The InSiGHT Database – utilizing 100 Years of insights into Lynch Syndrome

J. P. Plazzer[1], R. H. Sijmons[2], M. O. Woods[3], P. Peltomäki[4], B. Thompson[5, 6], J. T. Den Dunnen[7], F. Macrae[1]

InSiGHT Variant Interpretation Committee

(1) Dept of Colorectal Medicine and Genetics, The Royal Melbourne Hospital, Parkville, Australia

(2) Dept of Genetics, University of Groningen, University Medical Center Groningen, the Netherlands

(3) Discipline of Genetics, Faculty of Medicine, Memorial University, St. John's, NL, Canada

(4) Dept of Medical Genetics, Haartman Institute, University of Helsinki, Finland

(5) Queensland Institute of Medical Research, Herston, Brisbane, Australia.

(6) School of Medicine, University of Queensland, Brisbane, Australia.

(7) Center of Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands

John-Paul Plazzer
Email: johnpaul@variome.org

Acknowledgments
The interpretation committee is operationally funded through the Cancer Council of Victoria (Australia), and the curation generously funded by the George Hicks Foundation in Melbourne Australia (to September 2012) and then The Royal Melbourne Hospital Foundation

Abstract

This article provides a historical overview of the online database (www.insight-group.org/mutations) maintained by the International Society for Gastrointestinal Hereditary Tumours. The focus is on the mismatch repair genes which are mutated in Lynch Syndrome. *APC*, *MUTYH* and other genes are also an important part of the database, but are not covered here. Over time, as the understanding of the genetics of Lynch Syndrome increased, databases were created to centralise and share the variants which were being detected in ever greater numbers. These databases were eventually merged into the InSiGHT database, a comprehensive repository of gene variant and disease phenotype information, serving as a starting point for important endeavors including variant interpretation, research, diagnostics and enhanced global collection. Pivotal to its success has been the collaborative spirit in which it has been developed, its association with the Human Variome Project, the appointment of a full time curator and its governance stemming from the well established organizational structure of InSiGHT.

Keywords
Lynch Syndrome, InSiGHT database, microattribution, variant interpretation, variant classification

Definitions
A *database entry* is defined as a single record of a variant in the database. A *unique variant* is a summary account of all the entries of a variant.

Introduction

The International Society for Gastrointestinal Hereditary Tumours (InSiGHT) is the peak professional body representing the interests of healthcare workers interested in familial gastrointestinal cancer. It was formed by the merger of the Leeds Castle Polyposis Group and the International Collaborative Group on Hereditary Nonpolyposis Colorectal Cancer (ICG-HNPCC), which were both established around the time of the cloning of *APC* and the Mismatch Repair (*MMR)* genes. Its governance is democratic, and geographically and broadly representative of the stakeholders and disciplines in the field**.** InSiGHT was incorporated in 2010, in response to a need to be able to negotiate with commercial (e.g private DNA diagnostic laboratories) and public entities. InSiGHT's mission includes development of its database, facilitating international collaborative projects among its members, supporting the implementation of programs for research into familial gastrointestinal (GI) cancer, and applying a robust system to assign disease associations (pathogenicity) to variants.

Lynch Syndrome (previously known as Hereditary Non Polyposis Colorectal Cancer Syndrome) is due to the inheritable consequences of a variant in the mismatch repair system, leading to rapid somatic accumulation of secondary (sometimes also driver) variants in a tumour, unable to be corrected by the defective MMR machinery. As a consequence, the tumour spectrum is wide, but bowel cancer predominates. Lynch Syndrome has evolved in the literature from being defined by Amsterdam (or Bethesda) criteria, to being simply families where pathogenic MMR gene variants have been identified. Screening strategies have also evolved from ascertainment based on family history, to testing all cancers presenting <50 for MMR deficiency, through to testing all cancers regardless of age in the same manner. This means efforts to categorize pathogenicity based on all available evidence becomes crucial in an egalitarian world, where access to these diagnostic deliberations can, and should, be shared across the community.

The InSiGHT database is the primary store of public information of inherited gastrointestinal (GI) cancer gene variants. The intended uses of the database are to assist clinicians in providing accurate healthcare to their patients and to facilitate biological and clinical studies on variants associated with hereditary GI syndromes. This is primarily accomplished through sharing of variant information from individual clinics and regional or national organisations. InSiGHT is consolidating data from all such sources into one central and public web-accessible database powered by the Leiden Open Variation Database (LOVD) [1] platform. Although the database has traditionally been variant centric, data can be entered in a patient centric manner as well. A typical entry records the variant as well as data from the patient: age, sex, tumour microsatellite instability (MSI), disease type and family history. In addition, the database contains a considerable catalogue of assay results for variants tested

functionally (in vitro) and links to the relevant PubMed references. All of these data contribute to the InSiGHT variant interpretation process (see below).

Establishment and evolution of the database

At a meeting of the ICG-HNPCC in Milan, Italy in 1994, it was decided that a centralised database of HNPCC variants should be created. This database began from questionnaire results completed by the meeting participants. The first published analysis of the database in 1997 noted there were 156 reports on 134 unique variants [2]. By gene, there were 71 unique variants in *MLH1* and 55 unique variants in *MSH2*. Only 7 unique variants were listed in *PMS2* and 1 for *PMS1*. The relative proportions were 53% for *MLH1* and 41% for *MSH2*, with *PMS2* and *PMS1* comprising 6%. Although *MSH6* had not been widely associated with HNPCC at the time (1997), the ratio of *MLH1* to *MSH2* variants is still consistent with proportions in the database today. In addition to the variant and protein alteration, the database also included links to published reports, the geographic origins of patients, the family's Amsterdam I criteria (Amsterdam positive, negative or not specified), and a pathogenicity class. However, at this time, pathogenicity assessment of variants was limited to a basic Yes/No statement derived from a theoretical understanding of the variant effect, and from family information. In 2003, InSiGHT was formed through the merger of the ICG-HNPPC and the Leeds Castle Polyposis Group. Then in 2004, a second database analysis was published [3] showing that over the intervening years since 1994, the size of the database had gradually increased to 556 unique variants. Contributions from individual submitters formed the majority of entries, sent via electronic submission forms. A minority of reports were extracted from published literature. It was noted in the 2004 analysis that variants were increasingly submitted on families which were small or atypical.

Emergence of new public databases

Over a decade since the initial ICG-HNPCC database was formed, DNA sequencing techniques had improved and our understanding of inherited GI cancer increased dramatically. This was apparent from the expanding number of published reports relating to the MMR genes. In 2007, a public database was established at Memorial University in Newfoundland, Canada: the Mismatch Repair Genes Variant Database (MMRGVD Literature) (http://www.med.mun.ca/MMRvariants/) [4]. The curators of this database obtained variants solely from published articles, consisting of 6136 entries on 2260 unique variants. This would go on to represent the largest single component of the InSiGHT database with 49% of total entries and covering 74% of unique variants. Of these, 1366 unique variants had not been submitted directly by any other contributor to the InSiGHT database (Table 2). 580 disease phenotype descriptions were listed, associated with 372 unique variants. 259 reports had Amsterdam or Bethesda criteria (positive or negative) listed. All of this information was obtained from 1100 published articles.

In 2008, a public database was created at the University Medical Center Groningen, the Netherlands which detailed functional assay results and *in silico* predictions [5]. With this information available, the curators aimed to facilitate interpretation of MMR variants. Known as the Mismatch Repair Gene Unclassified Variants Database (MMRUV Functional) (www.mmruv.info), it also relied on information extracted from published articles. A total of 3175 entries for 534 variants were collected. 1014 of these entries were functional results for 339 mostly missense or in-frame variants. Where possible,

additional clinical, MSI, segregation and data on population frequency of the gene variants were also extracted. The database curators assigned a degree of pathogenicity to the entries, though no formal classification algorithm was used. The classification provided in the original report, as well as the expert opinion of the curators, resulted in a limited pathogenicity assignment, based only on the particular aspect which the functional assay was designed to assess. As a result, many variants had multiple functional results which lead to multiple classifications, often not concordant. An example is MLH1:c.731G>A p.Gly244Asp which had several conflicting results (Table 1). Indeed, this database clearly demonstrated how many different test outcomes had been generated by functional analysis of different protein properties. The curators intended these classifications for research purposes only, with clinical classification awaiting the expertise of an officially recognised scientific panel. InSiGHT has identified 118 variants as having multiple interpretations and is working to assign consistent classification to each one, through its variant interpretation committee.

Table 1 MLH1:c.731G>A p.Gly244Asp, a variant with discordant interpretation - extract derived from InSiGHT database, accessed May 2012

Test Method	Test Type	Result	Classification
dominant negative effect	reporter assays in yeast	DNE in 0 out of 3 tests	PATHOGENIC
expression level of mutant allele	in vivo assay in human cell line	comparable to WT	NEUTRAL
mutation rate at HPRT gene	in vivo assay in human cell line	increased mutation rate	PATHOGENIC
tolerance to methylating agents	in vivo assay in human cell line	tolerance as in WT	NEUTRAL
MMR activity assay	in vitro assay	19,4% compared to 0% in MLH1 deficient cell line	VUS
MLH1 expression	protein abundancy	>75% of WT level	NEUTRAL
comparison of mutation rate between haploid yeast	functional assays using yeast	mutation rate in a lys+ reporter gene comparable to a MLH1 deficient haploid yeast	PATHOGENIC
pSPL3 minigene	splicing assay	no change in exon inclusion	NEUTRAL
human cell extract+in vitro MMR assay	cell based in vitro MMR functional assay using a human expression system	reduced repair efficiency compared to the WT	PATHOGENIC

The InSiGHT database - a pilot project for the Human Variome Project

During the 2007 InSiGHT meeting in Yokohama, Japan, Professor Richard Cotton presented the work of the Human Variome Project (HVP), at which point the shared interests of InSiGHT and the HVP in increasing variant submissions to central databases became apparent. A pilot project was initiated between InSiGHT and the HVP to improve the systems and processes for sharing of variant information at the gene/disease specific level [6]. A major outcome of the Yokohama meeting was the establishment of the Gene/Disease Specific Databases (GDSDb) as an integral part of the HVP vision. Additionally, each GDSDb would integrate into the Country Node approach of the HVP, which sees individual country's co-ordinating the collection of all variant information into a central national database. InSiGHT, as a GDSDb will receive data from HVP Country Nodes, such as the Australian Node which is currently expanding to cover 15 laboratories. The main advantage of a GDSDb is to allow gene/disease specialists to "Value-add" to the data collected in Country Nodes, for example, by providing expert interpretation of variant effects, or through expert curation. A key aspect of any GDSDb is the that there should be a single repository of data, created through the sharing and combining of different sources into one central resource. This concept was embraced by InSiGHT which had 3 separate databases which could be combined. Thus, the InSiGHT database became a model system of the HVP for other gene/disease groups to emulate.

Databases merge to form the current database

Following the collaboration with the HVP, InSiGHT merged the large public databases into one centralised system based on version 2 of the LOVD platform for gene variant databases. By this time, the original ICG-HNPCC/InSiGHT database had grown to 956 reports of 570 variants. With the merger of the existing databases, and incorporation of new submissions from the German HNPCC consortium and the Netherlands, the InSiGHT database expanded to over 12,000 entries. All the information present in the different systems was combined into one database by adding new fields to the database to match those in the original systems. Thus, the InSiGHT database became a complex mix of data types (variant, functional assay, *in silico*, patient demographics,disease, tumour information), collected from a variety of sources (literature, individual submissions, national collections).

As of May 2012, there are 12,538 entries for 3,072 unique variants across the MMR genes in the InSiGHT database. Categorizing by gene indicates that *MLH1* has 42% of the unique variants, followed by *MSH2* with 33%. *MSH6* and *PMS2* comprise of 18% and 7.5% respectively. Frequency of the different variant types are as follows: nonsense/frameshift alterations are 36%, missense variants are 34%, followed by intronic variants at 11%. Large genomic rearrangements and splice-site variants are around 6% each. Silent and in-frame variants are 4.6% and 1.9% respectively, with splicing aberrations making up 1% of variants (Table 2).

Table 2 Variant Types as of May 2012 in the InSiGHT database (Number of Unique Variants)

Gene	MISSENSE	SILENT	IN-FRAME^	INTRONIC	NONSENSE/FRAMESHIFT	LGR#	SPLICE SITE#	SPLICING ABERRATION**	Total by Gene
MLH1	445	48	21	129	439	82	96	21	1281 (41.7%)

MSH2	272	42	22	84	431	88	64	6	1009 (32.8%)
MSH6	203	42	12	96	180	8	10	1	552 (18.0%)
PMS2	122	10	2	23	47	20	5	1	230 (7.5%)
Total by type	1042 (34.0%)	142 (4.6%)	57 (1.9%)	332 (10.8%)	1097 (35.7%)	198 (6.4%)	175 (5.7%)	29 (0.9%)	3072

^ In-frame variants refer to insertions, deletions or indels which do not affect the reading frame.
#Large Genomic Rearrangement.*The splice site variants are the untested variants in the canonical dinucleotides. **splicing aberrations are the variants that have been experimentally shown to cause splicing aberrations, regardless of their location (these would include some variants in the canonical splice sites).

To understand the breakdown of contributions, all contributors were divided into 6 groups corresponding to their organisational or geographic origin (Table 3). Their submissions were analysed to see which were solely attributed to their region or organisation, or which were shared with other submitting organisations.

Table 3 Novel contributions by submitters form a large proportion (67.2%) of the database, as accessed May 2012

Contributing centre	Unique Variants	Novel contributions	Contributions overlapping with other centres	Percent Novel
MMRGVD (literature)	2260	1366	894	60.4%
ICG-HNPCC	570	172	398	30.2%
Netherlands	382	138	244	36.1%
German HNPCC	445	126	319	28.3%
MMR UV (Functional)	534	115	419	21.5%
Other (small contributors)	286	138	148	48.3%
Whole database	3072	2064	1008	67.2%

Out of the 3,072 unique variants in the InSiGHT database, 2,064 were acquired from a single regional or centralised resource, as shown in Table 3. The remaining 1008 variants are common to 2 or more

centres. The largest source of novel variants is the MMRGVD literature database. The remaining sources had relatively small (<10%) numbers of new variants. However, if looking at each contribution individually, a large portion (37% on average) of their submissions are novel variants. This indicates the importance of each contributing centre - all have made distinct contributions to the database. With the combined information, a clearer understanding of the MMR variant spectrum is obtained. For example, the curators of MMRGVD reported on the proportion of variants found in their database in 2007 [4]. They then found that *MSH2* had the highest proportion of truncating and splicing variants, while *MLH1* had the most missense and silent substitutions. A similar analysis of the InSiGHT database now shows that *MLH1* now leads in both classes (Table 2). This example shows how a more complete picture is now possible since the merger of the databases.

Discordant data and Interpretation

In March, 2011 InSiGHT established a variant interpretation committee which is tasked with assigning pathogenicity to MMR variants. Over 45 experts from InSiGHT have volunteered their time on a regular basis to classify variants. The process of interpretation involves a 4 week literature review, informed by the curator, where members independently assess the literature and pathogenicity of variants, with each variant having 4 independent reviewers. A summary worksheet is produced detailing all the evidence and each reviewer's classifications. Finally, a teleconference is held where all participants discuss the reviewed variants to ensure a consensus classification is reached. A rotating panel of around 15-20 members are included on each call, alongside a core stable representation of the chairman, curator and InSiGHT secretary. A set of rules has been employed to ensure each classification is based on standard evaluation of the available evidence, and these rules have been modified as the experience of classification has progressed. The committee has adopted the 5-class system recommended by the International Agency for Research on Cancer (IARC)[7]. The InSiGHT interpretation committee is invited to send any unpublished information on the variants to be considered at each teleconference, to inform the discussion. Sometimes the unpublished information can be crucial to the interpretation (e.g. on co-occurrence or segregation analysis). Variants with discordant classifications known from literature were chosen as the first batch to be assigned pathogenicity. When this first round of classification is complete, results will be published on the InSiGHT website. Plans to further this approach include invitations to the InSiGHT community to submit unclassified variants to the curator, and a broader approach to "flush out" unpublished data that could be important in interpretation. As a control, variants considered pathogenic on biological grounds (e.g. truncating) are being submitted to the same process for internal validity purposes. This committee reports and is under the control of the InSiGHT database governance committee which reports to Council. It is also assisted by an Ethics Committee, formulated in 2012 to address ad hoc enquiries. Members provide their time and expertise pro bono.

Microattributions for database submissions

Due to the evolving requirements of the interpretation process, it has become apparent that a new way of providing attribution to database submitters is needed. Information reviewed by the interpretation committee is combined in a complex format for multifactorial analysis. This requires a stringent method to keep track of each data element and provide accurate attribution to the submitter. Microattributions are a way to recognise submitters for their data which is collected through clinical or research efforts

and stored in a database, but below the threshold for acceptance in the peer reviewed literature [8]. Microattribution envisions a publicly accessible central repository such as an National Center for Biotechnology Information (NCBI) or European Bioinformatics Institute (EBI) database. From this central database, submitters can have their contributions recognised by journals, academic institutions and other organizations. Though still a work in progress, it is hoped that microattribution will provide incentive for submitters to provide their data to InSiGHT which can also be cited in the same way as published articles. Considerable work has been done to define the unit of information which would receive a microattribution and which will enable a more sophisticated way to share data between databases. This will also ensure proper attribution (and recognition for curriculum vitae purposes) for submitters when their data is cited in publications. The submission process is likely to remain the same as in the past, except for the addition of an Author ID to be provided by submitters. An Author ID can be obtained from about.orcid.org or other ID schemes. The microattribution process may evolve to being a key performance indicator for diagnostic and research laboratories with funding associations. The InSiGHT database is one of a few LSDBs who is actively developing this provision in their database.

InSiGHT database: looking forward

There is ongoing activity in all aspects of the database: collection and submission of data, annotation and classification, data model improvement, microattribution, defining database disclaimers and updating of reference sequences. InSiGHT is collaborating with international partners to facilitate the free and open sharing of annotated and curated information. New systems and processes are under development with InSiGHT's experience and knowledge assisting the HVP, Gen2Phen and NCBI. InSiGHT is also negotiating with various national organisations to facilitate the incorporation of their variant information. Given the fact that the types and amount of data contained in the database have already expanded much over the recent years and are expected to continue to do so in the future, the database user interface will be further developed to allow for easier navigation and customized display of database query results. LOVDv3 is the next version of the database software that will become available early 2013, adding further flexibility and enabling even more complex data submissions. e.g. a patient's detailed disease history. The research potential of the InSiGHT database can be realised by detailed analysis of the variant spectrum, by gene and exon, especially when cross-referenced to patient demographics and disease information. Additionally, the results of the interpretation process will allow the calibration of automated methods of variant classification. These aims remain as future challenges, once the current goals are accomplished. InSiGHT continues to encourage contributions to the database from all laboratories, and will endeavour to improve upon the existing systems and processes for the benefit of all clinicians and patients. The familial bowel cancer community is invited to engage with the process, through submission of data, joining the variant Interpretation Committee, and/or joining InSiGHT. InSiGHT's next biennial meeting in Cairns, Australia, in August 2013 will be an opportunity for all stakeholders and interested healthcare workers to learn, engage, and contribute to the process which has to date been widely acclaimed.

References

1. Fokkema IF, Taschner PE, Schaafsma GC et al (2011) LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32:557-63

2. Peltomäki P, Vasen H (1997) Mutations predisposing to hereditary nonpolyposis colorectal cancer: database and results of a collaborative study. The International Collaborative Group on Hereditary Nonpolyposis Colorectal Cancer. Gastroenterology 113:1146-58

3. Peltomäki P, Vasen H (2004) Mutations associated with HNPCC predisposition -- Update of ICG-HNPCC/INSiGHT mutation database. Dis Markers 20:269–276

4. Woods MO, Williams P, Careen A et al (2007) A new variant database for mismatch repair genes associated with Lynch syndrome. Hum Mutat 28:669-73

5. Ou J, Niessen RC, Vonk J et al (2008) A database to support the interpretation of human mismatch repair gene variants. Hum Mutat 29:1337-41

6. Kaput J, Cotton RG, Hardman L et al (2009) Planning the Human Variome Project: The Spain Report. Hum Mutat 30:496-510

7. Plon SE, Eccles DM, Easton D et al (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 29:1282-91

8. Giardine B, Borg J, Higgs DR et al (2011) Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nat Genet 43(4):295-301

Author/s:
Plazzer, JP; Sijmons, RH; Woods, MO; Peltomaki, P; Thompson, B; Den Dunnen, JT; Macrae, F

Title:
The InSiGHT database: utilizing 100 years of insights into Lynch Syndrome

Date:
2013-06-01

Citation:
Plazzer, J. P., Sijmons, R. H., Woods, M. O., Peltomaki, P., Thompson, B., Den Dunnen, J. T. & Macrae, F. (2013). The InSiGHT database: utilizing 100 years of insights into Lynch Syndrome. FAMILIAL CANCER, 12 (2), pp.175-180. https://doi.org/10.1007/s10689-013-9616-0.