BMC Research Notes

BioMed Central

Correspondence

The need for genetic variant naming standards in published abstracts of human genetic association studies

Wei Yu*, Renée Ned, Anja Wulf, Tiebin Liu, Muin J Khoury and Marta Gwinn

Address: Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia 30341, USA

Email: Wei Yu* - WYu@cdc.gov; Renée Ned - RNed@cdc.gov; Anja Wulf - AWulf@cdc.gov; Tiebin Liu - TLiu@cdc.gov; Muin J Khoury - MKhoury@cdc.gov; Marta Gwinn - MGwinn@cdc.gov

* Corresponding author

Published: 14 April 2009

BMC Research Notes 2009, **2**:56 doi:10.1186/1756-0500-2-56

Received: 1 December 2008
Accepted: 14 April 2009

This article is available from: http://www.biomedcentral.com/1756-0500/2/56

Abstract

We analyzed the use of RefSNP (rs) numbers to identify genetic variants in abstracts of human genetic association studies published from 2001 through 2007. The proportion of abstracts reporting rs numbers increased rapidly but was still only 15% in 2007. We developed a web-based tool called Variant Name Mapper to assist in mapping historical genetic variant names to rs numbers. The consistent use of rs numbers in abstracts that report genetic associations would enhance knowledge synthesis and translation in this field.

Discussion

By identifying millions of single nucleotide polymorphisms (SNPs), high-throughput genotyping technology has dramatically boosted the yield of genetic association studies [1]. Translating these data into useful health information depends on systematic review and knowledge synthesis [2]. However, the inconsistent description of key data elements – such as gene names, gene variant names, and measures of association – makes retrieval of published information challenging. Names for genes and polymorphisms are particularly problematic because historical or common names have often been used instead of standard nomenclature [3,4], particularly in candidate gene association studies.

The National Library of Medicine (NLM) provides free access via PubMed [5] to the most comprehensive repository of biomedical literature abstracts in the world. Thus, the efficiency and sensitivity of scientific literature searches, as well as the robustness of computerized processes for data and text mining, depend closely on the way that information is presented in PubMed abstracts. By using standard names for genes and genetic variants in published abstracts, authors can increase the accessibility, utility, and influence of their findings.

The Human Genome Epidemiology (HuGE) Navigator is an integrated and searchable knowledge base of human genetic associations that have been extracted from PubMed weekly since 2001 by a combination of automatic and manual processes [6]. The curator indexes each new abstract with the relevant HUGO gene symbol(s) [4], so that users can perform gene-specific queries that can also accommodate gene aliases or protein names. For systematic review and synthesis of gene-disease associations, more specific data – at the level of the genetic variant – are required. The National Center for Biotechnology Information (NCBI) has developed the SNP database (dbSNP) [7] as a central repository for SNPs and other genetic variants, each of which is identified by a unique reference cluster number (rs number).

We examined with the HuGE Navigator trends in the reporting of gene variants and odds ratios in PubMed

abstracts that were published from 2001 through 2007 (N = 27,132). Overall, 6.3% of abstracts reported rs numbers; 27% reported odds ratios. The proportion of abstracts reporting rs numbers increased substantially (from 1% to 17%) during this period, while the proportion reporting odds ratios remained fairly steady (Fig. 1). Abstracts for genome-wide association studies were more likely than other genetic association studies to include rs numbers (42%) and odds ratios (40%). Conversely, we selected a random 2% sample of all of the extracted PubMed abstracts for hand searching and found that almost all (91%) included common or historical genetic variant names. Matching these common names to the corresponding rs numbers would greatly aid in retrieval and synthesis of genetic association data.

To facilitate the mapping of historical names for genetic variants to their rs numbers, we developed a searchable, web-based database called Variant Name Mapper [8]. This database contains historical names matched with their corresponding rs numbers. These data have been extracted from multiple open-access databases, including: SNP500Cancer [9], SNPedia [10], pharmGKB [11], ALFRED [12], AlzGene [13], PDGene [14], SZgene [15], and LSDBs [16], as well as from our own curated data from the HuGE Navigator. User submissions are also welcome. In the Variant Name Mapper, the user is able to search by historical (common) name of the polymorphism, by rs number, or by gene information (including gene symbol, gene name, and gene alias). The display information includes rs number, common/historical polymorphism names, gene-centered information, and a listing of the data sources [Figure 2]. We evaluated the tool's mapping capacity by entering the common names for genetic variants included in the 2% sample of abstracts described above. Overall, 62% of common names could be mapped to an rs number by using the Variant Name Mapper. This low return may be due to the heterogenous nature of the common names and limitations of the data sources. The content of the database will be continually improved and expanded as new data sources become available.

Genome-wide bioinformatics tools, such as HapMap [17] and the UCSC Genome Browser [18], are most useful to researchers for mining genomic information when data can be linked at the variant level. The Human Genome Variation Society (HGVS) has proposed a comprehensive and systematic nomenclature for the description of genetic variants [19]. The combination of dbSNP acces-

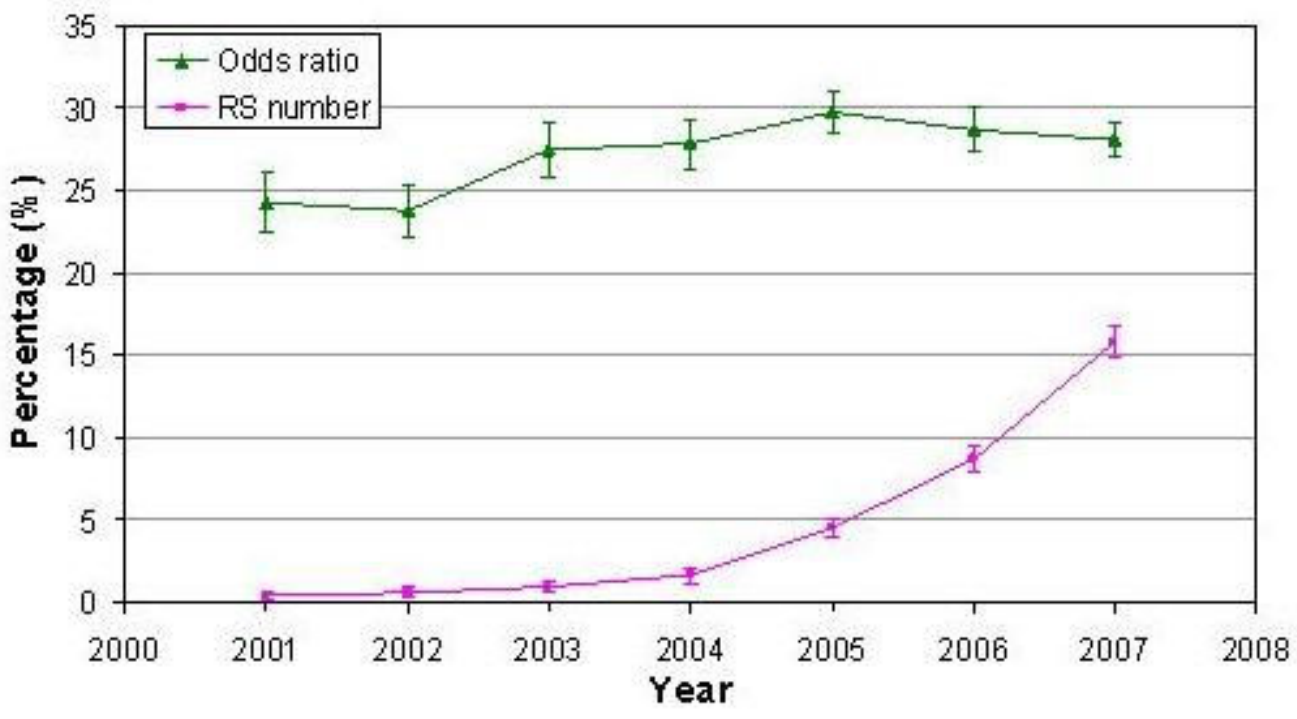
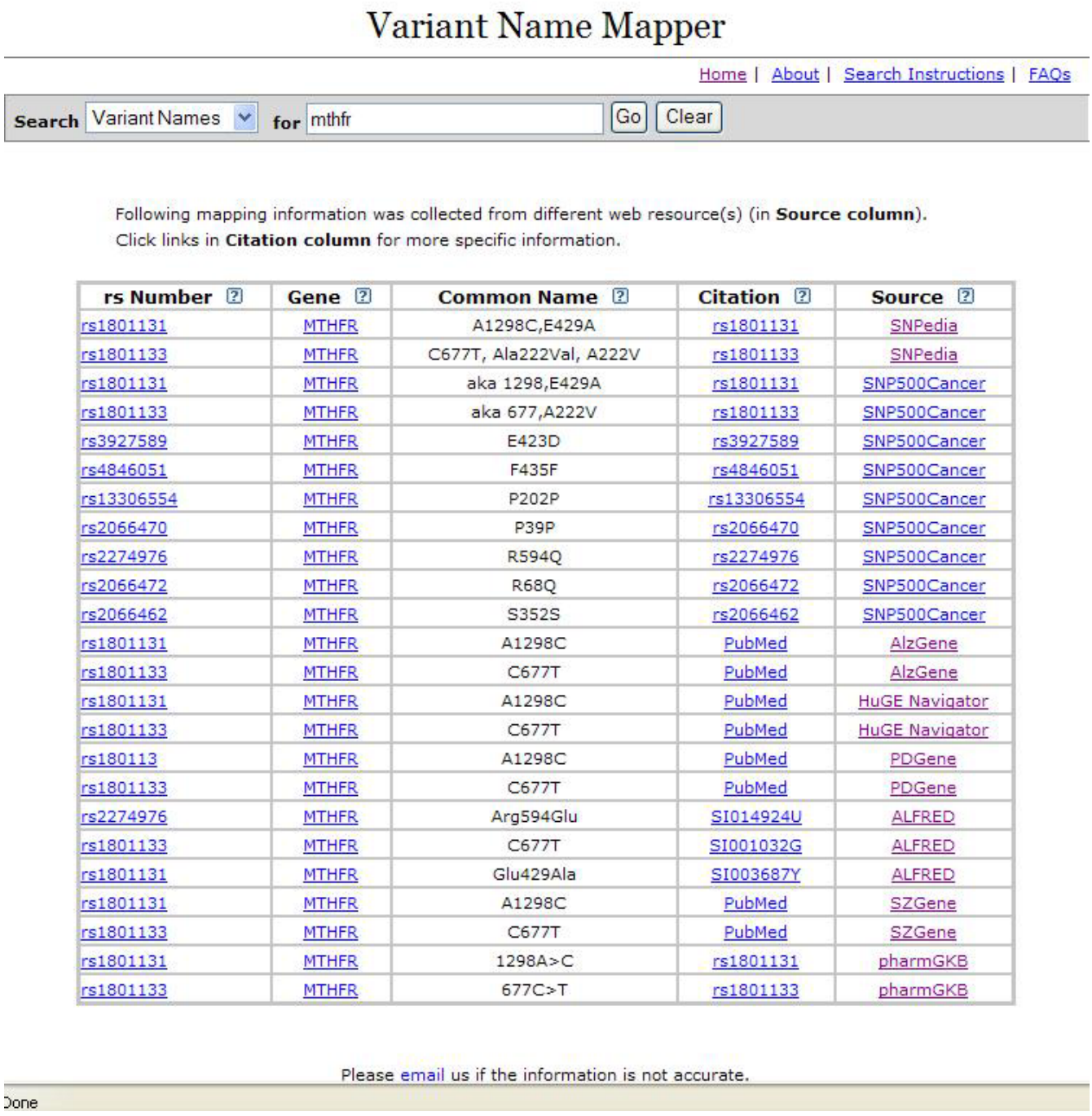


Figure 1
Trends in the percentage of abstracts reporting odds ratios and rs numbers for gene variants, HuGE Navigator database, 2001–2007.



Variant Name Mapper

Home | About | Search Instructions | FAQs

Search Variant Names for mthfr Go Clear

Following mapping information was collected from different web resource(s) (in **Source column**).
Click links in **Citation column** for more specific information.

rs Number	Gene	Common Name	Citation	Source
rs1801131	MTHFR	A1298C,E429A	rs1801131	SNPedia
rs1801133	MTHFR	C677T, Ala222Val, A222V	rs1801133	SNPedia
rs1801131	MTHFR	aka 1298,E429A	rs1801131	SNP500Cancer
rs1801133	MTHFR	aka 677,A222V	rs1801133	SNP500Cancer
rs3927589	MTHFR	E423D	rs3927589	SNP500Cancer
rs4846051	MTHFR	F435F	rs4846051	SNP500Cancer
rs13306554	MTHFR	P202P	rs13306554	SNP500Cancer
rs2066470	MTHFR	P39P	rs2066470	SNP500Cancer
rs2274976	MTHFR	R594Q	rs2274976	SNP500Cancer
rs2066472	MTHFR	R68Q	rs2066472	SNP500Cancer
rs2066462	MTHFR	S352S	rs2066462	SNP500Cancer
rs1801131	MTHFR	A1298C	PubMed	AlzGene
rs1801133	MTHFR	C677T	PubMed	AlzGene
rs1801131	MTHFR	A1298C	PubMed	HuGE Navigator
rs1801133	MTHFR	C677T	PubMed	HuGE Navigator
rs180113	MTHFR	A1298C	PubMed	PDGene
rs1801133	MTHFR	C677T	PubMed	PDGene
rs2274976	MTHFR	Arg594Glu	SI014924U	ALFRED
rs1801133	MTHFR	C677T	SI001032G	ALFRED
rs1801131	MTHFR	Glu429Ala	SI003687Y	ALFRED
rs1801131	MTHFR	A1298C	PubMed	SZGene
rs1801133	MTHFR	C677T	PubMed	SZGene
rs1801131	MTHFR	1298A>C	rs1801131	pharmGKB
rs1801133	MTHFR	677C>T	rs1801133	pharmGKB

Please email us if the information is not accurate.

Done

Figure 2
A screenshot of the Variant Name Mapper.

sion identifiers (rs numbers) with HGVS nomenclature will be beneficial for standardization. The use of standard nomenclatures (e.g., HUGO for genes, dbSNP for gene variants) and systematic reporting of statistics (e.g., odds ratios) in published abstracts would represent an evolutionary advance in information integration and retrieval, which are the first steps in translating genomic research.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WY drafted the manuscript, and designed and implemented the mapping tool, wrote the source codes. RN was involved in the data extraction and curation and helped in manuscript preparation. AW was involved in the data extraction and the data quality control. TL performed the data preparation and analysis. MJK oversaw the project and revised the draft manuscript. MG provided advice on the project and revised the draft manuscript and led the project. All authors read and approved the final document.

Acknowledgements

We appreciate valuable comments from Donna Maglott. Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of CDC.

References

1. Kim S, Misra A: **SNP genotyping: technologies and biomedical applications.** *Annu Rev Biomed Eng* 2007, **9(289–320):**289-320.
2. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L: **The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?** *Genet Med* 2007, **9:**665-674.
3. Smigielski EM, Sirotkin K, Ward M, Sherry ST: **dbSNP: a database of single nucleotide polymorphisms.** *Nucleic Acids Res* 2000, **28:**352-355.
4. **HUGO Gene Nomenclature** [http://www.gene.ucl.ac.uk/nomenclature]
5. **PubMed** [http://www.ncbi.nlm.nih.gov/entrez]
6. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: **A navigator for human genome epidemiology.** *Nat Genet* 2008, **40:**124-125.
7. **dbSNP** [http://www.ncbi.nlm.nih.gov/projects/SNP/]
8. **Variant Name Mapper** [http://www.hugenavigator.net/HuGENavigator/startPageMapper.do]
9. **SNP500Cancer** [http://snp500cancer.nci.nih.gov/home_1.cfm]
10. **SNPedia** [http://www.snpedia.com/index.php/SNPedia]
11. **pharmGKB** [http://www.pharmgkb.org/]
12. **ALFRED** [http://alfred.med.yale.edu/alfred/]
13. **AlzGene** [http://www.alzforum.org/res/com/gen/alzgene/default.asp]
14. **PDGene** [http://www.pdgene.org/]
15. **SZgene** [http://www.schizophreniaforum.org/res/sczgene/default.asp]
16. **LSDBs** [http://www.hgvs.org/dblist/glsdb.html]
17. **The International HapMap Project.** *Nature* 2003, **426:**789-796.
18. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, *et al.*: **The UCSC Genome Browser Database: update 2006.** *Nucleic Acids Res* 2006, **34:**D590-D598.
19. den Dunnen JT, Antonarakis SE: **Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion.** *Hum Mutat* 2000, **15:**7-12.