Establishment of an international database for genetic variants in esophageal cancer

Vihinen, Mauno

Published in:
Annals of the New York Academy of Sciences

DOI:
10.1111/nyas.13152

2016

Document Version:
Peer reviewed version (aka post-print)

ESTABLISHMENT OF AN INTERNATIONAL DATABASE FOR GENETIC VARIANTS IN ESOPHAGEAL CANCER

Mauno Vihinen

Protein Structure and Bioinformatics Group, Department of Experimental Medical Science, Lund University, Lund, Sweden

Address for correspondence: Mauno Vihinen, Protein Structure and Bioinformatics Group, Department of Experimental Medical Science, Lund University, BMC B13, SE-22184 Lund, Sweden. mauno.vihinen@med.lu.se

The establishment of a database has been suggested in order to collect, organize, and distribute genetic information about esophageal cancer. The World Organization for Specialized Studies on Diseases of the Esophagus and the Human Variome Project will be in charge of a central database of information about esophageal cancer–related variations from publications, databases, and laboratories; in addition to genetic details, clinical parameters will also be included. The aim will be to get all the central players in research, clinical, and commercial laboratories to contribute. The database will follow established recommendations and guidelines. The database will require a team of dedicated curators with different backgrounds. Numerous layers of systematics will be applied to facilitate computational analyses. The data items will be extensively integrated with other information sources. The database will be distributed as open access to ensure exchange of the data with other databases. Variations will be reported in relation to reference sequences on three levels—DNA, R NA, and protein—whenever applicable. In the first phase, the database will concentrate on genetic variations including both somatic and germline variations for susceptibility genes. Additional types of information can be integrated at a later stage.

Keywords: esophageal cancer; variation database; mutation database; disease-causing variations; mutation; data integration; OESO; the Human Variome Project

Introduction

Genetic methods are widely used for investigation and analysis of diseases, including cancers. Barrett's esophagus and esophageal adenocarcinoma have several risk factors, such as sex (male), ethnicity (Caucasian), obesity, and gastroesophageal reflux. In addition, a genetic component has been shown to be important. Several studies have indicated genetic variations in esophageal cancer.[1–5] There is evident genetic predisposition to the disease due to inherited variants and the effects of acquired somatic variants. To be able to understand the disease progress and genetic relevance, it is essential to collect, organize, distribute, analyze, and interpret genetic variations in esophageal cancer.

Genetic variation information has been collected in numerous open-access repositories, but a large part of the data remains private and is never published. One of the major goals is to collect the data and distribute it freely to the community. Several examples are available for databases, and recommendations have been published for establishing and curating such repositories. The esophageal cancer community needs to agree on the establishment and maintenance of a database that willserve in many ways in both research and clinical applications.

The World Organization for Specialized Studies on Diseases of the Esophagus (OESO; http://oeso.org/) acts as a forum, bringing together clinicians and researchers. Within OESO, the need for a dedicated database for esophageal cancer has been noticed. Here, the path for establishing and maintaining such a database is drafted.

Locus-specific variation databases

Locus-specific databases (LSDBs) are important sources of information for clinicians and researchers in assessing data and forming opinions on the clinical relevance of gene and protein variants. LSDBs list variants in specific genes/diseases and are typically annotated manually. Many LSDBs also contain other information, sometimes even very detailed clinical data. LSDBs are typically curated and maintained for single genes or groups of genes related to certain diseases. These resources are widely used for diverse purposes.

Another type of variation database is represented by central databases that include data from large studies covering many genes and proteins. Cancer genomic data are available from genomic projects in the Cancer Genome Project (https://www.sanger.ac.uk/research/projects/cancergenome/), the Cancer Genome Atlas (http://cancergenome.nih.gov/), and the International Cancer Genome Consortium (https://icgc.org/).

Despite a large number of studies of genetic variation in numerous cancers, relatively little is known about the true cancer driver variants. The Cancer Gene Census (CGC; http://www.sanger.ac.uk/genetics/CGP/Census/) lists genes implicated in cancer.[6] Variations have been identified in these genes and collected to theCOSMICdatabase.[7] However, the variations have seldom been experimentally validated to be involved in carcinogenesis. Even in the genes listed in the CGC, just a small fraction of variations are predicted to be harmful.[8]

A framework for the collection of somatic variations to promote standards for annotation of variation data and to promote and facilitate data integration with other data resources has been presented.[9] For the interpretation of cancer specific variation patterns, a minimum set of information is needed, including the variation, the method of detection, and details of the tumor sample. For that purpose, the Minimum Information About Somatic Mutation criteria were developed[9] (http://structure.bmc.lu.se/MIASM/miasm.html).

The proposed database is actually not a typical LSDB, as numerous genes need to be included. However, it is a disease-specific database, and most of the lessons learned from LSDBs also apply to this kind of repository.

OESO–HVP database

OESO aims to establish a central database that will collect information about esophageal cancer– related variations from publications, databases, and laboratories. In addition to genetic details, clinical parameters should also be included. The goal is to get all the central players in research, clinical, and commercial laboratories to contribute.

The database will be developed together with the Human Variome Project (HVP; http://www.humanvariomeproject.org/). The HVP is a world organization that works toward facilitating the collection, curation, interpretation, and free and open sharing of genetic variation information. A key component of HVP activities is the development of standards and guidelines.[10] The HVP has endorsed numerous guidelines and recommendations and has released quality assessment criteria to evaluate the quality of genetic variation databases and to stimulate database curators to make improvements where they are needed[11] (http://www.humanvariomeproject.org/finish/19/255.html). Special attention will be paid to the quality in the new database to be established.

InSiGHT variation database as a model

The International Society for GastrointestinalHereditary Tumours Incorporated (InSiGHT; http://insight-group.org/) provides an example of how a cancer variation database can be established, maintained, and analyzed as a community-wide effort. InSiGHT has established a working group for collection and interpretation of variations in gastrointestinal cancers. The mission of the society is to improve the quality of care of patients and their families with conditions resulting in hereditary gastrointestinal tumors.

Sequence variants of uncertain functional and clinical relevance are common in genetic testing. InSiGHT has brought together the research and clinical community to collect and interpret variations in genes coding for mismatch repair system proteins, which are responsible for the disease when they contain variations. Several separate databases were combined into one global variation database.[12] There are currently data for over 26,000 patients—altogether over 29,000 variants, of which 4608 are unique (http://chromium.lovd.nl/LOVD2/colon_cancer/home.php).

Different schemes have been presented for the classification of variants in genes associated with diseases, usually in relation to Mendelian conditions. The International Agency for Research on Cancer classification system can be used for standardized categorization in five classes ranging from pathogenic to likely pathogenic, likely not pathogenic, and not pathogenic. In addition, class 3 describes variations with uncertain classification. InSiGHT has classified almost 3000 variants and their relevance for cancer based on worldwide data.[13] The variant classification was done by a group representing international leading laboratories.

InSiGHT provides a model for howto collect, distribute, and interpret variations in a communitywide manner including both research and clinical experts. The produced database, along with the variationclassifications, is freely available.

Establishing an esophageal variation database

The esophageal cancer database does not have to start from scratch. There is plenty of available experience with numerous kinds of variation databases. Several guidelines and recommendations have been published for LSDBs, including how to establish a database,[14] their curation,[15] overall contents,[16] ethics,[17,18] data collection,[19,20] somatic variations,[21] interpretation and reporting of variants,[22,23] data sharing,[24] and nomenclature.[25,26]

Systematics should be used in variation databases whenever possible, including phenotype databases, such as the Human Phenotype Ontology (HPO)[27] and the Variation Ontology (VariO),[28] for a systematic description of effects, mechanisms, and consequences of variants. Several *de facto* standards are useful for variation databases. These include systematic gene names available from the HUGO Gene Nomenclature Committee (HGNC).[29] Reference sequences have to be specified, including version numbers, unless Locus Reference Genomic (LRG)[30] entries are used. For naming variations, the Human Genome Variation Society (HGVS) nomenclature[26] should be used. It is widely used in the literature and by several computer programs, including those interpreting variations.

Action plan

A dedicated database will be established to collect, organize, and distribute genetic information for esophageal cancers (Fig. 1). Data to be collected need to be made freely available in a public database on the Internet to allow efficient search and analysis of the data. The database will follow established recommendations and guidelines as published, for example, by the HVP and the HGVS.

The database will apply for HVP Gene/Disease Specific Database Council membership once established. The database will require a team of dedicated curators with backgrounds in genetics, medicine, computer science, and bioinformatics. The variation data have to be collected from various sources, including the literature, existing databases, and direct submissions from laboratories. An important activity will be to get the community engaged in providing data but also participating (e.g., in interpretation).

The database has to be constructed in a systematic way to allow integration with other resources and reuse of data. The database will apply numerous layers of systematics to facilitate computational analyses. The systematics will include HGNC gene names, establishment of LRGreference sequences (if not yet available), HGVS variation nomenclature, HPO and other ontologies for phenotype, VariO for effects of variations, and many other factors. The data items will be extensively integrated with other information sources (e.g., at gene, protein, structure, and interaction levels).

Variation data will be entered into the database in a systematic fashion. Genotype, variation and effects, and pathogenicity will be described in detail. The database will be distributed as open access to ensure exchange of the data with other databases. Variations will be reported in relation to reference sequences on three levels—DNA, RNA,and protein—whenever applicable, using HGVS recommendations. Submitters need consent from patients/parents/carers before submitting information. All patient data will be anonymized before submission and release.

In the first phase, the database will concentrate on variations, both somatic and germline variations for susceptibility genes. These data are readily available in large quantities. Additional types of information can be integrated at later stages, including epigenetics, transcriptomics, and proteomics data.

Dedicated work groups (e.g., for the interpretation of the pathogenicity of variations) could be established in relation to the database. The establishment, maintenance, curation, and distribution of the database will require substantial and continued amounts of work. Some funds need to be raised. An ethical overseer committee needs to be nominated to decide on ethical issues pertinent to the collection and distribution of data. HVP has published ethical guidelines for LSDBs,[18] which are currently being updated.

Once the database is established, training has to be organized for end users. This can be organized, for example, in association with major meetings, such as OESO conferences. Instructive training material can also be distributed online to allow users to familiarize themselves with it whenever is best suitable for them. It is also a way to reach for the largest audience.

Recommendations

- OESO, together with HVP, will establish an open-access database for collection of genetic variations relevant for esophageal cancer;

- expert curators will be nominated;

- the database will follow existing guidelines and standards;

- the community will be mobilized to support and help the database;

- extensive data integration will be performed.

- data will be made freely available;

- systematics will be used at all levels;

- data will be collected and distributed in the database;

- funding will be obtained;

- an ethics committee will be nominated; and

- education and training for users and data providers will be arranged.

Conflicts of interest

The author declares no conflicts of interest.

References

1. Hu, N., M. Kadota, H. Liu, *et al*. 2016. Genomic landscape of somatic alterations in esophageal squamous cell carcinoma and gastric cancer. *Cancer Res*. **76:** 1714–1723.

2. To, H., N.J. Clemons, C.P. Duong, *et al*. 2016. The genetics of Barrett's esophagus: a familial and population-based perspective. *Dig. Dis. Sci.* **61:** 1826–1834.

3. Gharahkhani, P., J. Tung, D. Hinds, *et al*. 2016. Chronic gastroesophageal reflux disease shares genetic background with esophageal adenocarcinoma and Barrett's esophagus. *Hum. Mol. Genet.* **25:** 828–835.

4. Wu, C., Z.Wang, X. Song, *et al*. 2014. Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nat.Genet.* **46:** 1001–1006.

5. Weaver, J.M., C.S. Ross-Innes, N. Shannon, *et al*. 2014. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46:** 837–843.

6. Futreal, P.A., L. Coin, M. Marshall, *et al*. 2004. A census of human cancer genes. *Nat. Rev. Cancer* **4:** 177–183.

7. Forbes, S.A., N. Bindal, S. Bamford, *et al*. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*. **39:** D945–D950.

8. Niroula, A. & M. Vihinen. 2015. Harmful somatic amino acid substitutions affect key pathways in cancers. *BMCMed. Genomics* **8:** 53.

9. Olivier, M., A. Petitjean, J. Teague, *et al*. 2009. Somatic mutation databases as tools for molecular epidemiology and molecular pathology of cancer: proposed guidelines for improving data collection, distribution, and integration. *Hum. Mutat.* **30:** 275–282.

10. Smith, T.D. & M. Vihinen. 2015. Standard development at the Human Variome Project. *Database (Oxford)* **2015**. doi:10.1093/database/bav024.

11. Vihinen, M., J. M. Hancock, D. R. Maglott, *et al*. 2016.Human Variome Project quality assessment criteria for variation databases. *Hum. Mutat.* **37:** 549–558.

12. Plazzer, J.P., R.H. Sijmons, M.O. Woods, *et al*. 2013. The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam. Cancer* **12:** 175–180.

13. Thompson, B.A., A.B. Spurdle, J.P. Plazzer, *et al*. 2014. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat. Genet.* **46:** 107–115.

14. Vihinen, M., J.T. den Dunnen, R. Dalgleish, *et al*. 2012. Guidelines for establishing locus specific databases. *Hum. Mutat.* **33:** 298–305.

15. Celli, J., R. Dalgleish, M. Vihinen, *et al*. 2012. Curating gene variant databases (LSDBs): toward a universal standard. *Hum. Mutat.* **33:** 291–297.

16. Kohonen-Corish, M.R., J.Y. Al-Aama, A.D. Auerbach, *et al*. 2010. How to catch all those mutations—the report of the third Human Variome Project Meeting, UNESCO Paris, May 2010. *Hum. Mutat.* **31:** 1374–1381.

17. Cotton, R.G., C. Sallee & B.M. Knoppers. 2005. Locus specific databases: from ethical principles to practice. *Hum. Mutat.* **26:** 489–493.

18. Povey, S., A.I. Al Aqeel, A. Cambon-Thomsen, *et al*. 2010. Practical guidelines addressing ethical issues pertaining to the curation of human locus-specific variation databases (LSDBs). *Hum. Mutat.* **31:** 1179–1184.

19. Cotton, R.G., A.D. Auerbach, A.F. Brown, *et al*. 2007. A structured simple form for ordering genetic tests is needed to ensure coupling of clinical detail (phenotype) with DNA variants (genotype) to ensure utility in publication and databases. *Hum. Mutat.* **28:** 931–932.

20. Cotton, R.G., A.I. Al Aqeel, F. Al-Mulla, *et al*. 2009. Capturing all disease-causing mutations for clinical and research use: toward an effortless system for the Human Variome Project. *Genet. Med.* **11:** 843–849.

21. Olivier, M., A. Petitjean, J. Teague, *et al*. 2009. Somatic mutation databases as tools for molecular epidemiology and molecular pathology of cancer: proposed guidelines for improving data collection, distribution, and integration. *Hum. Mutat.* **30:** 275–282.

22. Richards, C.S., S. Bale, D.B. Bellissimo, *et al*. 2008. ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. *Genet. Med.* **10:** 294–300.

23. Plon, S.E., D.M. Eccles, D. Easton, *et al*. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29:** 1282–1291.

24. den Dunnen, J.T., R.H. Sijmons, P.S. Andersen, *et al*. 2009. Sharing data between LSDBs and central repositories. *Hum. Mutat.* **30:** 493–495.

25. Taschner, P.E. & J.T. den Dunnen. 2011. Describing structural changes by extending HGVS sequence variation nomenclature. *Hum. Mutat.* **32:** 507–511.

26. den Dunnen, J.T. & S.E. Antonarakis. 2001. Nomenclature for the description of human sequence variations. *Hum. Genet.* **109:** 121–124.

27. Robinson, P.N., S. Kohler, S. Bauer, *et al*. 2008. The human phenotype ontology: a tool for annotating and analysing human hereditary disease. *Am. J. Hum. Genet.* **83:** 610–615.

28. Vihinen, M. 2014.Variation ontology for annotation of variation effects and mechanisms. *Genome Res*. **24:** 356–364.

29. Gray, K.A., B. Yates, R.L. Seal, *et al*. 2015. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* **43:** D1079–D1085.

30. Dalgleish, R., P. Flicek, F. Cunningham, *et al*. 2010. Locus reference genomic sequences: an improved basis for describing human DNA variants. *Genome Med*. **2:** 24.

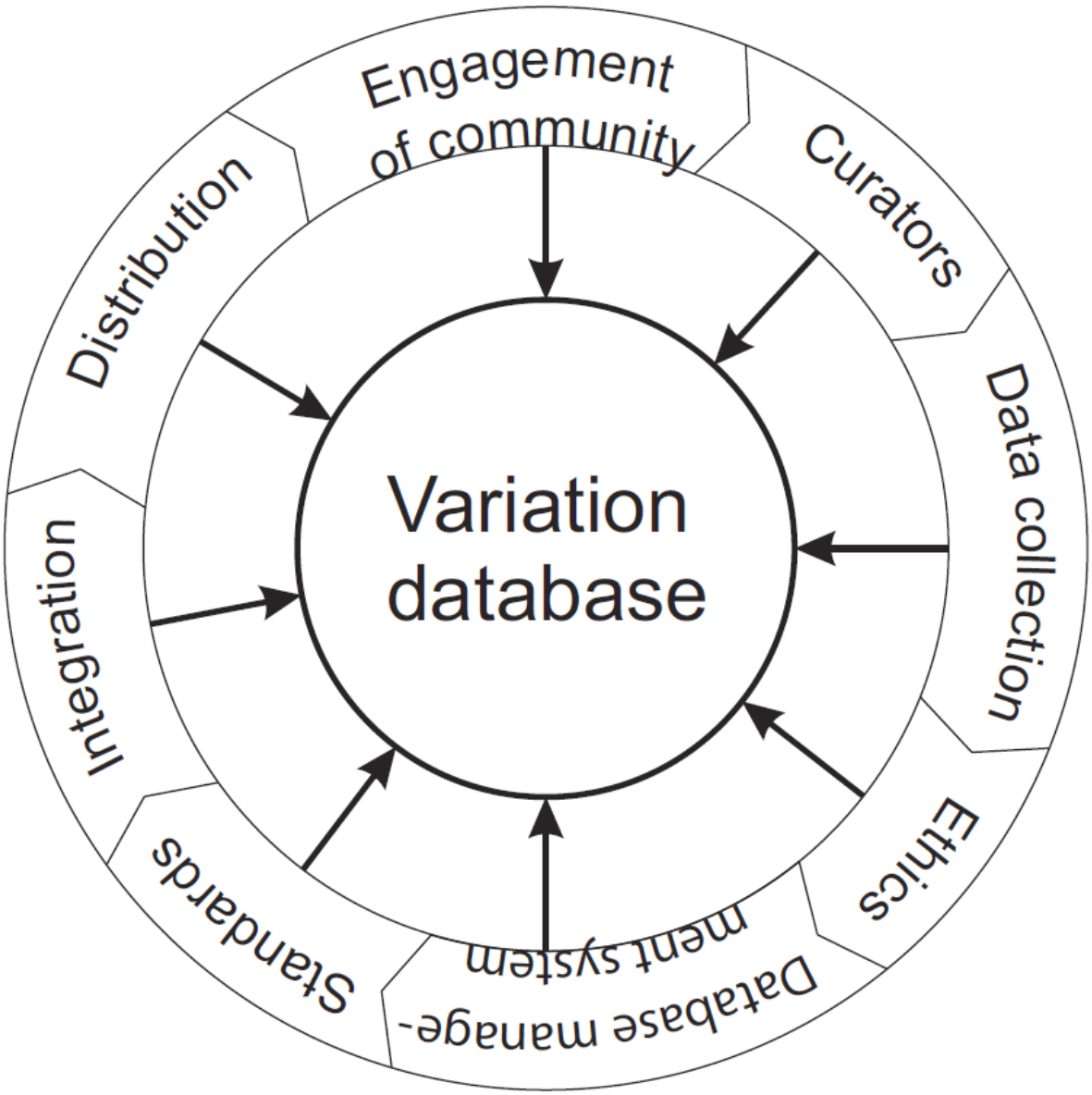

Figure 1. Aspects to be taken into account when building the OESO–HVP variation database for variants in esophageal cancer.