COMPARISON OF A LITERATURE SEARCH ALGORITHM AND CURATED PUBLICATION DATABASE WITH THE LITERATURE CONTENT OF OTHER LOCUS SPECIFIC DATABASES

Lisa Esterling, PhD; Dru DiFrancesco, MA; Paola Nix, PhD; Jean-Paul De La O, PhD; Bradford Coffee, PhD; Susan Manley, MS, CGC, MBA
Myriad Genetics, Inc., Salt Lake City, UT

BACKGROUND

- Effectively searching the scientific literature for publications providing evidence for the pathogenicity of a variant is critical in variant classification.
- Searches for relevant citations may be complicated by the use of alternative variant nomenclatures, gene names, and reference sequences.
- To ensure the most exhaustive search possible, we have developed an automated literature search algorithm coupled with a curated, searchable publication database linked to specific variants.
- The aim of this study was to validate the efficacy of our algorithm and database for the classification of variants included in a 25-gene hereditary cancer panel.

METHODS

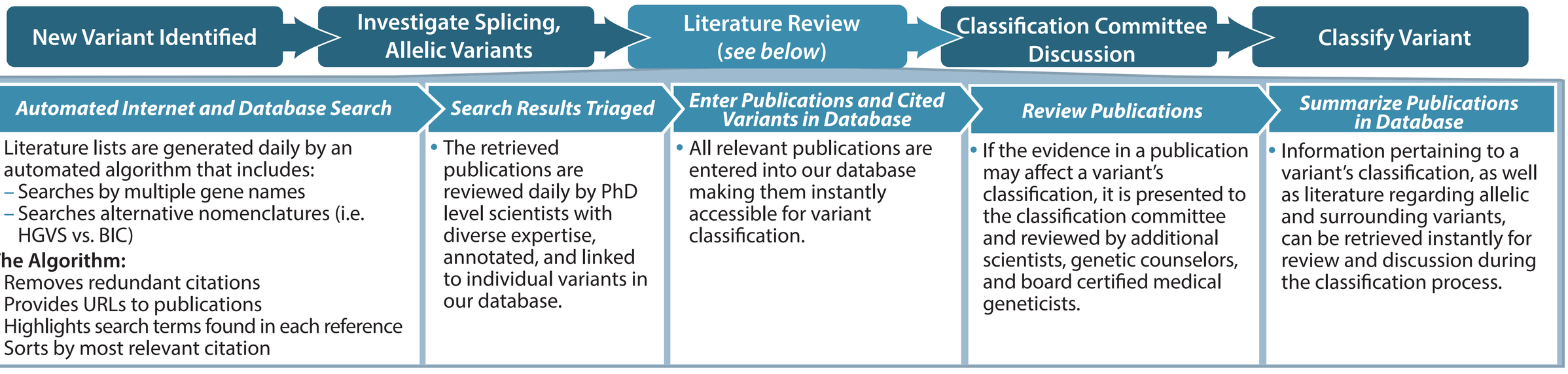
Literature Search in Variant Classification

- The overall process of variant classification and the utility of the literature search algorithm is shown in Figure 1.

Validating the Automated Literature Search Algorithm and Curated Database

- To assess how comprehensive and effective our method is compared to searching Locus Specific Databases (LSDBs) for identifying relevant literature, we compared the number of publications identified via our method to publications referenced in LSDBs (Table 1) for 1,553 variants seen during a 1 month period.
- The genes included in the pan-cancer panel are shown in Table 1. Variants in all genes were investigated here, including 755 previously classified and 798 novel variants.
- For *BRCA1* and *BRCA2*, the citations stored in our database were compared to those in HGMD (a commercial, curated database).

Figure 1. **Variant Classification and Literature Review Method**



RESULTS

Table 1. **Number of Citations by Gene and Database**

Gene	Our Database	HGMD[2]	LOVD[3]	LSDB Combined*	ClinVar[12]	Total	# of Variants
BRCA1/ BRCA2	294	42	54	13	80	461	300
APC	52	-	6	4	4	66	164
ATM	45	-	23	6	5	79	163
BARD1	1	-	1	0	0	2	44
BMPR1A	3	-	0	0	0	3	31
BRIP1	1	-	0	1	1	3	67
CDH1	14	-	3	0	0	17	61
CDK4	0	-	0	0	0	0	17
CHEK2	3	-	0	0	0	3	50
MLH1	78	-	39	6	31	154	60
MSH2	47	-	31	16	36	130	83
MSH6	28	-	13	1	13	55	110
MUTYH	16	-	17	0	1	34	55
NBN	1	-	0	0	0	1	37
CDKN2A	5	-	0	0	0	5	30
PALB2	17	-	3	9	9	38	61
PMS2	8	-	10	1	2	21	59
PTEN	4	-	23	30	30	87	18
RAD51C	1	-	0	0	0	1	27
RAD51D	4	-	0	0	0	4	24
SMAD4	0	-	0	0	0	0	25
STK11	0	-	1	0	0	1	40
TP53	190	-	13	17	29	249	29
Total	**790**	**42**	**237**	**104**	**241**	**1372**	**1553**

*Includes UMD[4], RAPID[5], COSMIC[6], FA Mutation Database[7], Memorial University[8], ARUP[9], IARC[10], Charles University in Prague[11]

- A total of 852 unique publications were identified in all databases, with 334% more publications identified in our database relative to the combined public databases.
  - This included references for a total of 1,372 (88.3%) of the 1,553 variants observed during the time of this study.
- Our method identified 36% more variant references than the other public databases combined (Table 1).
- For *BRCA1* and *BRCA2* variants, our method yielded 700% more references than HGMD for the variants examined (Table 1).
- 1,030 variant references referred to previously classified variants, while the remaining 342 referred to variants with novel classifications, which are presumably more rare.
- The majority of variant references were found for missense and nonsense variants (Table 2).

Table 2. **Number of Citations by Variant Type and Database**

Variant Type	Our Database	HGMD*	LOVD	LSDBs	ClinVar	Total	# of Variants
Missense	436	15	98	30	114	678	892
Nonsense	150	8	83	31	51	315	56
Frameshift	79	8	19	5	20	123	100
In-Frame Indel	10	1	3	0	15	28	30
Silent	33	0	15	7	4	59	274
Intronic	79	10	18	3	36	136	182
5'UTR	3	0	1	0	1	5	13
3'UTR	0	0	0	0	0	0	6

*Only literature pertaining to *BRCA1* and *BRCA2* variants were compared to HGMD.

DISCUSSION

- These results confirm that our literature search method and algorithm is more comprehensive than using what is available to the public as well as HGMD, a private curated database.
- Caution should be used when considering the evidence in literature and the search strategy, as all data should be subjected to scientific review representing a wide range of expertise.
- As expected, previously classified variants had significantly more citations than novel variants.
- The effectiveness of this method illustrates the significant amount of time and resources that need to be dedicated to variant classification to provide physicians and patients the most accurate test results for clinical decisions.

REFERENCES

1. Eggington et al. *Clin Genet*. 2014;86:229-37.
2. Stenson et al. *Hum Mutat*. 2003;21:577-81. (http://www.hgmd.cf.ac.uk/ac/index.php)
3. Fokkema et al. *Hum Mutat*. 2011;32:557-63. (http://www.lovd.nl/2.0/index_list.php)
4. Béroud et al. *Hum Mutat*. 2005;26:184-91. (http://www.umd.be/)
5. Hijikata et al. *DNA Res*. 2010;17:197-208. (http://web16.kazusa.or.jp/rapid/)
6. Forbes et al. *Nucleic Acids Res*. 2015;43:D805-11. (http://cancer.sanger.ac.uk/cosmic)
7. Levran et al. *Hum Mutat*. 2005;25:142-9. (http://www.rockefeller.edu/fanconi/genes/jumpa)
8. Woods et al. *Hum Mutat*. 2007;28:669-73. (http://www.med.mun.ca/mmrvariants/)
9. Millson et al. *J Mol Diagn*. doi: 10.1016/j.jmoldx.2015.05.005. (http://www.arup.utah.edu/database/SMAD4/SMAD4_welcome.php)
10. Petitjean et al. *Hum Mutat*. 2007;28:622-9. (http://p53.iarc.fr/)
11. Sedlacek et al. *Nucleic Acids Res*. 1998;26:214-5. (http://www.lf2.cuni.cz/projects/germline_mut_p53.htm)
12. Landrum et al. *Nucleic Acids Res*. 2014;42:D980-5. (http://www.ncbi.nlm.nih.gov/clinvar/)

Presented at ASHG - October 8, 2015