Comment

Pharmacology

Pharmacology 2013;92:319–323
DOI: 10.1159/000356324

Received: October 4, 2013
Accepted: October 9, 2013
Published online: December 12, 2013

Caveat Emptor: Single Nucleotide Polymorphism Reporting in Pharmacogenomics

Izel Tekin Kent E. Vrana

Department of Pharmacology, Penn State College of Medicine, Pennsylvania State University, Hershey, Pa., USA

Key Words
National Center for Biotechnology Information · Database · Single Nucleotide Polymorphisms

Abstract
While it is arguably the most comprehensive source of genetic information, the NCBI's dbSNP database (National Center for Biotechnology Information database of single nucleotide polymorphisms; http://www.ncbi.nlm.nih.gov/projects/SNP/) is imperfect. In this commentary, we highlight the issues surrounding this database, while considering the great importance and utility of this resource for those in the pharmacology and pharmacogenomics communities. We describe our experience with the information in this database as a cautionary tale for those who will utilize such information in the future. We also discuss several measures that could render it more reliable.

Caveat emptor – 'Let the buyer beware'. In this new era of exploding genomic data in the biomedical sciences, this catchphrase is as relevant now as when the maxim was first used in the 16th century. Indeed, it is derived from a larger context: Caveat emptor, quia ignorare non debuit quod jus alienum emit – 'Let a purchaser beware, for he ought not to be ignorant of the nature of the property which he is buying from another party' [1]. While we have ready access to tremendous genetic databases, use of these data requires vigilance. Since the first draft of the human genome project was released in 2001, we have had comprehensive information at our fingertips. One example of this is the NCBI (National Center for Biotechnology Information) genomic website that provides seemingly endless information on a given gene of interest. One can access the site (www.ncbi.nlm.nih.gov/), provide the name of a given gene, and retrieve such varied information as the following: the chromosomal location of the gene; the intron/exon structure; the complete sequence of the gene; the mRNA sequence, and the corresponding protein primary structure. Using the specific dbSNP database (National Center for Biotechnology Information database of single nucleotide polymorphisms; http://www.ncbi.nlm.nih.gov/SNP/), you can peruse all known SNPs – whether published in the peer-reviewed literature or simply directly deposited to this site by investigators throughout the world. These genetic tools are truly remarkable and are readily accessible from personal computers. Moreover, they provide insights, at the gene level, into health and disease. Such resources were almost unimaginable 20 years ago.

NCBI's dbSNP database is not the only source for information regarding genomic variations, although it is the most convenient 'one stop shop'. Other sources of genomic data come from the HapMap Project (the Haplotype

KARGER

E-Mail karger@karger.com
www.karger.com/pha

0031–7012/13/0926–0319$38.00/0

Kent Vrana, PhD, Elliot S. Vesell Professor and Chair
Department of Pharmacology, Penn State College of Medicine
500 University Drive, R130, PO Box 850
Hershey, PA 17033-0850 (USA)
E-Mail kvrana@psu.edu

Map Project; http://hapmap.ncbi.nlm.nih.gov/) and the 1000 Genomes Project (http://www.1000genomes.org/). The goal of the HapMap Project is to identify haplotypes specific to certain populations. Haplotypes are sets of genetic variants or polymorphisms that tend to be inherited together and hence can be associated with disease states. Such an analysis of haplotypes will allow the identification of genomic regions that correlate with different phenotypes in an attempt to obtain genetic markers that are associated with a variety of different conditions so as to predict risk or identify the responsible gene. This international project represents one of the largest efforts to understand the contribution of genetics to human health and disease. Instead of a whole-genome sequencing approach, the HapMap data focus on analyzing several gene regions at one time. As a result, this project has provided scientists with SNP frequencies and haplotypes for participants from several ethnic groups, and these data are available on the project website, as well as NCBI's dbSNP database. The 1000 Genomes Project was launched to obtain complete genomic sequences of at least 1,000 participants, and this number has already been surpassed (1,092 at the end of phase I, as of October 2012) [2]. This international consortium plans to provide the most detailed and highly validated database for the complete human genome sequence. The determination of the complete genome sequences for such a high number of subjects will not only provide us with statistically significant information, but it will provide insights into the consequences of rare to common variants via their presence in this ethnically diverse cohort. In the present context, one of the advantages of the dbSNP database is that genomic data from both HapMap and 1000 Genomes Projects are also deposited here, in addition to independent and individual submissions. This allows access to all of these data from a single source.

So, why is this relevant to the readers of *Pharmacology*? A simple survey of metabolic enzymes, receptors and signal transduction proteins highlights the scope of polymorphisms relevant to pharmacogenomics. Table 1 compiles a partial list of polymorphic proteins with known drug associations. Remarkably, analysis of the dbSNP database indicates that there are a total of 32,265 genetic variants embodied within these 15 genes. Moreover, some of these proteins (especially the cytochrome p450 family members) have known drug associations with up to 50 different drugs. CYP2D6, for instance, has several polymorphisms that determine the rate of drug metabolism in different individuals. These genetic variants will therefore provide a rich landscape upon which to develop personalized pharmacotherapeutic approaches.

A Genetic Cautionary Tale

All is not what it seems, however. As the managers of the dbSNP site note, the database is not curated or verified – nor could it be. In the interests of utility and academic freedom, individual investigators are permitted to submit sequence variant information directly to dbSNP without any independent verification or quality control information. In some cases this is justified in that the observation of a single SNP, in a single patient, might be impossible to replicate if the allelic frequency is very rare or if the SNP is an acquired mutation (in a neoplasia, for instance). Moreover, even if it is a rare or a singular observation in 1 patient or sample, such an SNP may provide important structure/function insights into a gene or corresponding protein. The problem arises when one acts on dbSNP information and it turns out to be incorrect.

Our recent characterization of the tyrosine hydroxylase gene *(TH)* may serve as a cautionary tale in this regard. Specifically, the dbSNP database and the published literature (as of March 2013) reported that there have been 65 different SNPs documented within the coding region of this rate-limiting enzyme in the synthesis of the catecholamines (dopamine, norepinephrine and epinephrine). In addition, there are 228 noncoding region polymorphisms, as well as the report of 1 patient with a complete gene deletion on 1 allele. Interestingly, there are only 6 genetic variants with appreciable allelic frequency within the general population (ranging from 0.05 to 40%); the remainder of the variants generally represents single 'case reports'. As part of a larger project, we sequenced (using traditional Sanger sequencing) all 14 exons of the human *TH* gene in 10 Parkinson's disease patients and 10 controls. Interestingly, during our initial sequencing of exon amplicons, we observed 7 previously reported SNPs. Unfortunately, this prevalence seemed much too high given that most had no reported population frequency. Subsequent agarose gel purification of the particular amplicons followed by resequencing failed to confirm any of the observed SNPs save for a common synonymous change (K240K, rs6357) and a common coding-region nonsynonymous SNP (V81M, rs6356). The telling point is that most of the polymorphisms we initially identified had all been previously reported in dbSNP.

The reason for false identification of SNPs is inherent in the process of Sanger sequencing and automated software 'calling' of sequence. As demonstrated for one of the phantom SNPs (fig. 1a, upper panel), part of the problem is the noisy nature of PCR amplicon sequencing using Sanger sequencing. However, once the same amplicon is

Table 1. Number of reported SNPs in genes that are known to have established associations with the pharmacokinetics of known drugs

Gene symbol	Gene name	Number of total variants in dbSNP	Number of total validated variants	Known drug associations
BRAF	v-raf murine sarcoma viral oncogene homolog B1	3,456	2,401	vemurafenib
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2	323	183	fluvoxamine, olanzapine
CYP2C19	cytochrome P450, family 2, subfamily C, polypeptide 19	2,458	1,820	amitriptyline, carisoprodol, citalopram, clobazam, clomipramine, clopidogrel, diazepam, doxepin, drospirenone, escitalopram, esomeprazole, ethinyl estradiol, fluvoxamine, imipramine, lansoprazole, moclobemide, modafinil, nelfinavir, omeprazole, pantoprazole, prasugrel, rabeprazole, sertraline, trimipramine, voriconazole
CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	1,423	996	acenocoumarol, celecoxib, flurbiprofen, fluvoxamine, glibenclamide, gliclazide, glimepiride, phenprocoumon, phenytoin, tolbutamide, warfarin
CYP2D6	cytochrome P450, family 2, subfamily D, polypeptide 6	474	217	acetaminophen, amitriptyline, aripiprazole, atomoxetine, carvediol, cevimeline, citalopram, clomipramine, clozapine, codeine, desipramine, dextromethorphan, doxepin, duloxetine, flecainide, fluoxetine, flupenthixol, fluvoxamine, galantamine, gefitinib, haloperidol, iloperidone, imipramine, metoprolol, mirtazapine, modafinil, notriptyline, olanzapine, oxycodone, paroxetine, perphenazine, pimozide, propafenone, propranolol, protriptyline, quinidine, risperidone, tamoxifen, terbinafine, tetrabenazine, thioridazine, timolol, tiotropium, tolterodine, tramadol, trimipramine, venlafaxine, zuclopenthixol
CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	715	458	aripiprazole, citalopram, fluvoxamine, gefitinib, nelfinavir, ticagrelor
CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5	718	446	tacrolimus
DPYD	dihydropyrimidine dehydrogenase	15,916	8,525	capecitabine, fluorouracil, tegafur
EGFR	epidermal growth factor receptor	4,507	3,184	cetuximab, erlotinib, gefitinib
G6PD	glucose-6-phosphate dehydrogenase	243	203	chloroquine, dapsone, glibenclamide, methylene blue, nalidixic acid, nitrofurantoin, norlofloxacin, pegloticase, primaquine, probenecid, rasburicase, sulfadiazine, sulfamethoxazole, sulfasalazine, sulfisoxazole, trimethoprim, vitamin C
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	1,021	722	cetuximab, panitumumab
SLCO1B1	solute carrier organic anion transporter family, member 1B1	2,741	1,999	simvastatin
TPMT	thiopurine S-methyltransferase	672	437	azathioprine, cisplatin, mercaptopurine, thioguanine
UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1	450	283	erlotinib, indacaterol, irinotecan, nilotinib
VKORC1	vitamin K epoxide reductase complex, subunit 1	147	93	acenocoumarol, phenprocoumon, warfarin
Total		35,264	21,967	

The data are compiled from the Pharmacogenomics Knowledgebase (http://www.pharmgkb.org/search/knownPairs.action) and the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/snp/), both accessed on September 23rd, 2013.

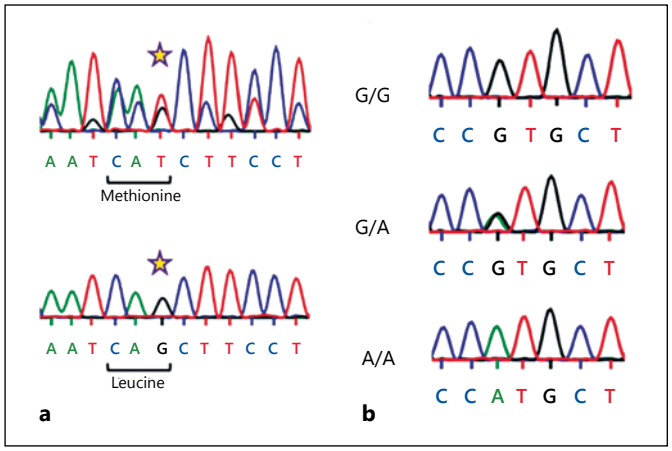

Fig. 1. Representative chromatograms from Sanger sequencing performed after PCR amplification of individual exons of the *TH* gene. **a** The upper figure shows the chromatogram of the amplified exon 6. This was obtained after direct PCR purification (using a commercially available kit). The area marked with the star was named as a polymorphism by both the computer and the experimenter. The bottom panel shows the chromatogram of the sequencing performed after the same amplicon was resolved on, and purified from, an agarose gel. The decrease in the background noise can clearly be seen in this figure. This decrease in the noise resulted in the consequent elimination of false positives that were observed otherwise. **b** These chromatograms were obtained from sequencing following agarose gel purification of independent amplicons of exon 3 of *TH*. In the case of the heterozygote sample, the computer named the residue as guanine, even though there is a comparable abundance of adenine. Careful observation of each chromatogram is required to assess the presence of multiple nucleotides in order to observe heterozygosity.

gel purified and resequenced the artifact disappears and the true sequence emerges. This is what happened for all of our observed variants (whether previously reported or not) except for V81M and the common K240K synonymous variant. As the background noise was clearly high in the original sequencing (fig. 1a), we propose that investigators should be very careful in interpreting Sanger sequencing data. The aforementioned issue can be a common occurrence as varying DNA structures, amplified materials or PCR components compromise the sequencing reactions. Another problem is the fact that automated sequence calling is less than perfect. For instance, whenever we identify a variant, we examine the primary sequence data. As shown in figure 1b, this very clearly discriminates homozygotes from heterozygotes. However, the automated calling software, in this case, chose a single sequence for the heterozygote (calling the slightly more abundant nucleotide). This could easily produce a spurious result if not interrogated at the level of the primary data output, although that process is very time consuming and becomes problematic when the sample size is high.

Finally, there is the nature of the reported polymorphism. In examining the NCBI dbSNP database for tryptophan hydroxylase (rate-limiting enzyme in serotonin biosynthesis) coding-region polymorphisms, we encountered an intriguing report of an SNP within the catalytic domain of the protein. The amino acid substitution would be predicted to significantly compromise the enzyme activity. We therefore contacted the submitting scientist to gather information on the nature of the subject. Much to our surprise, the SNP was identified as an acquired mutation in a tumor that would not express the neurotransmitter biosynthetic enzyme. The SNP was present in the tumor and not in the patient's normal tissue. Therefore, this was a case where the dysregulated neoplasm acquired a mutation that had no effect on its viability and would never have been expressed in vivo. This was a situation where we had confidence in the report (based on the underlying sequencing technology), but the observation probably has no relevance to human health and disease.

Source of the Issue

So what do these personal observations tell us about the nascent field of genomic sequencing and the impact of the wealth of genetic data at our disposal? First, in a totally appropriate egalitarian environment in which information is freely shared, the buyer must be wary. In the case of the NCBI dbSNP database, anyone can submit a nucleotide sequence variant. There is no curation or confirmation involved – there is no quality control required for reporting a genetic variant. In order to deposit a new variant into the dbSNP database, an investigator first obtains a handler ID by entering their information into the NCBI system. Then, guidelines are followed regarding the nomenclature, SNP distribution, method of discovery, population data, and information regarding the validation of the SNP. Among these, the only information that is absolutely required is the description of the particular discovery method. So, one can imagine that the system has the potential to disseminate incorrect information. Indeed, a recent report describes how problems inherent to the sequencing method can inadvertently introduce spurious information into dbSNP [5]. However, recent changes at NCBI allow the investigator to learn about the validation of a particular SNP and how this step has been performed (through icons listed under the relevant section). Returning to table 1 and consideration of genetic variation and pharmacogenomics, one

can gauge the true scale of the problem. Of the 35,264 polymorphisms reported in dbSNP for these 15 genes, approximately 40% have been validated. This should not be taken to mean that the other 60% are incorrect – merely that they have the potential to be artifactual.

Clearly, the dbSNP database is an extremely valuable resource. Because of the aforementioned issues, however, many investigators refuse to trust deposited data that are not accompanied by some measure of allelic frequency (reflecting some measurable presence in the general population and a sense of reliability and reproducibility). However, the literature is replete with examples of very rare variants (even acquired single individual mutations) that are directly associated with disease [3]. For instance, there are several examples of genetic variants in the *TH* gene that produce dopa-responsive dystonias [4, 6]. Several of these variants are limited to single individuals and hence would probably more accurately be considered a mutation rather than a population polymorphism. Therefore, the lack of a measureable population frequency does not necessarily lessen the impact of the variant on our understanding of health and disease.

A Partial Solution to the Problem

Given the value of allowing direct, unimpeded submission of data to dbSNP and the potential importance of the functional implications of sequence variants, the question is how best to solve the problem at hand – the unreliability of the data. First, we would suggest that each dbSNP entry be accompanied by the primary data (including replicate sequencing on both strands of the DNA). A simple submission of a PDF of the sequencing chromatogram of that section of the gene (perhaps encompassing 20 nucleotides) would suffice. The ancillary value of such a requirement would be to force the submitter to examine the quality of the sequencing data. In terms of next generation sequencing, the sequencing coverage, average quality score and forward/reverse balance should be submitted (accompanied by a FASTQ file of all mapped reads containing the SNP in question). A second change to the submission would be the inclusion of a simple description of the source of the material for the sequencing. This would protect the interested party, for example, from acting on a polymorphism in a CNS receptor that was identified in a squamous cell carcinoma (when that SNP was not present in normal somatic tissue).

On the other side of the equation, an interested scientist also needs to actively seek out these types of information from the submitting scientist. Clearly, before undertaking the use of dbSNP data to begin a characterization, one should know about the source of the polymorphism and any associated clinical characteristics. A good place to start is with an email query to the submitting scientist. Regardless, access to these necessary data from the beginning would save investigators a great deal of time.

In discussing these issues concerning the dbSNP database with geneticist colleagues, a common refrain is that there is no need to make changes at this late date. We should all just understand that the data are noisy at best and wholly unreliable at worst, and so no one should consider them as intrinsically informative. We, on the other hand, ask 'what is the purpose of compiling the data if they cannot be considered actionable?' In closing, as molecular pharmacologists, we find the dbSNP database to be a tremendous resource for mining the data embodied in the human genome. However, in its present state, we must caution the 'buyer' to beware for he/she ought not to be ignorant of the nature of the genetic property they are buying from another party.

Acknowledgments

The authors would like to thank Dr. Willard M. Freeman, Dr. Nurgul Carkaci-Salli, Dr. Carla Gallagher and Mr. Dustin R. Masser for their helpful contributions. This work was supported by grants from the National Institutes of Health (GM38931) and the Penn State Institute for Personalized Medicine (04-017-52 HY 8A1HO, under a grant with the Pennsylvania Department of Health using Tobacco CURE funds – the Department specifically disclaims responsibility for any analyses, interpretations or conclusions).

References

1 Robertson JG: Robertson's Words for a Modern Age: a Cross Reference of Latin and Greek Combining Elements. Los Angeles, Senior Scribe Publications, 1991.

2 The 1000 Genomes Project Consortium: An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.

3 Musumeci L, Arthur JW, Cheung FS, Hogue A, Lippman S, Reichardt JK: Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Hum Mutat 2010;31:67–73.

4 Foo JN, Liu JJ, Tan EK: Whole-genome and whole-exome sequencing in neurological diseases. Nat Rev Neurol 2012;8:508–517.

5 Haavik J, Blau N, Thony B: Mutations in human monoamine-related neurotransmitter pathway genes. Hum Mutat 2008;29:891–902.

6 Segawa M: Dopa-responsive dystonia. Handb Clin Neurol 2011;100:539–557.