



www.gen2phen.org

D1.2. Initial Report from the Project Assessment Pilot

WP1 – SCIENTIFIC COORDINATION

**V1.3
Final**

Lead beneficiary: ULEIC

Date: 06/02/2009

Nature: Report

Dissemination level: PU



 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

TABLE OF CONTENTS

DOCUMENT INFORMATION	3
DOCUMENT HISTORY	3
DEFINITIONS	4
1. EXECUTIVE SUMMARY	5
2. INTRODUCTION.....	5
3. DESCRIPTION OF WORK	6
4. LYNCH SYNDROME.....	6
5. SUMMARY OF INTERVIEWS.....	7
6. LYNCH SYNDROME DATA USE.....	8
7. DEVELOPMENT OF DATA STANDARDS.....	18
8. DEVELOPMENT OF GENOME-WIDE DATABASES.....	19
9. DATA FLOWS.....	19
10. DATA ACCESS TECHNOLOGIES.....	20
11. SUMMARY OF ISSUES RELATING TO GEN2PHEN	20
12. REFERENCES.....	22
13. SUPPLEMENTARY.....	26

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Document Information

Grant Agreement Number	HEALTH-F4-2007-200754	Acronym	GEN2PHEN
Full title	Genotype-To-Phenotype Databases: A Holistic Solution		
Project URL	http://www.gen2phen.org		
EU Project officer	Frederick Marcus (Frederick.Marcus@ec.europa.eu)		


Deliverable	Number	1.2	Title	Initial Report from Project Assessment Pilot
Work package	Number	1	Title	Scientific Coordination

Delivery date	Contractual	Month 12	Actual	Month 13
Status	Version 1.3		final <input checked="" type="checkbox"/>	
Nature	Report <input checked="" type="checkbox"/> Prototype <input type="checkbox"/> Other <input type="checkbox"/>			
Dissemination Level	Public <input checked="" type="checkbox"/> Confidential <input type="checkbox"/>			

Authors (Partner)	M. Cornell, A. Devereau (UNIMAN)			
Responsible Author	M. Cornell		Email	michael.cornell@cmmc.nhs.uk
	Partner	UNIMAN	Phone	+44 (0)161 2768716

Document History

Name	Date	Version	Description
M. Cornell (UNIMAN)	1/12/08	V0	Document creation
A. Devereau (UNIMAN)	1/12/08	V0	Document refinement
M. Cornell (UNIMAN)	8/12/08	V1	Document revision
A. Devereau (UNIMAN)	9/12/08	V1	Document refinement
M. Cornell (UNIMAN)	23/12/08	V1.1	Document revision
M. Cornell (UNIMAN)	4/02/09	V1.2	Document revision following internal review
	6/02/09	V1.3	Consortium review

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Definitions

- Partners of the GEN2PHEN Consortium are referred to herein according to the following codes:

ULEIC – University of Leicester (UK) – Coordinator

EMBL – European Molecular Biology Laboratory (Germany) – Beneficiary

FIMIM – Fundació IMIM (Spain) – Beneficiary

LUMC – Leiden University Medical Center (Netherlands) – Beneficiary

INSERM – Institut National de la Santé et de la Recherche Médicale (France) – Beneficiary

KI – Karolinska Institutet (Sweden) – Beneficiary

FORTH – Foundation for Research and Technology Hellas (Greece) – Beneficiary

CEA – Commissariat à l’Energie Atomique (France) – Beneficiary

EMC – Erasmus Universitair Medisch Centrum Rotterdam (Netherlands) – Beneficiary

UH.FGC – Helsingin Yliopisto (Finland) – Beneficiary

UAVR – Universidade de Aveiro (Portugal) – Beneficiary

UWC – University of the Western Cape (South Africa) – Beneficiary

CSIR – Council of Scientific and Industrial Research (India) – Beneficiary

SIB – Swiss Institute of Bioinformatics (Switzerland) – Beneficiary

UNIMAN – The University of Manchester (UK) – Beneficiary


BIOBASE – BioBase GmbH. (Germany) – Beneficiary

deCODE – Islensk Erfoagreining EH (Iceland) – Beneficiary

PHENO – Phenosystems S.A. (Belgium) – Beneficiary

BCP – Biocomputing Platforms Ltd. Oy (Finland) – Beneficiary

- Grant Agreement:** The agreement signed between the beneficiaries and the European Commission for the undertaking of the GEN2PHEN project (HEALTH-200754).
- Project:** The sum of all activities carried out in the framework of the Grant Agreement by the Consortium.
- Work plan:** Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out for the GEN2PHEN project, as specified in Annex I to the Grant Agreement.
- Consortium:** The GEN2PHEN Consortium, conformed by the above-mentioned legal entities.
- Consortium agreement:** Agreement concluded amongst GEN2PHEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties’ obligations to the Community and/or to one another arising from the Grant Agreement.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

1. Executive Summary

Purpose: By means of a "Pilot study into G2P database usage by the HNPCC community", we aimed to draw some general conclusions that would help guide the GEN2PHEN project.

Method: We conducted structured interviews with a large number of clinical and research individuals interested in HNPCC, to assess their use and opinion of G2P database resources.

Key Findings: The pilot project community is mostly interested in LSDB related databases, and not genome wide databases. For LSDB type resources, large deficiencies were identified. These partly reflect fundamental scientific limitations in defining pathogenicity, and partly reflect gaps in the information technology backdrop. Only the latter can be addressed by GEN2PHEN, and key issues here include:

- The need for robust nomenclature and reference sequences.
- The need for an extended range of information to be held in databases (data quality information, disease specific data, and method data).
- The need for a universally agreed list of 'pathogenicity categories'.
- The need for 'generic' GEN2PHEN databases, data models, and software to be flexible enough to handle differing use case requirements.
- Problems with analysis software provenance and reliability, which perhaps argues for these roles being delivered by the commercial sector.
- Challenges with data submission to public databases (single gene mutation scans through to GWAS). Such submissions need to become routine, but this is made difficult by the complexities of the process and the lack of sufficient incentive.
- The need for better search capabilities.
- The desirability, but uncertain feasibility, of web-services based pathogenicity interpretation.
- The need for improved advertising and deployment of GEN2PHEN solutions.
- The need for guidance on legal and ethical aspects of making public and storing patient G2P data.


Some aspects of GEN2PHEN are not covered in this report, such as:

- The integration of data across LSDBs, which would justify improved harmonisation of data models.
- Representing LSDB data in central browsers, a focus of WP6.
- The ethical implications of pushing core pathogenicity data into central browsers.
- Whether linking so much data could make individuals easier to identify.

Further elaborations of how the findings of this Pilot could or should impact GEN2PHEN are provided in Section 11.

2. Introduction

A priority for WP1 is to continually assess the effectiveness and utility of the items emerging from GEN2PHEN. This deliverable represents the first instance of a rolling 'Pilot Project' which will try to use the developing G2P database network for the purpose for which it was created, i.e. to explore G2P relationships in depth and across a range of species and situations and thus

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

fundamentally assess and potentially redirect efforts in all other work packages. However, at this early stage of the project, this type of investigation is not yet possible. Therefore this first report presents an overview of current research practices and notes problem areas in G2P research, and uses this to assess and steer the project plans.

The first phase of the pilot focuses on Lynch syndrome, also known as HNPCC (hereditary nonpolyposis colorectal cancer), and documents the requirements of a defined user community. Lynch syndrome is a complex inherited disease: there are multiple genes associated with the syndrome and multiple phenotypes. As well as causing colorectal cancer, mutations in Lynch syndrome-related-genes cause cancers in other tissues. The nature of the syndrome therefore means that conclusions should be applicable to other complex diseases. In addition, there is a large and established professional society representing those working on this disease and this helped to provide the defined user community for the study.

3. Description of work


In order to identify the needs of Lynch syndrome researchers a series of interviews with data users were conducted. Potential interviewees were selected either (i) from the membership list of InSiGHT (International Society for Gastrointestinal Hereditary Tumours); (ii) from recent publications on Lynch syndrome and/or genome wide association studies on colorectal cancer; and (iii) staff at UK regional genetic centres. Details of interviewees are listed in Supplementary Table 1. More than 90 individuals were contacted requesting an interview, 30 individuals were interviewed. Interviews were mostly conducted by telephone or face to face, with a few exceptions where the interviewee requested a list of questions by email. Interviews discussed the following topics: the interviewee's role and how Lynch syndrome data is being used; the software they are using and its limitations, data flow into public repositories, developing data standards and the development of genome-wide genotype to phenotype databases. A full list of the questions is provided in Supplementary data.

4. Lynch syndrome

Lynch syndrome is the form of hereditary colorectal cancer (CRC) caused by mutations in DNA mismatch repair (MMR) genes. It is responsible for 1–5% of all CRCs. Several MMR genes have been linked to Lynch syndrome. Those most frequently identified are MSH2 and MLH1, while others include MSH6, PMS1, MLH3 and PMS2. In addition, a large percentage of Lynch syndrome cases (~20-25%) have no apparent mutations in MMR genes, suggesting that further Lynch syndrome genes will emerge.

Lynch syndrome is associated primarily with colorectal cancer but also with cancers of the endometrium, ovary, stomach, small intestine, hepatobiliary tract, upper urinary tract, brain, and skin.

It is very important to distinguish CRC due to Lynch syndrome from sporadic cases (or other hereditary syndromes). Positive identification will affect the clinical management. It will require surveillance, both for the affected patient and also for mutation-positive relatives. Currently, germline sequencing of MMR genes is too time-consuming, difficult, and expensive to be feasible for all CRC patients. Therefore, there will be a screening process which determines whether sequencing should be carried out.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

5. Summary of interviews

5.1 Data Users

For the purposes of conducting interviews, the following classification of data users was used.

- Medical geneticists
- Genetic Counsellors
- Molecular biologists
- Clinical Scientists (Diagnostic DNA labs)
- Database curators
- Gene Association Researchers
- Interpretation committee members.

In addition, the interviews covered both lab heads, who might have a better understanding of future trends in G2P research; and lab researchers, who might have a better understanding of day-to-day problems.

5.2 Aims


Interviewees were questioned about how they used existing HNPCC resources.

Use	Number of interviewees
Clinical use (interpretation of variants)	15
Research	7
Don't use any (interviewee relies on the interpretation of variants supplied by another scientist)	2
Developing new systems (e.g. INFOBIOMED)	3

5.3 Resources used

Which online resource is used?

Resource	Number of interviewees
InSiGHT database	10
MMRVariants	7
DMuDB	3
LOVD	1
Unclassified Variants Database	1
dbSNP	1
PubMed	1
Google/Google Scholar	4

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

5.4 Reason for choice of LSDB

InSiGHT was the first publicly available HNPCC database. Some of those interviewed have continued using InSiGHT and not investigated other databases; while some use InSiGHT because it's the one that their colleagues use.

Others have migrated from InSiGHT and in most cases are using the MMRVariants database mainly because they felt it was more up to date and had links to publications.

Only one interviewee not employed at a UK diagnostic lab mentioned the diagnostic mutation database (DMuDB) and only one was aware of LOVD.

The number of interviewees using the InSiGHT database may reflect a bias in the selection of interviewees.

5.5 Are current resources suitable for users' needs?

Users were evenly split when asked whether current databases met their needs. Just over 55% said they were suited to their needs. However, when asked about positive and negative features of their LSDB of choice they had many criticisms.


6. Lynch Syndrome data use

In order to highlight problems with LSDBs, the software used in the analysis of HNPCC data was considered. For this data use was split into “clinical” and “research”. Clearly this split is somewhat artificial, since research into Lynch syndrome may well have a clinical impact. Therefore, “clinical” here represents data use for genetic testing of patients, while “research” represents data used for producing publications. At different stages of the research process, aspects of the data that may need to be modelled have been highlighted (“data issues”). There will be important issues concerning data quality. Quality can refer to both the completeness of the data (i.e. is there sufficient metadata to ensure that the entry will be useful); and accuracy of the data (i.e. has the variant been accurately identified). This second aspect is rather more difficult to deal with and it may be that there needs to be restrictions as to who can submit data.

6.1 Clinical use

Much of the Lynch syndrome data is being generated by genetic testing laboratories. In these cases a patient with cancer is tested to determine whether their cancer is caused by a germline mutation (i.e. Lynch syndrome). If so, this will lead to testing of relatives. It is not the case that all patients with colorectal cancer are tested for Lynch syndrome. Patients' backgrounds are examined for evidence of familial cancers. The Amsterdam and Bethesda criteria have been developed for evaluating a patient's suitability (see Supplementary data). These criteria have relaxed over time and further changes may occur (Byfield and Syngal, 2008). For example, Harris (2008) suggests that all young (under age 50) patients with colorectal cancer should be screened. Therefore the numbers of individuals likely to be tested will probably increase.

Data Issue: Interviewees at genetic testing labs have stated that they place more faith in data from other genetic testing labs than from non-diagnostic labs. They feel that the rigorous reviewing process that takes place before a report is signed off ensures that the results are accurate. Therefore the data source (institution) should be stored. There are also details about the patient that would be useful. For example, age of onset of disease and ethnicity.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

6.1.1. Supporting Evidence for Testing

Other evidence for the presence of Lynch syndrome can include micro satellite instability (MSI) and immunohistochemistry (IHC). These could be used as supporting evidence in the decision as to whether to sequence MMR genes. If such supporting evidence is to be stored in a database it is important to consider the extent to which the evidence is defined. For example MSI can be type A (alterations are defined as length changes of ≤ 6 bp) or Type B (modifications of ≥ 8 bp) (Oda *et al.*, 2005). For IHC there are different protocols involving different antibodies, with the result that there can be different IHC results for the same variant (Ian Frayling, personal communication). UK guidelines have been developed for IHC (<http://cmgsweb.shared.hosting.zen.co.uk/BPGs/pdfs%20current%20bpgs/hnpcc%20recommendations%20B.pdf>)

Other supporting evidence can also be available. For example, Muir-Torre Syndrome can be a good indicator of HNPCC, while BRAF mutations are often evidence against Lynch syndrome (Loughrey *et al.*, 2007).

Data Issue: What supporting data should be stored? In the case of Amsterdam criteria this could be done using a Boolean. Other data, e.g. details of Muir-Torre syndrome and BRAF mutations, could be stored as free text, taking into account that there may be no available data. In addition, it needs to be remembered that this is very much disease specific. The Amsterdam and Bethesda criteria only refer to Lynch syndrome; other diseases will have different criteria.

6.1.2. DNA source

Usually DNA is extracted from a blood sample. However, in some cases (e.g. patient is deceased) the DNA is extracted from tumour tissue. In such cases variants may not represent germline mutations.

Data: Tissue type should be stored.

6.1.3. Testing Methods


Genetic testing methods can vary. Often it consists of the PCR amplification and sequencing of exons from MLH1, MSH2 and in some cases MSH6. In other cases multiplex ligation-dependent probe amplification (MLPA) is used to detect genomic deletions and insertions. There are some advantages to non sequencing methods. For example, sequencing may not detect mutations involving whole exon deletions or insertions (Akrami *et al.*, 2005) Other methods, such as denaturing gradient gel electrophoresis (DGGE), denaturing high performance liquid chromatography (DHPLC), and single strand conformation analysis (SSCA), have been developed but no longer appear to be widely used.

It has been suggested in interviews that metadata about testing methods should be stored. The use of microarray based technology for variant identification (Beaudet and Belmont 2008), may require metadata about hybridisation and normalisation.

Data Issue: Should the identification method be stored? There are a limited number of these so it might be possible for users to choose from a list, rather than permitting free text. It might also be that multiple methods have been used.

6.1.4. Sequences Tested.

As discussed above, the current practice is to sequence exons from one or more MMR genes. This may vary in future, perhaps more genes will be sequenced or important intronic sequences

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

will be identified. The link constructed between genotype and phenotype is therefore based upon our knowledge of those regions which are sequenced.

Data Issue: For a patient, data on the genes and exons sequenced should be stored.

6.1.5. Sequence Analysis

When sequence data is generated, variations from reference sequences can be identified. Many labs are now using the commercial software Mutation Surveyor developed by Softgenetics (<http://www.softgenetics.com/ms/index.htm>) and feel its results are reliable. However, some felt that software such as Staden (<http://staden.sourceforge.net>) was less accurate.

Data Issue: Details of the analysis including the sequence analysis software used needs to be stored.

6.1.6. Variant Naming

Different systems are used for naming variants but the only widely accepted standard is that presented by the Human Genome Variation Society (HGVS; <http://www.hgvs.org>). For interoperability and to avoid mistakes variant nomenclature should be in accordance with these guidelines. In some instances, e.g. single base pair changes, naming is relatively straightforward, but for more complicated mutations the naming can become problematic. Software such as Mutalyzer (<http://www.lovd.nl/mutalyzer/1.0.1>) can be used.

Data Issue: The basis of the nomenclature, i.e. HGVS and the reference sequence used, and the software used for naming variants should be stored. In addition, the variant name may be affected by the reference sequence which is being used. Therefore, the name of the reference sequence should be stored and where possible the LRG sequence should be used as a reference (see Section 7).

6.1.7. Variant Classification


Following the identification of variants, a report is produced which includes a variant classification. These reports do not necessarily have the same format but broadly consist of a summary statement and a list of the analyses conducted to support this statement. For example, the London Kennedy-Galton Centre Regional Genetics Service provide a score of between 1 and 4 (1 = certainly benign, 2 = most likely benign, 3 = most likely pathogenic, 4 = certainly pathogenic), while Liverpool include a fifth class (“unknown variant”), plus an eight page report listing the analyses that were carried out to support this score. In general mutations which result in large deletions or frame shifts are easier to call as pathogenic. Far more problematic are missense mutations. Best practice guidelines have been developed for the interpretation of variants (see http://cmgsweb.shared.hosting.zen.co.uk/BPGs/Best_Practice_Guidelines.htm).

The analyses can be roughly divided into three parts:

6.1.7.1. Database searches

These are split into searches of LSDBs and sequences of other genomic databases. Several LSDBs for MMR genes exist:


- **InSiGHT** (International Society for Gastrointestinal Hereditary Tumours; <http://www.insight-group.org>) was established 1994 and collects information about variants from individual submitters (Peltomaki and Vasen, 1997).

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

- **MUN database** (Memorial University of Newfoundland; <http://www.med.mun.ca/MMRvariants/default.aspx>) collects published variants.
- **Missense Variant Database** (<http://www.mmrmissense.info>) is designed to improve the annotation of unclassified variants.
- **LOVD** (Leiden Open Variation Database; <http://chromium.liacs.nl/lovd2/home.php>) is not restricted to MMR data. Instead it is “LSDB in a box” software that users download and install to create their own LSDBs. LOVD databases have been created for MMR genes. Also, there is an ongoing merger of the InSiGHT, MUN and Missense databases into LOVD format.
- **DMuDB** (<http://ngri.man.ac.uk/dmudb/index.html>). Again this is not limited to MMR data. DMuDB is a relational database, managed by NGRI Manchester, used to store variant data for 13 genes (including MMR genes) for UK diagnostic labs.

Other databases cited in interviews include: COSMIC (Catalogue of Somatic Mutations in Cancer; <http://www.sanger.ac.uk/genetics/CGP/cosmic>), dbSNP, Pubmed, Omin, Google and Google Scholar. However, these were clearly not central to the current needs of the HNPCC diagnostic community, who are mostly interested in Mendelian type mutations. Therefore, although these 'genome wide' databases are part of the GEN2PHEN project, this section of the Pilot report focusses upon difficulties that users have identified with the present day LSDB type databases. These may be summarised as follows:

- **Lack of data.** According to interviewees, variants identified in genetic tests are generally not found in databases.
- **Lack of data submission.** Most variant data is generated by testing labs and tends not to be included in public repositories. This process is now changing, for example UK labs are now entering data into DMuDB; and data from the German HNPCC Consortium (<http://www.egms.de/en/meetings/dkk2006/06dkk072.shtml>) has been entered into LOVD. However, the majority of data in most LOVDs (except DMuDB) still comes from publications. Even in those cases where data is submitted into repositories, the entry tends to be done in bulk by the database administrator, rather than by the user as part of the analysis process.
- **Duplicated data submission.** Some LSDBs treat separate reports of the same variant as separate entities. This can lead to situations in which a variant has been described as “pathogenic” and “non pathogenic” by different contributors. For example, the MLH1 mutation c.350C>T has 44 entries in the LOVD database. Thirteen entries describe the mutation as “probably pathogenic” while the remainder describe it as “effect unknown”.
- **Different selection criteria.** It has been suggested in interviews that the types of mutations featured in publications involve large families with clear segregation and a high degree of penetrance. These tend not to be representative of the majority of Lynch syndrome patients.
- **Up-to-dateness.** There are two related problems: Firstly, because much of the data comes from publications, there will always be a time lag before variants are identified and added to the database, and some publications are likely to be missed. Secondly, because some databases do not provide information about when they were last updated users may perceive them as being out of date.
- **Lack of trust.** In those cases where a variant is already in a database, the genetic testing procedure requires that the original publication is checked in order to determine the

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

pathogenicity. Therefore, while the description of pathogenicity listed in a database might be noted, it is not considered to be sufficient for deciding upon pathogenicity. In effect, the LSDB becomes a means for identifying publications. To this end some interviewees have stated that they find Google and Google Scholar to be more useful because (i) they can search for the variant name, something that some of the LSDBs don't permit; and (ii) using Google means they are less likely to miss papers.

- **Lack of supporting data.** As detailed above, scientists will use background data on segregation, MSI, IHC etc in deciding upon pathogenicity. However, in many databases this supporting data is not included. In some (e.g. LOVD) there are fields for some of this data but they are usually not populated. For example, the origin field should contain either “somatic” or “germline” but in the vast majority of cases is “-“ or “?”.
- **Classification differences.** For example, most of the MLH1 variants in the InSiGHT database were classified as pathogenic, and this database only allows pathogenic and non-pathogenic classifications (not unknown or unclassified). In comparison, the LOVD reported classifications of 3467 MLH1 variants as follows:


no known pathogenicity	288
probably no pathogenicity	3
probably pathogenic	2
Pathogenic	893
effect unknown	2460

6.1.7.2. *In silico* analyses

Although GEN2PHEN has only a minimal effort towards the development, integration and use of variant analysis tools, the use of these tools influences variant classification which in turn adds to the metadata that may be collected for a variant. Therefore, an analysis of the tools that are currently being used is included. Clearly the tools used will change over time as new methods are developed, such as MAPP-MMR (Chao *et al.*, 2008)

Protein Structure/Function Tools

- **PolyPhen** (<http://genetics.bwh.harvard.edu/pph/>) – Predicts the functional effects of human nonsynonymous SNPs.
- **SIFT** (<http://blocks.fhcrc.org/sift/SIFT.html>) predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. Note that there is another website MutDB (Dantzer *et al.*, 2005; <http://mutdb.org/>) which is a resource for scientists to identify the likely underlying molecular effects of a non-synonymous SNP. MutDB has a web service API developed using the SOAP protocol.
- **Align GVGD** (http://agvgd.iarc.fr/agvgd_input.php) analyses amino acid substitutions. This software considers the conservation of different amino acids using an alignment of paralogs.
- **Russell Analysis.** (<http://www.russell.embl.de/aas/>) Looks at effect of amino acid substitution
- **Grantham Analysis.** (Grantham 1974) This categorizes codon replacements into classes of increasing chemical dissimilarity. Replacements are conservative (0-50), moderately conservative (51-100), moderately radical (101-150), or radical (151).

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Splice site prediction tools

- **Fruitfly** (http://www.fruitfly.org/seq_tools/splice.html) this is splice site prediction software which uses a neural network. It was developed by the Berkeley Drosophila Genome Project.
- **NetGene2** (<http://www.cbs.dtu.dk/services/NetGene2/>) also a neural network, Developed by Centre for Biological Sequence Analysis at Technical University of Denmark.
- **Human Splicing Finder** (<http://www.umd.be/HSF/>).

Problems with *in silico* analyses include the following:

- **Software provenance and reliability.** At present analysis involves using multiple software tools to perform essentially the same analysis (e.g. splice site analysis). Users complain that they don't know enough about the software, whether it's known to be more reliable, and whether it's been developed by a well established group or is "somebody's MSc project". Coupled with this there are problems of reliability, websites cease to exist with no warning. This perhaps argues for these roles being delivered by the commercial sector.
- **Software use and interpretation.** The interviewees complained that they were not always certain that they were using the software in the correct way. For example, Align GVGD requires a protein alignment as an input. Interviewees were not certain which species should be used to generate an alignment. There are in fact guidelines produced by CMGS which list the species required. However, in practice this will vary between proteins, the range of species required for the optimal MLH1 alignment might be different from BRCA2. In addition, for some species there are multiple sequences available and interviewees were not certain which should be used. There may also be problems produced by the software used to generate alignments. Will those generated by ClustalW necessarily be the best? These problems are fairly intractable. At present there is no expert tuition available to advise on how to, for example, produce the best alignment for MLH1.


Data Issues: If the results of *in silico* analyses are to be included in databases it is important that the software and settings are recorded. It is also important that the tools, or at least information about them, remain available in the future.

6.1.7.3. Functional Studies

Functional assays are considered to be a reliable means for confirming pathogenicity. They can be either RNA studies (testing splicing predictions) or protein studies (functional tests for protein activity). In some cases, these tests are performed as part of the testing protocol. However, there is also functional data made available via publications. These can contradict data from family studies. For example Ollila *et al.*, (2008) have shown that the mutation c.380A>G (N127) does not appear to affect protein functionality. However this mutation has been described as pathogenic in the InSiGHT database.

6.1.7.4. Reporting a variant

As discussed earlier the above analyses are combined to generate a report which will then be used, for example by genetic counsellors, in developing screening strategies for the patient. The report will give a summary of the analyses and a score describing the pathogenicity. Points to note:

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

- These reports tend to be conservative. A mutation, especially a missense mutation, is unlikely to be scored as definitely pathogenic or definitely non-pathogenic.
- Variants which are unclassified will subsequently be periodically reviewed. This means that some or all of the above analyses will need to be repeated.
- Although the production of a report generates a great deal of data about the variant, this data does not go anywhere. It is conceivable that some of the unclassified variants that are identified in patients have been previously seen in other labs.


6.2. Research Use

Five types of research use have been considered: case studies, functional studies, association studies, modifier studies and “other disease” studies (i.e. where researchers investigate phenotypes not normally associated with variants). In each of these five, the data types used and generated have been considered.

6.2.1. Case Studies

These are similar to the clinical use considered above, but the output is a paper discussing the genotype to phenotype data for a family. For example, Agostini *et al.*, (2005) reports an atypical Turcot syndrome family with small bowel cancer. The following data are included:

- **Phenotype data of patient:** (café-au-lait spots, early onset adenomas, duodenal cancer, glioblastoma).
- **Age of disease onset:** 17-yr-old son had undergone surgery for an infiltrating G2 adenocarcinoma.
- **Phenotype data of relatives:** (colonic adenoma (mother), jejunal (maternal grandfather), lung (father), and colorectal (paternal uncle) cancers).
- **Genetic testing result from patient:** Truncating 1951C>T (Q643X) and the missense 161C>T (S46I) MSI result PMS2 expression was absent in the proband's duodenal cancer with high microsatellite instability. The paper does not discuss any mutations in other MMR genes. IHC expression patterns were normal for MLH1 and MSH2 so these were not sequenced. The MSH6 and PMS2 genes were entirely sequenced (note – presumably this means all exons) sequenced. MSH6 did not contain any significant mutations.
- **Gene testing results from relatives:** Patient's mother and his brother carried only the nonsense mutation, suggesting that the two mutations were located in two different alleles of the proband, with the nonsense mutation inherited from his mother and the missense mutation probably inherited from his father (who had died from lung cancer). The presence of the truncating mutation could not be confirmed in the maternal grandfather due to the difficulty of obtaining a PMS2-specific PCR from DNA extracted from very old paraffin embedded tissue blocks.
- **Expression analysis:** The normal cells also displayed no PMS2 expression and some degree of instability.
- **In silico evaluation of pathogenicity:** Several points support pathogenicity role: (i) Serine 46 maps on an evolutionary conserved domain and is an evolutionary conserved amino acid; (ii) the amino acid change is not conservative and is predicted to be “probably damaging” (PolyPhen); (iii) this variant was absent in 118 normal control chromosomes from healthy blood donors.
- **Genotype to phenotype association:** PMS2 and Turcot syndrome.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

- **Clinical recommendation:** Results support the hypothesis that patients with a few polyps, small bowel tumours with a very early onset, glioblastoma, and café-au-lait spots should be considered as a variant of hereditary nonpolyposis colorectal cancer.

Points to note

- The paper makes no reference to any LSDB searches. This is not unusual; LSDBs are usually not cited in research papers. For example, an internet search using Google Scholar with the search terms “InSiGHT database” and HNPCC identified 28 papers, compared to 16,100 identified using HNPCC alone.
- These two mutations are not listed in either InSiGHT or Newfoundland databases.
- The same mutations are listed in another paper (Park *et al.*, 2006).

Data Issue: How should data from publications be included within LSDBs. Papers often state that variants are pathogenic, in contrast to genetic testing labs. If these variants had been identified by a testing lab would they have drawn the same conclusions?

6.2.2. Functional Studies


The aim of these studies is to investigate mutations that have been previously identified by other researchers.

For example, Korhonen *et al.*, (2008) investigated seven missense mutations (Q24E, R647C, S817G, G933C, W1276R, A1394T, E1451K) in MLH3. The seven mutations were found in colorectal or endometrial cancer patients and reported as pathogenic by previous authors (Wu *et al.*, 2001; Liu *et al.*, 2003) and were listed as being pathogenic in the InSiGHT database. The authors used an *in vitro* mismatch repair assay to study the effect of these mutations on MLH3 function. They concluded that their studies showed that the MLH3 mutations alone do not interfere with MMR and that further studies are needed to evaluate the pathogenicity of MLH3 mutations in compound with other MMR mutations.

In a second example, Takahashi *et al.*, (2007) conducted a functional analysis of 101 MLH1 variants, 99 of which were missense variants. In this case human MLH1 variants were expressed in the yeast *Saccharomyces cerevisiae*. The authors determined the % MMR activity and % protein expression (relative to wild type) and compared these results with SIFT predictions. The agreement between their results and SIFT predictions were felt to be generally good.

Data Issues: Both of these studies refer to variants which have been identified and published. They have both used LSDBs in their publications. Korhonen *et al.*, (2008) refer to the fact that the seven MLH3 variants are described as pathogenic in InSiGHT; while Takahashi *et al.*, (2007) used InSiGHT to obtain a list of MLH1 mutations for further analysis. The types of LSDB use are rather different, one searches using variant name, the other is searches for a type of mutation (i.e. missense). This second search is rather more complicated since variants are not specifically identified as being of a particular type. LOVD does allow the user to search for substitutions but theoretically this could also include premature stops and splicing variants.

There are also data submission issues. The information gathered from these studies could well contradict previous statements regarding pathogenicity. This raises the question of who has the rights to modify an existing entry. At present because much of the data is submitted by the database administrator, this is not necessarily a problem. However, moving to a system in which

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

more data is submitted by the user requires consideration of how the findings of subsequent studies can be stored.

6.2.3. Genome-wide association studies

Association studies involve the direct testing of genetic polymorphisms in large numbers of cases versus controls. It allows the identification of lower penetrance alleles that cannot be detected by genetic linkage analysis. Twin studies have shown that ~35% of colorectal cancer can be explained by inherited susceptibility (Lichtenstein *et al.*, 2000). However, Mendelian syndromes, such as Lynch Syndromes account for <5%.

Webb *et al.*, (2006) used array based technology to genotype 1467 non-synonymous SNPs mapping to 871 candidate cancer genes in 2575 cases and 2707 controls They generated a list of 44 SNPs showing significant association with risk of colorectal cancer. It is interesting to note that a subsequent analysis of ten of these SNPs (Frank *et al.*, 2008), found that the associations could not be replicated.

To date there have been two genome wide association studies (GWAS) of colorectal cancer which have identified a total of five statistically significant loci (Easton and Eeles, 2008). These studies are not specifically concerned with Lynch syndrome. Instead, they are focussed on identifying genetic loci linked to colorectal cancer which are not associated with MMR genes.

From interviews the following features of GWAS experiments were discussed:

6.2.3.1. Selection of control groups


The selection of affected and control groups is critical and is one of the reasons for differences in results between GWA studies. Recent GWAS (Zanke *et al.*, 2007; Houlston *et al.*, 2008; Tenesa *et al.*, 2008; Tomlinson *et al.*, 2007, 2008) have used matched control groups (i.e. individuals in the control group are selected so that their ages, genders, etc, match those of the control group). An alternative to using a matched control group is to use randomly selected control groups.

In some cases GWA studies are conducted in several phases. For example, Tenesa *et al.*, (2008) genotyped 555,510 SNPs in 1,012 early-onset CRC cases and 1,012 controls (Phase 1). In phase 2, they genotyped the 15,008 highest-ranked SNPs in 2,057 cases and 2,111 controls. In these cases the selection of the control group for the first phase is more critical than for the second. There could be more than two phases, Zanke *et al.*, (2007) list seven.

The sample sizes will be important in determining which SNPs are identified. In order to identify rare SNPs a larger sample size will be required.

A further consideration is how many patients are excluded. For example, have the obvious disease cases been excluded? In the case of colorectal cancer patients, including Lynch syndrome patients will bias the study towards identification of enhancers of Lynch syndrome. In order to identify new types of loci, these patients need to be identified and excluded.

Data Issues: Although this data is considered important in the interpretation of GWAS, it is not clear how data relating to specific individuals in control or affected groups could be stored, especially considering Homer *et al.*, (2007). Possibly a general description detailing the way in which an experiment was conducted could be modelled. However, there are as yet no universally accepted standards (c.f. MIAME) detailing how GWAS should be conducted or the required meta data. Instead, the rule seems to be “if you want to publish in a good journal, you have to perform the study in a certain way”.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

6.2.3.2. Platforms and genotyping

Multiple platforms exist for GWAS (e.g. there are multiple chips for both Affymetrix and Illumina). Cross comparisons across different types of array is not straightforward and may be “discouraged by reviewers”. One example of a screen where multiple platforms were used is Zanke *et al.*, (2007).

Data Issues: Storage the type of array, hybridisation conditions, etc. is needed. Standards and models already exist.

6.2.3.3. Data submission

None of the GWAS data for colorectal cancer have been submitted to public repositories such as HGVbaseG2P (<http://www.hgvbaseg2p.org>), dbGaP (<http://www.ncbi.nlm.nih.gov/gap>) or the EGA (<http://www.ebi.ac.uk/ega/page.php>). The authors intend to submit in the future but have not yet finished their analysis of the data. The amount of work required to submit the data was also a problem. This seems typical of GWAS experiments.


6.2.4. Modifier Studies

Modifier studies are investigations of factors influencing variance of disease across a population. For both breast and colorectal cancer, there is evidence that genetic modifiers are important factors in determining when (or if) a carrier develops the disease (Scott, 2008). These are also association studies; but unlike the GWA studies discussed above, they start with a set of individuals with Lynch syndrome related mutations and identify genetic factors modifying the disease. For example, insulin-like growth factor I (IGF1) has a role in the development of colorectal cancer and elevated plasma levels of IGF1 are linked to sporadic and hereditary colorectal cancer risk. Modifier studies have shown that there is a CA repeat in the IGF1 promoter. The length of the CA repeats influences the age of onset of colorectal cancer (Zecevic *et al.*, 2006; Reeves *et al.*, 2008). Individuals with less than 17 CA repeats tend to have earlier onset of disease compared with those having 18 or more.

Modifier studies are greatly influenced by the numbers of individuals used in the study. The IGF1 study (Reeves *et al.*, 2008) discussed above involved more than four hundred individuals. In other instances the use of smaller sample sizes has led to problems. For example, Niessen *et al.*, (2006) and Steinke *et al.*, (2008) dispute the role of MUTYH as a modifier of Lynch syndrome. As noted by Scott (2008) there have been several published associations which have not been substantiated by other reports. This can lead to a problem of publication bias, with a tendency to focus on positive associations rather than those publications which refute these claims (Scott, 2008). The view of an interviewee was that for an association to be considered real it must have been reproducible in several studies.

6.2.4.1. Selection of samples.

As discussed above, the number of individuals in a study is important. Unlike the GWA studies there is no requirement for a matched group of controls. Instead, the studies involve identifying a large number of Lynch syndrome patients. This will be done using a pre-existing register of patients. For example, the IGF1 study (Reeves *et al.*, 2008) involved 443 mutation carriers from Poland and Australia with confirmed causative MLH1 or MSH2 mutations. As with the GWA studies it is unclear how data relating to specific individuals in these studies could be stored.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

6.2.3.2. Platforms

To date these studies have involved DNA sequencing, e.g. of IGF1 promoter region. However, as with GWA studies, array technology will be used in the future.

6.2.5. Other Disease Studies

The above examples all focus on cancer. However, in addition to their role in Lynch syndrome, MMR genes have been identified as having a role in male fertility. For example, Avdievich *et al.*, (2008) have shown that MLH1^{G67R/G67R} mice are sterile. The G67R mutation is located in one of the ATP-binding domains of MLH1 protein and prevents MLH1 interacting with meiotic chromosomes at pachynema. The glycine residue at position 67 is conserved in human MLH1 but the G67R variant has not been reported, although the G67E has been observed.

6.2.5.1. Obtaining phenotype data for other diseases.

Patients who are selected for Lynch syndrome screening are selected on the basis of their meeting either Amsterdam or Bethesda criteria. As a result, the phenotype data that is obtained is likely to be associated with colon cancer rather than other phenotypes (such as male infertility), which might be associated with MMR variants.


6.2.5.2. Integration of model organism data.

The use of mice as a model organism for investigating functions raises the question of how GEN2PHEN will integrate G2P data for model organism with human data. The vast majority of Lynch syndrome research does not appear to involve the use of model organisms such as mice. This possibly reflects the fact that the phenotype associated with MMR mutations differ in mice and humans. Compound mutant of some MMR mouse genes can cause adenocarcinomas of the intestines. However, their histopathology is much milder than human colorectal cancer (Taketo, 2006).

There is also the question of how variants are mapped across model organism. The entries in LSDBs refer to specific positions in a human reference sequence and mapping them across to other species (or vice versa) may not be simple. For example, the mouse G67 residue discussed above maps to G67 in humans. However, 420D in mouse MLH1 maps to 418E in human MLH1.

7. Development of data standards

Interviewees were questioned about the development of data standards by GEN2PHEN; specifically the LRG gene reference sequences and the PAGE-OM data model. This part of the interviews proved rather problematic as it is too early for the interviewees to be aware of these projects yet. When discussing the development of reference sequences, all agreed that this was useful. Although the problem of changing reference sequences had not affected MMR genes, interviewees working with BRCA sequences had encountered problems due to their re-annotation. It was rather more difficult to obtain feedback concerning PAGE-OM. The interviewees were generally not familiar with UML models. It may be that to obtain input from data users on the development of these models, real world examples (case studies or story boards) need to be provided to help explain the models.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

As well as these standards, the ongoing development of data standards for LSDBs must be considered (Cotton *et al.*, 2008; Greenblat *et al.*, 2008). For example, a recent review of LSDBs (Greenblat *et al.*, 2008) made the following recommendations:

- 1) LSDBs should only report a conclusion related to pathogenicity if a consensus has been reached by an expert panel.
- 2) The system used to classify variants should be standardized. The Working Group encourages use of the five class system (Plon *et al.*, 2008) as follows: Class 5, pathogenic (>99% posterior probability); Class 4, likely pathogenic (95–99%); Class 3, uncertain (5–95%); Class 2, likely neutral (0.1–5%); and Class 1, neutral, no clinical significance (<0.1%)
- 3) Evidence that supports a conclusion should be reported in the database, including sources and criteria used for assignment.
- 4) Variants should only be classified as pathogenic if more than one type of evidence has been considered.
- 5) All instances of all variants should be recorded.

An InSiGHT meeting was held on 9th December, 2008. InSiGHT will be developing data standards specific to Lynch syndrome. These standards will be similar to those developed by INFOBIOMED (see below).


G2P software is being developed by INFOBIOMED (<http://www.infobiomed.org/>). They are currently implementing a pilot project to collect G2P data on Lynch syndrome patients using an XML database. This is a closed hospital system and a signed contract is required for access of data.

8. Development of Genome-wide databases

Interviewees were asked about the development of genome-wide databases, specifically HGVbaseG2P. Again, this section of the interview was problematic, in part because most of the interviewees were not yet aware of these projects but also because at present studies cannot be viewed. Interviewees who were most interested in its development were those involved in GWAS and database developments. They asked to be kept informed of developments (in particular when the cancer studies can be viewed). There was less immediate interest from the genetic testing community. This will change as genetic tests for loci identified by GWAS come on-line.

9. Data flows

Interviewees were asked whether they had submitted data to a public repository. Nine had submitted to a database, although for some of these this was not routine. When asked what the barriers were to data submission the answer was always “not enough time”. One also added that “this was not their job”. Although, there was agreement that confidentiality was an issue, it was felt that if the information provided in online databases did not exceed that given in publications there should be no problem.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

10. Data access technologies

Interviewees were asked about the use of Ensembl software for accessing GEN2PHEN data. Most interviewees had some familiarity with Ensembl, even if they had not necessarily used the software. All interviewees who expressed an opinion felt that a graphical model of genes would be very useful. It was felt that the users should be able to click on icons showing, for example, SNPs and that the protein motifs (e.g., Pfam) should be shown. Both of these issues appear to be dealt with in current versions of Ensembl. However, there were concerns raised by clinical scientists that the new version of Ensembl although faster is less straightforward to use and that changes to the website are not adequately explained.

11. Summary of issues relating to GEN2PHEN

11.1. Obtaining users. When questioned about which LSDBs they were using most interviewees said InSiGHT or MUN, rather than the more recently developed LOVD or DMuDBs. This might suggest a lack of awareness of new software in the Lynch syndrome community. Alternatively there might be some resistance to trying new software.

11.2. Lack of data input. Although there are many variants listed in LSDBs, particularly in LOVD, data is added by the database administrator rather than the user. It may be that this method is not workable if we want to have more of the data that is being generated entered into repositories. In order to get users to submit their own data, submission methods must be streamlined. The INFOBIOMED system may provide examples.

11.3. Data quality. A potential problem with having data submitted by users is that other users must be able to trust the data. It may be that submission of data is limited to “trusted sources” with password protection.


11.4. Assessment of data quality. It may be (at least in the early stages of the project) that submitted data needs to be checked prior to being accepted for entry in the database. This has been the case for DMuDB and INFOBIOMED.

11.5. Increased supporting data. In particular there needs to be storage of data supporting pathogenicity (or non-pathogenicity). Lynch syndrome experts need to reach agreement as to what is required. As with 10.2, capturing this data will require us to provide easy to use software. In addition, much of this data is very disease specific. Therefore it should be possible to customise the software to meet the needs of other diseases.

11.6. Developing data standards. Data standards will be produced by InSiGHT and INFOBIOMED for the reporting of Lynch syndrome mutations. GEN2PHEN software should be developed in accordance with these standards. However, in addition there are standards which are being developed by genetic testing laboratories, which are likely to vary between countries. Since most of the data will be generated by these laboratories it is also important that the software reflects these standards. It likely standards will vary over time as new experimental methods and analysis software are developed. There needs to be consideration as to how future changes can be accommodated within the GEN2PHEN data models.

11.7. Data standards and software flexibility. Clearly the meta data requirements developed for Lynch syndrome will differ from those required for genetic diseases. Software needs to be flexible enough to cope with these differences.

11.8. Duplication of entries. If data submission is increased then the number of duplicate entries will increase. It is important to capture the number of times a variant has been identified.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

However, it becomes a problem to users if duplicate entries have different descriptions of pathogenicity.

11.9. Evaluation of penetrance. If there are multiple reports for variants it may be possible to gain some insight into the penetrance associated with individual variants.

11.10. Expert classification. The recent paper by Greenblat *et al.*, (2008) states that LSDBs should only report a conclusion related to pathogenicity if a consensus has been reached by an expert panel. Since this only occurs some time after data submission there is a need to store both the assessment of pathogenicity made by the submitter and the assessment made by the expert panel.

11.11. Multiple submitters for a record. The software designed by INFOBIOMED requires that different submitters contribute to different parts of each record. If GEN2PHEN adopts a similar approach, software may be needed to keep track of records.

11.12. Permissions for altering records. The storage of more clinical data may mean that this data changes over time. A system of permissions is required to allow users to edit data entries for which they are submitters.

11.13. Linking records of related individuals. The INFOBIOMED software will link records for related individuals. Access to the database will be limited to within Denmark. For GEN2PHEN, this may be more problematic as access will be at least Europe-wide. Potentially it might be easier to identify individuals because of the records they are linked to. Will linking data for related individuals present different legal requirements in different countries?


11.14 Linking variants for a patient. For each patient it is necessary to link all the genes that have been tested and all the variants that have been identified. It is important that the difference between “gene tested but no variants identified” and “not tested” can be distinguished. In addition, it is likely that new targets will be developed over time, for example from modifier studies.

11.15 Improved searching. At present LSDBs often only allow the user to scroll through a list of variants (e.g. all variants for exon 1 of MLH1). With increased storage of supporting information there may be benefits from increased search options. For example “return all variants associated with individuals whose onset of disease occurred before age 35”; or “return patient ids for individuals who were tested positive for IHC for MLH1 but who lacked MLH1 variants”.

11.16 Duplication of work. Many of the analyses performed by diagnostic labs involve using different types of software to achieve the same task such as analysis of splice site variations. It might be that this is the sort of task that could be streamlined. An alternative might be that the user inputs a variant description and these analyses are run automatically, possibly using Web Services. This might have several benefits:

- The user does not have to perform repetitive tasks, perhaps making mistakes.
- Analyses are run using predefined standards.
- It provides a “carrot” to get users to use our software.
- It might be part of the submission process, encouraging the user to submit data themselves.
- The periodic repeat of analysis of unclassified variants could be automated.

11.17 Duplication of software. The INFOBIOMED project shares some of the objectives of GEN2PHEN. Because this pilot project focussed on Lynch syndrome it is possible that other relevant software focussing on other genes is also available.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

11.18 Changes in sequencing methods. Given that new sequencing technology is being developed it is likely that there will be further changes in the methods used to identify sequence variants. It has been suggested that in the not too distant future whole genome sequencing will become routine (for example see Feero *et al.*, 2008). We need to ensure that the GEN2PHEN software is designed in a way that allows data from these experiments to be incorporated.

11.19 Promoting GEN2PHEN developments. The majority of interviewees were not aware of PAGE-OM, LRG or HGVbaseG2P. What are the best ways of promoting these developments?

11.20 Legal implications of storing variant data. None of the interviewees were certain what data could be legally stored in publicly accessible databases. Bearing in mind those legal requirements might vary between different European countries (and may be more stringent in the US); there needs to be more information on the legal implications of GEN2PHEN.

12. References

Agostini M, Tibiletti MG, Lucci-Cordisco E, Chiaravalli A, Morreau H, Furlan D, Boccutto L, Pucciarelli S, Capella C, Boiocchi M, Viel A. 2005. Two PMS2 mutations in a Turcot syndrome family with small bowel cancers. *Am. J. Gastroenterol.* **100**:1886-91.

Akrami SM, Dunlop MG, Farrington SM, Frayling IM, MacDonald F, Harvey JF, Armour JA. 2005. Screening for exonic copy number mutations at MSH2 and MLH1 by MAPH. *Fam. Cancer.* **4**:145-9.

Avdievich E, Reiss C, Scherer SJ, Zhang Y, Maier SM, Jin B, Hou H Jr, Rosenwald A, Riedmiller H, Kucherlapati R, Cohen PE, Edelmann W, Kneitz B. 2008. Distinct effects of the recurrent Mlh1G67R mutation on MMR functions, cancer, and meiosis. *Proc. Natl. Acad. Sci. U S A.* **105**: 4247-52.


Beaudet AL, Belmont JW. 2008. Array-based DNA diagnostics: let the revolution begin. *Annu. Rev. Med.* **59**:113-29.

Byfield SA and Syngal S. 2008. Clinical guidelines versus universal molecular testing: are we ready to choose an optimal strategy for lynch syndrome identification? *Am. J. Gastroenterol.* **103**:2837-40.

Chao EC, Velasquez JL, Witherspoon MS, Rozek LS, Peel D, Ng P, Gruber SB, Watson P, Rennert G, Anton-Culver H, Lynch H, Lipkin SM. 2008. Accurate classification of MLH1/MSH2 missense variants with multivariate analysis of protein polymorphisms-mismatch repair (MAPP-MMR). *Hum. Mutat.* **29**:852-60.

Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehvaslaiho H, Liang P, Marsh S, Nebert DW, Povey S, Rossetti S, Scriver CR, Summar M, Tolan DR, Verma IC, Vihinen M, den Dunnen JT. 2008. Recommendations for locus-specific databases and their curation. *Hum. Mutat.* **29**:2-5.

Dantzer J, Moad C, Heiland R, Mooney S. 2005. MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res.* **33**(Web Server issue):W311-4.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Easton DF, Eeles RA. 2008. Genome-wide association studies in cancer. *Hum. Mol. Genet.* **17**:R109-15.

Frank B, Burwinkel B, Bermejo JL, Försti A, Hemminki K, Houlston R, Mangold E, Rahner N, Friedl W, Friedrichs N, Buettner R, Engel C, Loeffler M, Holinski-Feder E, Morak M, Keller G, Schackert HK, Krüger S, Goecke T, Moeslein G, Kloor M, Gebert J, Kunstmann E, Schulmann K, Rüschoff J, Propping P; German HNPCC Consortium. 2008. Ten recently identified associations between nsSNPs and colorectal cancer could not be replicated in German families. *Cancer Lett.* **271**:153-7.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**:862–4.

Greenblatt MS, Brody LC, Foulkes WD, Genuardi M, Hofstra RM, Olivier M, Plon SE, Sijmons RH, Sinilnikova O, Spurdle AB; IARC Unclassified Genetic Variants Working Group. 2008. Locus-specific databases and recommendations to strengthen their contribution to the classification of variants in cancer susceptibility genes. *Hum. Mutat.* **29**:1273-81.

Harris M 2008. Why all young bowel cancer patients should be screened for Lynch syndrome. *ANZ J. Surg.* **78**: 531-2.


Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**: e1000167.

Korhonen MK, Vuorenmaa E, Nyström M. 2008. The first functional study of MLH3 mutations found in cancer patients. *Genes, Chromosomes and Cancer.* **47**:803-9.

Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S; Colorectal Cancer Association Study Consortium, Carvajal-Carmona L, Howarth K, Jaeger E, Spain SL, Walther A, Barclay E, Martin L, Gorman M, Domingo E, Teixeira AS; CoRGI Consortium, Kerr D, Cazier JB, Niittymäki I, Tuupanen S, Karhu A, Aaltonen LA, Tomlinson IP, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Cetnarskyj R, Porteous ME, Pharoah PD, Koessler T, Hampe J, Buch S, Schafmayer C, Tepel J, Schreiber S, Völzke H, Chang-Claude J, Hoffmeister M, Brenner H, Zanke BW, Montpetit A, Hudson TJ, Gallinger S; International Colorectal Cancer Genetic Association Consortium, Campbell H, Dunlop MG. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**: 1426-35.

Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**:78-85.

Liu HX, Zhou XL, Liu T, Werelius B, Lindmark G, Dahl N, Lindblom A. 2003. The role of hMLH3 in familial colorectal cancer. *Cancer Res.* **63**: 1894-9.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Loughrey MB, Waring PM, Tan A, Trivett M, Kovalenko S, Beshay V, Young MA, McArthur G, Boussioutas A, Dobrovic A. 2007. Incorporation of somatic BRAF mutation testing into an algorithm for the investigation of hereditary non-polyposis colorectal cancer. *Fam. Cancer.* **6**:301-10.

Niessen RC, Sijmons RH, Ou J, Olthof SG, Osinga J, Ligtenberg MJ, Hogervorst FB, Weiss MM, Tops CM, Hes FJ, de Bock GH, Buys CH, Kleibeuker JH, Hofstra RM. 2006. MUTYH and the mismatch repair system: partners in crime? *Hum. Genet.* **119**: 206-11.

Oda S, Maehara Y, Ikeda Y, Oki E, Egashira A, Okamura Y, Takahashi I, Kakeji Y, Sumiyoshi Y, Miyashita K, Yamada Y, Zhao Y, Hattori H, Taguchi K, Ikeuchi T, Tsuzuki T, Sekiguchi M, Karran P, Yoshida MA. 2005 Two modes of microsatellite instability in human cancer: differential connection of defective DNA mismatch repair to dinucleotide repeat instability. *Nucleic Acids Res.* **33**:1628-36.

Ollila S, Dermadi Bebek D, Greenblatt M, Nyström M. 2008. Uncertain pathogenicity of MSH2 variants N127S and G322D challenges their classification. *Int. J. Cancer.* **123**:720-4.

Park JG, Kim DW, Hong CW, Nam BH, Shin YK, Hong SH, Kim IJ, Lim SB, Aronson M, Bisgaard ML, Brown GJ, Burn J, Chow E, Conrad P, Douglas F, Dunlop M, Ford J, Greenblatt MS, Heikki J, Heinimann K, Lynch EL, Macrae F, McKinnon WC, Möeslein G, Rossi BM, Rozen P, Schofield L, Vaccaro C, Vasen H, Velthuisen M, Viel A, Wijnen J; International Society for Gastrointestinal Hereditary Tumours. 2006. Germ line mutations of mismatch repair genes in hereditary nonpolyposis colorectal cancer patients with small bowel cancer: International Society for Gastrointestinal Hereditary Tumours Collaborative Study. *Clin Cancer Res.* **12**: 3389-93.


Peltomäki P, Vasen HF. 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.

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV; IARC Unclassified Genetic Variants Working Group. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**:1282-91.

Reeves SG, Rich D, Meldrum CJ, Colyvas K, Kurzawski G, Suchy J, Lubinski J, Scott RJ. 2008. IGF1 is a modifier of disease risk in hereditary non-polyposis colorectal cancer. *Int J Cancer.* **123**:1339-43.

Scott RJ. 2008. Modifier genes and HNPCC: variable phenotypic expression in HNPCC and the search for modifier genes. *Eur. J. Hum. Genet.* **16**:531-2.

Steinke V, Rahner N, Morak M, Keller G, Schackert HK, Görgens H, Schmiegel W, Royer-Pokora B, Dietmaier W, Kloor M, Engel C, Propping P, Aretz S; German HNPCC Consortium.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WPI: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

2008. No association between MUTYH and MSH6 germline mutations in 64 HNPCC patients. *Eur. J. Hum. Genet.* **16**: 587-92.

Takahashi M, Shimodaira H, Andreutti-Zaugg, C Iggo R, Kolodner R.D and Ishioka C. 2007. Functional Analysis of Human MLH1 Variants Using Yeast and In vitro Mismatch Repair Assays. *Cancer Res.* **67**: 4595-4064.


Taketo MM. 2006. Mouse models of gastrointestinal tumors. *Cancer Sci.* **97**:355-61.

Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N, Semple C, Clark AJ, Reid FJ, Smith LA, Kavoussanakis K, Koessler T, Pharoah PD, Buch S, Schafmayer C, Tepel J, Schreiber S, Völzke H, Schmidt CO, Hampe J, Chang-Claude J, Hoffmeister M, Brenner H, Wilkening S, Canzian F, Capella G, Moreno V, Deary IJ, Starr JM, Tomlinson IP, Kemp Z, Howarth K, Carvajal-Carmona L, Webb E, Broderick P, Vijayakrishnan J, Houlston RS, Rennert G, Ballinger D, Rozek L, Gruber SB, Matsuda K, Kidokoro T, Nakamura Y, Zanke BW, Greenwood CM, Rangrej J, Kustra R, Montpetit A, Hudson TJ, Gallinger S, Campbell H, Dunlop MG. 2008. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.* **40**:631–637.

Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K, Jaeger E, Fielding S, Rowan A, Vijayakrishnan J, Domingo E, Chandler I, Kemp Z, Qureshi M, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Penegar S, Barclay E, Wood W, Martin L, Gorman M, Thomas H, Peto J, Bishop DT, Gray R, Maher ER, Lucassen A, Kerr D, Evans DG; CORGI Consortium, Schafmayer C, Buch S, Völzke H, Hampe J, Schreiber S, John U, Koessler T, Pharoah P, van Wezel T, Morreau H, Wijnen JT, Hopper JL, Southey MC, Giles GG, Severi G, Castellví-Bel S, Ruiz-Ponte C, Carracedo A, Castells A; EPICOLON Consortium, Försti A, Hemminki K, Vodicka P, Naccarati A, Lipton L, Ho JW, Cheng KK, Sham PC, Luk J, Agúndez JA, Ladero JM, de la Hoya M, Caldés T, Niittymäki I, Tuupanen S, Karhu A, Aaltonen L, Cazier JB, Campbell H, Dunlop MG, Houlston RS. 2008. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**:623–630.

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, Barclay E, Lubbe S, Martin L, Sellick G, Jaeger E, Hubner R, Wild R, Rowan A, Fielding S, Howarth K; CORGI Consortium, Silver A, Atkin W, Muir K, Logan R, Kerr D, Johnstone E, Sieber O, Gray R, Thomas H, Peto J, Cazier JB, Houlston R. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39**:984–988.

Wu Y, Berends MJW, Sijmons RH, Mensink RGJ, Verlind E, Kooi KA, van der Sluis T, Kempinga C, van der Zee AGJ, Hollema H, Buys CHCM, Kleibeuker JH, Hofstra RMW. 2001. A role for MLH3 in hereditary nonpolyposis colorectal cancer. *Nat. Genet.* **29**:137-138.

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Zanke B.W., Greenwood C.M., Rangrej J., Kustra R., Tenesa A., Farrington S.M., Prendergast J., Olschwang S., Chiang T., Crowdy E., et al. 2007. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**:989–994.

Zecevic M, Amos CI, Gu X, Campos IM, Jones JS, Lynch PM, Rodriguez-Bigas MA, Frazier ML. 2006. IGF1 gene polymorphism and risk for hereditary nonpolyposis colorectal cancer. *J. Natl. Cancer Inst.* **98**: 139-43.

13. Supplementary

The Amsterdam criteria for HNPCC were:

- Three or more cases of colorectal cancer in a minimum of two generations.
- One affected individual should be first-degree to the other cases of colorectal cancer.
- One case of colorectal cancer should be diagnosed under age 50.
- A diagnosis of Familial Adenomatous Polyposis (FAP) should be excluded.

The criteria have since been modified to:

- Two cases of colorectal cancer where families are small (one age under 55).
- Two cases of colorectal cancer and one case of endometrial cancer, or other early onset cancer.

There is also a second set of guidelines, the **Bethesda criteria**. The revised version (2002) of these criteria are:

1. The patient is younger than age 50.
2. The patient has multiple HNPCC-associated tumours in the colon or in other areas known to be caused by the same mutations, either at the same time or occurring over a period of time.
3. A patient younger than age 60 has colorectal cancer that has microscopic characteristics that are often indicative of MSI.
4. A patient has one or more first-degree relatives who had an HNPCC-related tumour at age 50 or younger.
5. A patient has two or more first- or second-degree relatives who had HNPCC-related tumours at any age.




 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Table 1. Details of telephone interviewees.

Name	Organisation	Role
Alessandra Viel	National Cancer Institute, Aviano (PN) Italy	Molecular Biologist
David Gokhale	Molecular Genetics Laboratory, Liverpool Women's NHS Foundation Trust	Deputy of Lab and Registered Clinical Scientist.
Diana Cairns	Molecular Genetics Laboratory, Liverpool Women's NHS Foundation Trust	Clinical Scientist
Gabriela Moslein	St. Josefs-Hospital, Bochum-Linden	Colorectal Surgeon and Chairperson of InSiGHT
Georgina Hall	St Mary's Hospital, Manchester, UK	Genetic Counsellors
Ian Frayling	All Wales Medical Genetics Service, University Hospital of Wales, Cardiff, Wales	Laboratory Director and Consultant Geneticist
Inge Bernstein	Hvidove Hospital, Denmark	Colorectal Surgeon. Also leads the Danish HNPCC register
Jason Kennedy	St Mary's Hospital, Manchester, UK	Clinical Scientist
Jenny Myring	All Wales Medical Genetics Service, University Hospital of Wales, Cardiff, Wales	Clinical Scientist
John Burn	Institute of Human Genetics, Newcastle	Head of the Institute of Human Genetics in Newcastle
Julian Barwell	University of Leicester	Interest in familial cancers. Senior lecturer in Cancer Genetics.
Justo Lorenzo Berjemo	DKFZ (German Cancer Research Centre)	Statistician. Is using the Swedish Family Cancer Database to do analysis of 11 million individuals.
Katharina Wimmer	Department of Medical Genetics, Medical University of Vienna	Associate Professor, responsible for genetic diagnosis as part of service offered to hospitals and private doctors
Lene Juel Rasmussen	Roskilde University	Professor, Life and Sciences department. Interest in Genomic instability and DNA damage
Liliana Varesco	S.S. Centro Tumori Ereditari, Istituto Nazionale per la Ricerca sul Cancro, Genova, Italy	Responsible for unit offering genetic testing for colorectal and breast cancer and genetic counselling.
Laura Belvederesi	Università Politecnica delle Marche, Ancona Italy	Biologist

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Name	Organisation	Role
Marta Pereira	Regional Genetics Service, London Kennedy-Galton Centre, Harrow, Middlesex	Molecular Geneticist
Mary Porteous	SE Scotland Genetic Service	Consultant clinical geneticist. Also in charge of molecular lab and involved in research including research in bowel cancer.
Nancy Uhrhammer	Centre Jean Perrin, Clermont-Ferrand , France	Clinical Scientist
Robert Hofstra	Department of Genetics, University Medical Center, Groningen, Netherlands	Head of R&D section in department. Has recently published a paper on interpretation of missense variants.
Rodney Scott	School of Biomedical Sciences Faculty of Health University of Newcastle, New South Wales, Australia	Head of the Discipline of Medical Genetics
Rolf Sijmons	Department of Genetics University Medical Center Groningen, Netherlands.	Clinical geneticist, also involved in the development of the MMR Gene Unclassified Variants Database
Sheila Goff	St George's Hospital, London	Lead Genetic Counsellor has an interest in cancer genetics, mostly breast cancer but also bowel cancer.
Sheila Palmer-Smith	All Wales Medical Genetics Service, University Hospital of Wales, Cardiff, Wales	Clinical Scientist
Shirley Hodgson	St George's Hospital, London	Clinician
Stewart Payne	Regional Genetics Service, London Kennedy-Galton Centre, Harrow, Middlesex	Director of laboratory
Susan Farrington	Colon Cancer Genetics Group, MRC - Human Genetics Unit, Edinburgh	Senior Scientist
Tara Clancy	St Mary's Hospital, Manchester	Consultant genetic counsellor and lecturer
Victoria Murday	Royal Hospital for Sick Children, Glasgow, UK	Consultant clinical geneticist
Yves-Jean Bignon	Centre Jean Perrin, Clermont- Ferrand , France	Scientific Director

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

Example of interview questions. The following is the list of questions used for interviews. Since most interviews were conducted either by telephone or face-to-face, the list was not rigidly adhered to and we would divert from the questions to cover topics the interviewee was most interested in. In other cases the might declare at the start that they had not used any LSDBs, in which case there was little point in trying to discuss specifics of LSDB software and we focussed on their requirement for dealing with Lynch syndrome patients. However, where possible, the eight sections were all covered in interviews.

Section 1: User details

Name

Organisation

Which of the following describe your role (select more than one if necessary)?


- a) Medical geneticists
- b) Genetic Counsellors
- c) Molecular biologists
- d) Diagnostic DNA labs
- e) Database curators
- f) Database developer
- g) Other (please specify)

Which of the following best describes how you use the existing databases?

- a) Research
- b) Interpretation of variants
- c) Database development
- d) Sequence alignments
- e) Data deposition

Section 2: Choice of Locus Specific Database

Which of the following locus specific databases (LSDBs) have you used in the analysis of HNPCC genes?

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

If not, is there other data not currently included in the database that would help? If so, please specify.

Section 4: Data Quality

Do you consider the LSDB to be up to date?

Yes No Don't know

Do you believe that the data in the LSDB is accurate?

Yes No Don't know

Does the LSDB naming of sequence variation conform to the HGVS conventions?

Yes No Don't know

Do you feel that data is consistent between databases?

Yes No Don't know

Section 5: Improvements

Is there data, not currently stored in the database, which would make it more useful? If so, please specify.

Does the LSDB feature a graphical model of genes showing the position of known mutations?

Yes No


If not, do you believe that such a model would be useful?

Yes No Don't know

Section 6: Entering data into LSDBs/ Data flows

Have you entered the results of a screening into a public repository?

Yes No

 HEALTH-200754	D1.2 – Initial Report from the Project Assessment Pilot		
	WP1: Scientific Coordination		Security: PU
	Author(s): M. Cornell, A. Devereau		Version: v1.3 – Final

3/ Data on result

4/ Data on interpretation

Section 8: Genome-wide databases

Have you used HGVbaseG2P (it contains G2P information for groups (not individuals) i.e. gene association data?

Do you have any requirements for G2P databases?