Annotation and curation of human genomic variations: an ELIXIR Implementation Study [version 1; peer review: 2 approved with reservations]

Alessia David¹, Valérie Barbié², Marcella Attimonelli³, Roberto Preste⁴, Enni Makkonen⁴, Heidi Marjonen⁵, Mats Lindstedt⁴, Kati Kristiansson⁵, Sarah E. Hunt⁶, Fiona Cunningham⁶, Ilkka Lappalainen⁴, Michael J.E. Sternberg¹

¹Centre for Integrative Systems Biology and Bioinformatics, Imperial College London, London, SW7 2AZ, UK
²Clinical Bioinformatics, Swiss Institute of Bioinformatics, Geneva, Switzerland
³Department of Biosciences, University of Bari, Bari, Italy
⁴CSC-IT Center for Science, Espoo, Finland
⁵Finnish Institute for Health and Welfare, Helsinki, Finland
⁶European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

Abstract
Background: ELIXIR is an intergovernmental organization, primarily based around European countries, established to host life science resources, including databases, software tools, training material and cloud storage for the scientific community under a single infrastructure.

Methods: In 2018, ELIXIR commissioned an international survey on the usage of databases and tools for annotating and curating human genomic variants with the aim of improving ELIXIR resources. The 27-question survey was made available on-line between September and December 2018 to rank the importance and explore the usage and limitations of a wide range of databases and tools for annotating and curating human genomic variants, including resources specific for next generation sequencing, research into mitochondria and protein structure.

Results: Eighteen countries participated in the survey and a total of 92 questionnaires were collected and analysed. Most respondents (89%, n=82) were from academia or a research environment. 51% (n=47) of respondents gave answers on behalf of a small research group (<10 people), 33% (n=30) in relation to individual work and 16% (n=15) on behalf of a large group (>10 people). The survey showed that the scientific community considers several resources supported by ELIXIR crucial or very important. Moreover, it showed that the work done by ELIXIR is greatly valued. In particular, most respondents acknowledged the importance of key features and benefits promoted
by ELIXIR, such as the verified scientific quality and maintenance of ELIXIR-approved resources.

**Conclusions** ELIXIR is a “one-stop-shop” that helps researchers identify the most suitable, robust and well-maintained bioinformatics resources for delivering their research tasks.

**Keywords**
ELIXIR, survey, database, bioinformatics tools

---

**Corresponding author:** Alessia David (alessia.david09@imperial.ac.uk)

**Author roles:** David A: Conceptualization, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; Barbié V: Conceptualization, Writing – Review & Editing; Attimonelli M: Conceptualization, Writing – Review & Editing; Preste R: Conceptualization, Writing – Review & Editing; Makkonen E: Conceptualization, Writing – Review & Editing; Marjonen H: Conceptualization, Writing – Review & Editing; Lindstedt M: Conceptualization, Writing – Review & Editing; Kristiansson K: Conceptualization, Writing – Review & Editing; Hunt SE: Conceptualization, Writing – Review & Editing; Cunningham F: Conceptualization, Writing – Review & Editing; Lappalainen I: Conceptualization, Writing – Review & Editing; Sternberg MJE: Conceptualization, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This study received funding from ELIXIR: the research infrastructure for life-science data. In addition, we acknowledge funding from the European Molecular Biology Laboratory and Imperial College London.

**Copyright:** © 2020 David A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** David A, Barbié V, Attimonelli M et al. Annotation and curation of human genomic variations: an ELIXIR Implementation Study [version 1; peer review: 2 approved with reservations] F1000Research 2020, 9(ELIXIR):1207 https://doi.org/10.12688/f1000research.24427.1

**First published:** 08 Oct 2020, 9(ELIXIR):1207 https://doi.org/10.12688/f1000research.24427.1
Introduction
ELIXIR is an intergovernmental organization, primarily centred in Europe, established to host resources, such as databases, software tools, training material and cloud storage for the scientific community, under a single infrastructure. ELIXIR started in 2013 and now includes 22 state members and over 220 research Institutions. By providing a one-stop-shop for the scientific community, it aims to help researchers identify the most suitable bioinformatics resources (and the appropriate training material and workshops) to deliver their research task. Moreover, ELIXIR recognizes and facilitates sharing of data and exchange of expertise between its members, with the goal of agreeing on best practice.

In 2018, ELIXIR commissioned an international survey on the usage of databases and tools for annotating and curating human genomic variants with the aim of improving ELIXIR resources. Here we present the result of the survey.

Methods
A 27-question survey (Extended data) was designed and agreed upon by several ELIXIR members from various European countries. The survey was constructed around six main themes (detailed in “Questionnaire structure”) aimed at exploring the usage of ELIXIR resources and tools. Six ELIXIR nodes/partners informed the construction of the survey. Two researchers/research centres were identified by each of the six ELIXIR nodes/partners to participate in a pilot test. The survey was modified based on the feedback received. The final questionnaire was made available on-line using the Webropol survey website between September and December 2018. Responses were uploaded to a Finland-based university server. Members of the Finland ELIXIR node accessed the responses and stored them in a .csv file for analysis. Information about participation was displayed on the survey and completion of the survey was taken as confirmation of participation.

Considering the absence of identifying information in data published here, and the non-sensitive nature of the survey, no ethical approval was sought for this study. No information presented here can be used to identify survey participants.

Participants
For each country, an ELIXIR representative was asked to recruit prospective survey respondents from academia, clinical diagnostics, clinical research, industry and the Government within their country. No additional eligibility criteria were set. Participation to the survey was advertised using several means, such as the mailing lists of universities, research centers, private companies and research institutes within each European country. Prospective participants were asked to fill the anonymized on-line questionnaire. A three-month deadline was given for completing the on-line questionnaire. Reminders to complete the survey were published in the regular newsletters at the recruited institutions. This approach did not allow calculation of the response rate.

Data collected
The survey collected two types of data:
- quantitative data: this included category ranking metrics, as well as general frequency of use of the tools and databases surveyed. Whenever possible, a list of possible choices was provided for selection, to allow proper survey analysis and interpretation.
- qualitative data: this included participant comments.

Questionnaire structure
The survey was divided in six sections (the full questionnaire with the list of tools, databases and resources that were surveyed is presented in Extended Data):
Section 1 - Background information
Section 2 - Resources for annotating and curating human genomic variants (9 questions)
Section 3 - Next generation sequencing (3 questions, If no, skip this section)
Section 4 - Mitochondria (4 questions, If no, skip this section)
Section 5 - Proteins (5 questions, If no, skip this section)
Section 6 - ELIXIR (3 questions)

Sections 1, 2 and 6 were open to all participants, whereas sections 3 to 5 were only open to participants who worked or had an interest in working in these specific fields. Accordingly, respondents were asked to skip the sections that were not pertinent to their work or that of their group.

The complete list of tools and databases surveyed is presented in Extended data.

Data analysis
Data analysis was performed using RStudio (version 0.98.1062) and R (R version 3.4.4). Quantitative data are presented as absolute and relative frequency. Qualitative data were analysed per theme.

Results
A total of 92 questionnaires were available for analysis. Data were collected from 18 European countries and the Unites States (Extended data: Figure S1). Finland contributed to the majority of answers (30%, n=28), followed by the United Kingdom (11%, n=10). The large majority of respondents (89%, n=82) identified themselves as belonging to academia or a research environment (Extended data: Table S1). The large majority (51%, n=47) of respondents gave answers to the questionnaire in relation to their work and that of a small group of (<10) people. This was followed by answers given in relation to individual work (33%, n=30) and the work of a large group of (>10) people (16%, n=15). As the survey was sent to colleagues by the organisers, the responders may well not be representative of the wider community. Additionally, most
of the responders were from European countries and the responses could be substantially different elsewhere, such as from the USA.

**Key resources**

Section 2 of the survey aimed at identifying key data resources and tools used by the scientific community primarily in ELIXIR member states. Respondents were asked to rank 52 resources (listed in Extended data: Table S2). A total of 2055 responses were collected from 88% (82) responders. In total, 22 resources were considered critical/very important by more than 50% of respondents who knew the resources and considered them relevant to their work (Figure 1). Three ELIXIR resources, Ensembl\(^2\), the Ensembl Variant Effect Predictor (VEP)\(^3\) and UniProt\(^4\), were within the top ten. Several additional resources were listed by respondents and are presented in Extended data: Tables S3 and S4.

**Guidelines**

A total of 78 responses were collected from 50 (54%) respondents on the standards and guidelines used for the interpretation of sequence variants. 66% (n=33) of respondents followed the American College of Medical Genetics (ACMG) standards and guidelines for the interpretation of sequence variants\(^5\), 22% (n=11) national guidelines, and 8% (n=4/50) in house guidelines. 4% (n=2) answered “other guidelines” and listed the European Leukemia Net (ELN) guidelines\(^6\) and the “Sequence Ontology base classification for variants”. With regards to the human genome reference assembly version, approximately one third (38%, n=31/81) of respondents used both GRCh37 and GRCh38, 44% (n=36/81) GRCh37 only, 15% (n=12/81) GRCh38 alone, and 2% (n=2/81) GRCh37, GRCh38 and older versions (Figure 2).

**Data/data formats**

Four types of data/data formats used in human genome variation analysis (VCF/BCF, BAM, FASTQ, FASTA) were considered critical/very important by the majority of respondents (Extended data: Table S5). Additional data formats that were listed as critical/very important were: BED (cited 4 times), gVCF (cited twice), PLINK\(^8\), IMPUTE\(^9\) and Hail MatrixTable. Genome-wide association studies (GWAS) and rare disease studies were listed as the most important topics and analytical operations in human genome variation analysis performed by respondents to this survey (Extended data: Table S6).

**Important resource features**

One of the main goals of this survey was to identify user requirements in the use of resources for human genome variation annotation and curation. A total of 777 responses on the importance of basic features of databases and tools were analysed. The majority of respondents (>83%) considered critical/very important a resource’s free/low cost, open license to academia, the ability to assess the quality of its data, good user documentation and its availability in English. An easy-to-use web browser and privacy policy were deemed critical/very important by >77% of respondents (Table 1).

A total of 1173 responses were analysed on the importance of technical features in variation analysis, annotation tools and databases. Good curation of the database was the top scoring technical feature, ranked critical/very important by 92% of respondents. The following five technical features were considered critical/very important by >75% of respondents: i) fit for purpose data used by the resource, ii) the scientific coverage and comprehensiveness of the resource; iii) scalability to high-throughput analysis; iv) availability of datasets for download; and v) ability to analyse large datasets/queries (Table 2).

**Next Generation Sequencing**

The next section of the survey focused on Next Generation Sequencing (NGS). 66 (72%) respondents worked with NGS and ranked 6 sequencing methods listed in Extended data: Table S7. Whole exome, panel DNA, whole genome and RNA sequencing were used by more than 60% of respondents. Methylation sequencing was the least used technique (28% of respondents) but was the technique the majority of respondents would like to use (23% of respondents). Several other techniques were listed among “Other”, such as single-cell RNA sequencing, ribosome profiling, chromatin conformation techniques (3C, 4C, HiC, capture HiC), assay for transposase-accessible chromatin using sequencing (ATAC-SEQ), targeted RNA sequencing, fusion genes on cDNA, microbiome sequencing, GRO-seq.

![Figure 1. List of 22 resources considered critical/very important by >50% of respondents (Q4).](image_url)
Table 1. Importance of basic features in variation analysis and annotation databases and tools (Q10).

| Feature                                         | Critical N (%) | Very important N (%) | Useful N (%) | Total N |
|-------------------------------------------------|----------------|----------------------|--------------|---------|
| Language, in English                            | 45 (52)        | 30 (35)              | 11 (13)      | 86      |
| Cost-free / Low cost of use                     | 37 (43)        | 34 (40)              | 15 (17)      | 86      |
| Easy to assess the quality of each data resource| 35 (40)        | 42 (48)              | 10 (11)      | 87      |
| Open licenses or public statement of open terms of use for academia | 31 (37) | 40 (48) | 13 (15) | 84 |
| Good user documentation                        | 30 (35)        | 38 (45)              | 17 (20)      | 85      |
| Easy-to-use web/browser access                  | 25 (33)        | 27 (36)              | 23 (31)      | 75      |
| Publicly available privacy policy (security around personal data and cookies are described) | 17 (24) | 32 (44) | 23 (32) | 72 |
| Availability for commercial use                 | 5 (19)         | 9 (33)               | 13 (48)      | 27      |
| Widely used                                     | 14 (18)        | 34 (44)              | 30 (38)      | 78      |
| Training and customer service available for use | 10 (14)        | 24 (33)              | 39 (53)      | 73      |
| Language, in local languages (non-English)     | 0 (0)          | 1 (5)                | 21 (95)      | 22      |
| TOTAL                                           | 251            | 311                  | 215          | 777     |

The next section of the survey explored the use of resources related to the study of mitochondria. Only 15 (16%) respondents worked with mitochondrial DNA (mtDNA) alone or in combination with nuclear genes involved in mitochondrial functionality. Of these, 7 used NGS alone, 2 NGS and Sanger sequencing and 6 could not specify. Additional methods listed under “Other methods” were whole genome sequencing (WGS), whole exome sequencing (WES), RNA-seq and mtDNA included in WGS or WES.

Mitochondrial resources
Among the mitochondria-dedicated databases, MITOMAP\(^{10}\) was ranked as critical/very important by 65% of respondents, followed by HmtDB (Human Mitochondrial Database)\(^{11}\) by 44% and HmtVar (Human Mitochondrial Variants Database)\(^{12}\) by 33%. With the exception of MITOMAP, MitoCarta\(^{13}\) and HmtDB, more than half of the respondents were unaware of all other databases (Extended data: Table S8).

Among the six tools for mitochondria-related research that were ranked, MToolBox\(^{14}\), HaploGrep2\(^{15}\), MitoTip\(^{16}\) and MitImpact2\(^{17}\) were equally reported as critical/very important by 18% of respondents (Extended data: Table S9). The following mitochondrial databases/tools not included in the survey were listed among “Other” and ranked critical/very important: MitoFates\(^{18}\) and an in house built annotation pipeline. mvTool
### Table 2. Importance of technical features in variation analysis and annotation tools and databases.

| Feature                                      | Critical N (%) | Very important N (%) | Useful N (%) | Total N |
|----------------------------------------------|----------------|----------------------|--------------|---------|
| Good curation of database                    | 38 (46)        | 38 (46)              | 7 (8)        | 83      |
| Ability to analyse large data sets/queries   | 33 (39)        | 31 (36)              | 21 (25)      | 85      |
| Data fit for purpose (e.g. complete set, clinically relevant, up-to-date data) | 32 (38)        | 40 (48)              | 12 (14)      | 84      |
| Scalability to high-throughput analysis      | 28 (35)        | 35 (44)              | 16 (20)      | 79      |
| Datasets available for download              | 26 (33)        | 35 (44)              | 18 (23)      | 79      |
| Scientific coverage and comprehensiveness of the resource | 25 (31)        | 43 (53)              | 13 (16)      | 81      |
| Programmatic access (e.g. linux/unix)        | 21 (30)        | 25 (36)              | 24 (34)      | 70      |
| Run locally                                  | 20 (28)        | 22 (31)              | 29 (34)      | 71      |
| Data in large number of global populations available | 20 (24)        | 35 (42)              | 29 (35)      | 84      |
| Database links to variants-related scientific literature | 18 (21)        | 35 (41)              | 32 (38)      | 85      |
| Population-specific data available           | 14 (18)        | 28 (36)              | 35 (45)      | 77      |
| Response time of key web pages and search functions | 13 (18)        | 29 (39)              | 32 (43)      | 74      |
| Open interfaces (e.g. REST)                  | 10 (16)        | 14 (23)              | 38 (61)      | 62      |
| Good data visualization options              | 12 (15)        | 27 (33)              | 42 (52)      | 81      |
| Multiple data sharing formats (e.g. plain text, FASTA, XML, RDF, Dublin Core, tsv, JSON) | 9 (12)           | 28 (36)              | 41 (53)      | 78      |
| TOTAL                                        | 319            | 465                  | 389          | 1173    |

(part of MSeqDR.org) and MitoMaster (part of MITOMAP) were listed among “Other tools” used for mitochondria-related research.

### Protein structure

The next section of the survey explored the use of resources related to the study of protein structures. 21 out of 92 (22.8%) respondents worked with protein structure, whereas 5 (5.4%) did not but would like to. The results of the survey are presented based on 182 answers from these 26 respondents. Databases of experimentally determined protein structure and complexes, such as PBD[^19] and PDBe[^20], were considered critical/very important by 65% (n=17/26) of respondents, followed by tools that report on the structural consequences of variants (62%, n=16/26). All 7 tools/databases that were surveyed were ranked critical/very important by over 45% of respondents (Table 3). The following tools/databases were not surveyed but cited by respondents as key resources for structural modelling of variants: Pymol[^21], Rosetta[^22], KiNG[^23], Modeller[^24], PolyPhen[^25], SNP3D[^26], SNPs&GO[^27], SAAPdab/SAAPpred[^28], HOPE[^29] and Yasara[^30].

Key limiting factors to the use of tools for modelling the structural consequences of variants were the lack of expertise in the area (54%, n=13/24), difficulties in using the tools (42%, n=10/24), and difficulties in interpreting the results (38%, n=9/24) (Figure 3). Lack of high throughput capability was listed among “Other limitations”. The difficulty to translate protein dynamics into structural models was also cited.

### ELIXIR

The last section of the questionnaire covered questions related to ELIXIR as a platform that coordinates the tools and databases surveyed in this study and the benefits that ELIXIR, as a European intergovernmental organization, offers to the scientific community. The majority of respondents (52%, n=42/92) were not aware that the tools and databases surveyed in Q25 are part of ELIXIR (Extended data: Figure S2). However, the long-term sustainability of ELIXIR core resources and the verified scientific quality of datasets were considered critical/very important by >76% (n=68) of participants who answered this question. The international standards for describing and saving data, the verified quality and maintenance of ELIXIR resources and facilities to find the right research tools were also considered critical/very important by the majority of respondents (>60%) (Table 4). Additional challenges that were identified by respondents regarding the annotation and curation of human genetic variants were related to the quality of data: incorrect entries, conflicting annotation, limited access to expert curation and classification of variants.
Table 3. List of protein-related resources for variant analysis there were ranked in the survey (Q21).

| Resource                                                                 | Critical N (%) | Very important N (%) | Useful N (%) | Not relevant N (%) | Not known N (%) | Total N | Critical/Very important N (%) |
|-------------------------------------------------------------------------|----------------|----------------------|--------------|-------------------|----------------|---------|-------------------------------|
| Databases of experimentally-determined protein structure including complexes (e.g. PDB/PDBe) | 14 (54%)       | 3 (12%)              | 7 (27%)      | 0                 | 2 (8%)         | 26      | 17 (65%)                     |
| Tools to predict protein structure from sequence (e.g. Phyre2/ SWISSMODEL) | 6 (23%)        | 6 (23%)              | 10 (38%)     | 3 (12%)           | 1 (4%)         | 26      | 12 (46%)                     |
| Molecular graphics viewers (e.g. PyMol, Chimera)                        | 5 (19%)        | 9 (35%)              | 7 (27%)      | 1 (4%)            | 4 (15%)        | 26      | 14 (54%)                     |
| Tools that report the structural consequences of variants (e.g. loss of a salt bridge) | 4 (15%)        | 12 (46%)             | 8 (31%)      | 2 (8%)            | 0             | 26      | 16 (62%)                     |
| Databases of predicted protein structures                               | 3 (12%)        | 10 (38%)             | 10 (38%)     | 3 (12%)           | 0             | 26      | 13 (50%)                     |
| Databases of predicted complexes                                        | 2 (8%)         | 9 (35%)              | 10 (38%)     | 4 (15%)           | 1 (4%)         | 26      | 11 (42%)                     |
| Tools to predict the structure of protein complexes (i.e. protein docking) | 2 (8%)         | 10 (38%)             | 8 (31%)      | 4 (15%)           | 2 (8%)         | 26      | 12 (46%)                     |
| TOTAL                                                                   | 36             | 59                   | 60           | 17                | 10            | 182     | 95                            |

Figure 3. Limitations to the use of tools to model the structural consequences of variants were ranked in Q23.

obtained in routine diagnostic setting and limited number genotype-phenotype annotations. An important caveat is the ascertainment bias that the survey was primarily sent to colleagues of the authors and this probably is not a representative sample of the community involved in the annotation and curation of human genomic variations.

Conclusions
This survey shows that the scientific community in ELIXIR member states considers several resources supported by ELIXIR crucial or very important. Moreover, it shows that the work done by ELIXIR is greatly valued. In particular, most respondents acknowledged the importance of key features and...
### Table 4. Importance of the benefits that ELIXIR offers to the scientific community (Q26).

| Benefit                                                                 | Critical N (%) | Very important N (%) | Useful N (%) | Not relevant N (%) | Don't know N (%) | Total N | Critical/Very important N (%) |
|------------------------------------------------------------------------|----------------|----------------------|--------------|-------------------|-----------------|---------|-------------------------------|
| Long-term sustainability of ELIXIR core resources                      | 38 (43%)       | 30 (34%)             | 11 (13%)     | 4 (5%)            | 5 (6%)          | 88      | 68 (77%)                      |
| Verified scientific quality of datasets                                | 32 (36%)       | 36 (40%)             | 14 (16%)     | 2 (2%)            | 5 (6%)          | 89      | 68 (76%)                      |
| Verified quality and maintenance of ELIXIR resources                   | 31 (36%)       | 32 (37%)             | 15 (17%)     | 3 (3%)            | 6 (7%)          | 87      | 63 (72%)                      |
| International standards for describing and saving data                  | 28 (32%)       | 38 (44%)             | 14 (16%)     | 3 (3%)            | 4 (5%)          | 87      | 66 (76%)                      |
| Facilities to find the right research tools                            | 24 (27%)       | 37 (42%)             | 16 (18%)     | 7 (8%)            | 5 (6%)          | 89      | 61 (69%)                      |
| Training in data analysis and current best practices                   | 18 (20%)       | 30 (34%)             | 33 (37%)     | 3 (3%)            | 6 (7%)          | 89      | 48 (54%)                      |
| Network of supercomputing services                                     | 10 (12%)       | 30 (35%)             | 28 (33%)     | 11 (13%)          | 6 (7%)          | 85      | 40 (47%)                      |
| Pan-European infrastructure of knowledge and support                    | 9 (10%)        | 31 (36%)             | 29 (34%)     | 7 (8%)            | 10 (12%)        | 86      | 40 (47%)                      |
| Promotion of industry collaboration                                     | 5 (6%)         | 15 (17%)             | 39 (45%)     | 17 (20%)          | 10 (12%)        | 86      | 20 (23%)                      |
| **TOTAL**                                                              | **195**        | **279**              | **199**      | **57**            | **61**          | **791** | **474**                       |

benefits promoted by ELIXIR, such as the verified scientific quality and maintenance of ELIXIR-approved resources.

**Data availability**

**Underlying data**

Open Science Framework: ELIXIR, [https://doi.org/10.17605/OSF.IO/SWX4](https://doi.org/10.17605/OSF.IO/SWX4).

**Extended data**

Open Science Framework: ELIXIR, [https://doi.org/10.17605/OSF.IO/SWX4](https://doi.org/10.17605/OSF.IO/SWX4).

This project contains the following extended data:
- Copy of the online survey
- Figure S1. Number of responses from the countries that participated in the survey
- Figure S2. Results for Q25: “Before the survey, were you aware that the resources listed below are part of ELIXIR?”
- Tables S1. Answers to Q2: Place of work. More than one answer was allowed per participant
- Table S2. List of Databases and Tools surveyed in Q4
- Table S3. Additional Resources listed by Respondents under “Other” in Q4 and considered Critical or Very Important
- Table S4. Other resources listed by Respondents in Q5
- Table S5. The importance of data formats in human genome variations analysis (Q8)
- Table S6. List of the topics and analytical operations in human genome variation analysis considered most important in the Respondents’ work (Q9)
- Table S7. Answers to Q14 (Do you use these sequencing methods?)
- Table S8. Mitochondrial databases (Q18)
- Table S9. Mitochondrial tools (Q18)

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).
References

1. R Core Team: R: A language and environment for statistical computing. Vienna, Austria; 2012.
2. Yates AD, Achuthan P, Akanni W, et al.: Ensemble 2020. Nucleic Acids Res. 2020; 48(D1): D682–D688. PubMed Abstract | Publisher Full Text | Free Full Text
3. McLaren W, Gil L, Hunt SE, et al.: The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1): 122. PubMed Abstract | Publisher Full Text | Free Full Text
4. UniProt Consortium: UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019; 47(D1): D506–D515. PubMed Abstract | Publisher Full Text | Free Full Text
5. Richards S, Aziz N, Bale S, et al.: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17(3): 405–424. PubMed Abstract | Publisher Full Text | Free Full Text
6. Döhner H, Estey E, Grimwade D, et al.: Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017; 129(4): 424–447. PubMed Abstract | Publisher Full Text | Free Full Text
7. Mungall CJ, Batchelor C, Eilbeck K: Evolution of the Sequence Ontology terms and relationships. J Biomed Inform. 2011; 44(1): 87–93. PubMed Abstract | Publisher Full Text | Free Full Text
8. Chang CC, Chow CC, Tellier LC, et al.: Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4: 7. PubMed Abstract | Publisher Full Text | Free Full Text
9. Howe BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5(6): e1000529. PubMed Abstract | Publisher Full Text | Free Full Text
10. Brandon MC, Lott MT, Nguyen KC, et al.: MITOMAP: a human mitochondrial genome database–2004 update. Nucleic Acids Res. 2005; 33(Database issue): D611–D613. PubMed Abstract | Publisher Full Text | Free Full Text
11. Clima R, Presto R, Calabrese C, et al.: HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor. Nucleic Acids Res. 2017; 45(D1): D698–D706. PubMed Abstract | Publisher Full Text | Free Full Text
12. Presto R, Vitala O, Clima R, et al.: HmtVar: a new resource for human mitochondrial variations and pathogenicity data. Nucleic Acids Res. 2019; 47(D1): D1202-D1210. PubMed Abstract | Publisher Full Text | Free Full Text
13. Calvo SE, Clauser KR, Mootha VK: Mitocarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2016; 44(D1): D1251–1257. PubMed Abstract | Publisher Full Text | Free Full Text
14. Calabrese C, Simone D, Dioma MA, et al.: MitoBoc: a highly automated pipeline for heteroplasmic annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics. 2014; 30(21): 3115–3117. PubMed Abstract | Publisher Full Text | Free Full Text
15. Weissensteiner H, Pacher D, Kloss-Brandstätter A, et al.: HaploGrep 2: mitochondrial haplgroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016; 44(W1): W58–63. PubMed Abstract | Publisher Full Text | Free Full Text
16. Sonney S, Leipzig J, Lott MT, et al.: Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. PLoS Comput Biol. 2017; 13(12): e1005867. PubMed Abstract | Publisher Full Text | Free Full Text
17. Castellana S, Rónai J, Mazza T: MitImpact: an exhaustive collection of pre-computed pathogenicity predictions of human mitochondrial non-synonymous variants. Hum Mutat. 2015; 36(2): E2413–2422. PubMed Abstract | Publisher Full Text
18. Fukasawa Y, Tsuji J, Fu SC, et al.: Mitofates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol Cell Proteomics. 2015; 14(4): 1113–1126. PubMed Abstract | Publisher Full Text | Free Full Text
19. Berman HM, Battistuz T, Bhat TN, et al.: The Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 2002; 58(Pt 6 No 1): 899–907. PubMed Abstract | Publisher Full Text | Free Full Text
20. PDBe-KB consortium: PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res. 2020; 48(D1): D344-D353. PubMed Abstract | Publisher Full Text | Free Full Text
21. Schrödinger: The PyMOL Molecular Graphics System. LLC.
22. Lyskov S, Chou FC, Conchur SÖ, et al.: Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIIE). PLoS One. 2013; 8(5): e63906. PubMed Abstract | Publisher Full Text | Free Full Text
23. Chen VB, Sali A: Protein Structure Modeling with MODELLER. Methods Mol Biol. 2017; 1654: 39–54. PubMed Abstract | Publisher Full Text | Free Full Text
24. Webb B, Sali A: Protein Structure Modeling with MODELLER. Methods Mol Biol. 2017; 1654: 39–54. PubMed Abstract | Publisher Full Text | Free Full Text
25. Adzhubei IA, Schmidt S, Peshkin L, et al.: A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4): 248–249. PubMed Abstract | Publisher Full Text | Free Full Text
26. Yue P, Melamud E, Moult J: SNP3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006; 7: 166. PubMed Abstract | Publisher Full Text | Free Full Text
27. Yue P, Melamud E, Moult J, et al.: SNP3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006; 7: 166. PubMed Abstract | Publisher Full Text | Free Full Text
28. Calabrese R, Capriotti E, Faresli P, et al.: Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009; 30(4): 1237–1244. PubMed Abstract | Publisher Full Text | Free Full Text
29. Al-Numair NS, Martin AC: The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. BMC Genomics. 2013; 14 Suppl 3(Suppl 3): 54. PubMed Abstract | Publisher Full Text | Free Full Text
30. Venselaar H, Te Beek TAH, Kuipers RKP, et al.: Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. 2010; 11: 548. PubMed Abstract | Publisher Full Text | Free Full Text
31. Land H, Hümble MS: VASAAR: A Tool to Obtain Structural Guidance in Biocatalytic Investigations. Methods Mol Biol. 2018; 1688: 43–67. PubMed Abstract | Publisher Full Text | Free Full Text
32. David A: Elixir. 2020. http://www doi.org/10.17605/OSF.IO/SWX42
Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 18 August 2022

https://doi.org/10.5256/f1000research.26946.r145730

© 2022 Ogasawara O. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Osamu Ogasawara
Bioinformation and DDBJ Center, National Institute of Genetics, Shizuoka, Japan

The authors conducted a systematic survey of the usage of databases and tools for annotating and curating human genomic variants with the aim of improving ELIXIR resources. Eighteen European countries and the United States participated in the survey and a total of 92 questionnaires were collected and analyzed. The survey questions were carefully designed and consisted of a series of questions about what kind of databases and tools users need, and what kind of technical requirements they would like to have in their computing infrastructure. The results of the survey are quite interesting and will be useful to researchers in many fields, not just those working on computational infrastructure and tools. In addition, the questions themselves and the survey results are provided as Extended Data which is very useful for ensuring the reproducibility of the paper's analysis results and for readers to verify any questions they may have.

Major comments:

1. Some items may be underestimated due to the mixing of questionnaire results from researchers with different project sizes and other different characteristics. A brief discussion of this effect should be added.

For example, for Question 11, "Run locally," the text indicates that 28% of the respondents chose "critical" as the 8th item from the top of Table 2, however, if this calculation is done (by using Extended Data) only for the large group (>10 people), 40% of the respondents chose "critical" as their response, and it comes in the upper part of Table 2. Thus, not all of the respondents to the questions are highly interested, but researchers with certain characteristics may be highly interested in some of the items. (On the other hand, as far as I can tell from my research using Extended Data, most of the questions in Questions 10, 11, and 26 seem to have almost the same ratio of critical, very important, and useful, even if the size of the research project varies. The few exceptions include “Data in large number of global populations available”, “Open interfaces (for example REST)”, and “Response time of key web pages and search functions”. These are all items in Q11, and the large group had a high percentage of "critical" responses.)
2. Since the purpose of the questionnaire survey is stated in the Abstract as "with the aim of improving ELIXIR resources", it would be better to briefly explain in the conclusion the idea of how to reflect the results of this survey in the management of ELIXIR.

Minor comment:
1. In the Abstract, the authors wrote “Results: Results: Eighteen countries participated in the survey and a total of 92 questionnaires were collected and analysed”, but as shown at the beginning of the Results (page 3) and in Figure S1, 19 countries (Eighteen European countries and the United States) participated.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, high performance computing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
The authors systematically analysed the results and statistics of the questionnaire for an international survey on the usage of ELIXIR databases and tools, which are mainly devoted to annotating and curating human genomic variants. This survey would benefit both ELIXIR and researchers in helping them to identify the most suitable and robust bioinformatics resources, and thus make those resources well-maintained. I have two major concerns regarding this work.

1. It is not clear how the 52 resources for Q4 (How important the following resources in your work) in Section 2 (Resources for annotating and curating human genomic variants) were selected. As not all of the resources are related with human genomic variants annotation and curation, the results or the question itself may be misleading. For example, ArrayExpress may be critical/important for gene expression studies, and Europe EMC may be critical/important for literature retrieval. However, they may not be critical/important for variation analysis. This should be addressed in the manuscript to avoid misunderstanding.

2. A small number of questionnaires are available for some questions. Especially, for sections 4 and 5, there are only about 20 participants. I am afraid some conclusions may not be well-supported by the statistics. MitoCarta, which is considered to be one of the most popular mitochondrial databases according to its high citations and z-index, ranks the first among the 53 related resources collected by Database Commons ([https://bigd.big.ac.cn/databasecommons](https://bigd.big.ac.cn/databasecommons)). However, only 24% of participants considered MitoCarta to be critical/important. Therefore, the authors should conclude with caution and discuss this in the manuscript.

Also, there is one typo in row 7 of ‘Protein structure’. ‘PBD’ should be ‘PDB’.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.
Reviewer Expertise: Big Data Integration and Analytics, Non-coding RNA Annotation and Curation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com