Published in final edited form as:
Hum Mutat. 2010 January ; 31(1): 67–73. doi:10.1002/humu.21137.

Single Nucleotide Differences (SNDs) in the dbSNP Database May Lead to Errors in Genotyping and Haplotyping Studies

Lucia Musumeci[1,6,*], **Jonathan W Arthur**[2,3,*], **Florence SG Cheung**[1], **Ashraful Hoque**[4], **Scott Lippman**[4], and **Juergen KV Reichardt**[1,5]

[1] Plunkett Chair of Molecular Biology (Medicine), Bosch Institute, The University of Sydney, Medical Foundation Building (K25), 92–94 Parramatta Road, Camperdown, NSW 2006, Australia

[2] Discipline of Medicine, Sydney Medical School, The University of Sydney, Camperdown, NSW 2006, Australia

[3] Sydney Bioinformatics, The University of Sydney, Camperdown, NSW 2006, Australia

[4] The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

Abstract

The creation of single-nucleotide polymorphism (SNP) databases (such as NCBI dbSNP) has facilitated scientific research in many fields. SNP discovery and detection has improved to the extent that there are over 17 million human reference (rs) SNPs reported to date (Build 129 of dbSNP). SNP databases are unfortunately not always complete and/or accurate. In fact, half of the reported SNPs are still only candidate SNPs and are not validated in a population. We describe the identification of SNDs (Single Nucleotide Differences) in humans, that may contaminate the dbSNP database. These SNDs, reported as real SNPs in the database, do not exist as such, but are merely artifacts due to the presence of a paralogue (highly similar duplicated) sequence in the genome. Using sequencing we showed how SNDs could originate in two paralogous genes and evaluated samples from a population of 100 individuals for the presence/absence of SNPs. Moreover using bioinformatics, we predicted as many as 8.32% of the biallelic, coding SNPs in the dbSNP database to be SNDs. Our identification of SNDs in the database will allow researchers to not only select truly informative SNPs for association studies, but also aid in determining accurate SNP genotypes and haplotypes.

Keywords

single nucleotide polymorphism; SNP; paralogue; single nucleotide difference; SND; alignment

Introduction

Single-nucleotide polymorphisms (SNPs) are the most common genetic variations in humans and occur approximately once in every 100 to 300 base pairs (http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml and http://www.maik.ru/abstract/biophyss/3/biophyss0081_abstract.pdf). The importance of SNPs and haplotypes - composed of multiple SNPs statistically associated - has been

[5]Corresponding Author: Prof. Juergen Reichardt, University of Sydney, Plunkett Chair of Molecular Biology (Medicine), Medical Foundation Building (K25), 92-94 Parramatta Road, Camperdown, New South Wales, 2006, Australia, jreichardt@med.usyd.edu.au.
[6]Current Affiliation: Immunology and Infectious Diseases Unit, GIGA-R, Liège University, Liège, Belgium
*LM and JWA contributed equally to this work

Supporting Information for this preprint is available from the *Human Mutation* editorial office upon request (humu@wiley.com)

documented in many complex disease association studies. SNPs are genetic markers likely to cause phenotypic differences between individuals (Suh et al., 2005; Sachidanandam et al., 2001). Therefore, SNPs, as well as haplotypes, are the genetic markers of choice in genome-wide association studies (GWAS) (Kruglyak, 2008). The creation of SNP databases (such as dbSNP at the National Center for Biotechnology Information [NCBI]) has facilitated scientific research by providing useful information for complex disease association studies. Although SNP discovery and detection in the last several years has improved to the extent that the NCBI dbSNP database (Sherry et al., 2001) reports 17,103,982 reference SNPs (rs) for *Homo sapiens* to date (October 28, 2008, Build 129), which accounts for 83% of all 20,569,318 sequence variations, only 63% are validated in populations (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). Previous studies (Mitchell et al. 2004) have highlighted that dbSNP has a false positive rate of approximately 15–17% due to sequencing errors arising from base calling software. Therefore, there is a tremendous need for objective evaluation of the online SNP databases (Reich et al., 2003).

In this report, we provide further experimental and computational evidence that not all SNPs listed in the NCBI database are reliable. In fact, we show that as many as 8.32% of the biallelic, coding SNPs in the NCBI dbSNP database appear to be artifacts due to the presence of highly similar genes in the human genome. This kind of problem arises when the DNA sequence similarity between two or more genomic regions is greater than 90%, *i.e.* duplicons (Gut and Lathrop, 2004). Since many human genes derive from ancient duplication (Britten, 2006), we believe a substantial portion of the SNPs, reported in the SNP-GeneView page of the NCBI dbSNP, may not be real inter-individual DNA sequence variations, but rather are sequence variations between two paralogous (*i.e.*, highly similar duplicated sequences) genes (Fredman et al., 2004) (Fig. 1).

We propose the term SND for Single Nucleotide Difference in a parallel to SNPs for these artifactual polymorphisms present in the NCBI dbSNP database due to paralogue loci and identify them using bioinformatics and experimental methods as well (Fig. 1).

Material and Methods

Bioinformatic procedure for the identification of SNDs

Preparation of genomic and SNP data—All 24 assembled chromosomes were downloaded from the NCBI Genome FTP site (Build 36, Version 3; Release date: March 24, 2008; current release at September 22, 2008) and formatted into a BLAST database. The rs SNP sequences associated with each of the 24 chromosomes were downloaded from the NCBI dbSNP FTP site along with the chromosome report files on September 22, 2008 (Build 129 of NCBI dbSNP; Release date: April 14, 2008). Each SNP sequence contains a header line with information about the SNP plus a variable length of sequence both up- and down-stream from the actual SNP location. The Ensembl BioMart tool was used to generate a list of unique coding SNPs drawn from dbSNP.

The SNP data was parsed to identify biallelic SNPs on the basis of the allele information in the header line of each sequence. Comparison of rs numbers with the list obtained from BioMart allowed the identification of those biallelic SNPs located in coding regions. Using these conditions, 119,932 SNPs were selected for analysis (Fig. 2).

As we expect no fundamental difference between coding and non-coding SNPs in terms of their propensity to be SND or not, coding SNPs were chosen as a convenient, representative sample of the whole population of SNPs. This sample is of appropriate size to be tractable with the computational resources available to the authors while being large enough to allow generalization of the findings to all SNPs.

Alignment of SNPs and identification of paralogues—Each SNP and its surrounding sequence were aligned to the human genome using BLASTN (Altschul et al., 1997). The genome was filtered for human repeats using Repbase (Jurka et al., 2005) and also for low complexity sequence to eliminate spurious matches. An expectation threshold of 0.001 was used to eliminate very short, but near identical spurious matches. The BLAST output file for each sequence was parsed to determine the genomic locus or loci to which the SNP could be mapped. Each high scoring pair (HSP) was considered and those alignments, containing the SNP position, with sequence identity greater than 90% over at least 20% of the full length of the SNP sequence, *i.e.*, both left and right flanking sequences, were selected as mapped loci. The parameters (90% identity and 20% sequence coverage) were chosen as the maximum sequence identity and coverage possible to maximize sensitivity and specificity of the bioinformatics technique, using the results of the experimental analysis of 31 SNPs reported in the dbSNP. This was done by using, as a test set, the 31 SNPs from *AKR1C1* (aldo-keto reductase family 1 member C1, which used to be referred to *DD1* as or *DDH1*; MIM# 60049) and *AKR1C2* genes (aldo-keto reductase family 1 member C1, also known as *DD2* or *DDH2*; MIM# 600450).

The two paralogous genes *AKR1C1* and *AKR1C2* are 97% similar, they are both located on 10p15-p14 in opposite transcriptional orientation. Their high sequence identity is probably due to a recent gene duplication event. The *AKR1C1* gene consists of 9 exons distributed over a 15 kbp of genomic sequence, while the *AKR1C2* gene has 11 exons distributed over 28 kbp. We used resequencing of 100 individuals as a validation method for the coding SNPs reported in the NCBI SNP-gene page of the *AKR1C1* and *AKR1C2* genes and showed that most of them are SNDs (Table 1).

The experimental determination of the status as SNDs or not of *AKR1C1* and *AKR1C2* genes was accepted as correct. The bioinformatics analysis of these 31 SNPs was completed using thresholds close to 100% for both sequence identity and coverage. False positives and false negatives were manually examined and the thresholds were gradually reduced until maximum specificity and sensitivity were obtained.

A small number of SNPs (146, 0.12% of the biallelic, coding SNPs) gave no BLAST results. This was due to the length of the SNP region and the proportion of the SNP region filtered by the two filters mentioned above. The remaining unfiltered region was insufficient to seed a BLAST alignment. These SNPs were excluded from further analysis. A second group of SNPs (117, 0.10% of the biallelic, coding SNPs) could not be mapped using this choice of parameters.

For each mapping, the chromosome number, the direction of alignment of the SNP, the genomic position of the SNP, and the allele at the position of the SNP were recorded. The number of mappings for each SNP differs from the *mapweight* of SNPs provided by the Chromosome Reports, which are also available on the dbSNP FTP site. These reports arise from an attempt to specifically match each SNP to a single location on the genome as described at: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.section.ch5.ch5-s8. In this process, the real location of the SNP is distinguished from locations in paralogous regions wherever possible. Thus as many SNPs as possible are mapped to only one location (mapweight = 1) with higher mapweights (corresponding to 2 or more positions on the genome) only being used when it is impossible to distinguish the true location from paralogues. In contrast, our method seeks to identify and retain these paralogous regions as well as the true SNP location, in order to assess the potential for paralogues to give rise to SNDs.

Identification of SNDs

For SNPs with more than one genomic loci, both the set of alleles at the position of the SNP and the base aligned to the actual SNP position in the rs record sequence were examined. Where all aligned bases were the same as one of the two alleles listed in the rs record for the SNP, the SNP was considered "real". Where one or more of the aligned bases corresponded to each of the two alleles listed in the rs record, the SNP was considered "SND. All other cases where only one (or neither) of the two reported alleles was present along with other, not reported alleles were classed as "undetermined" (Fig. 3).

For example, if the SNP is reported in dbSNP as a C/T variation and is mapped to two genomic locations, then such SNP is classed as:

- real, if the aligned residue in both mapping loci is C or the aligned residue in both mapping loci is T
- SND, if the aligned residue in one mapping locus is C while the corresponding aligned residue in the second mapping locus is T
- undetermined, if the aligned residue in one mapping is C or T while the aligned residue in the other mapping is A or G, or the aligned residue in both mapping loci is A or G.

Classification of SNDs—The chromosome report files were parsed to extract the heterozygosity and validation code associated with the SNP. The heterozygosity of each SND with exactly 2 mapped genomic loci was examined. If the heterozygosity of the SNP was over 0.4 and the validation code less than 4, the SND was classed as "very strong SND", indicating a higher likelihood the SNP is a SND. If the heterozygosity of the SNP was over 0.4 and the validation code more than 4 the SND was classed as "strong SND".

Identification of SNDs in AKR1C1 and AKR1C2 genes

Gene-specific primer design—*AKR1C1* and *AKR1C2* exons were amplified using oligonucleotide primers designed from flanking intronic or untranslated sequences (Supp. Table S1). For the design of PCR primers in the intron/exon boundaries, we used the genomic sequence obtained from GenBank [National Center for Biotechnology Information (NCBI)]. Usually the design of gene-specific primers (primers that anneal only to the target gene) is a simple task achieved by blasting the primer against the whole genome database (NCBI database). However, when dealing with genes that have been duplicated during evolution, such as pseudogenes, tandem repeats and paralogues, then SNP assays must be carefully designed to avoid mixed PCR products, resulting from the amplification of different genomic regions. All the primers for *AKR1C1* and *AKR1C2* genes were designed in regions where at least one mismatch (highlighted in red in Supp. Table S1) between the two genes was present. Usually the mismatch was positioned at the 3′ end of the oligonucleotide. The alignments between the two genes were performed using the sequence alignment program ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html). Exon 1 of *AKR1C1* aligned with exon 3 of *AKR1C2*.

PCR reaction optimization—*Taq* polymerase (Fermentas; USA) was used for all intron/exon PCR reactions. Reactions were optimized individually with variations of annealing temperatures and $MgCl_2$ concentration to yield specific PCR products.

Fluorescence-based DNA sequencing of PCR amplicons—DNA sequencing was carried out using an ABI 3730 automated sequencer and the ABI PRISM Dye Terminator Cycle Sequencing Kit, V.3 (Applied Biosystems, Foster City, CA). PCR amplicons were

purified using Marligen's PCR purification kit (Marligen Bioscience Inc. Ijamsville, MD) according to the manufacturer instructions, and sequenced with both forward and reverse primers.

SNP discovery was performed by aligning the electropherogram of each individual using both visual analysis and the demo version Sequencer software program (Gene Codes Corporation, Ann Arbor, MI).

Results

We propose the term "Single Nucleotide Difference" or SND for these artifactual polymorphisms (Fig. 1) and suggest that up to 8.32% of the biallelic coding SNPs in the NCBI dbSNP database are likely to be SNDs (Table 2 and Supp. Table S2).

We used a comprehensive bioinformatic approach (Fig. 2 and Fig. 3) to identify potential SNDs in the NCBI database (Build 129). The 14,319,123 reported SNPs were parsed to select a total of 119,932 biallelic coding SNPs for SND analysis. The sequence surrounding the SNP, as reported in the rs record of dbSNP, was used to align each SNP against the human genome. Only those alignments including the SNP position with at least 90% sequence identity (SID) over 20% of the full length of the SNP sequence or sequence coverage (SC) were selected as mapped loci.

A small number of SNPs (0.22% of the biallelic, coding SNPs) could not be mapped to the genome for reasons discussed in the Material and Methods section. SNPs producing only one alignment (82.12% of the biallelic, coding SNPs) with 90% SID and 20% SC were considered correct SNPs since they could be mapped unambiguously to only one chromosomal position. SNPs with two or more alignments (17.67% of the biallelic, coding SNPs) were considered to be potential SNDs and examined further.

Of the 21,184 SNPs, aligning with multiple loci, 9,979 were SNDs, 10,357 correct SNPs, and 848 undetermined SNPs (Fig. 3). In total, 9,979 SNPs were determined to be SNDs (8.32% of the biallelic, coding SNPs), 108,842 SNPs were determined to be accurate (90.75%), and the remainder 1,111 SNPs were undetermined (0.93%).

The SNDs were examined further for additional evidence by comparison with heterozygosity (H) data. In particular, when the heterozygosity value of the SNP is between 0.4 and 0.5, then the SNP is more likely to be a SND. In fact, in many SNP discovery procedures the co-amplification by polymerase chain reaction (PCR) of two paralogous genes can be misinterpreted as a SNP with H>0.4. In our analysis there were 3,121 SNDs with exactly two genomic mappings; of these 456 (14.61%) had heterozygosity scores in excess of 0.4. We further classified these 456 SNPs into two groups depending on the validation code associated with the SNP (Table 2 and Supp. Table S2). The first group of 189 SNDs are classified as "very strong SNDs" as they have validation code <4, while the second group of 267 SNDs are "strong SNDs" with validation code >4. A lower validation code indicates less evidence for the existence of the SNP and validation code 4 indicates that at least one of the cluster of submitted SNPs (ssSNPs) associated with the reference SNP (rsSNP) has been experimentally validated.

The list in Table 1 includes all the reported SNPs in the NCBI database for *AKR1C1* and *AKR1C2* genes. We classified as SNDs all the SNPs reported in the SNP database that could not be found in a population of 100 non-related individuals and that had as variant allele the same nucleotide as the contig of the paralogue in the corresponding position (*i.e.* a SND corresponds to a mismatch in the alignment of the two genes). Of the 31 SNPs experimentally analyzed, we found 22 to be SNDs and 9 non-SNDs (Table 1).

The experimental and bioinformatic analyses were compared using the 31 SNPs experimentally analyzed (Table 1). If the experimental data above is used as a reference for the classification of SNPs as SNDs, the bioinformatic predictions have a low false positive rate (2 false positives among 9 experimentally confirmed non-SNDs) and a very low false negative rate (1 false negative among 22 experimentally confirmed SNDs).

PCR amplifications of exons containing the SNPs reported in Table 1 were performed using gene-specific primers (Supp. Table S1). PCR conditions used to produce gene-specific target regions were obtained through optimization of PCR reactions. The parameters adjusted were annealing temperature, $MgCl_2$ concentration, number of cycles, and annealing time.

In general, specific amplification is complicated when the target gene has one or more highly similar genes (paralogues) in the genome. Non-specific products, commonly known as bias, are usually due to annealing of the primers to regions in the genome different from the target. Promiscuous primers anneal non-specifically when their sequence has 100% identity with two or more chromosomal locations and/or when PCR conditions are such to favor non-specific primer annealing (*e.g.* high $MgCl_2$ and low annealing temperatures). The success of gene-specific amplification relies on primer design. In Fig. 4 we show how SNDs could originate when designing non-specific or “promiscuous” primers (Supp. Table S1), *i.e.* primers that do not discriminate between the two paralogous genes. In another experiment, we used specific primers for the *AKR1C1* gene and showed that even with good primer design one could obtain co-amplification (mixed PCR products) if PCR conditions are sub-optimal (Fig. 5).

In order to assess the potential practical implications of SNDs, we also examined the presence of SNDs for Illumina and Affymetrix SNP submissions to NCBI. The list of SNPs and their flanking sequence was downloaded from the dbSNP database. We found 35,785 SNPs from our set of biallelic coding SNPs on the Illumina array. Of these 582 were SNDs (1.6%). Similarly, we found 8,647 SNPs from our set on the Affymetrix array and, of these, 215 (2.5%) were SNDs. There are thus real, practical implications associated with the presence of SNDs in the database.

Moreover to identify specific examples of SNDs having been associated with disease, we examined several potential sources of information associating SNPs with disease: Online Mendelian Inheritance in Man (Amberger et al., 2009), the database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap), MedRedSNP (Rhee and Lee, 2009), the Office of Population Genetics (OPG) catalog of genome wide association studies (Hindorff et al., 2009), and SNPedia (www.snpedia.com). Of these, only the latter two provided information associating individual SNPs with disease in a format suitable for automated analysis. The OPG catalog contains SNP-trait associations with p-values $< 1 \times 10^{-5}$ drawn from PubMed literature searches and other sources. SNPedia provides information on associations between SNPs and disease drawn from a range of sources, including literature reports and manual extracts from other databases such as OMIM.

Of the 6,344 SNPs available for download from SNPedia (www.snpedia.com/files/gbrowse/snpedia) and the 1,979 SNPs available for download from the OPG catalog, there were 6,486 unique SNPs, indicating a significant overlap between the two sources of data. Of these 6,486 SNPs, 1,849 were among the biallelic, coding SNPs examined in this study. Of these, 50 were SNDs and 1799 were not SNDs. Thus 0.5% of SNPs identified as SNDs have a disease association in SNPedia compared with 1.7% of those SNPs identified as not SNDs.

The fifty SNDs were only found in the SNPedia database. Manual examination of the SNPedia records for the fifty SNDs, showed 32 with either cross-references to OMIM or direct references to the scientific literature for range of "disorders" varying from the benign (e.g. blue eye color) to severe, highly prevalent disorders such as melanoma, schizophrenia, age-related macular degeneration, and diabetes. Thus, even in this small set of disease-associated SNPs, practical issues resulting from the presence of SNDs in the database do occur.

Discussion

We have systematically investigated the SNP database (dbSNP at NCBI, Build 129) for contamination by "Single Nucleotide Differences" or SNDs (as a parallel to SNPs) which are artifacts due to the presence of a paralogue (highly similar duplicated) sequence in the genome (Fig. 1).

The presence of such spurious SNDs in the database at a surprisingly high frequency of up to 8.32% (Table 2) is a concern. We also provide experimental validation in two very similar paralogues, *AKR1C1* and *AKR1C2*, for SNDs. Fredman et al. (2004) previously reported the presence of "complex SNP-related sequence variation in segmental genome duplications". We not only experimentally confirm their findings in duplicons, but extend the analysis to whole-genome. Furthermore, we highlight the issue by proposing to name these recurrent SNPs as SNDs (Single Nucleotide Differences) to aid further discussions of this phenomenon. We think it is of fundamental importance to identify SNDs in paralogus genes, since, as was previously reported by Yandell (2008), the probability of finding disease-related SNPs is higher in paralogous genes.

We propose that SNDs may have found their way into the online SNP database by at least two distinct ways: 1. by uncritical bioinformatic alignments, e.g. BLAST, of highly similar but distinct DNA sequences and 2. by PCRs using "promiscous" primers that do not discriminate properly between very similar yet different DNA segments. Thus, investigators are encouraged to use appropriately stringent methods when identifying, submitting and validating SNPs.

The validation code in the SNP record provides some limited information on how the SNP was identified and on the experimental evidence confirming its existence. The validation code is a number from 0 to 31 indicating an increasing level of experimental evidence for the SNP, broadly divided into 6 validation methods (none, by cluster, by frequency, by submitter, by double hit, and by HapMap). If the SNPs are grouped according to their validation method, the percentage of SNDs in each group is *higher* than the overall SND percentage in most groups. It rises from 9% in the "by frequency" group to 34% in the "by submitter" group. Only the HapMap validation method shows a reduced prevalence of SNDs, with only 3% of the 49,261 SNPs in this group being SNDs. Thus, there is little correlation of validation code with SND status. Furthermore, while restricting the use of SNPs to those validated by HapMap will reduce the prevalence of SNDs, it comes at the cost of ignoring 56% of the real SNPs in the database whose validation code falls into one of the other categories.

Another potential method for eliminating SNDs is to remove those SNPs associated with the "Lee" submissions. Anecdotally, it has been assumed that many of these submissions to dbSNP are false, although we were unable to identify a published study verifying this assumption. However, as a number of the SNDs identified in *AKR1C1* and *AKR1C2* genes are based on data from the "Lee" submissions, we decided to explore this possibility further.

In order to do so, we searched dbSNP using the handle associated with the Lee submissions. There are three such sets of submitted SNPs comprising a total of 99,992 submitted SNPs. These correspond to 58,983 reference SNPs of which 10,344 were among the biallelic coding SNPs examined in this study. Of these, 1,991 were SNDs and 8,353 were not SNDs. Thus, 20% of the SNPs identified as being SNDs are Lee SNPs compared with 8% of those SNPs identified as being not SNDs.

This indicates that, while a substantial proportion of the SNDs are due to the “Lee” submissions, eliminating these will still leave 80% of the SNPs identified as being SNDs. Further, eliminating the “Lee” submissions would also remove almost 10% of SNPs we have identified as legitimate (not SND).

A manual examination of other non-structured text in the dbSNP records or the associated biomedical literature may help to reveal the source of specific SNDs, but it is not feasible to undertake such an analysis in a systematic way for all SNDs.

We note here that the presence of such SNDs in the database at a significant frequency of up to 8.32% can create serious problems in SNP-based association studies of candidate genes (Dvornyk et al., 2004) as well as in haplotype-based investigations (de Bakker et al., 2005). The former may be affected as one is not really investigating a SNP at all, the latter as SNDs would break haplotype blocks as well. One method of eliminating SNDs is through a careful sifting of paralogous gene SNPs to determine their validity. A thorough “housecleaning” of online SNP databases that will aid in the selection of correct SNPs and haplotype tag SNPs (htSNPs) that allow accurate association and linkage data.

We believe SNDs may systematically bias association studies toward a false (probably mostly negative) conclusion. Furthermore, if SNDs are used as tagSNPs they may disrupt haplotype blocks as noted above and also affect association studies, although the extent of this possibility will depend on a number of factors including the number of tagSNPs involved. Thus, SNDs should be carefully considered and eliminated from any such studies. Furthermore, there may also be clinical implications should SNDs be used as genomic biomarkers in personalized medicine or drug development. Such studies should pay great attention to choosing only legitimate SNPs, and regulatory submissions for diagnostic, therapeutic, or preventative methods and technologies based on SNP data should be thoroughly examined for the presence of erroneous SNPs.

In conclusion, significant efforts should be made toward purging such SNPs from all online databases. SNPs in gene families with high frequencies (*e.g.* above 0.3), that are not in Hardy-Weinberg equilibrium in association studies ought to be considered as “suspect” or candidate SNDs. In the meantime, researchers, especially with negative association studies, may wish to consider carefully whether they may have inadvertently genotyped SNDs. Finally, we note that association results should be supported by functionally relevant data on such genotyped SNPs and/or haplotypes to confirm that the results are truly causal (Mehrian-Shai and Reichardt, 2004).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NCI grant P01 CA108964 (project 1)to JKVR who is also a Medical Foundation Fellow at the University of Sydney. We thank Sebastien Gerega (USyd) for helpful contributions regarding the bioinformatic procedure; Francine Marques (USyd), Sarah Curtis (USyd), and Raffaele Ottaviano for

their excellent assistance in revising the manuscript; Robin Leach (UTHSCSA) as well as Danielle Dye and Nigel Laing (WAIMR) for helpful discussions; the editors and reviewers of *Human Mutation* for their most gracious and helpful input; and David Handelsman and Gideon Sartorius (USyd) for providing genomic DNA samples and population related information.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 2009; 37:D793–D796. [PubMed: 18842627]

Britten RJ. Almost all human genes resulted from ancient duplication. PNAS. 2006; 103:19027–19032. [PubMed: 17146051]

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005; 37:1217–1223. [PubMed: 16244653]

Dvornyk V, Long JR, Xiong DH, Liu PY, Zhao LJ, Shen H, Zhang YY, Liu YJ, Rocha-Sanchez S, Xiao P, Recker RR, Deng HW. Current limitations of SNP data from the public domain for studies of complex disorders: a test for ten candidate genes for obesity and osteoporosis. BMC Genet. 2004; 5:4. [PubMed: 15113403]

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet. 2004; 36:861–866. [PubMed: 15247918]

Gut IG, Lathrop GM. Duplicating SNPs. Nat Genet. 36:861–866. [PubMed: 15247918]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci. 2009; 106:9362–9367. [PubMed: 19474294]

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005; 110:462–467. [PubMed: 16093699]

Kruglyak, Leonid. The road to genome-wide association studies. Nat Genet. 2008; 9:314–318.

Mehrian-Shai R, Reichardt JKV. A Renaissance of "Biochemical Genetics"? SNPs, Haplotypes, Function and Complex Diseases. Molec Genet Metabol. 2004; 83:47–50.

Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. Bioinformatics. 2004; 20:1022–32. [PubMed: 14764571]

Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. Nat Genet. 2003; 33:457–458. [PubMed: 12652301]

Rhee H, Lee JS. MedRefSNP: a database of medically investigated SNPs. Hum Mutat. 2009; 30:E460–E466. [PubMed: 19105187]

Sachidanandam R, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409:928–933. [PubMed: 11237013]

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]

Suh Y, Vijg J. SNP discovery in associating genetic variation with human disease phenotypes. Mutat Res. 2005; 573:41–53. [PubMed: 15829236]

Yandell M, Moore B, Salas F, Mungall C, MacBride A, White C, Reese MG. Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. PLoS Comput Biol. 2008; 4:e1000218. [PubMed: 18989397]

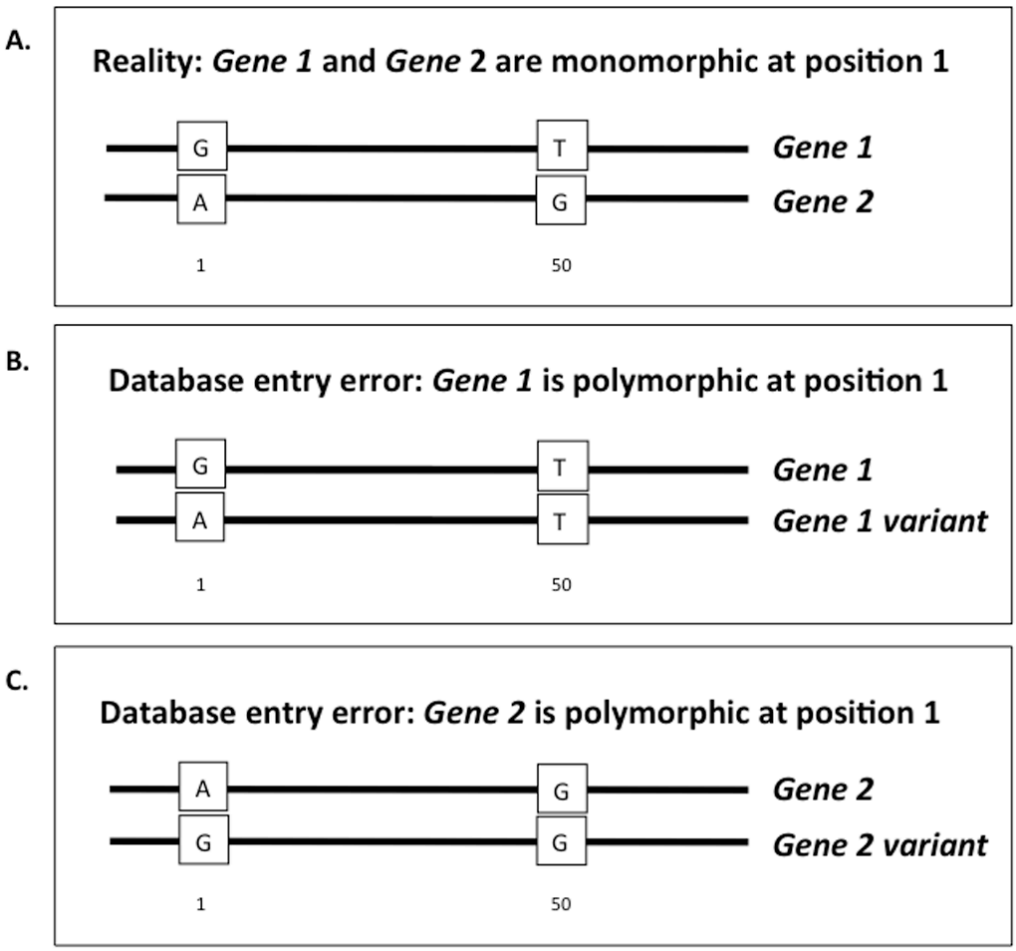

Figure 1.

The origin of "Single Nucleotide Differences" or SNDs in paralogous genes.

In panel **A.** *gene 1* and *gene 2* are highly similar paralogous genes, they differ in one or more nucleotides. In this simple example, the two genes differ in two positions: position 1 and position 50. These positions refer to two residues occupying the corresponding positions in both genes when the genes are aligned. Whenever the two genes are co-amplified because of poor primer design or suboptimal PCR conditions, or whenever uncritical sequence alignments are performed, the sequence of the mixed PCR product or the mismatch in the alignment could be mistakenly interpreted as a SNP for *gene 1*, panel **B**, or as a SNP for *gene 2*, panel C and therefore incorrectly reported in the database.

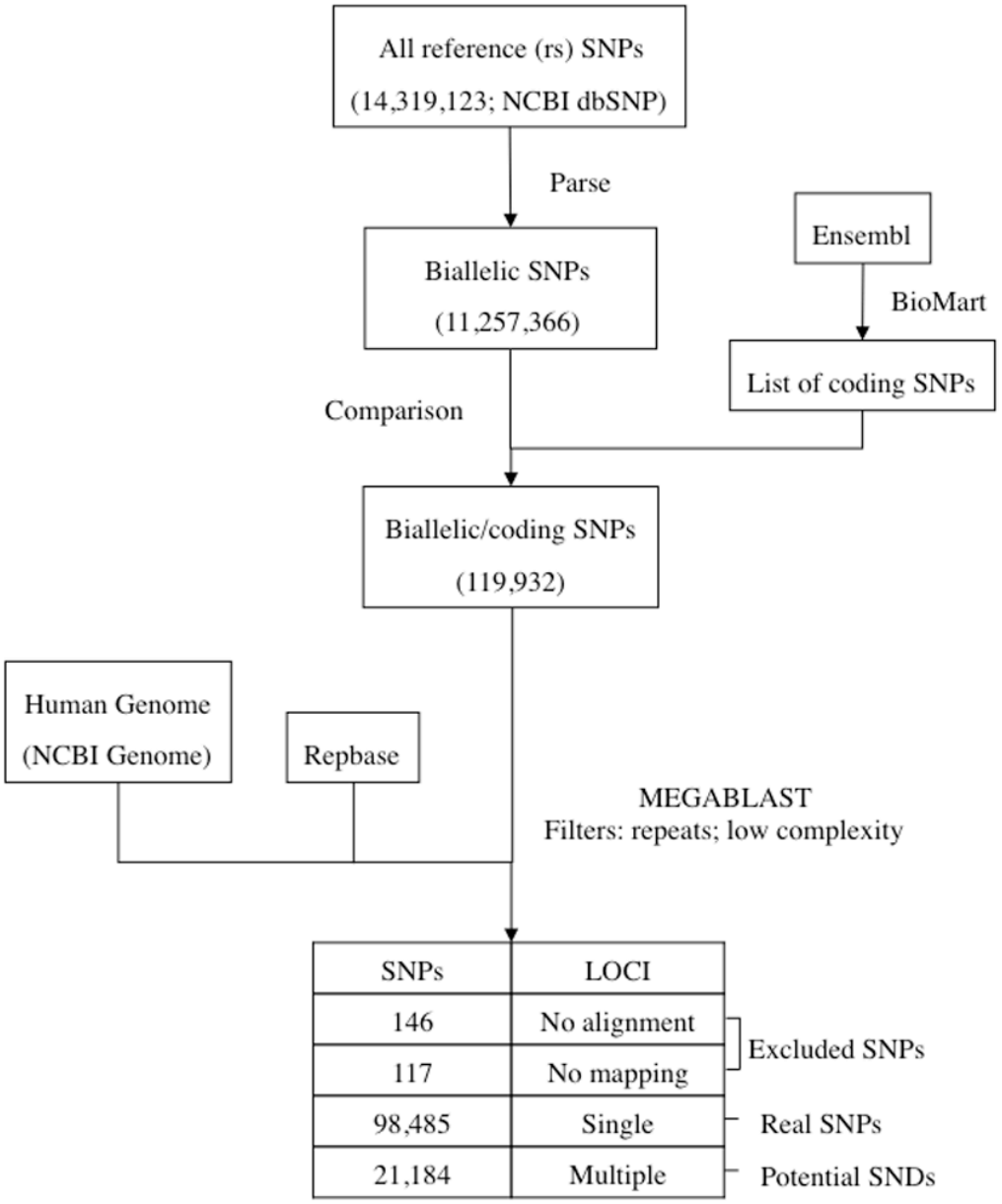

Figure 2.

Bioinformatic procedure for the identification of SNDs.

The chart diagram describes step-by-step the procedure used to identify SNDs in the NCBI dbSNP database, Build 129. Only BLAST hits with 90% SID and 20% SC containing the SNP were selected as mapped loci.

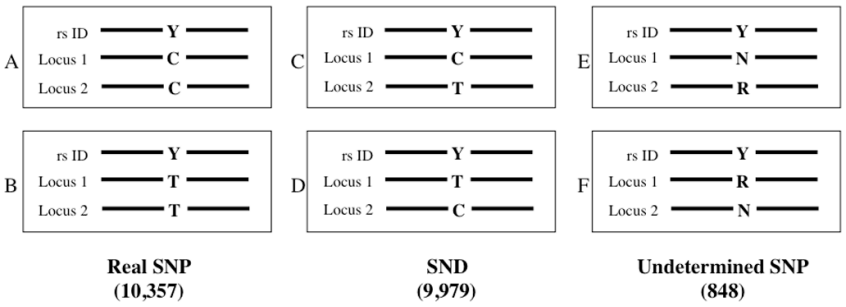

Figure 3.

Analysis of the 21,184 SNPs aligning with multiple mapping loci to determine the SND status. The figure shows an example in which the sequence of the reported SNP (rs ID) aligns to two mapped loci (Locus 1 and Locus 2) with at least 90% SID and 20% SC. The reported SNP is Y (C/T). The left part of the figure (panels A and B) shows two mappings resulting in the SNP being classed as “real”. The middle part (panels C and D) shows two mappings resulting in the SNP being classed as “SND”. The right part (panels E and F) shows two examples of mappings resulting in the SNP being classed as “undetermined”. Other combinations are possible as well.

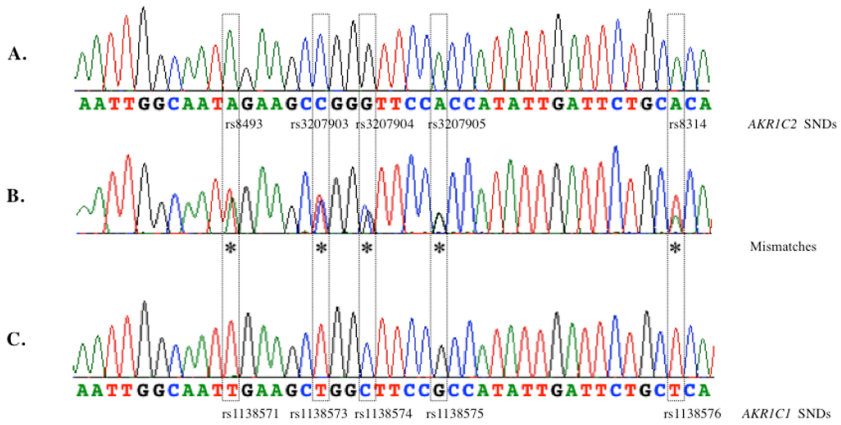

Figure 4.

SNDs in exon 2 of the *AKR1C1* gene and exon 4 of the *AKR1C2* gene.

The sequence in panel **A** and **C** was obtained using *AKR1C2* and *AKR1C1* specific primers, respectively. The two sequences differ at nucleotide positions indicated with an asterisk in panel **B.** The sequence in panel **B** was obtained using non-specific primers *i.e.* primers amplifying both exon 2 and exon 4 of the two genes, producing a mixed PCR product. We also show that *AKR1C2* and *AKR1C1* SNDs are located exactly where the mismatches between the two genes are. These SNPs are included in Table 1.

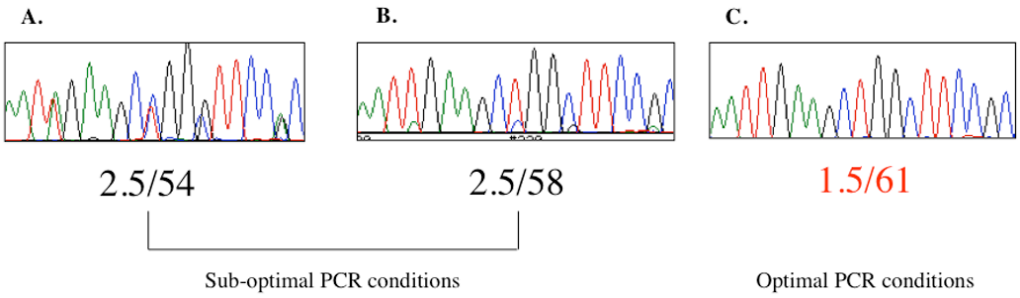

Figure 5.

AKR1C1 and *AKR1C2* co-amplification of exon 2 and exon 4 respectively in suboptimal PCR conditions using *AKR1C1* specific primers for *AKR1C1* exon 2 (Supp. Table S1). Panel A and B show electropherograms of mixed *AKR1C1/AKR1C2*-PCR products using sub-optimal PCR conditions: 2.5mM $MgCl_2$ and 54°C annealing temperature (A) and 2.5mM $MgCl_2$ and 58°C annealing temperature (B). On the other hand panel C shows the *AKR1C1*-specific PCR product produced using 1.5mM $MgCl_2$ and 61°C annealing temperature. Thus, with decreasing $MgCl_2$ concentration and increasing annealing temperature the double peaks become less prominent (A and B) and even completely eliminated (C).

Hum Mutat. Author manuscript; available in PMC 2011 January 1.

Table 1

Reported SNPs (as of 6/13/08) and SNDs for *AKR1C1* and *AKR1C2* paralogous genes.

Paralogue genes	rs#	Nucleotide Variation	Contig Nucleotide of Paralogue	SND Exp	SND Bio
AKR1C1	rs1138566	$\vec{T}C$	C	YES	YES
	rs1138567	$\vec{T}C$	C	YES	YES
	rs1138569	$\vec{A}G$	G	YES	YES
	rs1138570	$\vec{C}T$	T	YES	YES
	rs1138571	$\vec{T}A$	A	YES	YES
	rs1138573	$\vec{T}C$	C	YES	YES
	rs1138574	$\vec{C}G$	G	YES	YES
	rs1138575	$\vec{G}A$	A	YES	YES
	rs1138576	$\vec{T}A$	A	NO	YES
	rs11548049	$\vec{T}C$	C	YES	YES
	rs15986	$\vec{C}T$	T	YES	YES
	rs1138600	$\vec{G}A$	A	YES	NO
	rs59728642				
	rs56276325	$\vec{C}T$	C	NO−	NO
	rs17295755* rs17354222	$\vec{G}A$	A	NO	YES
	rs17354444#	$\vec{A}T$	T	YES	YES
	rs7097713	$\vec{G}A$	G	NO−	NO
AKR1C2	rs17341906∞ rs8625	$\vec{C}T$	T	YES	YES
	rs3207898	$\vec{G}A$	A	YES	YES
	rs3207901 rs17409350	$\vec{T}C$	C	YES	YES
	rs17341878 rs8493	$\vec{A}T$	T	YES	YES
	rs3207903	$\vec{C}T$	T	YES	YES
	rs3207904	$\vec{G}C$	C	YES	YES

Hum Mutat. Author manuscript; available in PMC 2011 January 1.

Paralogue genes	rs#	Nucleotide Variation	Contig Nucleotide of Paralogue	SND Exp	SND Bio
	rs2854482	$\vec{T}A$	T	NO+	NO
	rs3207905	$\vec{A}G$	G	YES	YES
	rs8314	$\vec{A}T$	T	YES	YES
	rs2854486	$\vec{T}C$	T	NO−	NO
	rs17344137[°] rs17356858	$\vec{T}C$	C	YES	YES
	rs56259037	$\vec{A}G$	G	NO	NO
	rs10618	$\vec{A}G$	G	YES	YES
	rs2518042	$\vec{A}G$	A	NO−	NO
	rs2518043	$\vec{A}G$	A	NO−	NO

The “SND Exp” column indicates if the reported SNP was found to be a SND by experimental data:

YES: the reported SNP is a SND because the variant allele is the same as the contig nucleotide of the paralogue in the corresponding position and the SNP is not found in a population of 100 non-related individuals.

NO: the reported SNP is not a SND because the variant allele is the same as the contig nucleotide of the paralogue in the corresponding position, suggesting the possibility for the SNP to be a SND, but the SNP is found (at a low frequency) in a population of 100 non-related individuals. The low frequencies of these real SNPs, which are also mismatches with the paralogous gene, are probably due to the fact that the SNP existed before the ancestral gene-duplication event.

NO−: the reported SNP is not a SND because the allele is not the same as the contig nucleotide of the paralogue in the corresponding position and the SNP is not found in a population of 100 non-related individuals. The SNP could be present in other populations or could have ambiguous chromosomal location.

NO+: the reported SNP is not a SND because the allele is not the same as the contig nucleotide of the paralogue in the corresponding position and the SNP is found in a population of 100 non-related individuals.

[*] Heterozygosity reported in NCBI is 0.499, while we found 1 heterozygote individual among 100 people.

[#] Heterozygosity reported in NCBI is 0.435. The SNP was not found in our population.

[∞] Heterozygosity reported in NCBI is 0.495. The SNP was not found in our population.

[°] Heterozygosity reported in NCBI is 0.500. The SNP was not found in our population.

The “SND Bio” column indicates if the reported SNP was found to be a SND using the bioinformatics procedure.

Table 2

Summary of the SNPs analyzed using bioinformatics and those found to be SNDs

	Number of SNPs	% of SNPs
SNPs analyzed	119,932	100
SNDs	9,979	8.32
Strong SNDs	456	0.38
Very Strong SNDs	189	0.16