1221	NT_006713	20 425
Cystathionine beta-synthase	236200	*CBS*	4.2.1.22	21q22.3	NM_000071	1656	NT_030188	23 170
Cystathionine gamma-lyase	607657	*CTH*	4.4.1.1	1p31.1	NM_001902	1218	NT_004464	28 301
Folate hydrolase 1 (glutamate carboxypeptidase II)	600934	*FOLH1*	3.4.17.21	11p11.2	NM_004476	2253	NT_033232	62 034
Folate receptor – adult	136430	*FOLR1*	—	11q13.3–14.1	NM_016725	774	NT_033927	32 383
Folate receptor – fetal	136425	*FOLR2*	—	11q13.3–q13.5	NM_000803	768	NT_033927	5171
Folate receptor – gamma	602469	*FOLR3*	—	11q13	NM_000804	732	NT_033927	4164
Glutamate-cysteine ligase	606857	*GCLC*	6.3.2.2	6p12	NM_001498	1914	NT_007592	46 987
Gastric intrinsic factor (cobalamin binding protein)	261000	*GIF*	—	11q13	NM_005142	1254	NT_033903	16 225
Glycine *N*-methyltransferase	606628	*GNMT*	2.1.1.20	6p12	NM_018960	888	NT_007592	3114
Methionine adenosyltransferase	250850	*MAT1A*	2.5.1.6	10q22	NM_000429	1188	NT_033890	18 137
Mitochondrial folate transporter/carrier	N/A	*MFTC*	—	8q22.3	NM_030780	948	NT_008046	16 618
5,10-methylenetetrahydrofolate dehydrogenase	172460	*MTHFD1*	1.5.1.5	14q24	NM_005956	2808	NT_026437	71 723
5,10-methenyltetrahydrofolate cyclohydrolase			3.5.4.9					
10-formyltetrahydrofolate synthetase			6.3.4.3					
5,10-methylenetetrahydrofolate reductase	607093	*MTHFR*	1.5.1.20	1p36.3	NM_005957	1971	NT_004488	12 708
Methionine synthase	156570	*MTR*	2.1.1.13	1q43	NM_000254	3798	NT_004836	105 308
Methionine synthase reductase	602568	*MTRR*	2.1.1.135	5p15.3–15.2	NM_002454	2097	NT_006576	32 017
Pyridoxal kinase	179020	*PDXK*	2.7.1.35	21q22.3	NM_003681	939	NT_011515	37 161
Plasma glutamate carboxypeptidase	N/A	*PGCP*	3.4.17.21	8q22.2	NM_016134	1419	NT_008046	498 224
Folate transporter (reduced folate carrier, *RFC*)	600424	*SLC19A1*	—	21q22.3	NM_003056	1776	NT_011515	28 733
Serine hydroxymethyltransferase 1 (cytoplasmic)	182144	*SHMT1*	2.1.2.1	17p11.2	NM_004169	1452	NT_030843	48 537
Serine hydroxymethyltransferase 2 (mitochondrial)	138450	*SHMT2*	2.1.2.1	12q12–q14	NM_005412	1515	NT_029419	28 628
Transcobalamin I	189905	*TCN1*	—	11q11–q12	NM_001062	1302	NT_033903	13 765
Transcobalamin II	275350	*TCN2*	—	22q12.2	NM_000355	1284	NT_011520	19 887

N/A, not available.

Table 2 Summary of all identified SNPs in coding regions

						Frequency of rare allele (# of tested alleles)			
						Mixed/non-Caucasians	Caucasians		
Gene symbol	*SNP subset*	*Nucleotide change*[a]	*Amino-acid change*	*NCBI rs # or reference*	*SNP frequency source*	*NCBI*	*NCBI*	*Published*	*Czech*
AHCY		c.954g>A	K318K	6088457					
BHMT		c.595g>A	G199S	Heil *et al*[13]	Heil *et al*[13]			0.01 (1292)	
	b	c.656T>g	F219C	672347					0 (306)
	b	c.657T>g	F219L	672346					0 (304)
	b	c.716g>A	Q239R	3733890	NCBI, Heil *et al*[13]	0.231 (1502)		0.32 (1382)	0.23 (128)
		c.792C>T	L264L	4703772					
	b	c.1114g>T	G372C	1050825					0 (300)
		c.1218g>T	Q406H	Heil *et al*[13]	Heil *et al*[13]			<0.01 (1582)	
CBS	b	c.209C>T	P70L	2229413		0.016 (64)	0 (28)		0 (310)
	b	c.636C>T	N212N	2298758		0.031 (1090)			0 (302)
	b	c.699C>T	Y233Y	234706	NCBI, Lievers *et al*[14]	0.29 (664)	0.42 (62)	0.35 (728)	0.32 (400)
	b	c.939G>A	T313T	2228298	NCBI	0.013 (72)	0 (30)		0 (314)
	b	c.1080T>C	A360A	1801181	NCBI, Lievers *et al*[14]	0.29 (408)	0.37 (62)	0.37 (742)	0.42 (400)
CTH	b	c.1208g>T	S403I	1021737	Wang and Hegele[15]			0.33 (120)	0.31 (1178)
FOLH1		c.223T>C	Y75H	202676	NCBI	0.360 (72)	0.17 (24)		
		c.333A>T	A111A	202680					
		c.395A>G	N132S	7128652					
		c.616g>A	G206R	2851529					
		c.732T>C	D244D	182169	NCBI	0.35 (1728)			
		c.976C>T	P326S	2851557					
		c.1059A>g	T353T	202716					
		c.1423C>T	H475Y	Devlin *et al*[16]	Devlin *et al*[16]			0.04 (150)	0.05 (1190)
		c.1838g>A	S613N	2988341					
		c.1879g>T	V627L	2988342					
		c.2181A>T	E727D	1803128					
		c.2198A>C	Y733S	1803127					
FOLR1		c.82T>C	W28R	7928649					
		c.480g>C	W160C	1801932					0 (112)
FOLR2		c.103g>A	E35K	13908					
		c.419T>g	F140C	1803569					
		c.660T>g	A220A	1803567					
FOLR3		c.76C>A	R26R	1802609					
		c.530C>T	A177V	2229185					
		c.550C>T	R184C	2229186					
		c.574C>T	P192S	637609					
		c.577T>C	F193L	1802608					
GCLC		c.164T>C	L55S	2066512	NCBI	0.012 (84)			
		c.234G>T	L78L	2066508	NCBI	0.032 (94)			
		c.1563C>T	D521D	2066509	NCBI	0.01 (96)			

Table 2 Continued

						Frequency of rare allele (# of tested alleles)			
						Mixed/non-Caucasians		*Caucasians*	
Gene symbol	*SNP subset*	*Nucleotide change*[a]	*Amino-acid change*	*NCBI rs # or reference*	*SNP frequency source*	*NCBI*	*NCBI*	*Published*	*Czech*
GIF		c.990g>A	N330N	2867802					
MAT1A		c.357G>T	Q119H	1143693	NCBI	0.324 (34)			
		c.426C>T	A142A	1143694	NCBI	0.10 (32)			
		c.1131C>T	Y377Y	2993763	NCBI	0.39 (1496)			
MTHFD1	b	c.401g>A	R134K	1950902	NCBI, Brody *et al*[17]	0.219 (1494)		0.18 (6062)	0.19 (120)
	b	c.485C>T	P162L	4902283					0 (314)
	b	c.1958g>A	R653Q	2236225	Brody *et al*[17]			0.45 (6100)	0.44 (110)
	b	c.2282C>T	T761M	10813	Brody *et al*[17]			0 (260)	0 (314)
	b	c.2380g>T	G794C	1803951	Brody *et al*[17]			0 (260)	0 (304)
	b	c.2777C>T	P926L	1803950	Brody *et al*[17]			0 (260)	0 (312)
MTHFR	b	c.117T>C	P39P	2066470	NCBI	0.118 (68)			0.08 (208)
	b	c.203g>A	R68Q	2066472	NCBI	0.015 (66)			0 (300)
	b	c.345C>A	T115T	2066461	NCBI	0.013 (76)			0 (230)
	b	c.417g>A	T139T	2066466	NCBI	0.026 (78)			0 (310)
	b	c.665C>T	A222V	1801133	NCBI, Kahleova *et al*[18]	0.40 (1484)		0.30 (346)	0.34 (1194)
		c.945g>A	V315V	6664734					
	b	c.1056C>T	S352S	2066462	NCBI	0.024 (82)			0.08 (216)
		c.1269g>T	E423D	3927589					
	b	c.1286A>C	E429A	1801131	van der Put *et al*[19]			0.33 (806)	0.33 (1194)
		c.1293C>T	F435F	4846051	van der Put and Blom[20]			0.003 (900)	
	b	c.1697g>A	G566E	2274974					0 (322)
	b	c.1781g>A	R594Q	2274976	Rady *et al*[21]	0.096 (1494)		0.07 (318)	0.06 (108)
MTR	b	c.764A>G	Y255C	1140598					0 (302)
	b	c.940G>A	D314N	2229274	NCBI	0.026 (38)			0.06 (110)
	b	c.1485G>A	M495I	2229275	NCBI	0.026 (38)			0 (304)
	b	c.2756A>G	D919G	1805087	NCBI, Kahleova *et al*[18]	0.19 (1494)	0.20 (84)	0.22 (346)	0.19 (1194)
	b	c.3144A>G	A1048A	2229276	NCBI	0.48 (1280)			0.40 (112)
	b	c.3576C>T	L1192L	1131449	NCBI	0.40 (50)			0.44(C) (198)
MTRR	b	c.54C>T	I18I	6413426	NCBI	0.005 (816)			0 (304)
	b	c.66A>G	I22M	1801394	NCBI, Gaughan *et al*[22]	0.355 (1558)	0.50 (62)	0.44 (1202)	0.41(A) (1194)
		c.481A>g	N161D	7728621					
	b	c.524C>T	S175L	1532268	NCBI, Kahleova *et al*[18]			0.43 (346)	0.37 (1194)
	b	c.537T>C	L179L	161870	NCBI	0.40 (56)	0.29 (14)		0.14 (200)
	b	c.769T>A	S257T	2303080	NCBI	0.120 (728)			0.10 (194)
	b	c.1049A>G	K350R	162036					0.14 (200)
	b	c.1155A>G	L385L	2287779	NCBI	0.174 (1498)			0.05 (200)
	b	c.1243C>T	R415C	2287780	NCBI	0.172 (1482)			0.05 (200)
		c.1464A>G	V488V						0.04 (200)
		c.1536C>T	S512S						0.04 (200)
		c.1653G>A	P551P						0.02 (200)
		c.1761T>C	Y587Y	6874544					0.04 (200)
	b	c.1783C>T	H595Y	10380	NCBI	0.12 (1486)	0.20 (188)		0.11 (200)

Table 2 Continued

						Frequency of rare allele (# of tested alleles)			
						Mixed/non-Caucasians	*Caucasians*		
Gene symbol	*SNP subset*	*Nucleotide change*[a]	*Amino-acid change*	*NCBI rs # or reference*	*SNP frequency source*	*NCBI*	*NCBI*	*Published*	*Czech*
	b	c.1875G>A	V625V	12347	NCBI	0.115 (414)	0.19 (168)		0.14 (200)
	b	c.1911G>A	A637A	1802059					0.31 (200)
PDXK		c.639C>T	S213S	8127335					
		c.799C>T	L267L	1129461					0 (110)
		c.780 g>C	V260V	762399	NCBI	0 (72)	0 (24)		0 (114)
		c.782C>T	S261F	1140133					0 (114)
RFC		c.80g>A	R27H	1051266	NCBI, Chango *et al*[23]	0.13 (54)		0.47 (338)	0.446 (1182)
		c.246g>C	P82P	1051269					
		c.696T>C	P232P	12659	NCBI	0.16 (58)	0.453(T) (188)		
		c.1406C>T	A469V	7278825					
SHMT1		c.1018g>C	E340Q	7215148					
		c.1181g>A	S349N	Heil *et al*[24]					
		c.1420C>T	L474F	1979277	NCBI, Heil *et al*[24]	0.25 (400)	0.32 (60)	0.32 (1298)	
SHMT2		c.798g>A	S266S	2229715	NCBI	0.038 (26)			
		c.813g>A	A271A	2229716	NCBI	0.077 (26)			
		c.850C>T	R284W	Heil *et al*[24]					
		c.906T>g	silent	Heil *et al*[24]					
		c.969g>T	L323L	2229717	NCBI	0 (26)			
		c.1356C>T	V452V	2229718	NCBI	0.037 (54)			
		c.1464T>g	R488R	14201					
TCNI		c.664A>g	K222E	1062607	NCBI	0 (414)	0 (62)		
		del 694A	K232?	4987226	NCBI	0.005 (404)	0 (62)		
		c.719A>g	N240S	4987227	NCBI	0.005 (404)	0 (62)		
		c.846C>T	S282S	1042613					
TCNII		c.67A>C	I23V	Li *et al*[25]	Lievers *et al*[26]			0.13 (1582)	
		c.280g>A	G94S	Lievers *et al*[26]	Lievers *et al2*[26]			0.008 (1582)	
		c.701A>g	Q234R	Li *et al*[27]	Lievers *et al*[26]			0 (1582)	
		c.776g>C	R259P	1801198	NCBI, Lievers *et al*[26]	0.453 (1478)		0.47 (1582)	
		c.1043C>T	S348F	Lievers *et al*[26]	Lievers *et al*[26]			0.11 (1582)	
		c.1127T>C	L376S	3178000					
		c.1196g>A	R399Q	4820889	Lievers *et al*[26]			0.002 (1582)	

(i) Original entries of dbSNPs build 110 (January 2003), which were updated from dbSNP build 117 (November 2003); (ii) newly observed SNPs from our laboratory; and (iii) published entries, which were not deposited in the database. The frequencies of rare alleles are presented for variant allele in most cases. If the wild-type allele is less frequent than the variant allele, the frequency of rare allele is marked with appropriate nucleotide within parentheses.

[a] All SNPs were numbered using the GenBank reference sequence and starting with adenosine in the first ATG. The use of our numbering system led to discrepancies to some previously published SNPs (eg MTHFR 665C>T, 1286A>C and 1293C>T correspond to the usual description 677C>T, 1298A>C and 1305C>T).

[b] SNPs available in dbSNP in build 110 (February 2003), which were validated experimentally.

Results

In this study, we collected information on SNPs in 24 genes relating directly or indirectly to homocysteine metabolism. First, by *in silico* analysis, we scanned almost 1200 kbp of sequence in the NCBI database (build 110) and we detected 1353 putative SNP DNA variations, of which 85 were contained in the coding regions. The SNP density varied considerably for individual genes reaching a median of 1:683 for the genic regions and 1:412 for the coding regions (for details see Table II in web supplement). The median SNP densities in genes relevant to homocysteine metabolism are similar to the published estimates of 1:567 for the entire genome (dbSNP Summary build 117, as of November 2003).

As other researchers may utilize in their genetic studies data on polymorphisms in genes relating to homocysteine metabolism, we collected data on additional cSNPs, which were not subject of the below described experimental validation, and we also updated cSNPs frequencies from all available sources as of November 2003 (including literature and dbSNP build 117). Table 2 shows data on 112 putative or confirmed cSNPs, experimentally determined frequency of the rare allele was available for 47 and 67 entries employing non-Caucasian/mixed and Caucasian samples, respectively.

To evaluate the applicability of the NCBI database to the Czech population, we selected a subset of 42 putative dbSNP entries in seven genes of folate and homocysteine metabolism for experimental validation (see Figure 1). As the first step in assessing the positive predictive value of this NCBI database subset for our population, we determined the frequency of all 42 cSNPs in at least 100 Czech control chromosomes using PCR-RFLP or ARMS-PCR (frequencies are given in Table 2). We than examined whether each of the putative database SNP entries meets the definition criterion, that is, frequency of the rare allele at a locus higher than 1%. As only 25 variants out of 42 putative cSNPs met the definition criteria while the remaining 17 variants were false positives, the positive predictive value of this NCBI SNP subset for the studied Czech population is 60%. Consequently, the median density of experimentally validated cSNPs (ie 1:950) is about half of that predicted from *in silico* searches (ie 1:412, for details see Table II in web supplement), which corresponds well to other studies.[29] Interestingly, the false-positive cSNP entries were either rare variants in the NCBI database (eight entries with frequency <3% in mixed samples) or the frequency was not available in the dbSNP (nine entries). These data suggest that dbSNP entries with low or missing frequency are more likely to be false positives in Caucasians.

It is possible that the failure of NCBI database in predicting some cSNPs in the Czech population may be a consequence of largely different SNP frequencies in samples used to create the NCBI entries. To test this hypothesis, we examined the role of ethnicity on frequency estimates. Of the 42 analyzed cSNPs, the NCBI database contained frequency information in non-Caucasian/mixed populations for 27 entries and in Caucasians for nine entries. In addition, literature contained frequency

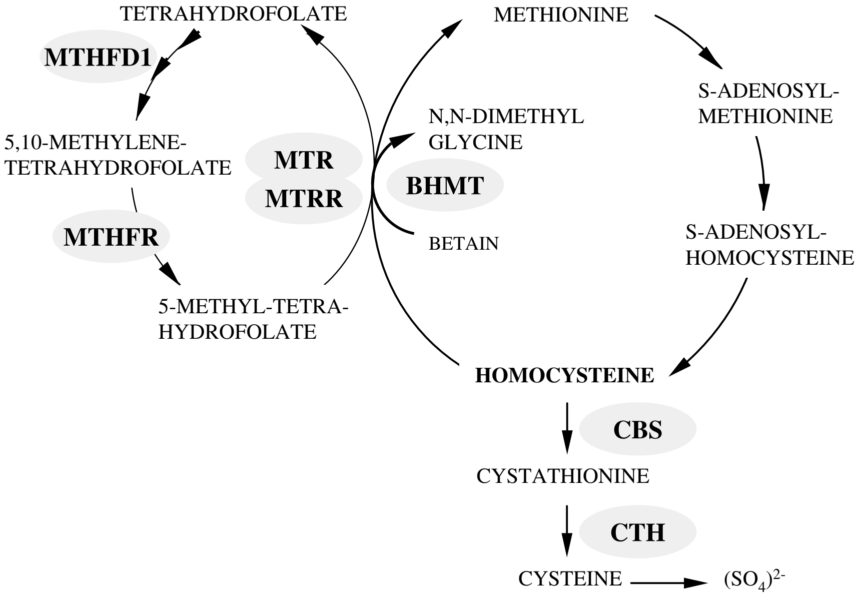

Figure 1 Homocysteine metabolism. Selected genes relating to homocysteine metabolism, which were selected for the experimental validation of cSNPs in this study, are shown in shaded ellipses (for abbreviations of gene names see Table 1).

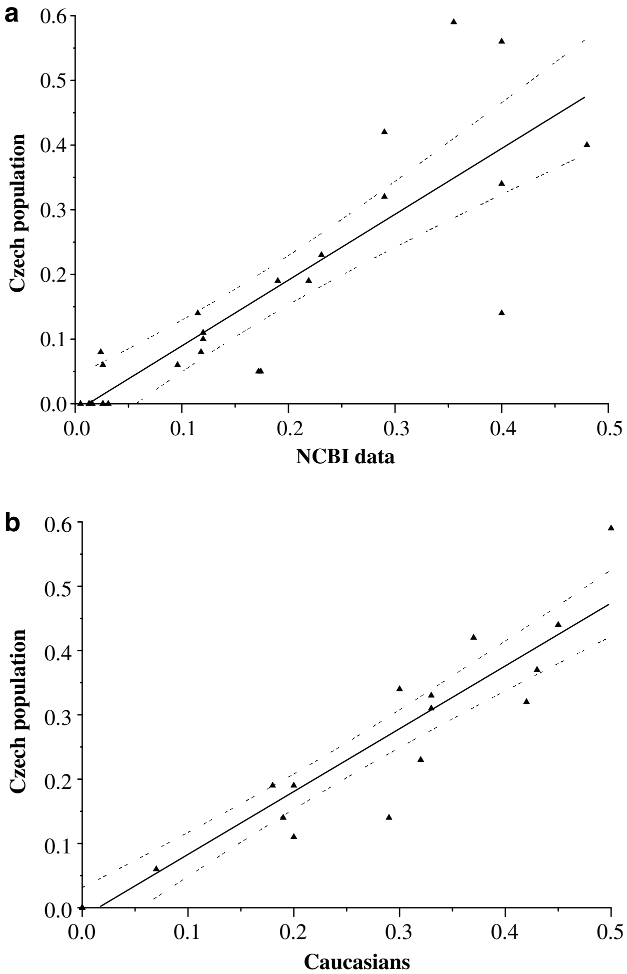

Figure 2 (**a**) Correlation of frequencies determined in the Czech population with frequencies found in NCBI database regardless of ethnicity. (**b**) Correlation of frequencies determined in the Czech population with frequencies among Caucasians found in NCBI database or literature. Dashed curves define the 95% confidence intervals of the regression lines ($f(x) = 1.018x - 0.01228$ and $f(x) = 0.9773x - 0.01479$ for (**a**) and (**b**), respectively).

data on 15 cSNPs for several European populations. The correlation of SNP frequencies between Czechs and other Caucasians ($r^2 = 0.90$, $P = 0.0001$, Figure 2b) was substantially stronger than between Czech controls and the general NCBI data set (see Figure 2a, $r^2 = 0.73$, $P = 0.0001$). Moreover, frequency of all 20 putative cSNPs, for which data in Caucasians were available, were congruently below or above the 1% frequency threshold both in the Czech population and in other Caucasians. In summary, these data suggest that for genes relating to homocysteine metabolism the cSNPs validated in one Caucasian population may be truly polymorphic in other Caucasians.

Discussion

To assess the applicability of dbSNPs in the public domain to one of European populations, we collected and evaluated allele frequency data for 42 variant alleles relating to homocysteine metabolism. The positive predictive value of the NCBI data set for a typical Caucasian population was 60%, which is intermediate between the study of Cox *et al*[6] and Reich *et al.*[9] Cox *et al* have found that 55% of *in silico* detected polymorphisms in coding sequence were indeed found by experimental method, while Reich *et al* confirmed in independent resequencing over 88% of SNPs that were available in three different public and commercial databases. In summary, our study suggests that about two-thirds of NCBI dbSNP entries may be truly polymorphic in European populations, which corresponds very well to the conclusions of Marth *et al* 'if a researcher uses the publicly available candidate SNPs for a study in a population, there is only a 66–70% chance that the SNPs have appreciable minor allele frequency'.[7]

The applicability of dbSNPs to study genetic variants in specific populations may be obscured by the presence of false-positive entries, which may constitute about one-third of database data.[7] Two types of false positivity may exist due to errors either at the step of entry generation or by errors in validation of the SNPs in a given population sample. At the step of entry generation, the false positives may be generated by technical problems such as sequencing errors or by errors during the computational data mining procedure,[6] or by analysis of patient samples and misclassification of pathogenic mutations as SNPs.[5] When validating the frequency of putative SNPs in a population sample, false positivity may originate from insufficient methods for their detection or insufficient sample size. In our study, the genotyping errors were quite unlikely as we employed quality control. Moreover, we screened at least 300 alleles for SNPs appearing as monomorphic, which gave us a power of 95.1% to classify them as truly false positive. All these data strongly suggest that these putative SNPs are indeed absent in the studied Czech population sample. Another and the most likely source of false positivity may be the different ethnicity of samples, from which the respective entry was generated, and of samples in the studied population. Indeed, the comparison of cSNP frequency data between Czechs on one side and other Caucasians or non-Caucasian or mixed samples on the other side show that SNP frequency from unrelated populations are less correlated than between closely related populations. The larger distance between Czechs and general NCBI datapool corresponds well to the observations of others,[10] who showed that frequencies between Koreans and other Asians correlated more strongly than between Koreans and general NCBI data set.

The databases may not contain all existing SNPs, which are reflected in another characteristic of the database, namely its sensitivity. The study of Cox *et al*[6] suggested that the sensitivity of database may be quite low as he detected by *in silico* search only 27% of those polymorphisms, which were in his study discovered experimentally. Since we did not sequence the seven genes of interest using multiple control samples, we were unable to evaluate accurately the sensitivity of dbSNP. However, these genes were systematically analyzed by other researchers in numerous clinical samples obtained from patients disturbed of homocysteine metabolism. Indeed, by searching literature and by our own experimental work, we were able to find only three additional cSNPs, which were lacking in the NCBI dbSNP (they were discovered experimentally by sequencing clinical samples). In summary, it is conceivable that most of the genetic variation in the coding regions of these seven genes has been already detected owing to the systematic analysis of these genes by the community of homocysteine researchers.

In our study, we collected structural and frequency data on polymorphisms in selected genes relating to sulfur amino-acid metabolism. This set of data suggests that about two-thirds of SNPs found in database are indeed polymorphic in our population, and that majority of existing cSNPs in genes relating to homocysteine metabolism have already been deposited in the NCBI database. However, the data from our study should be interpreted with caution as the number of genes was quite small and confounding since the interest of the scientific community in these selected genes may exist. Nevertheless, our study shows that the NCBI dbSNP is a valuable tool for selecting markers for genetic studies, and that experimental validation of cSNPs should be performed, especially if frequency on the candidate polymorphism is low or lacking.

Acknowledgements

We thank Mrs Jitka Sokolová for genotyping, Mrs Eva Richterová for excellent technical assistance and Dr Joseph D Terwilliger for critical reading of an earlier version of this manuscript. This study was supported by the Grant NM 6548-3 from Grant Agency of Ministry of Health of the Czech Republic. VK is supported by the Wellcome Trust International Senior Research Fellowship in Biomedical Science in Central Europe.

Databases

http://www.ncbi.nlm.nih.gov/SNP/;http://www.ncbi.nlm.nih.gov/Genbank/.

References

1 Brattstrom L, Wilcken DE: Homocysteine and cardiovascular disease: cause or effect? *Am J Clin Nutr* 2000; **72**: 315–323.

2 Ueland PM, Refsum H, Beresford SA, Vollset SE: The controversy over homocysteine and cardiovascular risk. *Am J Clin Nutr* 2000; **72**: 324–332.

3 Carmel R, Jacobsen D: *Homocysteine in Health and Disease*. Cambridge: Cambridge University Press, 2001.

4 Aerts J, Wetzels Y, Cohen N, Aerssens J: Data mining of public SNP databases for the selection of intragenic SNPs. *Hum Mutat* 2002; **20**: 162–173.

5 Sherry ST, Ward MH, Kholodov M *et al*: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**: 308–311.

6 Cox D, Boillot C, Canzian F: Data mining: efficiency of using sequence databases for polymorphism discovery. *Hum Mutat* 2001; **17**: 141–150.

7 Marth G, Yeh R, Minton M *et al*: Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* 2001; **27**: 371–372.

8 Sachidanandam R, Weissman D, Schmidt SC *et al*: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; **409**: 928–933.

9 Reich DE, Gabriel SB, Altshuler D: Quality and completeness of SNP databases. *Nat Genet* 2003; **33**: 457–458.

10 Lee JK, Kim HT, Cho SM *et al*: Characterization of 458 single nucleotide polymorphisms of disease candidate genes in the Korean population. *J Hum Genet* 2003; **48**: 213–216.

11 Lee SG, Yoon Y, Hong S, Yoo J, Yang I, Song K: Allele frequency determination of publicly available cSNPs in the Korean population. *Genet Med* 2002; **4**: 49S–51S.

12 Lazarus R, Klimecki WT, Palmer LJ *et al*: Single-nucleotide polymorphisms in the interleukin-10 gene: differences in frequencies, linkage disequilibrium patterns, and haplotypes in three United States ethnic groups. *Genomics* 2002; **80**: 223–228.

13 Heil SG, Lievers KJ, Boers GH *et al*: Betaine-homocysteine methyltransferase (BHMT): genomic sequencing and relevance to hyperhomocysteinemia and vascular disease in humans. *Mol Genet Metab* 2000; **71**: 511–519.

14 Lievers KJ, Kluijtmans LA, Heil SG *et al*: Cystathionine beta-synthase polymorphisms and hyperhomocysteinaemia: an association study. *Eur J Hum Genet* 2003; **11**: 23–29.

15 Wang J, Hegele RA: Genomic basis of cystathioninuria (MIM 219500) revealed by multiple mutations in cystathionine gamma-lyase (CTH). *Hum Genet* 2003; **112**: 404–408.

16 Devlin AM, Ling EH, Peerson JM *et al*: Glutamate carboxypeptidase II: a polymorphism associated with lower levels of serum folate and hyperhomocysteinemia. *Hum Mol Genet* 2000; **9**: 2837–2844.

17 Brody LC, Conley M, Cox C *et al*: A polymorphism, R653Q, in the trifunctional enzyme methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase/formyltetrahydrofolate synthetase is a maternal genetic risk factor for neural tube defects: report of the Birth Defects Research Group. *Am J Hum Genet* 2002; **71**: 1207–1215.

18 Kahleova R, Palyzova D, Zvara K *et al*: Essential hypertension in adolescents: association with insulin resistance and with metabolism of homocysteine and vitamins. *Am J Hypertens* 2002; **15**: 857–864.

19 van der Put NM, Gabreels F, Stevens EM *et al*: A second common mutation in the methylenetetrahydrofolate reductase gene: an additional risk factor for neural-tube defects? *Am J Hum Genet* 1998; **62**: 1044–1051.

20 van Der Put NM, Blom HJ: Reply to Donnelly. *Am J Hum Genet* 2000; **66**: 744–745.

21 Rady PL, Szucs S, Grady J *et al*: Genetic polymorphisms of methylenetetrahydrofolate reductase (MTHFR) and methionine synthase reductase (MTRR) in ethnic populations in Texas; a report of a novel MTHFR polymorphic site, G1793A. *Am J Med Genet* 2002; **107**: 162–168.

22 Gaughan DJ, Kluijtmans LA, Barbaux S *et al*: The methionine synthase reductase (MTRR) A66G polymorphism is a novel genetic determinant of plasma homocysteine concentrations. *Atherosclerosis* 2001; **157**: 451–456.

23 Chango A, Emery-Fillon N, de Courcy GP *et al*: A polymorphism (80G->A) in the reduced folate carrier gene and its associations with folate status and homocysteinemia. *Mol Genet Metab* 2000; **70**: 310–315.

24 Heil SG, Van der Put NM, Waas ET, den Heijer M, Trijbels FJ, Blom HJ: Is mutated serine hydroxymethyltransferase (SHMT) involved in the etiology of neural tube defects? *Mol Genet Metab* 2001; **73**: 164–172.

25 Li N, Seetharam S, Seetharam B: Genomic structure of human transcobalamin II: comparison to human intrinsic factor and transcobalamin I. *Biochem Biophys Res Commun* 1995; **208**: 756–764.

26 Lievers KJ, Afman LA, Kluijtmans LA *et al*: Polymorphisms in the transcobalamin gene: association with plasma homocysteine in healthy individuals and vascular disease patients. *Clin Chem* 2002; **48**: 1383–1389.

27 Li N, Sood GK, Seetharam S, Seetharam B: Polymorphism of human transcobalamin II: substitution of proline and/or glutamine residues by arginine. *Biochim Biophys Acta* 1994; **1219**: 515–520.

28 Janosikova B, Pavlikova M, Kocmanova D *et al*: Genetic variants of homocysteine metabolizing enzymes and the risk of coronary artery disease. *Mol Genet Metab* 2003; **79**: 167–175.

29 Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E: Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* 2003; **312**: 207–213.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (www.nature.com/ejhg)