ELECTRONIC LETTER

A survey of locus-specific database curation

Richard G H Cotton, Kate Phillips, Ourania Horaitis

J Med Genet 2007;44:e72 (<http://www.jmedgenet.com/cgi/content/full/44/4/e72>). doi: 10.1136/jmg.2006.044081

It is widely accepted that curation of variation in genes is best performed by experts in those genes and their variation. However, obtaining funding for such variation is difficult even though up-to-date lists of variations in genes are essential for optimum delivery of genetic healthcare and for medical research. This study was undertaken to gather information on gene-specific databases (locus-specific databases) in an effort to understand their functioning, funding and needs.

A questionnaire was sent to 125 curators and we received 47 responses. Individuals performed curation of up to 69 genes. The time curators spent curating was extremely variable. This ranged from 0 h per week up to 5 curators spending over 4 h per week. The funding required ranged from US\$600 to US\$45000 per year. Most databases were stimulated by the Human Genome Organization-Mutation Database Initiative and used their guidelines. Many databases reported unpublished mutations, with all but one respondent reporting errors in the literature. Of the 13 who reported hit rates, 9 reported over 52 000 hits per year.

On the basis of this, five recommendations were made to improve the curation of variation information, particularly that of mutations causing single-gene disorder:

1. A curator for each gene, who is an expert in it, should be identified or nominated.
2. Curation at a minimum of 2 h per week at US\$2000 per gene per year should be encouraged.
3. Guidelines and custom software use should be encouraged to facilitate easy setup and curation.
4. Hits per week on the website should be recorded to allow the importance of the site to be illustrated for grant-giving purposes.
5. Published protocols should be followed in the establishment of locus-specific databases.

Accurate and complete collection and public display of variation causing single-gene disorder (referred to here as mutation) is vital for research and efficient delivery of genetic healthcare. Curation of mutations in individual genes by experts in these genes began with the globin gene¹ and has been accelerating ever since. Today there are 571 gene-specific databases with expert curators on the internet (<http://www.hgvs.org/dblist/glsdb>). It is well known that these locus-specific databases (LSDBs) contain almost as many unpublished mutations as published ones.² Also, expert curation of published mutations in one gene has shown 10% errors on the reporting of mutations in the literature (David Ravine, personal communication). This indicates the importance of LSDBs and their curators. It has long been known that funding for mutation databases has been lacking³ yet the importance of such databases has been highlighted.⁴ This has been underlined recently.⁵ An attempt to facilitate curation led a group of geneticists in 1994 to agree that mutations were best collected by a Federation of LSDB curators.⁶ This led to the formation of the Human Genome Organization-Mutation Database Initiative

(HUGO-MDI),⁷ from which recently the Human Genome Variation Society (<http://www.hgvs.org>) have been created.⁸ The current situation regarding mutation collection and databases of mutation has recently been reviewed (submitted).

Recommended content for LSDBs was made⁹ and a survey performed that led to further recommendations.¹⁰ However, there has been no survey of key factors such as currency, viability and funding mechanisms. This survey is also indicated because there is not usually any direct funding curation available for LSDBs and this is carried out in the curators' spare time. This tenuous situation has led to many curators discontinuing their activities (http://benaroyaresearch.org/investigators/concannon_patrick/atm.htm). Thus, future prospects for this large number of databases need to be established to ensure the accuracy and substantiality of data for patient care and research purposes.

METHOD

A complete up-to-date listing of LSDBs and their curators was prepared (<http://www.hgvs.org/dblist/glsdb>). The number of LSDBs was 571. A questionnaire was developed as shown in Box 1. That questionnaire was sent in January 2006 to each of the 125 responsible curators. Response was requested in 2 weeks and a reminder was sent at 3–4 weeks.

RESULTS

Of 125 questionnaires sent, we received 47 replies in 4 weeks. Table 1 summarises the relevant data.

Curation

A total of 27 curators curated more than one gene ranging from 2 to 69. Such a situation makes for economy in time as long as the curators are expert in the genes. Only three curators reported not being expert in the genes.

In all, 19 reported as co-curating with others. This is an important characteristic as it not only increases the skill set curating the data but also allows back-up if one curator is busy or loses funding.

The time involved in curation was noted (28 curators). This was extremely variable, with five curators reporting over 4 h per week and most reporting around 1 h a week.

Over half (28 of 47) reported difficulty in maintaining their database. Where 18 reported that funds were required for curation, the amount needed ranged from US\$600 to US\$45000 per gene per year. Naturally some genes have more mutations and some curation is performed with more detail.

Funding sources mentioned, if there was no problem in curation, were diverse, including receiving no funding at all (7) to personal research grants, government grants and funding from their institution. Funding ideas ranged from obtaining funds from the diagnostic labs using the databases, support groups and government health agencies.

Abbreviations: HUGO-MDI, Human Genome Organization-Mutation Database Initiative; LSDB, locus-specific database

Box 1: Survey of the locus-specific-database questionnaire (data, whether or not there was a complete response, were compiled (unpublished) and a summary was prepared (table 1))

Please type into form and return via email to hgvsdb@gmail.com

Please note if you require details to be kept confidential.

1. Your name:
2. Name of database(s) (attach list if necessary):
3. URL of database(s) (attach list if necessary):

Part A. Database curation

4. Do you curate more than one database?
5. If so, how many?
6. What gene(s) do you curate?
7. Are you an expert in this/these genes?
8. If you co-curate, what are the names of the other curators?
9. How many hours per week are spent curating the database(s)?
10. Are you having difficulty maintaining your database? Y/N
11. If yes, what amount of funds in euro or US\$ would you need to maintain it?
12. If no, who funds it?
13. Do you have any suggestions for the global or individual funding of locus-specific databases (LSDBs)?

Part B. Database characteristics

14. In which year was your database(s) established?
15. When was/were your LSDB(s) last updated?
16. Was the database(s) created on Human Genome Organization (HUGO)/Mutation Database Initiative (MDI)/Human Genome Variation Society (HGVS) guidelines? Y/N
17. Was the database encouraged by HUGO/MDI/HGVS? Y/N
18. How many variations in your LSDB(s):
 - a. cause disease?
 - b. don't cause disease?
 - c. are unclassified?
19. What percentage of your mutations is unpublished?
20. Do you find errors in the literature reports of mutations?
21. What are the number of hits /per week on your site?
22. Comments and suggestions please.

Database characteristics

Most databases have been established since 1995 (42 of 47), with most being last updated in 2005 or later (41 of 46). Many of the responders (27 of 44) used the HUGO-MDI guidelines in developing their databases and a majority (24 of 43) were encouraged by the HUGO-MDI.

A majority (30 of 45) reported $\leq 1\%$ of unpublished mutations, with 10 reporting 11–95%. Of the 45, 44 reported errors in the literature.

Only 13 databases were able to report hit rates, with nine reporting 1000 hits per week per gene or more—over 52 000 hits per year.

A total of 15 respondents reported mutations curated in a single gene in three categories—that is, whether they were shown to cause disease, not cause disease or uncertain (unclassified). The number of disease-causing mutations per database (besides one with three) ranged from 82 to 2982, with variants per database not causing disease (besides 0) ranging

from 12 to thousands. Unclassified variants (besides 0) ranged from 8 to 165 per database.

A range of useful suggestions was made, mainly referring to the need for ease of collection and curation.

DISCUSSION

This survey reinforces earlier unsubstantiated comments on the state of curation of mutations in genes by experts in those genes. The survey results were made less easy to interpret in some cases where more than one gene was being curated by a single curator (or team). In fact, over half the curators curated more than one gene, with one curating up to 65.

Most curators are experiencing difficulty in maintaining their databases, and the range of curation is from 0.1 to 27 h per week. This precisely reflects the magnitude of the task. If we take 1 h per week as our approximate median, this translates at US\$20 per hour to US\$1040 per year. This compares with the approximated median of US\$5000 estimated by respondents as required to maintain the database. Naturally there are establishment expenses and it is presumed the figures relate to update costs.

The rate of establishment decreased in the past 5 years in the responding group, but almost all were updated in 2005 and 2006.

Most used the HUGO-MDI guidelines and were encouraged by HUGO-MDI.

The need for expert curation is emphasised as curators found an enormous error rate in the literature. The literature also contained many unpublished mutations.

Hit rates (the number of times a site is accessed) indicate the importance of mutations in the genes available. Thus, even a lower rate of 1000 per week in the survey translates to 52 000 per year, which is substantial. For complete and accurate data on mutations in genes to be available to those who need it, the following recommendations are suggested:

1. An expert curator for each gene should be identified or nominated.
2. Curation at a minimum of 2 h per week at US\$2000 per gene per year should be encouraged.
3. Guidelines and custom software use should be encouraged to facilitate easy setup and curation.
4. Hits per week on the website should be recorded to allow the importance of the site to be illustrated for grant-giving purposes.
5. Published protocols should be followed in the establishment of LSDBs.¹¹

Recently, concerted efforts on a general scale have begun to promote the collection of variation. This includes a meeting in March 2006 called by the European Commission to discuss possible action (<http://cordis.europa.eu/lifescihealth/genomics/home.htm>), and a name for a central or linked database was coined: HUGObase;¹² a second meeting on a global scale was convened in June 2006 (<http://www.humanvariomeproject.org>), where the Human Variome Project was formally initiated. It is hoped that these initiatives will provide the impetus for better variation collection and curation.

ACKNOWLEDGEMENTS

The study was funded by the NHMRC, the GDRC and the Helen Smibert Vacation Studentship. We thank Lauren Hardman for the preparation of the manuscript. We also thank the curators who responded to this survey.

Authors' affiliations

Kate Phillips, Ourania Horaitis, Genomic Disorders Research Centre, Melbourne, Australia

Table 1 Summary of responses to the questionnaire

Question	Question (see the full question in box 1)	Summary of responses	Number of responses to this question
4	Curation of more than one gene	27	47
5	Range of genes curated	2–69	47
7	Non-expert curators	3	46
8	Do you co-curate with others—Yes	19	47
9	Hours/week/gene for 1–5 gene curation	0 = 1, 0.1–0.9 = 13, 1–3 = 9, 4–15 = 3, >20 = 2	28
10	Difficulty maintaining	26	47
11	Funds needed per gene per year for 1–5 gene curation (US\$)	1–1000 = 3, 1001–5000 = 5, 5001–10 000 = 3, 10 001–20 000 = 4, 20 001–50 000 = 3	18
12	If no, who funded it	See text	
13	Suggestions	See text	
14	Established	2001–2006 = 15, 1995–2000 = 27, 1990–1994 = 4, 1985–1989 = 2	47
15	Last updated	2006 = 21, 2005 = 20, 2004 = 3, 2003 and before = 2	46
16	HUGO-MDI guidelines	Yes = 27, Part = 4, No = 13	44
17	HUGO-MDI encouraged	Yes = 24, No = 19	43
18	Types of variation in database	See text	15
19	Percentage of unpublished mutations	0 = 15, <1% = 15, 1–10% = 5, 11–20% = 2, 21–30% = 3, 50% = 2, 75% = 1, 95% = 2	45
20	Errors in literature reports	Yes = 44, No = 1	45
21	Hits per week per gene for 1–5 genes curated	1–100 = 1, 100–1000 = 3, 1000–5000 = 6, 5000–10 000 = 1, 10 000–20 000 = 2	13

HUGO-MDI, Human Genome Organization-Mutation Database Initiative.

The figures are from 47 respondents except where noted, representing 276 of 571 genes with LSDBs.

Richard G H Cotton, Department of Medicine, The University of Melbourne, Melbourne, Australia

Competing interests: None declared.

Correspondence to: Professor R G H Cotton, Genomic Disorders Research Centre, 7th Floor, Daly Wing, St Vincent's Hospital, Fitzroy, VIC 3065, Australia; cotton@unimelb.edu.au

Received 18 May 2006

Revised 23 July 2006

Accepted 23 August 2006

REFERENCES

- 1 **Huisman THJ**, Carver MFH, Baysal E. *A Syllabus of thalassemia mutations*. Augusta: The Sickle Cell Anemia Foundation, 1997.
- 2 **Cotton RG**. Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. *Hum Mutat* 2000;**15**:4–6.
- 3 **Krawczak M**, Cooper DN. Core database. *Nature* 1995;**374**:402.
- 4 **Maurer SM**, Firestone RB, Sriver CR. Science's neglected legacy. *Nature* 2000;**405**:117–20.
- 5 **Patrinis GP**, Brookes AJ. DNA, diseases and databases: disastrously deficient. *Trends Genet* 2005;**21**:333–8.
- 6 **Sriver CR**, Cotton RGH, Antonarakis S, McKusick VA. Mutation Database Initiative to go forward under HUGO. *Genome Digest* 1997;**4**:12–15.
- 7 **Cotton RG**, McKusick V, Sriver CR. The HUGO Mutation Database Initiative. *Science* 1998;**279**:10–11.
- 8 **Cotton RG**, Horaitis O. Human Genome Variation Society. In: Cooper DN, eds. *Nature encyclopedia of the human genome*. Vol 3. London: Nature Publishing Group, 2000:361–2.
- 9 **Sriver CR**, Nowacki PM, Lehtvaslainen H. Guidelines and recommendations for content, structure, and deployment of mutation databases: II. Journey in progress. *Hum Mutat* 2000;**15**:13–15.
- 10 **Claustres M**, Horaitis O, Vanevski M, Cotton RGH. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res* 2002;**12**:680–8.
- 11 **Horaitis O**, Cotton RGH. Human mutation databases. *Current protocols in human genetics*. Vol 2. New York: Wiley-Liss, 1999 (Updated in 2003).
- 12 WayStation to HUGOBase. *Nat Genet* 2005;**37**:783.