Online Mendelian Inheritance in Animals (OMIA): inclusion of a hyperlinked table of likely causal variants for inherited disorders and traits

I. Tammen¹ & F.W. Nicholas²

¹ The University of Sydney, Sydney School of Veterinary Science, 425 Werombi Road, Camden, NSW 2570, Australia
imke.tammen@sydney.edu.au (Corresponding Author)
² The University of Sydney, School of Life and Environmental Sciences, NSW 2006 Australia

Summary

Online Mendelian Inheritance in Animals (OMIA) is a comprehensive, annotated catalogue of inherited disorders and other familial traits in animals other than humans, rats and mice. OMIA is modelled on and is complementary to Online Mendelian Inheritance in Man (OMIM). Starting with the very first report of a likely causal variant (for inherited goitre in cattle, published in 1987), the total number of likely causal variants in OMIA is now nearly 1000, with new variants being included at a rate of more than one per week.

Until recently, information on these likely causal variants was imbedded in text fields in the OMIA database/website – a source of increasing frustration for researchers and chip designers who require a tabulated list in standard format. In 2013, one of us (FN) commenced placing information on likely causal variants from different sources in standard format in an Excel spreadsheet. In August 2017, work commenced incorporating the contents of the spreadsheet into appropriate new fields in the OMIA database, along with development of curation tools to enable new variants to be added as they are published. This will enable the presentation of all likely causal variants on the OMIA website in a standard tabular format that is readily sortable, and freely downloadable for one or more species.

The other major challenge with documenting likely causal variants is that very few (around 4%) of those relevant to OMIA traits have been submitted to any variant database. Primarily this is because variant database submission tools were designed for batch submission of thousands of variants from genome-wide studies, rather than for the submission of one or two variants. Commencing in September 2017, all animal variants will reside only in the European Variant Archive (EVA). Accordingly, one of us (FN) has been collaborating with EVA personnel to devise an easy protocol for the submission of likely causal variant to EVA. When this protocol is fully operational, journal editors will be requested to require submission of any new likely causal variant to EVA as a condition of publication. A major advantage of submitting a variant to EVA is that its genomic location will automatically be updated to the latest version of the relevant genome assembly. For each variant submitted to EVA, OMIA will be able to extract all relevant information (including current genomic location and flanking sequence) via the variant ID, by means of an Application Programming Interface (API).

The up-to-date table of likely causal variants that will be downloadable from OMIA will include hyperlinks from each variant to the relevant full entries in OMIA and EVA and to the original publication. This downloadable table should be of benefit to researchers, chip designers, diagnostic service providers, veterinarians and breed societies.

Keywords: OMIA, likely causal variants
Introduction

Online Mendelian Inheritance in Animals (OMIA) is a comprehensive, annotated catalogue of inherited disorders and other familial traits in animals other than humans, rats and mice. OMIA is modelled on and is complementary to Online Mendelian Inheritance in Man (OMIM), and provides links to PubMed and Gene records at the National Center for Biotechnology Information (NCBI), OMIM and the European Bioinformatics Institute (EBI)’s Ensembl.

The history and structure of OMIA has been summarised by Nicholas (1998, 2003) and Lenffer et al. (2006). Starting as a mainframe database in 1980 — the pre-Internet era — OMIA was launched on the Internet in 1995, with major enhancements in 2004 and 2011. In 2015 a simplified and global URL was acquired: http://omia.org.

OMIA has a hierarchical structure of traits (“phenes”) in one or more species. This structure is reflected in the OMIA ID, which has the form xxxxxx-yy...yy, where xxxxxx is the 6-digit phene ID and yy...yy is NCBI’s species taxonomy ID (usually four digits, but sometimes longer). Thus an OMIA ID signifies a phene-species, i.e. a particular phene in a particular species. At the time of writing, OMIA contains information on 3363 phene-species, and includes 24883 references to mostly peer-reviewed articles.

The curation of this ever-expanding corpus of knowledge has been mainly done by one of us (FN), ably assisted by colleagues such as Professor Tosso Leeb and by colleagues at the University of Sydney1. The curation process is now being automated as far as possible, but it will always require informed human input. FN is now offering online tutorials to help to train the next generation of curators. Volunteers are welcome! In addition, donations are being sought for a perpetual OMIA fund that will (hopefully) enable the part-time employment of a head curator (when FN can no longer fulfil this role) and of a software engineer to maintain the database and website.

Discovery of Mendelian traits

As detailed in OMIA’s Landmark-paper section (http://omia.org/key_articles/landmarks/), the first documented Mendelian traits in non-laboratory animals were described very soon after the rediscovery of Mendelism: pea comb, rose comb, polydactyly, shank colour, and white plumage (dominant white) in chickens (Bateson, 1902); and polled in cattle (Bateson and Saunders, 1902). One hundred and fifteen years later, the summary table on OMIA’s home page lists 1337 phene-species with Mendelian (single-locus) inheritance.

Discovery of likely causal variants

The history of the discovery of likely causal variants was described by Nicholas and Hobbs (2014). Fig. 1 summarises this history up to the time of writing.

The early discoveries were Herculean efforts with the very limited technological resources then available. The upsurge in the 1990s reflects the increasing availability of genome-wide microsatellite maps; the further increase in the following decade reflects the advent of genome-wide SNP chips; the current discovery rate of more than one per week reflects the ever-increasing availability of genome sequences and bioinformatics tools to handle huge sequence datasets.
Figure 1. The number of likely causal variants published per year across all animal species recorded in OMIA.

A table of OMIA likely causal variants

The standard OMIA practice has been that when a likely causal variant was published, it was entered into a text field in OMIA, often as a quote from the original paper. There are two obvious disadvantages of this strategy: the format is frustratingly variable, and the information is not readily available. Good models for the documentation of variants existed, e.g. the Human Gene Mutation Database (HGMD) (Stenson et al., 2009), but until recently, no funding was available to create something similar in OMIA.

Conscious of the ever-increasing need for variants to be tabulated, one of us (FN) commenced in 2013 entering new variants, as well as variants previously entered in text fields in OMIA, in a spreadsheet in a manner modelled on that used by HGMD. The information recorded included (where available): species, breed, gene, phenotype, type of mutation, location (reference sequence, chromosome, gDNA, cDNA, peptide and ID in a variant database). The need for such a table was well illustrated by Reeb et al. (2016), who spent “several person-months” extracting variant information from OMIA text fields for their bioinformatics study.

Other researchers had also become aware of the serious need for summarising information on likely causal variants. For example, McClure and McClure (2016) assembled a wealth of information on bovine variants included in the International Dairy and Beef (IDB) genotyping chip. In addition to providing information that was not previously in the OMIA spreadsheet (especially genomic locations), the IDB information also provided a useful check for information that was already included in the spreadsheet.

In August 2017, the University of Sydney generously provided the services of a software engineer for two days per week until the end of the year to incorporate the contents of the spreadsheet in a standard format in new fields in the OMIA database and website.

The result is:

1. A table of likely causal variants known for each phene-species OMIA entry, appearing on the web page for that entry
2. The entire table of likely causal variants in all species (or for one or more chosen species), freely downloadable and sortable by each column

Table 1 summarises the entire contents of the OMIA variant table at the time of writing.

Table 1. Summary of information of likely causal variants in OMIA by species.

<table>
<thead>
<tr>
<th></th>
<th>Dog</th>
<th>Cattle</th>
<th>Cat</th>
<th>Horse</th>
<th>Chicken</th>
<th>Sheep</th>
<th>Pig</th>
<th>Other</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number:</td>
<td>311</td>
<td>176</td>
<td>100</td>
<td>74</td>
<td>58</td>
<td>56</td>
<td>31</td>
<td>120</td>
<td>926</td>
<td></td>
</tr>
<tr>
<td>- genomic location</td>
<td>120</td>
<td>143</td>
<td>32</td>
<td>50</td>
<td>17</td>
<td>21</td>
<td>7</td>
<td>5</td>
<td>395</td>
<td>43%</td>
</tr>
<tr>
<td>- variant ID²</td>
<td>7</td>
<td>39</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>63</td>
<td>7%</td>
</tr>
<tr>
<td>Type of variant:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- missense</td>
<td>97</td>
<td>60</td>
<td>36</td>
<td>32</td>
<td>18</td>
<td>28</td>
<td>12</td>
<td>44</td>
<td>327</td>
<td>35%</td>
</tr>
<tr>
<td>- nonsense</td>
<td>32</td>
<td>29</td>
<td>11</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>1</td>
<td>14</td>
<td>106</td>
<td>11%</td>
</tr>
<tr>
<td>- extension</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>- splicing</td>
<td>25</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>78</td>
<td>8%</td>
</tr>
<tr>
<td>Type</td>
<td>Count</td>
<td>dbSNP</td>
<td>dbVar</td>
<td>OMIA</td>
<td>total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulatory</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indels</td>
<td>140</td>
<td>64</td>
<td>34</td>
<td>21</td>
<td>345</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rearrangement</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeat variation</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 Combined information for 49 species
2 This row indicates the number of variants known to FN to have been submitted to any one of the variant databases.

**Variants not in variant databases**

Anyone who had been assembling lists of likely causal variants in domesticated animals very quickly became aware that very few of these variants had been submitted to any of the standard variant databases, such as NCBI’s dbSNP or dbVar and EBI’s European Variant Archive (EVA). Primarily, this is because submitting single variants to such databases is a daunting and time-consuming task: the submission tools for these databases were designed for batch submission of thousands of variants from genome-wide studies, rather than for the submission of one or two variants. In addition, for likely causal variants in species for which genome assemblies were, or are still not, available, the genomic location (an obligatory item of information for submission to a variant database) may not yet have been worked out, or is still not available.

This problem was highlighted by Sharma et al. (2017), the title of whose paper says it all: “Limited representation of OMIA causative mutations for cattle in SNP databases”.

A development very relevant to the resolution of this dilemma is that in mid-2017, NCBI and EBI jointly announced that, in order to rationalise their ability to handle the ever-increasing number of variants being discovered, from September 2017 the NCBI databases (dbSNP and dbVar) will receive only human variants, and the EBI database (EVA) will receive only non-human variants. As part of this rationalisation, the NCBI databases are ensuring that they have all the historic human data that were submitted to EVA; and EVA is ensuring that it has all the historic non-human data that were submitted to dbSNP or dbVar. A favourable consequence of this development is that the challenge of incorporating all OMIA likely causal variants into a variant database involves collaborating with just one organisation (EBI) rather than two. A collaboration between OMIA and EVA commenced early in 2017, before the rationalisation announcement, and is now well advanced in developing a submission protocol that should make it far easier for the discoverers of a likely causal variant to submit that variant to EVA. When this protocol is fully operational, journal editors will be requested to require submission of any new likely causal variant to EVA as a condition of publication.

A major advantage of submitting a variant to EVA is that its genomic location will automatically be updated to the latest version of the relevant genome assembly.

It is realised, of course, that having this simplified protocol will not automatically enable the submission of all historic variants, and still excludes all those lacking genomic coordinates. In the short term, whatever information is available for such variants will be stored in, and will be made freely available from, the OMIA database and website. Over time, relevant authors will be contacted in the hope that genomic coordinates can be worked out (at least for those species with genome assemblies) for their variants, and submissions made to EVA.

In the meantime, genomic coordinates worked out by third parties, such as by McClure and McClure (2016) for cattle, have been generously made available to OMIA and have been incorporated (with acknowledgement) into OMIA.
For each variant that is submitted to EVA, OMIA will be able to extract all relevant information (including current genomic location and flanking sequence) via the variant ID, by means of an Application Programming Interface (API).

Conclusion

By the time of this conference, it will be possible to freely download from the OMIA web site a table of likely causal variants for one or several or all the major domesticated animal species included in OMIA, including (in an ever-increasing number of cases) the genomic location of each variant in the latest genome assembly. This table will include hyperlinks from each variant to the relevant full entries in OMIA and EVA and to the original publication. This downloadable table should be of benefit to researchers, chip designers, diagnostic service providers, veterinarians and breed societies.

Acknowledgements

The Acknowledgements tab on the OMIA home page provides a detailed account of the contributions of the many people who have contributed to the development of OMIA since its inception. Special thanks are due to the many researchers who advise FN of new likely causal variants; to Matt and Jennifer McClure for generously providing their bovine variant information; to the staff of Information and Communications Technology (ICT) at the University of Sydney for their support of the master database and website; to Matthew Hobbs for his impressive enhancements in past years; to Cali Willet, Jared Berghold and Josh Stretton from the University’s Informatics Hub, for their invaluable contribution to current enhancements; to Professor Rosanne Taylor, Dean of the Sydney School of Veterinary Science, for her moral and financial support of current enhancements and future developments; to fellow curators and colleagues Tosso Leeb, Bobbie Cansdale, Tracy Chew, Hamutal Mazrrier, Claire Wade, Bianca Waud, Cali Willet, and Shernae Woolley; to Zhiliang Hu and Jim Reecy, Iowa State University, for the extraordinary time and effort they have devoted to attempts at securing the long-term future of OMIA; to Zhiyong Lu (NCBI PubMed) and Juanmi Cejuela (Rostlab, Munich) for collaboration in text mining; to Gary Saunders and Cristina García (EVA) for their enthusiastic and constructive facilitation of submission of single variants; to Melissa Haendel (Oregon) and Chris Mungall (Lawrence Berkeley lab) for their enthusiastic determination to ontologise OMIA; and to Harkeet Pavia and Will Yaxley of Sydney University’s Division of Alumni & Development, together with Claire Wade, for invaluable help and guidance in aiming to secure the long-term future of OMIA.

List of References

Bateson, W., 1902. Experimental studies in the physiology of heredity. Part II. Poultry.
  Reports to the Evolutionary Committee of the Royal Society 1: 87-124.
  Reports to the Evolution Committee of the Royal Society 1: 125-160.
Lenfjer, J., F.W. Nicholas, K. Castle, A. Rao, S. Gregory, M. Poidinger, M.D. Mailman & S.
  Ranganathan, 2006. OMIA (Online Mendelian Inheritance in Animals): an enhanced
  platform and integration into the Entrez search interface at NCBI. Nucleic Acids Res. 34:


