Genetics inMedicine | BRIEF REPORT

Inferred inheritance of MorbidMap genes without OMIM clinical synopsis

Aamina Shakir, BA[1], Michael Ripperger[2], Zhijie Jiang PhD[3] and Klaas J. Wierenga, MD[4]

Purpose: The Genomic Oligoarray and SNP Array Evaluation Tool 3.0 matches candidate genes within regions of homozygosity with a patient's phenotype, by mining OMIM for gene entries that contain a Clinical Synopsis. However, the tool cannot identify genes/disorders whose OMIM entries lack a descriptor of the mode of (Mendelian) inheritance. This study aimed to improve the tool's diagnostic power by building a database of autosomal recessive diseases not diagnosable through OMIM.

Methods: We extracted a list of all genes in OMIM that produce disease phenotypes but lack Clinical Synopses or other statements of mode of inheritance. We then searched PubMed for literature regarding each gene in order to infer its inheritance pattern.

Results: We analyzed 1,392 genes. Disorders associated with 372 genes were annotated as recessive and 430 as dominant. Autosomal genes were ranked from 1 to 3, with 3 indicating the strongest evidence behind the inferred mode of inheritance. Of 834 autosomal genes, 158 were ranked as 1, 228 as 2, and 448 as 3.

Conclusion: The 372 genes associated with recessive disorders will be contributed to the SNP array tool, and the entire database to OMIM. We anticipate that these findings will be useful in rare disease diagnostics.

Genet Med advance online publication 24 August 2017

Key Words: Clinical Synopsis; OMIM; MorbidMap; regions of homozygosity; SNP array tool

INTRODUCTION

When evaluating a patient with a known or suspected genetic disorder, determining the associated gene mutation(s) is essential to establishing or confirming the diagnosis. The Genomic Oligoarray and SNP Array Evaluation Tool 3.0 (http://firefly.ccs.miami.edu/cgi-bin/ROH/ROH_analysis_tool.cgi) is an online program that provides a novel means of expediting this process for autosomal-recessive disorders.[1,2] First, if a patient presents with a genetic disease in the setting of known or suspected consanguinity, a single-nucleotide polymorphism (SNP) array is obtained. Generally, the more closely related a patient's parents are, the more regions of homozygosity (ROH) that patient's genome will contain, and the longer they will be. If ROH are found, the tool identifies candidate recessive disorders that (i) are associated with genes that map to any of the ROH and (ii) match the patient's phenotype. For example, inputting the clinical feature "microcephaly," using OMIM terminology or Human Phenotype Ontology[3] terms, will generate a list of genes associated with disorders having microcephaly as a characteristic.[4] Multiple search terms can be entered simultaneously, using Boolean operators "AND," "OR," and "NOT," to most accurately capture the patient's clinical presentation. The tool thus narrows down the list of genes mapped to the patient's ROH to only those that produce the phenotypes described by the search terms. The results can be further refined to include only genes that cause autosomal-recessive disorders. The final list of candidate genes/disorders serves as the differential diagnosis.

An important feature of the SNP array tool is that it relies on the Clinical Synopsis feature of the OMIM database.[5] Many genes have OMIM profiles that describe gene structure and function, and relevant peer-reviewed publications referenced in PubMed. If a gene causes a disorder when mutated, the disorder is assigned its own OMIM entry. However, because of OMIM's slow curation and documentation process, not all genes with associated disease phenotypes are documented as having them. Furthermore, only some of the entries on documented genetic disorders have an additional Clinical Synopsis link. The Clinical Synopsis is a summary of the most pertinent information about a gene, including locus, associated phenotype/disease, and inheritance pattern. Depending on the search criteria, a gene is often "visible" to the SNP array tool only if its OMIM profile has a Clinical Synopsis. Thus, most genes that lack a Clinical Synopsis will not appear in the tool's list of candidate disease loci.

An example of an autosomal-recessive disorder without a Clinical Synopsis is distal renal tubular acidosis (OMIM 602722).[5] This disorder's OMIM profile describes its clinical features and causative gene mutation (*ATP6V0A4*), but the lack of a Clinical Synopsis means that the SNP array tool would not identify it as a possible diagnosis if *ATP6V0A4* fell

[1]Department of Pediatrics, Section of Genetics, University of Oklahoma College of Medicine, Oklahoma City, Oklahoma, USA; [2]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA; [3]Center for Computational Studies, University of Miami, Miami, Florida, USA; [4]Department of Pediatrics, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA. Correspondence: Aamina Shakir (Aamina-Shakir@ouhsc.edu)

Submitted 7 March 2017; accepted 27 June 2017; advance online publication 24 August 2017. doi:10.1038/gim.2017.131

within a patient's ROH. When this project was completed in July 2015, the most striking example of a well-known autosomal-recessive disease without a Clinical Synopsis was sickle cell anemia. Thus, although responsible for one of the most classic, well-researched recessive diseases, the *HBB* mutation theoretically would have gone undetected by the SNP array tool at that time. (A Clinical Synopsis was added to the *HBB*/sickle cell anemia profile on 27 May 2016.[6] It is important to note that the date of last edit is listed at the end of every OMIM entry.) Although OMIM continues to update and edit its database, it is clear that the incomplete annotations diminish the SNP array tool's ability to identify all possible disease loci within a given ROH. This decreases the tool's diagnostic power, and thus utility. Therefore, the purpose of this project is to build a database of autosomal-recessive diseases that the SNP array tool would fail to diagnose due to their causative genes' lack of Clinical Synopses in OMIM, and to contribute this data set to the tool—and, ultimately, to OMIM.

MATERIALS AND METHODS

Information about all the genes in the human genome without Clinical Synopses was extracted from the OMIM database. The genes known to produce a clinical disease phenotype ("MorbidMap genes") were then compiled into a spreadsheet. This list was generated 12 June 2015. It contained 2,273 entries, each consisting of the name of a gene, the associated disease, and a hyperlink to the gene's OMIM profile. Every MorbidMap gene was then individually researched by searching through the PubMed articles cited in its OMIM profile. This research process was conducted solely by both authors of this report. The information gleaned from the researched articles was then synthesized to infer the gene's inheritance pattern. For example, if the literature on a particular gene showed that multiple patients in various studies were homozygous for mutations in that gene, had the same clinical presentation, and had consanguineous ancestors, it would be reasonable to infer that a mutation in this gene produces an autosomal-recessive disease. Any disagreements on inferred inheritance patterns were resolved by consensus meeting between the authors.

Two scripts were developed to facilitate the researching process. The first script scraped the text of relevant OMIM Web pages; searched for preselected relevant keywords including "homozygous," "heterozygous," "compound heterozygous," "dominant," and "recessive"; searched for cited PubMed hyperlinks; and repeated this process of text scraping and keyword finding for the extracted PubMed Web pages. The second script highlighted the keywords on the relevant PubMed Web pages by pulling keyword arguments from hyperlinks created by the first script. The first script was written in Python and utilized BeautifulSoup and Pandas.[7,8] The second script was written in JavaScript for the Tampermonkey Chrome extension with the help of published highlighting code.[9]

These scripts greatly improved the efficiency of the research process, as the highlighted keywords quickly identified relevant articles. Each publication containing pertinent keywords was searched through PubMed initially, and then, if necessary, through other online journal databases, including Elsevier, Wolters Kluwer, and JSTOR. Relevant information (such as pedigrees, whole-exome scans, and linkage analyses) was consolidated for every MorbidMap gene, and used to infer the inheritance pattern of each associated disease.

As each MorbidMap gene was studied, three additional columns of information were added to its spreadsheet entry: one denoting the inferred inheritance pattern, one for supporting evidence, and one with the PubMed ID of every researched publication. The supporting evidence column listed the number of studies in the cited publications, the number of patients and families researched across those studies, genome and linkage study results, and patterns found in pedigrees and/or patient history (e.g., vertical transmission, consanguinity). Genes that conferred susceptibility rather than specific clinical disease were excluded from analysis.

After all the MorbidMap genes were annotated with a putative inheritance pattern, the genes determined to be autosomally inherited were ranked from 1 to 3 based on the strength of the supporting evidence, with a ranking of 1 demonstrating the weakest evidence, and 3 demonstrating the strongest. A gene received a ranking of 1 if it caused disease in three or fewer unrelated individuals, or in two or fewer families; a ranking of 2 if it caused disease in four to 10 unrelated individuals or three to five families; and a ranking of 3 if it caused disease in more than 11 unrelated individuals or six families. However, a gene received a ranking of 2 if it caused disease in 10 or more relatives within one to two families, and a ranking of 3 if it caused disease in 10 or more relatives within three to five families. Genes associated with diseases that had multiple modes of inheritance were separately ranked for each inheritance pattern. These criteria were arbitrarily set, with the purpose of helping OMIM and SNP array tool users exercise clinical judgment when adding MorbidMap genes to their databases.

RESULTS

Of the 2,273 MorbidMap genes, 879 conferred susceptibility only, and were excluded from analysis. Of the 1,392 remaining genes that were researched and annotated, 372 were inferred to be associated with autosomal-recessive diseases, and 430 with autosomal-dominant diseases. Sixteen genes were found to be associated with both autosomal-dominant and autosomal-recessive diseases. The remaining 590 genes either lacked enough information to infer an inheritance pattern or were associated with disorders annotated as having nonautosomal inheritance. These results are summarized in **Table 1**.

The next step in analyzing the results was to gauge what proportion of autosomally inherited diseases the SNP array tool would fail to identify. First, the search terms "autosomal dominant" and "autosomal recessive" were run through the

Table 1 Analysis of MorbidMap genes

	Researched genes	Clinical synopsis genes	% Increase in total number of genes recognizable by SNP array tool
AR only	372	1,524	24.4%
AD only	430	1,135	37.9%
AD and AR	16	N/A	N/A
X-linked and Y-linked	144	N/A	N/A
Other (e.g., susceptibility, multifactorial, undetermined)	1,310	N/A	N/A
Total	2,273	12,822	17.7%

AD, autosomal dominant; AR, autosomal recessive; N/A, not available; SNP, single-nucleotide polymorphism.

SNP array tool, with the number of results representing the number of disease-causing genes “visible” to the tool. These numbers were then compared to the numbers of MorbidMap genes without Clinical Synopses that were inferred to have autosomal-dominant and -recessive inheritance, and would thus be “invisible” to the tool. Adding the 372 MorbidMap genes inferred to cause autosomal-recessive disease to the 1,524 genes annotated by OMIM as causing recessive disease resulted in a total of 1,896 recessive diseases diagnosable by the SNP array tool—an increase of 24.4%. Likewise, adding the 430 MorbidMap genes inferred to cause autosomal-dominant disease to the 1,135 genes annotated in OMIM as causing dominant disease resulted in a total of 1,565 diagnosable dominant diseases—an increase of 37.9%. These results are summarized in **Table 1**.

Next, the genes inferred to have autosomal inheritance were ranked according to the system outlined in Materials and Methods. Of the 430 genes annotated as having autosomal-dominant inheritance, 61 received a ranking of 1, while 108 received a ranking of 2 and 261 received a ranking of 3. Of the 372 genes annotated as having autosomal-recessive inheritance, 81 received a ranking of 1, 113 received a ranking of 2, and 178 received a ranking of 3. The 16 genes annotated as having both dominant and recessive inheritance were ranked twice, once for each mode of transmission. For autosomal-recessive inheritance, 11 of these genes received a ranking of 1, two received a ranking of 2, and three received a ranking of 3. For autosomal-dominant inheritance, five genes received a ranking of 1, five received a ranking of 2 and six received a ranking of 3. These results are summarized in **Table 2**.

Table 2 Rankings of annotated genes

		Strength of evidence		
		1	2	3
Diseases with single mode of inheritance	AR	81	113	178
	AD	61	108	261
Diseases with both AR and AD inheritance	AR	11	2	3
	AD	5	5	6
	Total	158	228	448

AD, autosomal dominant; AR, autosomal recessive.

DISCUSSION

The results indicate that a significant percentage of MorbidMap genes’ OMIM profiles lack a Clinical Synopsis or other statement regarding Mendelian inheritance pattern. These profiles are therefore unusable by most diagnostic tools when inheritance pattern is employed as a search strategy. At the time this study was performed, 1,524 OMIM gene profiles had Clinical Synopses that described an association with an autosomal-recessive disorder, and 1,135 had Clinical Synopses that described an association with an autosomal-dominant disorder. Of the MorbidMap genes analyzed in this study, 372 were inferred to be associated with autosomal-recessive disorders, and 430 with autosomal-dominant disorders. Thus, combining the results of this study with the number of genetic disorders known to have autosomal inheritance per their Clinical Synopses increases the number of autosomal-recessive disorders by 24.4%, and the number of autosomal-dominant disorders by 37.9%. These results are especially significant because the Clinical Synopsis pages on OMIM comprise the only database available to most software programs that diagnose genetic disease. It is clear that building on this database would greatly improve the diagnostic power of these programs.

Evidently, the availability of databases that contain details on genes, their structure and function, and their associated diseases is of immense importance. Many challenges exist for the individuals maintaining and curating these databases, and for the institutions financing them. We can imagine that providing details on inheritance pattern with a limited number of reported cases is suboptimal from a curation point of view. On the other hand, omission of such details in the interest of completeness results in delayed documentation, and compromises the utility of a database. In our view, this conflict can be overcome by providing a standardized measure of evidence to the stated determination of inheritance for each disorder. The method we employed in our study is somewhat arbitrary, and may need to be optimized by consensus. Of course, as accumulating cases increase evidence over time, the evidence score can be adjusted.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 2015;43: D789–D798.
2. Wierenga KJ, Jiang Z, Yang AC, Mulvihill JJ, Tsinoremas NF. A clinical evaluation tool for SNP arrays, especially for autosomal recessive conditions in offspring of consanguineous parents. *Genet Med* 2013;5:354–360.
3. Köhler S, Vasilevsky N, Engelstad M, et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Res* 2017;45(D1):D865–D876.
4. Cohen R, Gefen A, Elhadad M, Birk OS. CSI-OMIM—Clinical Synopsis search in OMIM. *BMC Bioinformatics* 2011;12:65.
5. OMIM. #602722 Renal tubular acidosis, distal, autosomal recessive; RTADR. http://omim.org/entry/602722. Accessed 6 March 2017.
6. OMIM. #603903 Sickle cell anemia. http://omim.org/entry/603903. Accessed 6 March 2017
7. Beautiful Soup. 2015. https://www.crummy.com/software/BeautifulSoup/. Accessed 6 March 2017.
8. Pandas. 2015. http://pandas.pydata.org/index.html. Accessed 6 March 2017.
9. TamperMonkey. https://tampermonkey.net/. Accessed 2015.