HUMAN MUTATION 20:162–173 (2002)

DATABASE

Data Mining of Public SNP Databases for the Selection of Intragenic SNPs

Jan Aerts,[1] Yves Wetzels,[1] Nadine Cohen,[2] and Jeroen Aerssens[1*]

[1]Department Pharmacogenomics, Janssen Research Foundation, Beerse, Belgium; [2]Department Pharmacogenomics, R.W. Johnson Pharmaceutical Research Institute, Raritan, New Jersey

Communicated by Pui-Yan Kwok

Different strategies to search public single nucleotide polymorphism (SNP) databases for intragenic SNPs were evaluated. First, we assembled a strategy to annotate SNPs onto candidate genes based on a BLAST search of public SNP databases (Intragenic SNP Annotation by BLAST, ISAB). Only BLAST hits that complied with stringent criteria according to 1) percentage identity (minimum 98%), 2) BLAST hit length (the hit covers at least 98% of the length of the SNP entry in the database, or the hit is longer than 250 base pairs), and 3) location in non-repetitive DNA, were considered as valid SNPs. We assessed the intragenic context and redundancy of these SNPs, and demonstrated that the SNP content of the dbSNP and HGBASE/HGVbase databases are highly complementary but also overlap significantly. Second, we assessed the validity of intragenic SNP annotation available on the dbSNP and HGVbase websites by comparison with the results of the ISAB strategy. Only a minority of all annotated SNPs was found in common between the respective public SNP database websites and the ISAB annotation strategy. A detailed analysis was performed aiming to explain this discrepancy. As a conclusion, we recommend the application of an independent strategy (such as ISAB) to annotate intragenic SNPs, complementary to the annotation provided at the dbSNP and HGVbase websites. Such an approach might be useful in the selection process of intragenic SNPs for genotyping in genetic studies. Hum Mutat 20:162–173, 2002. © **2002 Wiley-Liss, Inc.**

KEY WORDS: SNP; database; bioinformatics; pharmacogenomics; dbSNP; HGBASE; HGVbase; ISAB; computational biology

DATABASES:
http://www.ncbi.nlm.nih.gov/SNP/ (dbSNP); http://hgvbase.cgb.ki.se (HGVbase);
http://genome.ucsc.edu (Human Genome Project Working Draft); www.girinst.org (REPBASE);
http://www.imm.ki.se/CYPalleles/cyp2a6.htm (CYP2A6 allele nomenclature)

INTRODUCTION

As a result of large-scale projects aiming to discover sequence variations in the human genome, the number of publicly available single nucleotide polymorphisms (SNPs) has increased enormously over the past few years, and offers new opportunities in genetic research. A high abundance of genetic markers, in casu SNPs, facilitates association studies on complex multifactorial diseases, both based on single SNPs and haplotypes [Lander and Schork, 1994; Gray et al., 2000; Kao et al., 2000; Johnson et al., 2001]. Several public SNP databases exist, among which dbSNP and HGBASE are the largest, together comprising several million SNPs. The dbSNP database is a central repository for newly discovered genomic and cDNA sequence variations, both single base changes and short deletions and insertions, from all species [Sherry et al., 2001]. The Human Genic Bi-Allelic SEquences (HGBASE) database is gene-oriented; it supports the candidate gene association study principle and is

The Supplementary Material referred to in this article can be found at http://www.wiley.com/humanmutation/suppmat/2002/v20.html

Received 28 November 2001; accepted revised manuscript 3 May 2002.

*Correspondence to: Jeroen Aerssens, Genome Center Maastricht, University of Maastricht, Universiteitssingel 50 (postvak 16), Postbus 616, 6200 MD Maastricht, The Netherlands.
E-mail: jeroen.aerssens@pandora.be

Current address for Jan Aerts: Department of Animal Breeding and Genetics, Wageningen University and Research Centre, 6709 PG Wageningen, The Netherlands.

DOI: 10.1002/humu.10107
Published online in Wiley InterScience (www.interscience.wiley.com).

therefore a catalog of intragenic sequence variants [Brookes et al., 2000]. Recently, the HGBASE database has adopted the new name HGVbase (Human Genome Variation database). For each SNP entry, both databases comprise at least a unique accession number, the nucleotide variation, and the sequence context of the SNP. When available, additional information is provided on the chromosomal location, the gene that comprises the SNP, the effect of an SNP on the amino acid sequence of an encoded protein, the allele frequencies in different ethnic populations, and/or the methods of assay and discovery. As a consequence, these databases represent a highly valuable resource of information for the selection of SNPs to be analyzed in a genotyping facility.

Here, we report on the retrieval of SNPs from public databases by standard bioinformatics tools in 24 genes with high relevance for the expanding field of pharmacogenomics. We report on the genomic distribution of SNPs among these 24 genes based on the obtained annotation results using an in-house developed in silico method: intragenic SNP annotation by BLAST (ISAB). This method semi-automatically annotates SNPs onto genes based on extensive BLAST analysis [Altschul et al., 1997] of gene sequences against public SNP databases. For the present study, we limited our working definition of "annotation" only to the assignment of the name and/or symbol of the gene that comprises the SNP and the position of the SNP within that gene. The public SNP database websites from dbSNP and HGVbase also perform such annotation, which allows the direct retrieval of the SNPs within a candidate gene. To our knowledge, no reports have been published so far on the quality of this annotation by dbSNP or HGVbase. Yet, it is important to know whether the selection of SNPs in candidate genes to be used in a genotyping facility can be based on these annotations, or whether alternative strategies are required to obtain high quality intragenic SNPs. Many of the SNPs submitted to these databases have been identified by automated sequence data analysis, largely through multiple alignment of expressed sequence tags (ESTs), resulting in so-called in silico SNPs [Taillon-Miller et al., 1998; Buetow et al., 1999; Board et al., 2000]. The gene annotation by the SNP databases is mainly done automatically as well, which might occasionally lead to an inaccurate or wrong annotation [Brookes et al., 2000]. In this report, we assessed the validity of the intragenic SNP annotation provided on the dbSNP and HGVbase websites by comparison with the intragenic SNP annotation results based on the ISAB strategy.

MATERIAL AND METHODS

Databases

The data for this analysis was comprised of the DNA sequences of 24 genes available in the public domain. Detailed information on the genes is provided in Tables 1 and 2. Genomic and cDNA sequences were obtained from the December 12, 2000 freeze of the Human Genome Project Working Draft at UCSC. Each genomic sequence comprised a 1 kb region upstream of the 5′ end of the first exon expected to

TABLE 1. List of the 24 Genes Used in This Study*

Gene	gDNA working draft location	mRNA Genbank accession	OMIM reference
ABCC2	chr10:107186139-107255907	NM.000392	601107
AHR	chr7:16497123-16545269	NM.001621	600253
COMT	chr22:16868600-16896734	NM.000754	116790
CYP1A2	chr15:71476562-71485318	NM.000761	124060
CYP2A6	chr19:48317128-48325023	NM.000762	122720
CYP3A5	chr7:100914305-100947094	NM.000777	605325
DIA1	chr22:39531318-39560515	NM.007326	250800
FIGN	chr2:166707091-166710185	NM.018086	605295
FMO3	chr1:192780592-192808359	NM.006894	136132
GSTM4	chr1:120957118-120963406	NM.000850	138333
GSTP1	chr11:71702072-71705901	NM.000852	134660
GSTT1	chr22:21022190-21031280	NM.000853	600436
GSTT2	chr22:20945658-20950432	NM.000854	600437
GSTZ1	chr14:75768981-75780511	NM.001513	603758
MTHFR	chr1:12195380-12209036	NM.005957	236250
NAT2	chr8:19782053-19784227	NM.000015	243400
NOS2	chr17:29252826-29316071	NM.000625	163730
NOS3	chr7:157454503-157476139	NM.000603	163729
NR3C1	chr5:156499128-156534720	NM.000176	138040
RXRB	chr6:36178177-36185380	NM.021976	180246
RXRG	chr1:186726743-186771545	NM.006917	180247
STE	chr4:72685927-72705121	NM.005420	600043
UGT1A1	chr2:240466042-240480062	NM.000463	191740
UGT2B15	chr4:71768275-71793261	NM.001076	600069

*Indicating the location of the gene on the Human Genome Working Draft sequence (Dec 12, 2000 Freeze), the mRNA GenBank Accession number, and the OMIM Reference number.

TABLE 2. Total Length of Each Genomic Region and the Length of the Nonrepetitive DNA in This Region for All 24 Genes Analysed in This Study

	Length (bp)											
	Total		Promoter		5′UTR		CDS		Intron		3′UTR	
Gene	Total	Nonrepeat	Total	Nonrepeat	Total	Nonrepeat	Total	Nonrepeat	Total	Nonrepeat	Total	Nonrepeat
ABCC2	69,769	50,104	1,000	789	37	37	4,638	4,638	63,909	44,455	185	185
AHR	48,147	44,421	1,000	1,000	643	643	2,547	2,547	41,654	38,177	2,303	2,054
COMT	28,135	20,394	1,000	1,000	203	203	817	817	25,844	18,103	271	271
CYP1A2	8,757	6,997	1,000	861	64	64	1,548	1,548	4,631	4,073	1,514	451
CYP2A6	7,896	7,581	1,000	1,000	9	9	1,485	1,485	5,146	4,831	256	256
CYP3A5	32,790	25,250	1,000	1,000	87	87	1,509	1,509	30,083	22,543	111	111
DIA1	29,198	22,338	1,000	612	175	175	837	837	26,228	19,756	958	958
FIGN	3,095	3,095	1,000	1,000	162	162	1,920	1,920	0	0	13	13
FMO3	27,768	21,451	1,000	1,000	93	93	1,599	1,599	24,855	18,538	221	221
GSTM4	6,289	6,019	1,000	730	263	263	657	657	4,208	4,208	161	161
GSTP1	3,830	3,398	1,000	568	29	29	633	633	2,099	2,099	69	69
GSTT1	9,091	6,761	1,000	836	0	0	723	723	7,086	4,920	282	282
GSTT2	4,775	4,198	1,000	1,000	64	64	735	735	2,674	2,097	302	302
GSTZ1	11,531	10,680	1,000	1,000	103	103	651	651	9,376	8,525	401	401
MTHFR	13,657	11,683	1,000	1,000	12	12	1,971	1,971	10,469	8,495	205	205
NAT2	2,175	2,175	1,000	1,000	107	107	873	873	0	0	195	195
NOS2	63,246	44,247	1,000	822	194	194	3,462	3,462	58,391	39,570	199	199
NOS3	21,637	19,777	1,000	1,000	20	20	3,612	3,612	16,947	15,087	58	58
NR3C1	35,593	27,641	1,000	1,000	132	132	2,334	2,334	31,119	23,167	1,008	1,008
RXRB	7,204	6,961	1,000	1,000	179	179	1,602	1,602	4,193	3,950	230	230
RXRG	44,803	42,052	1,000	1,000	27	27	1,392	1,392	42,239	39,488	145	145
STE	19,195	17,228	1,000	1,000	106	106	885	885	17,150	15,183	54	54
UGT1A1	14,021	11,703	1,000	773	15	15	1,602	1,602	10,670	8,579	734	734
UGT2B15	24,987	15,613	1,000	637	21	21	1,593	1,593	21,908	13,037	465	325
Total	537,589	431,767	24,000	21,628	2,745	2,745	39,625	39,625	460,879	358,881	10,340	8,888

include (at least part of) the promoter, until and including the 3′ untranslated region (3′UTR). The boundaries of the intragenic regions (5′UTR, coding sequence, introns, 3′UTR) were derived from the same website. Blastable SNP databases were downloaded from public ftp sites and installed locally. These included the August 6, 2001 download of the non-redundant reference SNP data of dbSNP, version 10.0 of HGVbase, and version 6.6 of the repeat database REPBASE [Jurka, 2000].

Intragenic SNP Annotation by BLAST (ISAB)

We applied an in-house assembled intragenic SNP annotation strategy based on a BLAST search (ISAB) of the genes' genomic and cDNA sequences against local copies of the SNP databases. An overview of the ISAB strategy is presented in Figure 1. The strategy consists of four major steps that are applied for each individual gene.

In a first step, the genomic and cDNA sequences are chopped into subsequences of 1,500 base pairs (bp), with an overlap of 250 bp between flanking subsequences. In a second step, these subsequences are blasted against local copies of the downloaded SNP databases, with a maximum number of description lines set at 2,000 SNP hits per subsequence in the reported output file. The dbSNP and HGVbase databases are analyzed successively and independently. An automated in-house developed algorithm then selects the relevant SNPs from these BLAST output files. According to this algorithm, a BLAST hit is considered a valid SNP if it complies with the following criteria: 1) the actual SNP is located within the boundaries of the BLAST hit, 2) its expectation value is below 10E-12, 3) a minimum of 98% identity between SNP hit and query sequence, and 4) the hit length exceeds 250 bp, or alternatively equals a minimum of 98% of the context length of that SNP in the public SNP database (defined as relative hit length). The threshold of 98% on the relative hit length was included to ensure that the whole database SNP entry would match the query sequence. The initial chopping procedure, with overlapping fragment size set at 250 bp, necessitated dropping this requirement in case of BLAST hit lengths larger than 250 bp. In such case, the limited size of the chopped subsequence that was used as query sequence in the BLAST searches might result in a BLAST hit length that is much smaller than the context sequence in the database SNP entry. As the third step of the ISAB strategy, the outcomes of the individual BLAST analyses of each of the chopped subsequences are integrated into the original cDNA or genomic gene sequence. Subsequently, the SNPs in the cDNA sequence are further integrated with the results of the genomic sequence, resulting in one summary table with all intragenic SNPs for each individual gene. The fourth step in the annotation strategy consists of a search for all repeat regions. To this end, the genomic sequences are blasted against the REPBASE database, with an upper limit on the expectation value of 10E-3. Only SNPs located in non-repeat regions are finally retained by the ISAB strategy. A detailed description of the scripts used in the ISAB strategy is available online (see the Supplementary Material for this article, available at http://www.wiley.com/humanmutation/suppmat/2002/v20.html).

Genetic Context and Redundancy of the SNPs

Following annotation by ISAB, we assessed the genetic context and redundancy of these SNPs. As the nucleotide positions were calculated by ISAB, different SNP IDs that referred to the same physical SNP could be matched, and the genomic region (i.e., repeat, promoter, coding sequence, etc.) was defined.

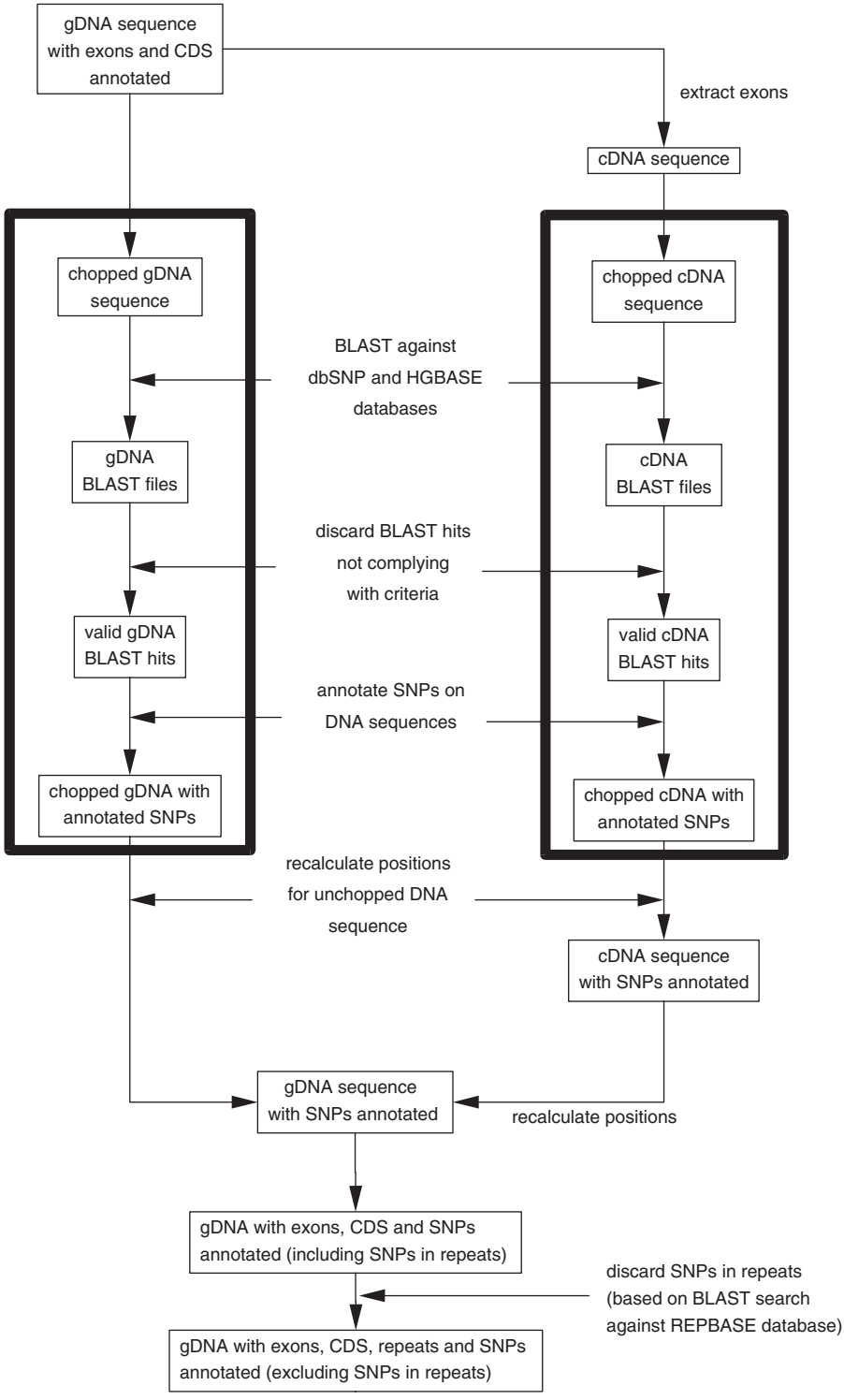

FIGURE 1. **Schematic overview of the Intragenic SNP Annotation by BLAST (ISAB) strategy. Criteria used to retain a BLAST hit: 1) the actual SNP is located in the BLAST hit, 2) a minimum of 98% identity between SNP hit and query sequence, 3) the hit length exceeds 250 bp or, alternatively, equals a minimum of 98% of the context length of that SNP in the public SNP database.**

Annotation Quality at dbSNP and HGVbase Websites

The dbSNP and HGVbase websites were searched by gene name to yield a list of SNPs annotated to that gene by these websites. All 24 genes from our list were searched on the websites on the same day as when the SNP databases were downloaded (August 6, 2001). Two analyses were performed on these SNP entries. First, the list of SNPs annotated by the websites was compared to the list of SNPs annotated by the

ISAB algorithm. Second, the flanking regions of all retrieved SNPs annotated by the websites were extracted from the downloaded databases and aligned with the chopped genomic and cDNA sequences of the corresponding gene using the BL2SEQ algorithm (part of the BLAST package).

RESULTS

The first part of our analysis verified the sequence context and redundancy of the high quality SNPs that were selected based on the ISAB annotation. Figure 2 shows the number of SNPs identified in the set of 24 genes stratified by intragenic region (promoter, 5′UTR, coding sequence, intron, and 3′UTR). The individual data for each of the 24 genes are presented in Table 3. The large majority (91%) of the initially selected SNPs from HGVbase was annotated on (mostly intronic) repeat regions. For dbSNP, this fraction was only 22% of the selected SNPs. The ISAB algorithm automatically discarded SNPs that were annotated in repeat regions for all further analyses. Overall, the number and distribution of intragenic SNPs in non-repeat regions is similar in dbSNP and HGVbase, regardless of the intragenic region. A total of 377 and 327 SNPs were identified in non-repeat regions for dbSNP and HGVbase, respectively. A merger of identified SNPs from both databases resulted in a total of 471 unique SNPs residing in non-repeat regions. Of those, 86 (18%) were located in the coding sequence of the genes. Forty-eight of these SNPs (56%) in the coding sequence caused non-synonymous changes (n = 46) or stop codons (n = 2) in the amino acid sequence of the encoded protein. Based on the combined analysis of dbSNP and HGVbase, we found overall SNP densities of one SNP/901 bp in the promoter, one SNP/457 bp in the 5′UTR, one SNP/460 bp in the coding sequence, one SNP/1078 bp in the introns, and one SNP/404 bp in the 3′UTR in the non-repeat regions.

Figure 3 shows the degree of redundancy according to database source. Redundancy is defined as the number of SNP IDs annotated at the same physical location. The 377 unique SNP positions identified by dbSNP were covered by 387 different SNP IDs. For HGVbase, 327 unique SNP positions were covered by 437 SNP IDs. Despite a large overlap of SNPs present in both dbSNP and HGVbase, a non-negligible number of SNPs was found in only one of them. More specifically, 204 SNPs (43%) were covered only once by one of the databases, 224 (48%) were covered once by both databases, and 43 SNPs (9%) had redundant annotations within at least one of the databases. We assessed the added value of both the

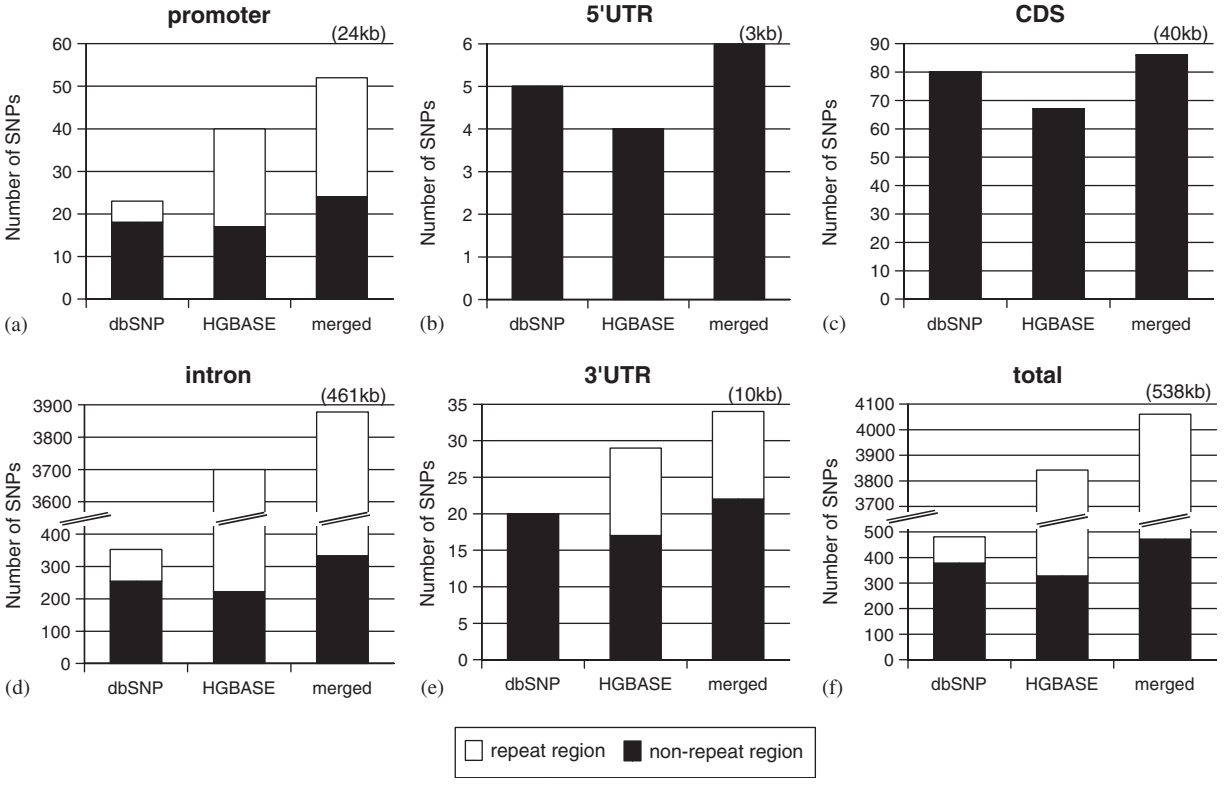

FIGURE 2. **Distribution of SNPs annotated to specific genes by the ISAB strategy stratified by intragenic region. Data are shown for SNPs obtained from dbSNP and HGVbase, and from a merged analysis. Each bar is divided into the share annotated in repeat regions (upper, white) and the share annotated in non-repeat regions (lower, black).**

TABLE 3. Overview of Number of SNPs Per Gene as Annotated by the ISAB Algorithm, Grouped by Database Source and Genomic Region

	dbSNP						HGBASE						dbSNP and HGBASE merged					
Gene	Total	Promoter	5′UTR	CDS	Intron	3′UTR	Total	Promoter	5′UTR	CDS	Intron	3′UTR	Total	Promoter	5′UTR	CDS	Intron	3′UTR
ABCC2	10	0	1	1	8	0	34	0	1	1	32	0	41	0	1	1	39	0
AHR	5	0	0	1	3	1	3	0	0	0	1	2	8	0	0	1	4	3
COMT	44	1	1	8	31	3	32	0	1	9	20	2	45	1	1	9	31	3
CYP1A2	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
CYP2A6	9	0	0	5	2	2	1	0	0	1	0	0	9	0	0	5	2	2
CYP3A5	4	0	0	0	3	1	2	0	0	0	1	1	5	0	0	0	4	1
DIA1	28	1	0	1	23	3	33	0	0	1	29	3	35	1	0	1	30	3
FIGN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FMO3	25	0	0	10	15	0	15	0	0	8	7	0	26	0	0	10	16	0
GSTM4	15	2	0	3	9	1	12	0	0	3	9	0	18	2	0	4	11	1
GSTP1	7	0	0	3	3	1	9	2	0	4	2	1	11	2	0	4	4	1
GSTT1	13	0	0	4	8	1	14	0	0	6	7	1	17	0	0	6	10	1
GSTT2	20	4	0	5	10	1	18	6	0	3	8	1	23	7	0	5	10	1
GSTZ1	16	0	3	6	7	0	9	0	2	4	3	0	21	0	4	7	10	0
MTHFR	19	1	0	7	11	0	18	0	0	7	11	0	19	1	0	7	11	0
NAT2	11	2	0	8	0	1	11	2	0	8	0	1	11	2	0	8	0	1
NOS2	7	0	0	3	4	0	1	0	0	0	1	0	8	0	0	3	5	0
NOS3	19	1	0	3	15	0	38	2	0	2	34	0	47	2	0	3	42	0
NR3C1	18	0	0	6	10	2	8	0	0	5	1	2	18	0	0	6	10	2
RXRB	3	1	0	2	0	0	4	1	0	2	1	0	4	1	0	2	1	0
RXRG	67	1	0	2	64	0	55	1	0	2	52	0	67	1	0	2	64	0
STE	12	0	0	0	12	0	1	0	0	0	1	0	13	0	0	0	13	0
UGT1A1	9	3	0	1	2	3	9	3	0	1	2	3	9	3	0	1	2	3
UGT2B15	15	1	0	1	13	0	0	0	0	0	0	0	15	1	0	1	13	0
Total	377	18	5	80	254	20	327	17	4	67	222	17	471	24	6	86	333	22

The right part displays the number of SNPs after merging of the data from dbSNP and HGBASE. These numbers do not include SNPs located in repeat regions.

genomic or cDNA sequence, focusing on the intragenic regions comprised in the cDNA (5′UTR, CDS and 3′UTR) and determined how many of the 105 (5+80+20) dbSNP and 88 (4+67+17) HGVbase SNPs could be picked up using either the genomic or cDNA sequence only. We found that application of the ISAB strategy on the genomic DNA alone yielded 90% (95 out of 105) of the dbSNP and 97% (85 out of 88) of the HGVbase SNPs that were identified using both the genomic and cDNA sequences. If only the cDNA sequences were used, 65% (68 out of 105) and 84% (74 out of 88) of all SNPs identified in this study were found for dbSNP and HGVbase, respectively. Our analysis therefore suggested that the genomic DNA is the most valuable source sequence to search with the ISAB strategy for intragenic SNPs in dbSNP and HGVbase.

In order to assess the validity of the SNPs picked up by the ISAB strategy, we compared our results for some genes with wet laboratory SNP screening studies reported in the literature. Cauchi et al. [2001] performed a polymorphism screening of the *AHR* gene in 30 individuals, with a focus on the promoter and exons. They identified three SNPs, one of which was also annotated by the ISAB strategy. The ISAB strategy found three additional SNPs in these intragenic regions. A review by Raunio et al. [2001] reported on the polymorphisms found in the *CYP2A6* gene, and refers to the *CYP2A6*-specific website for a full list of polymorphisms. The website (updated June 11, 2001) listed six SNPs in the *CYP2A6* gene, one of which was also annotated by the ISAB strategy. Eight SNPs annotated by ISAB were not listed on the website. Two independent studies, performed by van der Put et al. [1998] and Weisberg et al. [1998], respectively, document three polymorphic sites in the coding region of the *MTHFR* gene. Two of these polymorphisms were also found by the ISAB strategy. Five polymorphisms annotated by ISAB in the coding region were not reported in these articles. Importantly, none of the SNPs reported in these publications but not detected by ISAB were present in the dbSNP or HGVbase databases.

In the second part of our analysis, the quality of the annotation performed by the public SNP databases was evaluated. In this context, annotation of an SNP is defined as its genomic location within a gene. Therefore, we assessed the similarity between the annotation by the SNP database websites and the ISAB strategy. Figure 4 summarizes the results of intragenic SNP selection performed either by searching the database websites by gene name or by applying the ISAB algorithm. In contrast to Table 3, the data in Figure 4 indicate the number of SNP IDs rather than the number of unique SNP positions, because the degree of redundancy of the SNPs annotated only by the SNP database websites could not be evaluated. Furthermore, the data based on the ISAB strategy include only SNPs that are not located in repeat regions; this could not be verified for the SNPs

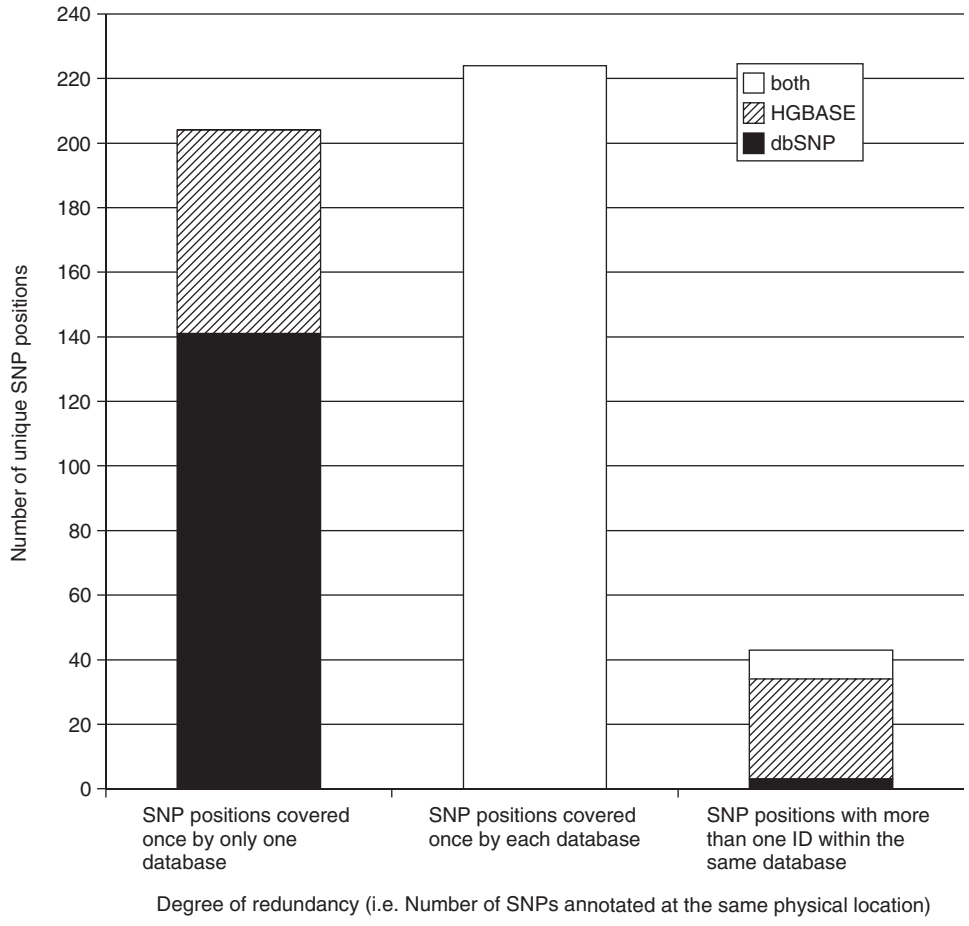

FIGURE 3. **Number of physical SNP positions, stratified by degree of redundancy of the annotation. Redundancy is defined as the number of independent SNP IDs annotated onto a unique physical location. The figure also shows the share of SNP positions with only SNP IDs from dbSNP (black), the share with only SNP IDs from HGVbase (shaded), and the share with SNP IDs from both databases (white).**

annotated by the website only. Respectively 50% (306 out of 607) (dbSNP) and 11% (54 out of 477) (HGVbase) of all annotated SNPs were in common between the public SNP database website annotation and the ISAB annotation strategy. A total of 220 (36%) of all the intragenic SNPs in dbSNP were annotated exclusively by the website, because the ISAB criteria for valid SNPs were not fulfilled or because of the synchronization delay of the database version between the download (ftp) versus the website (see below). Of all 387 (306+81) dbSNP hits annotated by ISAB, only 81 SNPs (21%) were not annotated by the dbSNP website. For HGVbase, 40 SNPs annotated by the website did not comply with the ISAB criteria. Of all 437 (383+54) HGVbase hits annotated to the specified genes, 383 SNPs (80%) were annotated exclusively by the ISAB algorithm.

Remarkably, as low as 58% (306 out of (220+306)) and 57% (54 out of (40+54)) of all the intragenic SNPs annotated by the dbSNP and HGVbase websites, respectively, fulfilled the ISAB criteria for a valid SNP. We therefore assessed why the intragenic SNPs annotated exclusively by dbSNP and HGVbase websites did not comply with the criteria of our ISAB strategy. For each SNP annotated by the websites, the flanking regions were extracted from the downloaded databases and were aligned with the genomic and cDNA sequences of the corresponding genes using the BL2SEQ algorithm.

Of the 526 SNPs annotated by the dbSNP website, 61 (i.e., more than 10% of the SNPs annotated by the dbSNP website) were not present in the database version that was downloaded by ftp on the same day (August 6, 2001). It turned out that later versions of the downloadable dbSNP database (i.e., from the version of August 20, 2001, onward) comprised all 61 SNPs, indicating that updates of the database used for the dbSNP website and the database available for download on the ftp website are not fully synchronized. The 465 SNPs that were present in the

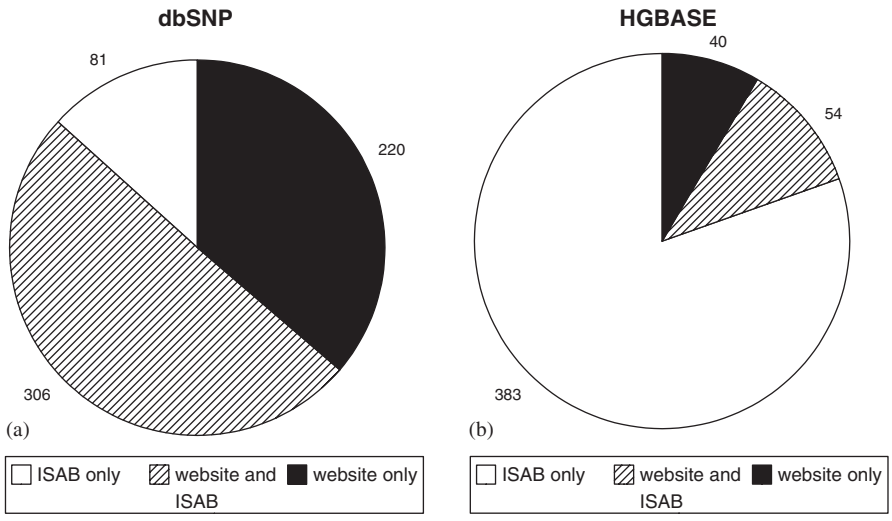

FIGURE 4. **Summary of the number of SNPs annotated onto one of our list of 24 genes. A differentiation is made between SNPs annotated exclusively on the public SNP database websites (black), SNPs annotated both on the website and by the ISAB strategy (shaded), and SNPs annotated exclusively by using the ISAB strategy (white). Results are shown for dbSNP (A), and HGVbase (B). Only SNPs located in non-repeat regions are included for the parts annotated with the ISAB strategy (white and shaded), but this could not be verified for SNPs annotated solely by the websites (black).**

downloaded database were blasted individually against the corresponding cDNA and genomic gene sequences (BL2SEQ algorithm). Sixty-nine of these SNPs did not produce a BL2SEQ hit with an expectation value lower than 10E-12, and therefore were discarded for further annotation analysis. The BL2SEQ algorithm, however, produced significant hits for the other 396 SNPs. The downloaded HGVbase database comprised all 94 SNPs annotated by the website, but no significant BL2SEQ hit (i.e., expectation value below 10E-12) was found for 29 HGVbase SNPs.

For all SNPs for which a significant BL2SEQ hit was identified, the best hit was selected based on percentage identity and length (in that order). Figure 5 shows the distribution of the SNP hits according to percentage identity and hit length. According to this stratification, only the SNPs that fulfilled the 98–100% criterion for percentage identity and that were longer than 250 bp (part 1) or showed a relative hit length higher than 98% (part 2) would have been selected by the ISAB algorithm. Relative hit length was defined as the ratio of BL2SEQ hit length versus the context length of the SNP in the public database. Only a very limited number of the significant BL2SEQ hits have a low percentage identity, as a consequence of the fact that the BLAST and BL2SEQ algorithms optimize primarily the percentage identity and only secondarily the hit length. Of note, the BL2SEQ search produced seven SNPs that complied with the ISAB criteria that were not found using the BLAST algorithm.

DISCUSSION

A prerequisite for the implementation of high-throughput SNP genotyping as a tool in genetic research projects is the availability of databases comprising high quality annotation data on known SNPs. Such a resource is especially important when the selection of SNPs to be assayed in a genotyping facility is based on SNP database information rather than on expensive in-house SNP discovery studies. Obviously, the SNP annotation quality in such databases should be high in order to avoid costly SNP assay development and genotyping of SNPs that later turn out not to be valid SNPs or not located at the expected chromosomal region according to the database annotation. Here, we present a strategy (ISAB) to annotate intragenic SNPs to specified genes, that makes use of the BLAST algorithm and predefined criteria to select valid SNPs from the public dbSNP and HGVbase databases. Obviously, an annotation strategy of SNPs available in silico cannot judge the experimental validity of selected intragenic SNPs. Yet, our study demonstrates that there is room for improvement of the SNP annotation quality in public SNP databases.

Annotation of Intragenic SNPs Using the ISAB Strategy

Both the genomic and cDNA sequences of each gene were blasted against the downloaded dbSNP and HGVbase databases as part of the ISAB strategy. Our analysis showed that the genomic DNA is the most

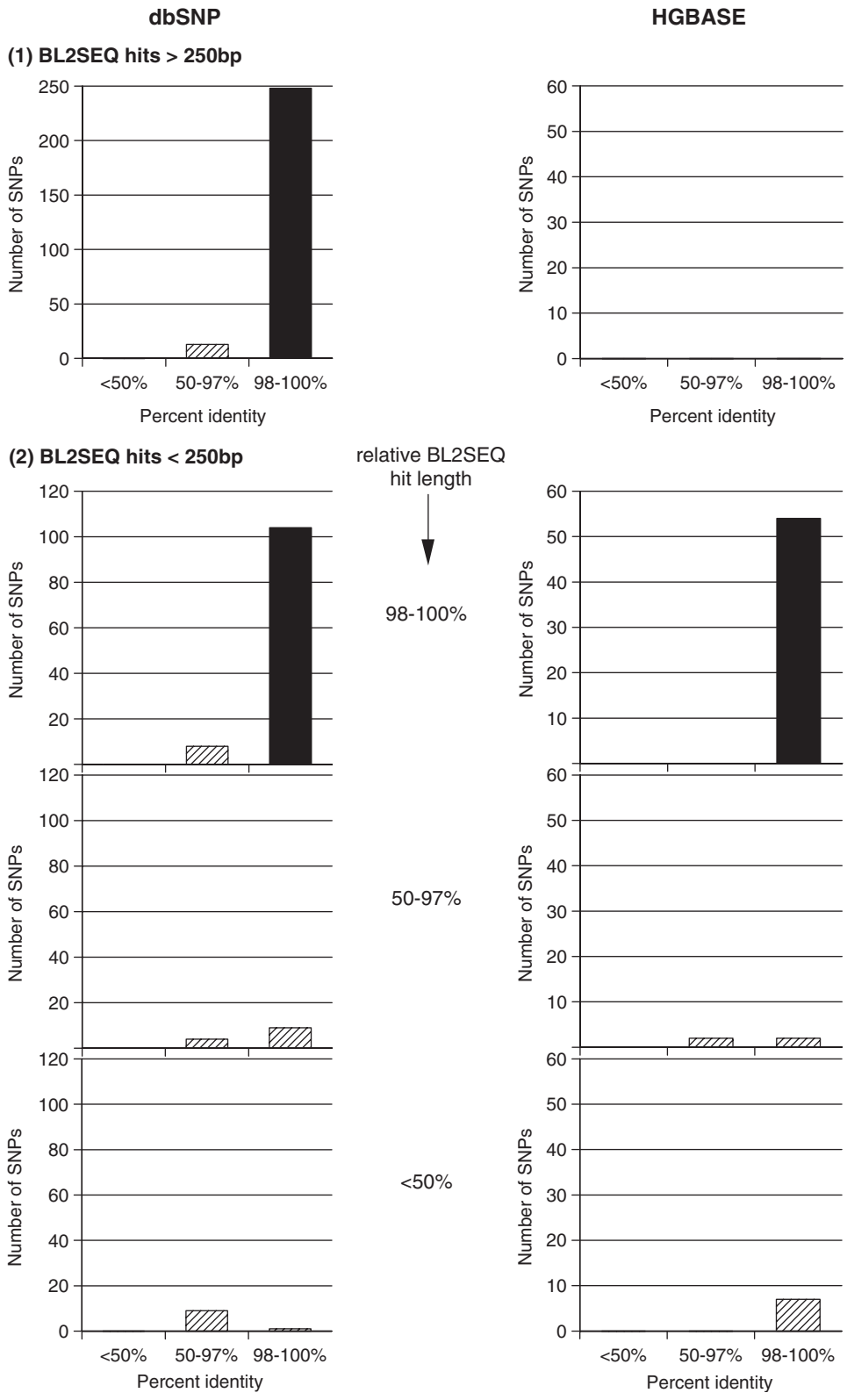

FIGURE 5. **Distribution of BL2SEQ hits according to hit length and percentage identity of SNPs annotated by the websites to map in one of the listed genes. For each SNP, the best BL2SEQ hit was selected based on percent identity and length (in this order). Part 1 shows the number of hits that are longer than 250 nucleotides. Hits shorter than 250 nucleotides are displayed in part 2, stratified by relative BL2SEQ hit length (defined as the relative length of the BL2SEQ hit to the length of the SNP database entry). Bars in black represent BL2SEQ hits that comply with the ISAB criteria to be selected for annotation.**

valuable sequence to perform a BLAST search on the databases, both for dbSNP and HGVbase. However, the cDNA is still a valuable source for blasting HGVbase when the genomic DNA sequence is not available. More specifically, four out of five SNPs that were found in HGVbase could be identified using the cDNA sequence only. Both the gDNA and cDNA sequences were chopped into smaller subsequences before they were blasted. The BLAST algorithm, originally designed to search for amino acid sequences, is not really suited to process large sequences. More specifically, the algorithm tries to spread its hits over the full length of the queried sequence, and is therefore likely to report less relevant hits that are located in an area with fewer SNPs over more relevant hits in a region with many SNPs. In the presented ISAB strategy, we tried to circumvent this problem by chopping the query sequence in smaller parts (1,500 bp), and increasing the number of BLAST hits to be returned. This is, however, no absolute guarantee that all relevant hits are reported, as can be deduced from the fact that a few SNPs were not found using the ISAB strategy, but nevertheless complied with the ISAB criteria when their flanking sequences were aligned to the specific gene sequence using BL2SEQ. To reduce the chances of erroneous annotation, we introduced the parameter "relative BLAST hit length" as a selection criterion in the ISAB algorithm. The stringent 98% threshold prevents BLAST hits that have a high percentage identity, but include only a (small) part of the SNP database entry to be accepted as valid SNPs. However, this threshold could not be used for BLAST hits longer than 250 bp, because SNPs that are mapped in the overlapping region of two subsequences would intrinsically not reach this threshold. We therefore accepted all BLAST hits longer than 250 nucleotides when all other criteria were fulfilled. A hit longer than 250 nucleotides provides some assurance that the SNP's flanking regions are mapped to the correct genomic or cDNA sequence. Yet we recognize that this might be a problem with highly homologous genes or pseudogenes. To further reduce the chance that the latter problem would arise (e.g., in a gene known to be a member of a large gene family), the length of the overlapping subsequence fragments could be increased (e.g., 500 bp rather than 250 bp). This would allow for a more stringent criterion on the BLAST hit length: the hit length should exceed 98% of the database SNP entry length (same as before), unless it is longer than 500 bp. Finally, the ISAB algorithm filters SNPs that mapped into regions of repeat sequences, which might have discarded a number of intragenic SNPs that are actually annotated by the public SNP databases. We anticipated that such SNPs are usually less relevant to be selected for genotyping purposes.

The ISAB strategy showed that the SNP content of the public SNP databases HGVbase and dbSNP are highly complementary but also overlap significantly. It is therefore not sufficient to search only one of these databases in order to identify all intragenic SNPs in a candidate gene. Moreover, when available, other gene-specific mutation databases or literature might contain additional information on mutations or SNPs in the candidate gene. The combined analysis of dbSNP and HGVbase yielded SNP densities in the same order of magnitude as the most often quoted figure of one SNP per 1,000 bp, and estimates that go up to one SNP per 350 bp [Taillon-Miller et al., 1998; Cargill et al., 1999; Semple, 2000]. Interestingly, the in silico SNP density we found is higher in the coding regions than in intronic regions. This may be attributed to the fact that many SNPs have been identified by alignment of EST sequences, resulting in an artificial increase of the number of SNPs found in coding regions.

A proper validation of a strategy such as ISAB is difficult for several reasons. It is clear that the discrepancy between the SNPs identified by mutation screening experiments of candidate genes in the laboratory and the SNPs retrieved from HGVbase and dbSNP by the ISAB strategy has multiple causes. First, not all SNPs reported in the literature reports based on mutation screening experiments are found by ISAB. Either these SNPs were not retained by the ISAB strategy because they did not comply with the defined criteria, or the SNPs were not present in the databases. The latter was the case for all the SNPs in *AHR*, *CYP2A6*, and *MTHFR* that were reported in literature [van der Put et al., 1998; Weisberg et al., 1998; Cauchi et al., 2001; Raunio et al., 2001] but not found by ISAB. Second, ISAB identified additional SNPs that were not found in the reported mutation screening experiments. The mutation screening experiments might have missed some SNPs because of the selection of individuals used in the analysis (e.g., population size, ethnicity, sex), the low allele frequency of SNPs, or the technology used for mutation screening. It is likely that such SNPs have been found in other experimental studies and have been entered in dbSNP and/or HGVbase. Alternatively, it can not be excluded that some SNPs have been erroneously annotated by ISAB, because the SNP fulfilled all criteria but is actually located in a homologous gene or pseudogene. Only experimental confirmation of SNPs can overcome this inherent shortcoming of any in silico SNP searching method. The stringent criteria applied by the ISAB strategy for the selection of SNPs should minimize the chances of such erroneous annotation.

SNP Annotation by Database Websites

The large discrepancy between the annotation of SNPs to specific genes reported by the public SNP websites and the annotation based on interpretation of BLAST results is sobering. Approximately half of

the SNPs annotated by the public SNP databases to map in one of the genes under study could not be retrieved in the output of a BLAST search of these genes against the SNP databases. In addition, a large number of SNPs that fulfilled our rather stringent criteria to accept an SNP for annotation to a specific gene was not (yet) annotated by the public SNP database websites. This illustrates that the annotation of SNPs by dbSNP and HGVbase is a work in progress. Recently, Johnson et al. [2001] reported that dbSNP contained no more than 25% of the SNPs that they identified using wet experiments. Unfortunately, the method they used to search the dbSNP database was not specified. This might suggest that the dbSNP database currently contains only a small part of all SNPs in the human genome, and/or that the annotation available at the dbSNP website can be improved. Concerning the HGVbase database, Brookes et al. [2000] stated that 25 bp at either end of the SNP is effective to define a SNP, because such a string of 51 bp is highly likely to be unique within a given gene. However, the link with the gene has to be preserved. When using the HGVbase database as a blastable resource of SNPs without keeping this link, the short flanking regions cause many SNPs to be mapped in repeat regions, homologous genes, or pseudogenes. This might at least partially explain the high number of apparently valid SNPs according to the ISAB strategy that were not annotated by the HGVbase website. Together, these observations support our approach to complement the SNP database website gene annotations with an independent annotation strategy.

Multiple factors might contribute to the difference in the annotation of intragenic SNPs at the SNP database websites and the ISAB annotation strategy. The genomic reference sequence of a specific gene that was used for the annotation might have been different. We used the genomic DNA sequences of the genes as they are present and annotated in the so-called "golden path" of the human genome, whereas the public SNP databases might have used other sources of gene sequence. The origin and nature of the SNP might also play a role: the annotation quality of SNPs that have been identified in silico by comparison of sequences of EST clones is likely to be different from SNPs that have been identified experimentally in well delineated pieces of genomic DNA from a documented group of individuals. Furthermore, at least part of the annotation discrepancy in dbSNP might be due to the delay in synchronization between the database version that can be queried on the website and the database version that is available for download. Another legitimate remark might be whether our predefined criteria to accept an SNP for annotation to a specific gene, based on percentage identity and relative hit length in the BLAST output of the gene sequence against the SNP database, are appropriate. The 98% threshold values for both parameters were chosen, as we believe these are likely to result in high quality data. It can, however, easily be demonstrated from the presented data that somewhat less stringent threshold values (e.g., 95%) would have lead to very similar conclusions.

It is important to realize that data mining of the public SNP databases for the selection of intragenic SNPs is only a first step toward the final selection of a set of validated SNPs in a candidate gene. Our bioinformatics analysis does not take into account whether or not the SNPs deposited in dbSNP or HGVbase have been validated by wet experiments or represent only in silico SNPs. It is well recognized that in silico SNP discovery is prone to so-called false positive SNPs that can not be confirmed experimentally in the laboratory [Cox et al., 2001]. Moreover, it was reported that only 66–70% of the publicly available SNPs have appreciable minor allele frequencies, and approximately 50% of the SNPs have alleles that are common in a given population [Marth et al., 2001]. Our evaluation of SNP annotation quality focused only on the gene to which a SNP is annotated. Thus, the allele frequency nor the ethnic population in which a SNP was identified has been taken into account, although this is important information in order to select SNPs for further use in association studies. The ALFRED database [Cheung et al., 2000] provides allele frequencies for SNPs, but cannot (yet) be searched in an automated way. Because no appropriate bioinformatics tools are currently available to select only truly validated SNPs, we included all SNPs from the dbSNP and HGVbase databases in our analysis regardless of ethnicity and allele frequency.

In conclusion, our analysis suggests that the identification of intragenic SNPs comprised in dbSNP and HGVbase can be done efficiently by applying the BLAST algorithm with appropriate threshold settings to these databases. Such a data mining approach is likely to reveal additional and high quality SNPs compared to the SNPs that are annotated by the respective websites. However, problems might arise in case of short flanking regions in the database (e.g., HGVbase). We suggest this approach might be the first step in the process to select validated intragenic SNPs from public SNP databases that are to be used in candidate gene association studies.

ACKNOWLEDGMENT

The authors thank Steven Osselaer for informatics and database support.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST:

a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Board P, Tetlow N, Blackburn A, Chelvanayagam G. 2000. Database analysis and gene discovery in pharmacogenetics. Clin Chem Lab Med 38:863–867.

Brookes AJ, Lehväslaiho H, Siegfried M, Boehm JG, Yuan YP, Sarkar CM, Bork P, Ortigao F. 2000. HGBASE: a database of SNPs and other variations in and around human genes. Nucleic Acids Res 28:356–360.

Buetow KH, Edmonson MN, Cassidy AB. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. Nat Genet 21:323–325.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238.

Cauchi S, Stucker I, Solas C, Laurent-Puig P, Cenee S, Hemon D, Jacquet M, Kremers P, Beaune P, Massaad-Massade L. 2001. Polymorphisms of human aryl hydrocarbon receptor (AHR) gene in a French population: relationship with CYP1A1 inducibility and lung cancer. Carcinogenesis 22:1819–1824.

Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. 2000. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res 28:361–363.

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

Gray IC, Campbell DA, Spurr NK. 2000. Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 9:2403–2408.

Johnson GC, Esposito L, Barrat BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29: 233–237.

Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16:418–420.

Kao SL, Chong SS, Lee CG. 2000. The role of single nucleotide polymorphisms (SNPs) in understanding complex disorders and pharmacogenetics. Ann Acad Med Singapore 29: 376–382.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. Science 265:2037–2048.

Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, Davenport R, Miller RD, Kwok PY. 2001. Single-nucleotide polymorphisms in the public domain: how useful are they? Nat Genet 27:371–372.

Raunio H, Rautio A, Gullsten H, Pelkonen O. 2001. Polymorphisms of CYP2A6 and its practical consequences. Br J Clin Pharmacol 52:357–363.

Semple S. 2000. In silico identification of transcripts and SNPs from a region of 4p linked with bipolar affective disorder. Bioinformatics Discovery Note 16:735–738.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Taillon-Miller P, Gu Z, Li Q, Hillier LD, Kwok PY. 1998. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res 8:748–754.

van der Put NM, Gabreels F, Stevens EM, Smeitink JA, Trijbels FJ, Eskes TK, van den Heuvel LP, Blom HJ. 1998. A second common mutation in the methylenetetrahydrofolate reductase gene: an additional risk factor for neural-tube defects? Am J Hum Genet 62:1044–1051.

Weisberg I, Tran P, Christensen B, Sibani S, Rozen R. 1998. A second genetic polymorphism in methylenetetrahydrofolate reductase (MTHFR) associated with decreased enzyme activity. Mol Genet Metab 64:169–172.