1. Final publishable summary report

1.1 Executive summary

GEN2PHEN was launched in early 2008, at a time when few genotype-phenotype (G2P) databases existed, characterised by immense structural diversity, little semantic or syntactic consistency, no inter-resource federation, and (therefore) no ability to search across the whole domain for research or healthcare purposes. This pitiful situation was further compounded by a rapidly increasing rate of G2P data production, both in single gene/disease contexts (rare genetic disease) and from whole genome analysis of complex disease. And during the project's 5.5 year lifetime the G2P data generation rate increased even further, due to technical advances in next generation sequencing (NGS) and whole genome genotyping.

GEN2PHEN aimed to address all the above challenges and deficiencies in a comprehensive manner, to enable G2P data to be maximally exploited. The project was highly successful in this ambitious mission, in that it developed a broad range of operational and technical solutions, and used these (alone and with many external partners) to create and populate a large and high quality internet-based holistic 'Knowledge-Environment'. The architecture of this system is partially centralised and partially federated, supports direct and automated data submission, enables powerful comprehensive searching, and hence brings substantial new utility to many research and healthcare user groups.

The project's success was a consequence of the strategic approach taken. This involved initially and repeatedly analysing and aligning with the field's needs (by extensive community engagement), working with many other groups in a 'middle out' approach towards standards and policy development (ethical principles, data models, exchange formats, ontologies, mutation nomenclature, reference sequences, entity ID systems), and creating modular and interoperable database and data management components, tools and services. These were used to create and/or enhance, and extensively populate, several thousand new gene/disease specific databases (e.g., LOVD, DMuDB, UMD) and comprehensive GWAS and genome annotation databases (e.g., GWAS Central, Ensembl), plus several commercial systems and tools. These resources were brought together via web-services and new data aggregation and discovery approaches, to provide local and holistic search and presentation solutions suitable for research and healthcare users. As an integral part of all this, issues such as training, support services, platform effectiveness, and system sustainability, were professionally dealt with, led by synergistic scientific coordination and project management activities. The totality of these advances, plus extensive G2P community resources, were assembled and provided or linked to via the main project 'Knowledge Center' (available at www.gen2phen.org).

GEN2PHEN thus demonstrated, both in principle and in practice, how a consolidated, community level initiative can achieve dramatic progress in organising and enhancing a new data domain. Of course, this work can never be deemed to be 'complete', as the scale and type of data being produced will always continually change and the potential users and uses of this information will similarly

evolve. This is abundantly true for the G2P domain, and therefore GEN2PHEN made a special effort in its latter months and years to map out emerging challenges and opportunities.

The over-riding conclusions is that there is a need for better ways to effectively bridge the IT gap between research and healthcare – to unify the diverse categories of data produced in these two realms, and to exploit this information to distil out new knowledge of direct clinical utility. Progress will require a new focus on 'Knowledge Engineering' (informaticians that work across research and healthcare), evidence based analytics to infer mutation pathogenicity, and instantiating a layer of federated 'integration' databases to engage domain experts and protect subject privacy. Furthermore, such advances themselves imply the need for global unique ID systems (resolving patients, scholars, data entities, biobanks, databases), and a radical broadening of data sharing notions to include knowledge sharing and data discovery strategies as priorities. Rare disease then stands out as an ideal sub-domain where all of this can be rapidly developed and applied, and where the existing political desire for progress should logically translate into suitably scaled, community based, trans-national initiatives.

In summary, GEN2PHEN has been highly effective in helping to covert the G2P data field from a nascent topic to a data-rich, multi-purpose online ecosystem. The project also provides a valuable model and roadmap for taking the important next steps towards truly personalised genomic medicine - starting with rare disease and expanding outwards to more complex disorders and more diverse sorts of information, to ultimately engage fully with the emerging 'Big Data' era that is rapidly approaching.

1.2 Description of project context and objectives

The GEN2PHEN project aims to unify human and model organism genetic variation databases towards increasingly holistic views into Genotype-To-Phenotype (G2P) data, and to link this system into other biomedical knowledge sources via genome browser functionality. Target user communities include researchers, healthcare professionals, policymakers, and general citizens. The project is tackled by a program of 10 workpackages (WPs) and extensive collaboration with others, addressing community networking and consensus building, standards development, database development (modular, centralised, and federated), data integration efforts, and the provision of search capabilities that are comprehensive across the domain. All this work is coupled to a range of ethics, training, and dissemination actions. Specific objectives by WP are as follows:

WP1: SCIENTIFIC COORDINATION - is concerned with providing top-level oversight and scientific coordination to make GEN2PHEN a success, operating closely with WP10. Activity #1 concerns **'Project Coordination**', to monitor and optimize the project and organize a Steering Committee, a Scientific Advisory Board, and other ad-hoc Boards/Committees. Activity #2 concerns **'Project Quality and Assessment**', not least by running sequential system-wide utility assessment 'Pilots'. Activity #3 concerns **'Ethical Issues**', by establishing suitable oversight, analysis and guidance procedures to 'hard code' good ethical principles into the IT systems that GEN2PHEN creates. Two **new goals** emerged from the project's mid-term review. The first is a theoretical exploration of how to best connect research data directly into healthcare ('**I4Health**' - Integration and Interpretation of Information for Individualised Healthcare). The second concerns working with the Open Researcher Contributor ID (**ORCID**) initiative to develop the use of online digital identities for researchers.

WP2: DOMAIN ANALYSIS AND COMMUNITY RELATIONS – seeks to analyze the existing G2P database field, and its findings will guide the strategic evolution of the project. Completion in a timely manner is therefore vital, and so WP2 tasks are all due for completion in the first few years.

WP2 is also tasked with helping to develop good community relations. Activity #1 concerns **'Community Consultations**', for which a detailed action plan will be devised for community consultation, targeting multiple G2P field stakeholders, to create a complete systems requirements document. Activity #2 concerns **'Technical Domain Analysis**', via which we will assess and document the technical state-of-the-art for each sub-type of G2P database (LSDBs, Diagnostics DBs, and Genomics DBs). Data models and data exchange formats will also be considered.

WP3: STANDARD DATA MODELS AND TERMINOLOGIES - tackles the lack of interoperability between public databases, spanning general to specialized resources and central to federated approaches. The approach involves defining partner use cases and then starting standard development work, typically in collaboration with many other expert stakeholders. Activity #1 concerns **'Core Data Model Development**', which via use case evaluation will establish a priority list for syntax, semantic and technical standards creation. A 'Core Data Model' will be devised, followed by 'Specific Data Models', including data exchange formats. Activity #2 concerns **'Advanced Data Modeling Issues**', which will build on Activity #1 to develop models for new concepts and challenges. Activity #3 concerns **'Other Standards Development**', wherein we anticipate the focus will shift from syntax models and technical aspects more towards structured nomenclature for the G2P field, not least by working on ontology terms, mappings, and tools for building and using ontologies. Additionally, one predefined standards gap is the lack of a completely stable reference gene structure for genes/regions of interest, and so this will be tackled in Activity #3 early on the project and in close coordination with all relevant stakeholders globally.

WP4: GENETICS G2P DATABASES - will develop generic database solutions for the management of G2P information relating to particular genes/regions/diseases (Locus-Specific DataBases, LSDBs), and promote the creation and deployment of such databases in research and diagnostic settings. This will entail creating tools for local data management, curation, submission, and collection. The resulting interoperable databases will allow a federated network to be created, to underpin holistic data unification. Activity #1 '**LSDB-In-A-Box Solutions**', will devise modular components for use in building and running single gene level G2P genetics databases, based on a core data model generated by WP3. This will facilitate data input/output/exchange, search functionality, and APIs for grid/network integration. Activity #2 '**LSDB Creation**', will involve local deployment of the technologies developed in Activity #1, including foundational databases for all genes involved in Mendelian disorders, and populating these (with WP7). Additionally, a database archiving service will be provided to back-up existing LSDB databases upon request. Activity #3 '**Solutions for Diagnostic Labs**', targets the capture of G2P data generated by diagnostic laboratories. Components to be developed include data submission and curation solutions, designed in compliance with ethical guidelines from WP1. Commercial and open source software will be investigated. Activity #4 '**Ontologies in Genetics Databases**', will be conducted during the latter years of the project and guided by WP2 and WP3, to increase the consistency of description of G2P related features, improve annotations, and facilitate advanced searches. Activity #5 '**Testing and Validation**' imposes a quality assurance policy for every piece of software developed in WP4, based upon guidelines from WP1. Activity #6 '**Deployment Partnerships**' entails enlisting curators to 'adopt' the foundational LSDBs designed and built in activity #2. These will be domain experts, from diverse sources.

WP5: GENOMICS G2P DATABASES - will develop generic G2P database solutions relating to any gene/region or the whole genome, and implement specific examples of these. This will entail creating tools for local data management, curation, submission/collection, and online interrogation. Enabling this to be replicated elsewhere will encourage a federated network to emerge. Activity #1 concerns developing '**Genomics Database Solutions**' for summary level G2P genomics databases, building upon pre-existing Partner platforms HGVbaseG2P (GWAS Central) and IGVdb. These will be significantly improved with respect to data import, validation, text and graphical outputs, search

options, and APIs for grid/network integration. Activity #2 is about local '**Creation of Central Genomics Database(s)**', using technologies from Activity #1. These databases will be increasingly populated with datasets, in partnership with WP7. Activity #3 concerns '**Tools For Data Submission**', targeting the collection of small to medium sized datasets provided by the broad community. Submission procedures and curation pipelines will leverage standards from WP3, and ethical guidelines from WP1. Activity #4 concerns '**Tools For Data Harvesting**', which will progress by adapting many items from Activity #3 to handle larger datasets. Activity #5 involves building '**Tools for Local Data Management and Output**', to help others transfer their G2P data into the database world. Solutions will be open source and commercial. Activity #6 will tackle '**Ontologies in Genomics Databases**' during the latter years of the project, guided by WP2 and WP3. It will increase the consistency of description of G2P related features, improve annotations, and facilitate advanced searches. Activity #7 on '**Testing and Validation**' will impose a quality assurance policy for all the software developed in WP5, guided by WP1 policies. Activity #8 on '**Deployment Partnerships**', will progress from a 'central database' approach towards the emphasis of 'database federation', by deploying WP5 solutions to help others set up genomics G2P databases. One **new objective** recommended by the mid-term review seeks to devise a standalone tool ('**Omics Connect'**) to enabled researchers to visually explore and mine local and remote multi-omics datasets in a convenient and integrative manner.

WP6: INTEGRATION AND DATA ACCESS TECHNOLOGIES - seeks to integrate database content generated and collated by the other WPs, to benefit many classes of user. Activity #1 deals with larger scale '**Cross-Domain Integration**', centered principally upon the Ensembl and Uniprot/SwissprotKB platforms. Once integrated into central facilities, information can be easily connected to other constellations of biomedical data. Activity #2 provides the supporting '**Data Processing**' work, employing XML and other technologies to produce pipelines that automatically parse G2P datasets from the community into Ensembl and Uniprot/SwissprotKB, including validation and curation efforts. One aspect of this involves routinely computing the likely functional consequences of imported variants. Activity #3 exploits '**GRID Technologies**', as the main means for data integration in the foreseeable future. Therefore, we will investigate GRID technologies including both established approaches (e.g., BioMart) and more exploratory alternatives. Activity #4 develops '**Query and Display Options for Research Biologists**' to allow personalization of search and display functions for G2P data made available via Ensembl, and the inclusion/integration of local datasets. In parallel, Activity #5 develops '**Search & Display Options for Other Users**' to provide search routes into G2P data that are of particular utility for alternative stakeholders, such as clinically oriented groups. Their needs will grow with time, as diagnostics and personalized medicine become increasingly commonplace. Activity #6 concerns '**Data Presentation Technologies**', which primarily relates to the DiseaseCard and WAVe software, substantially exploiting webservices emerging from GEN2PHEN. Activity #7 on '**Testing and Validation**' imposes a quality assurance policy for every piece of software developed in WP6, based upon guidelines from WP1.

WP7: DATA FLOWS - acts to co-ordinate, facilitate and undertake the collection and population of G2P data into GEN2PHEN databases, with increasing intensity as the project proceeds. Activity #1 concerns delivering a '**Populated LSDB Federation**', which means federating larger existing LSDBs by providing data exchange formats, archiving data, and facilitating multi-database searches. Foundational LSDBs from WP4 will be populated with public domain polymorphism data. Activity #2 concerns the '**Population of Genomics Databases**', from WP5, using G2P data submitted by researchers, or harvested from other public resources. Additionally, 'dbSNP-lite' will be created to automatically extract and incorporate core elements from each new dbSNP build, including details of record changes. Activity #3 will establish '**Population-Specific Datasets**' by a focused effort targeting local and remote data resources, to identify, ascertain, structurally harmonise, and integrate variant data that includes population or ethnic specific frequency information. Activity #4 concerns

'**Quality Control for Datasets**', by developing standardized guidance for data submitters and for receiving databases. One **new objective** arising from the mid-term review concerned moving beyond 'data sharing' as a strategy (where access control and privacy can be challenging), to explore the novel concept of '**Open Data Discovery**' (wherein the existence rather than the substance of the data are made available). This should provide a rapid and viable way to catalog and expose all G2P data, enabling direct Knowledge Sharing and suitable data access to sapidly follow.

WP8: GEN2PHEN KNOWLEDGE CENTER – will undertake the internet-based exchange of GEN2PHEN information to and from the community, via a dedicated web portal, or 'Knowledge Center' (KC). It will encompass G2P data (via hyper-links and via direct access) and knowledge about GEN2PHEN and G2P databasing in general. Information exchange will be bidirectional – facilitating a discourse with the community. A related activity is the provision of training information, to support GEN2PHEN Partners and the general community. Activity #1 takes care of the **'Knowledge Center Construction and Use**', to; i) promote and explain GEN2PHEN, ii) distribute systems and results created by the project, iii) list upcoming G2P meetings, iv) give access to the G2P data query systems we develop, v) host internet chats and scientific debate, vi) enable users to update database entries, and vii) connect database searches with discussion fora. Activity #2, the **'GEN2PHEN Diary**', will routinely gather and post G2P field announcements on the KC, including a calendar of meetings, a news section, and a diary documenting GEN2PHEN progress. Activity #3 concerns the provision of **'Training Activities**', by leveraging the KC to disseminate and announce these activities. Internal training will encompass preparing electronic media, literature recommendations, guidance texts, workshops, and conference calls. External training will entail documentation about GEN2PHEN tools and systems, FAQ pages, and user support services.

WP9: DISSEMINATION, USE AND FUTURE SUSTAINABILITY - is responsible for disseminating information about GEN2PHEN to relevant stakeholders. It also optimizes the use made of the project's results both during and beyond the project. Activity #1 produces the **'Communication Plan and Tools**', to effectively publicise the GEN2PHEN project and its results. This will involve work on webpages, newsletters, brochures, handouts, etc. Activity #2 concerns **'Dissemination Activities**', spanning all activities that use the items produced by activity #1. Activity #3 addresses **'Incentive & Reward Issues**', by establishing a dialogue with relevant stakeholders. Areas of study include citation of databases and biobanks used in biomedical research, possibly leading to some kind of a Bio-Resource Research Impact Factor (BRIF). We will also consider the release of diagnostic lab data, and the underlying ethical, financial, incentivisation, and practical issues. Activity #4 concerns **'Long-Term Sustainability and Business Models**'. The aim here is to explore business models and other funding paradigms for making G2P databases and GEN2PHEN activities durable in the long-term and financially self-sustaining.

WP10: PROJECT MANAGEMENT - is devoted to project management, working closely with WP1. Activities will ensure that the project is appropriately managed in all relevant aspects, and that work is implemented according to the plan so that results are delivered on time, with high quality, and within budget. Activity #1 is about '**Day-to-day Management'**, covering work plan control, liaison with stakeholders, support to decision-making and conflict resolution, facilitating communication among partners, organization of meetings, risk management and coordination of quality control. Activity #2 concerns '**Reporting and Administration'**, which particularly relates to periodic reporting and financial management, and supporting Partners in these areas. Activity #3 on **'Contract and Legal Management'** encompasses maintenance of the Grant and Consortium Agreements, dealing with partnership evolution and providing support to global knowledge and IPR management.

1.3 Description of the main S&T results/foregrounds

The over-arching goal of the GEN2PHEN project was to improve the Genotype-To-Phenotype (G2P) data domain, specifically by developing and deploying technical and ELSI solutions, and a federated set of local and central databases and tools, to underpin a more unified and useful arrangement of G2P related information and knowledge. The project targeted human and model organism genetic variation databases towards increasingly holistic views into G2P data, linked into other biomedical knowledge sources via genome browser functionality, to serve both research and healthcare use cases. Stakeholders considered by this effort include researchers, healthcare professionals, policymakers, and general citizens.
The project goals were tackled by a program of work that spanned 66 months, and included community networking and consensus building, standards development, database development work (modular, centralised, and federated), data integration efforts, and the provision of search capabilities that are comprehensive across the domain. Traditional and advanced/emerging technologies were used and evaluated by the project, with all activities tightly coupled to a range of related ethics, training, and dissemination actions. Many external collaborators were worked with during the project, which greatly benefited our work in terms of progress achieved and its relevance to other developments in the field.
GEN2PHEN activities were organised into ten Work Packages (WPs), as described below.

WP1: SCIENTIFIC COORDINATION

Work Package 1 (WP1) was responsible for all aspects of higher-level coordination of the science in GEN2PHEN. Key tasks included forming and operating the project's committees and Boards, enabling and promoting dynamic intra-project communications, overseeing progress on all activities and Deliverables, controlling quality within the project, and ensuring all things were done with full regard to ethical considerations.
Intense work was started in WP1 right from the first year of the project (the first Reporting Period), to ensure GEN2PHEN quickly gained good momentum, and to foster a highly collaborative working atmosphere. Hence, within 12 months, involving all the project Partners, we:

- Published a review of the G2P databasing field, enunciating the GEN2HEN plan
- Formed all required Committees and Boards
- Began producing monthly 'Progress Reports' (a practice which continued throughout the project), based on regular updates submitted from all Partners describing all significant areas of progress. These were disseminated to the Consortium and beyond (e.g., compiled into 4-monthly summaries for the EC Officer) to help keep everyone informed, to guide strategy, to help with report and Deliverable production, and to identify potential synergies as early as possible.
- Compiled and issued formal guidance on QC/QA for software development
- Completed a comprehensive 'Project Assessment Pilot', which was then repeated three times during the life of GEN2PHEN, to evaluate our progress comprehensively against the changing needs of the field, to identify key challenges and emerging opportunities, and thereby to continually refine and optimize the work being conducted.

The good early momentum was maintained in the second year (the second Reporting Period), enabling many areas of technical development to become highly productive and effectively targeted for maximal impact and collaborative relevance to external developments in the field. This was facilitated, not least by engaging strategic contacts and partnerships, such as with the P3G biobanking project, with the UK Select Committee report on Genomic Medicine, with Cameron Neylon regarding linked-data and social networking issues, and with John Barber regarding the ECARUCA database. With practical work well under way, we were able to start to increase the

intensity of our work on ethics in terms of exploring Consortium views and awareness of issues relevant to the systems they were developing, and devising concrete operational policies for the project.

During the third Reporting Period (months 25-42 inclusive) WP1 led the consortium in undertaking our mid-term review. This review was extremely successful, and based on the excellent progress apparent across all the WPs we decided to stretch the project goals in some key emerging areas, such as starting to work with the Open Researcher and Contributor Identifier (ORCID) project. We also strengthened our clinical utility program, where the challenges were turning out to be greater than anticipated, by initiating extensive community discussion (international workshops, surveys, publications) on the question of how to bridge the IT gap between research and healthcare, and by beginning to develop a completely new databasing platform for pathogenicity inference of DNA mutations ('PathoKB').

Better interoperability and consolidation between knowledge management systems in research and healthcare is essential if personalised medicine is to become a reality. The work we did in this area after the mid-term review was fundamental and of high impact. We worked out, and catalysed a broad recognition of the fact, that there needs to be a major push (in terms of funding, training, innovation, commercialisation) to develop a new discipline that emphasizes 'Knowledge Engineering' as its core paradigm for Integration and Interpretation of Information for Individualised Healthcare ('I4Health'). These essential concepts then became part of a successful bid for Pilot Project funding towards an ICT FET Flagship called 'Information Technology, Future of Medicine ('IT-FoM'). And as our ideas firmed up still further, and were shared and published during the final Reporting Period, we explored their particular design and implementation in close partnership with the global 'IRDiRC' initiative. We thereby defined a Rare Disease ecoSYsteM for PATHogenicity Inference (RD-SymPathI workshop) which provides a clear roadmap for much of the global work now being undertaken in RD informatics. That valuable initiative continues beyond GEN2PHEN, by efforts to publish the concept in a high impact journal, by many GEN2PHEN partners having full or associate Partnership roles in the recently funded RD-CONNECT EU project, and by adoption of the ideas within various IRDiRC WorkGroups.

Regarding the transversal ethics dimension, this was addressed throughout the project in line with the state of the art, and a specific discussion was organised on various aspects at each annual GAM so that awareness was enhanced regularly. The first period concentrated on the internal survey and educational needs in ethics and law. The second period concentrated on context and other initiatives that could enlighten the establishment of the project ethics policy that was set up after the second period, with special contributions to the LSDB international ethics guidelines that were published in Human Mutation. The third period concentrated on patients' rights in the context of the openness of the project to managing and using clinical data as well as research data. Finally the fourth period concentrated on the follow up of the Data Protection Directive revision and the Regulation proposal and its comments and proposed amendments that were relevant for GEN2PHEN related issues. GAM9 in Toulouse was transversally organised around ethical aspects of the project, to critically review the ethics policy implementation in the project. This enabled us to successfully prepare an ethics policy as a sustainable instrument for future of G2P activities.

Most critically, the success of WP1 in leading and forming a durable and productive community of teams is reflected by the Consortium agreeing to continue working together on the goals of GEN2PHEN and their evolution, as part of a coalition called the 'GEN2PHEN-Alliance'.

Scheduled Deliverables from WP1 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D1.1	Specification of Procedures for Quality Testing of Software	M9	PU
D1.2	Initial Report from Project Assessment Pilot	M12	PU
D1.3	Report on General Ethical Issues in G2P Database Work	M12	PU

D1.4	Report on External ELSI Developments	M24	PU
D1.5	Intermediate Report from Project Assessment Pilot	M30	PU
D1.6	Report on Specific Issues Related to Patient Rights	M48	PU
D1.7	Update on data protection European Legal Framework	M60	PU
D1.8	Summary Report on GEN2PHEN Ethics	M66	PU
D1.9	Final Report from Project Assessment Pilot	M66	PU

Scientific papers that were primarily the result of work done in WP1 are as follows:

- M.Cases, L.I.Furlong, J.Albanell, R.B.Altman, R.Bellazzi, S.Boyer, A.Brand, A.J.Brookes, S.Brunak, T.W.Clark, J.Gea, P.Ghazal, N.Graf, R.Guigó, T.E.Klein, N.López-Bigas, V.Maojo, B.Mons, M.Musen, J.L.Oliveira, A.Rowe, P.Ruch, A.Shabo, E.H.Shortliffe, A.Valencia, J.VanDerLei, M.A.Mayer and F.Sanz. Improving Data And Knowledge Management To Better Integrate Healthcare And Research. Journal of Internal Medicine 2013; 274:321–28.
- T.Beck, S.Gollapudi, S.Brunak, N.Graf, H.U.Lemke, D.Dash, I.Buchan, C.Díaz, F.Sanz and AJ.Brookes. Knowledge Engineering for Health: A new discipline required to bridge the 'ICT Gap' between research and healthcare. Hum Mutat 2012; 33(5):797-802.
- Thorisson GA, Muilu J, Brookes AJ. Genotype-Phenotype Databases: Challenges and Solutions For The Post-Genomic Era. Nat Rev Genet 2009;10(1):9-18.
- Povey S, Al Aqeei Ai, Cambon-Thomsen A, Dalgleish R, Den Dunne JT, Firth HV, Greenblatt M, Barash CI, Parker M, Patrinos GP, Savige J, Sobrido MJ, Winship I, Cotton RGH. Practical guidelines addressing ethical issues pertaining to the curation of human locus specific variation databases (LSDBs) Hum Mutat 2010 Nov;31(11):1179-84.

WP2: DOMAIN ANALYSIS AND COMMUNITY RELATIONS

Work Package 2 (WP2) is focused upon i) ascertaining general field needs through community consultations, and ii) determining specific technology requirements from exhaustive domain analysis. This work must be undertaken early on in the GEN2PHEN project, as it instructs activities in many other Work Packages. In fact, this WP was scheduled for completion within 18 months, and it was completed successfully on time, after which only low-level additional work was undertaken for tasks that are likely to significantly benefit GEN2PHEN.

During the first year of the project, WP2 teams co-organised an international conference with the Human Variome Project (HVP) to look at the status and needs of many aspects of the G2P field, and also held a joint workshop with WP3 teams to review contemporary G2P database technologies, emphasizing data models and related issues. The initial WP2 plan only called for us to hold one such meeting, but we instead ran these two events, and reported them both in deliverable D2.1: "Workshop to Review the G2P Database Field and Current Data Models".

Subsequently, we extended our community consultation efforts by interacting with various experts and G2P field stakeholders, including G2P data creators, database technologists, biobank teams, G2P data end-users, LSDB curators, and members of several initiatives, e.g. HVP, GAIN, WTCCC, P3G consortium, human genetics societies and genetic journals, etc. As part of this we undertook a more formalized and extensive comparison of existing LSDBs, to understand the different data models and standards upon which these databases were developed. This enabled us to propose minimum content requirements and data model features for an optimum LSDB, and pass this guidance on to WP4 (Genetics G2P Databases) and to the general community. This provided the basis for improved uniformity of new LSDBs, and helped bring about their ever-deeper integration.

Progress was reported in Deliverable D2.2 ("General G2P Field System Requirements Report"), which represents a meaningful systems requirements document that compares and contrasts GEN2PHEN plans with what others judge to be the main needs and trends in the field.
We also produced D2.3 ("Technical State-Of-The-Art Document for G2P Databases"), the scope of which spans each sub-type of G2P database of interest to GEN2PHEN (LSDBs, Diagnostics DBs, and Genomics DBs holding individual and summary level datasets), plus current integration systems. A particular emphasis herein was placed upon technical aspects of current LSDBs, since these very important components of the G2P domain are plagued by great heterogeneity of content and structure.
Scheduled Deliverables from WP2 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D2.1	Workshop to Review the G2P Database Field and Current Data Models	M12	PU
D2.2	General G2P Field System Requirements Report	M18	PU
D2.3	Technical State-Of-The-Art Document for G2P Databases	M18	PU

WP3: STANDARD DATA MODELS AND TERMINOLOGIES

Work Package 3 (WP3) was tasked with developing standards to underlie the G2P database development and data exchange within GEN2PHEN. These standards include data models, data formats and terminologies by which the data are represented in repositories, formatted for exchange and annotated for analysis. WP3 covers a wide range of databases within GEN2PHEN and in the wider community, including the LSDBs, high throughput data generators, and summary level databases. Therefore, our objectives relate to use case development, iterative data modeling and format development to support WP4, WP5 and WP6.
Activities undertaken for WP3 from the outset concentrated on data structure/syntax issues. Specifically, we achieved the following:

- Comprehensive use case development for the G2P databasing field, aligned with the needs and ambitions of the Consortium
- Relationship development with related projects ENGAGE, EGA, BBMRI, CASIMIR to engage with the community in development of standards
- Co-development (with many international groups external to GEN2PHEN) and publication of the generic 'PaGE-OM' reference model for G2P data. To validate this model we successfully loaded diverse test data sets into a Reference Implementation (a database) of the model. This was subsequently accepted as an official OMG standard, providing a robust generic reference model for G2P data.
- Devised and validated (by extensive use with complex datasets) basic and advanced versions of the Observ-OM data model, designed to bring immense flexibility of scope. This represents a completely new approach to data modeling for the field, where all entities are abstracted into just 4 concepts (Observable Features, Observable Entities, Protocols, and Observations (Values)). It provides a powerful basis for databasing and integrating all the diverse information relevant to the G2P field, including any phenotype and any evidence underlying pathogenicity inference.
- Specification of a G2P data exchange formats (MAGETAB & Vario-ML) derived from high-level object modeling activities for each of the sub-domains: LSDBs, high-throughput data and phenotypic descriptions, with metadata and supporting evidence. The formats were tested, implemented and positively adopted by several others, e.g. EGA and the ENGAGE community.

- Creation of the Locus Reference Genomic (LRG) standard for the reporting of sequence variant positions. This standard is unique and powerful in that it is the only non-versioning reference structure for key genomic regions of interest. This feature brings long term utility and clarity for specifying variant coordinates (nucleotide and amino acid residue positions) in genes. A lot of effort has been put into establishing and issuing particular LRGs, based on requests from the community, such that ~1,000 LRGs had been adopted for routine use in diagnostic and research applications by the end of the project. EBI will continue this activity longer term.

In the latter one third of the project, as originally scheduled, WP3 extended its focus to semantic standardisation, in terms of developing, cross-mapping and evaluating ontologies, and by providing improved support tools for their examination and use. Extension of this into the realms of the semantic web was also begun. Work included collaborative efforts towards amalgamation and improvement of HPO, MPO, EFO, and Orphanet Ontologies, and extending this beyond rare to common disease. A new DNA Variation Ontology (VariO) was also finalised and released to the community.

Scheduled Deliverables from WP3 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D3.1	Identification of Consortium Use Cases	M9	PU
D3.2	Development of High-Level Domain Model Version 1	M12	PU
D3.3	Standard Reference Sequences, Made Available from Ensembl	M18	PU
D3.4	Scope and Range Requirements of Specialized Domain Models	M18	PU
D3.5	A High-Level Domain Model Version 2, with Sample/Phenotype Focus	M18	PU
D3.6	A High-Level Domain Model Version 3	M24	PU
D3.7	Derivation and Specification of Exchange Format	M24	PU
D3.8	DAS Implementation for GEN2PHEN Data	M24	PU
D3.9	Iterative Specialized Domain Modelling Complete	M48	PU

Scientific papers that were primarily the result of work done in WP3 are as follows:

- Byrne M, Fokkema IF, Lancaster O, Adamusiak T, Ahonen-Bishopp A, Atlan D, Béroud C, Cornell M, Dalgleish R, Devereau A, Patrinos GP, Swertz MA, Taschner PE, Thorisson GA, Vihinen M, Brookes AJ, Muilu J. VarioML framework for comprehensive variation data representation and exchange. BMC Bioinformatics. 2012 Oct 3;13:254.
- Adamusiak T, Parkinson H, Muilu J, Roos E, van der Velde KJ, Thorisson GA, Byrne M, Pang C, Gollapudi S, Ferretti V, Hillege H, Brookes AJ, Swertz MA. Observ-OM and Observ-TAB: Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. Hum Mutat 2012;33(5): 867–873.
- Swertz MA, Dijkstra M, Adamusiak T, van der Velde JK, Kanterakis A, Roos ET, Lops J, Thorisson GA, Arends D, Byelas G, Muilu J, Brookes AJ, de Brock EO, Jansen RC, Parkinson H. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. BMC Bioinformatics 2010 Dec 21;11 Suppl 12:S12.
- Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Beroud C, Dobson G, Lehvaslaiho H, Taschner PEM, den Dunnen JT, Devereau A, Birney E, Brookes AJ, Maglott DR. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med 2010 Apr 15;2(4):24-30.
- Brookes AJ, Lehvaslaiho H, Muilu J, Shigemoto Y, Oroguchi T, Tomiki T, Mukaiyama A, Konagaya A, Kojima T, Inoue I, Kuroda M, Mizushima H, Thorisson GA, Dash D, Rajeevan

H, Darlison MW, Woon M, Fredman D, Smith AV, Senger M, Naito K, Sugawara H. The Phenotype and Genotype Experiment Object Model (PaGE-OM): A robust data structure for information related to DNA variation. Hum Mutat 2009;30(6):968-77.

WP4: GENETICS G2P DATABASES

Work Package 4 (WP4) concerns creating modular and generic G2P database components for single gene analysis (typically for Mendelian disease related data). It also sought to then use these 'building blocks' to construct self-contained software for Locus Specific Database construction: the 'LSDB-in-a-box' concept. These solutions were used to host community-operated LSDBs on GEN2PHEN servers, and deployed for others to set up LSDBs on their own servers. The databases were supplemented by various tools to enable data input and exchange, especially in the context of transferring data from diagnostics laboratories into LSDBs. Improved graphical displays, updates to the Mutalyzer software, and improved search capabilities were also developed.

In terms of LSDB software, WP4 began by bringing together teams working on the UMD, LOVD, and Findis platforms. The DMuDB diagnostic lab database was also included in this activity. Each of these platforms was significantly improved during the course of the project, by incremental and major new releases bringing extensive new functionality - not least to handle omics scale data, to allow pathogenicity and phenotype data/predictions to be handled, and to allow federation of datasets. UMD and LOVD in particular proved very popular with the community, and were widely adopted.

A new database platform was also devised, called the 'Pathogenicity Knowledge Base' (PathoKB). This is now in alpha-testing, and it is designed to support the management of highly diverse and complex data pertaining to pathogenicity inferences about genetic mutations and the underlying experimental evidence and phenotypic details.

Deploying these systems, overall we have launched >200 new expert-curated gene variant databases, helped others to transfer >60 gene variant databases to LOVD format from their original basic systems, set up >2000 foundational databases for human genes, and recruited expert-curators for almost 300 genes in these foundational databases.

More generally, using LOVD3 we have tested a 'Whole Genome Datasets' installation containing all 22 thousand genes (populated with 2 million variants from next generation sequencing, copied from the Exome Variant Server), and created a 'Shared Genes' installation covering all human genes currently containing 27180 polymorphic variants.

A particularly important collaboration has occurred with the International Society for Gastrointestinal Hereditary Tumours (InSiGHT), resulting in the merger of colon cancer gene variant data from several different databases into one new LOVD database, and providing challenging pathogenicity datasets for testing the utility of PathoKB. We also collaborated with WikiProfessional, WikiPeople, and the Concept Web Alliance to connect LOVD systems to the Wiki environment.

To provide web-service accessibility to these increasingly heavily used resources, we built API webservices on top of these databases. Extensive work was done to test the practicality of moving data between systems via this mechanism, especially between research databases and the DMuDB platform, using standards from WP3. We also used the webservices to provide holistic searching across LSDB databases, via the 'OneSearch' system at the main GEN2PHEN Knowledge Center (see WP8), through the WAVe platform (see WP6), and by means of a completely new 'open data discovery' paradigm, implemented as the 'Cafe Variome' software (see WP7).

A nice demonstration of how GEN2PHEN tools could work together to bring new functionality was provided by creating a new version of the Finnish Disease (Findis) database. This exploited existing LOVD installations and data, plus their webservices, to create a virtual database displaying country-specific and world-wide variants using a federated approach.

Thus, WP4 was extremely successful in improving the state of the art for single gene/disease mutation databasing. Our efforts have radically increased the amount, quality, uniformity, and completeness of gene-related G2P information available for integration and use in research and healthcare settings.

Scheduled Deliverables from WP4 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D4.1	User-Manual for LSDB-in-a-box V1 Software	M18	PU
D4.2	Graphical Software for the Presentation of LSDB Data	M24	PU
D4.3	A Validated Code-Base for Checking Mutation Nomenclature	M24	PU
D4.4	Foundational LSDBs for all disease-related genes	M30	PU
D4.5	User Manual for a Stand-Alone Application for Checking Mutation Nomenclature	M63	PU
D4.6	Software for Automated Molecular Diagnostic Data Integration	M66	PU
D4.7	User-Manual for LSDB-in-a-box V2 Software	M66	PU

Scientific papers that were primarily the result of work done in WP4 are as follows:

- Polvi A, Linturi H, Varilo T, Anttonen AK, Byrne M, Fokkema IF, Almusa H, Metzidis A, Avela K, Aula P, Kestilä M, Muilu J. The Finnish Disease Heritage Database (FinDis) Update-A Database for the Genes Mutated in the Finnish Disease Heritage Brought to the Next-Generation Sequencing Era. Hum Mutat. 2013 Jul 31 [Epub ahead of print].
- Celli J, Dalgleish R, Vihinen M, Taschner PE, den Dunnen JT. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 2012 Feb;33(2):291-7.
- Vihinen M, den Dunnen JT, Dalgleish R, Cotton RG. Guidelines for establishing locus specific databases. Hum Mutat 2012 Feb; 33(2):298–305.
- Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutation 2011;32:557-63.
- Dalgleish R, Oetting WS, Auerbach AD, Beckmann JS, Cambon-Thomsen A, Devereau A, Greenblatt MS, Patrinos GP, Taylor GR, Vihinen M, Brookes AJ. Clarity and Claims in Variation/Mutation Databasing. [A response to: "MutaDATABASE: a centralized and standardized DNA variation database" by Bale et al., 2011]. Nat Biotechnol 2011 Sep 8;29(9):790–2; author reply 292-4.
- Mitropoulou C, Webb AJ, Mitropoulos K, Brookes AJ, Patrinos GP. Humu-2010-0187: Locus-specific database domain and data content analysis: Evolution and content maturation towards clinical use. Hum Mutat 2010 Oct;31(10):1109–16.
- 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. Recommendations for locus-specific databases and their curation. Hum Mutat. 2008 Jan;29(1):2-5.

WP5: GENOMICS G2P DATABASES

Work Package 5 (WP5) was tasked with creating modular and generic G2P database components for analyses involving any part, or the whole of, the human genome, plus the use of these 'building blocks' to construct major G2P databases needed by the field. These databases were first launched as 'central' databases, and then the software will be adapted and disseminated to enable a federated database network to naturally emerge. Work also sought to provide a range of related software components to enable the community to add data to, and extract information from, these databases:

such as tools for data submission, tools for data harvesting, tools for local data management and output, and relevant ontologies.
Construction work initially focused on enhancing the software and data content of two major public databases, HGVbaseG2P and IGVdb, providing GWAS data (global) and polymorphism frequency data (Indian sub-populations). Creating core functionalities in these systems involved addressing some basic problems, such as establishing a foundational layer of information on human sequence variation (via 'dbSNP-lite', see WP7), and the identification and correction of significant flaws in large public GWAS datasets. We also had to devise important enhancements to the Biomart data mining platform and the GBrowse browser technology to deploy them in support of our databases. After 1-2 years of such work, we combined efforts to rationalize all our advances into one new fully comprehensive and powerful GWAS database platform, called GWAS Central (www.gwascentral.org).
GWAS Central has now been through several stages of development, involving frequent new software and content releases, such that it now contains orders of magnitude more data than any other open public GWAS database, and highly sophisticated search and display options. It now includes pupose built data submission and curation software to aid data submission and entry, an intuitive search layout (divided into Study, Phenotype, Marker, and Gene/Region sections); semantic standardization of all phenotype content using MESH and HPO terms to enable intelligent searches, an integrated GWASMart data mining tool with canned queries; semantic Web/Linked Data support, and a resource for 'nano-publications' and 'micro-attributions' collaboration with Thomson Reuters). Usage of this database has grown to ~7,000 page views per week (excluding bots and webcrawlers). Data now gathered and prepared for the next imminent release will contain >1600 studies. The system includes extensive core data layers of mutation content provided by SME Partner BIOBASE from their Human Gene Mutation Database (HGMD®), the UCSC gene list, Human Genome Build 37 coordinates, a core representation of the latest content from dbSNP, and HapMap haplotype and LD information. A Data Sharing Statement and policy underpin access to the database, devised in conjunction with ethics work from WP1. Allied developments include devising new curation functions for employing VariO ontology (see WP3), and a standard for the submission of DNA variants to databases (seeWP6).
Beyond core database construction work we have conducted a 'Publicity project', involving emails to corresponding authors of all GWAS publications, requesting they check and enhance their published content in GWAS Central. Response rates and patterns are now being examined. We developed a system for viewing BCP platform data via GWAS Central, also allowing temporary upload of data via the BCP SNPmax publishing tool into the GWAS system for private viewing and integration along with all other stored records, with the option of permanently uploading their study into the database. In collaboration with James Reecy, ALSoD, DistiLD, epiGAD, Genetic Association Database, GWAS DB, GWASdb, Human Phenotype Ontology, and PharmGKB) we launched a GWAS phenotype ontology network called 'PhenoMap' (www.gwascentral.org/gwasphenomap) to employ unified phenotype annotations across all genome-wide association study resources.
All the above software advances were assembled into a 'Virtual Machine' so that others can easily install the full range of GWAS Central capabilities in other project settings. Using this, the IGVdb team created 'GWAS Central - India' (www.vigeyegpms.in/gwascentralindia). Initial datasets have been added comprising 15 Studies, and data requests have been issued to many other Indian groups. The Indian team branded this implementation in their own way, and integrated it via data links with the existing IGVdb resource. Separate discussions are underway regarding the creation of a psychiatric disorders GWAS database, and a pharmaco-genomics GWAS database. The latter possibility has been triggered by interactions with Arthur Holden & the International Serious Adverse Events Consortium, who recently chose to use GWAS Central as the platform via which

they will announce and share extensive GWAS data on adverse drug reactions, jointly created via a multi-pharm consortium.
Based on the positive outcome of the mid-term review, we extended the above scheduled workplan to tackle other challenges as well. First, based on our early realization that research subject identification was a risk when sharing GWAS databasing (which GWAS Central addresses by only providing non-directional content), we started to explore issues around using digital user-IDs to mediate equitable, convenient and secure access to sensitive data in federated systems. This led on to us co-organizing a workshop on this subject (IRBW2009), and a deep relationship with the ORCID project, not least in terms of joining their technical working group and being involved in drafting their initial system specification. Second, we have created the 'Omics Connect' platform. This extends the GWAS databasing and browsing capabilities to cover many other types of omics data, using the Observ-OM++ data model devised in WP3. It involves modular software that one installs locally to allow more involved data mining, optionally also using private/identifiable information that could not be openly shared. The system exploits data input and output via DAS, and as such is highly flexible in the types and sources of information it can handle (local and remote).
Thus, overall, WP5 has been a great success and had a significant impact in improving the quality, uniformity, and completeness of genome-wide G2P information available for access by various routes, not least providing the basis for a federated network of such databases.
Scheduled Deliverables from WP5 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D5.1	Summary Document for Genomics Database V-1 Software	M18	PU
D5.2	First Report on Tools for Data Collection for Genomics G2P Databases	M24	PU
D5.3	Second Report on Tools for Data Collection for Genomics G2P Databases	M48	PU
D5.4	Summary Document for Genomics Database V-2 Software	M66	PU
D5.5	A Fully-Functional and Data-Rich Genomics G2P Database	M66	PU

Scientific papers that were primarily the result of work done in WP5 are as follows:

- Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ. GWAS Central: A Comprehensive Resource For The Comparison And Interrogation Of Genome-Wide Association Studies. Eur. J. Hum Genet (2013) (in press)
- Beck T, Free RC, Thorisson GA, Brookes AJ. Semantically enabling a genome-wide association study database. J Biomed Semantics 2012 Dec 17;3(1):9.
- Beck T, Thorisson GA, Brookes AJ. Applying ontologies and exploring nanopublishing in a genome-wide association study database. In SWAT4LS'11 Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences 2012; 1-2.
- Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, Dash D, Brahmachari S, Brookes AJ. HGVbaseG2P: a Central Genetic Association Database. Nucleic Acids Res 2009;37:D797-802.
- Murtagh MJ, Thorisson GA, Wallace SG, Kaye J, Demir I, Fortier I, Harris JR, Cox D, Deschenes M, Laflamme P, Ferretti V, Sheehan N, Hudson T, Cambon-Thomsen A, Stolk R, Knoppers BM, Brookes AJ, Burton PR. On behalf of the P3G Consortium, GEN2PHEN and BioSHARE-EU Navigating the perfect [data] storm. Norsk Epidemiologi 2012; 21(2):203-9.

WP6: INTEGRATION AND DATA ACCESS TECHNOLOGIES

Work Package 6 (WP6) has responsibility for integrating G2P information in a variety of ways. Partly this information is gathered, generated, or created by GEN2PHEN Work Packages, whilst a large part of it originates outside our Consortium. Integration strategies involve technology development (e.g., SNP-DAS, GRID-based solutions), connections into large scale portals (such as Ensembl and UniProt), and the incorporation of more specialized resources (such as 'DiseaseCards'). The overall goal is to provide integration of this information for communities wishing to use and mine G2P data.

WP6 activities span both pragmatic data flow objectives and broader, technology based objectives. Briefly, these include; i) cross-domain integration, ii) data processing to Ensembl and UniProt, iii) exploring GRID technologies, iv) display development for research biologists, v) search and display options for other users, vi) data presentation, and vii) testing and validation of the integrated technologies.

WP6 was also centrally involved in working closely with WP3 & WP7 to develop and then arrange community support for the LRG standard for gene sequences (see WP3 for further details).

During the course of the project, considerable work has been directed towards further developing and growing the central Ensembl platform. Individual items of development are too numerous to mention, but for example we added DAS (Distributed Annotation System) functionality to the variation pages, integrated large amounts of variation content from 16 species, annotated this content with various ontologies and on structural variants, improved technologies for processing next-generation sequence data files, deployment of an on-line 'Variant Effect Predictor' for all variants, extended the Ensembl API which allows for programmatic access to Ensembl, incorporated initial LSDB data through infrastructure developments that support the storage and display LRG sequences and data. In particular, this work proceeded by close partnership with SwissProt whose systems were similarly improved and with whom much data exchange and harmonisation occurred.

Recent work on the Variant Effect Predictor (VEP) brought several notable improvements. It now provides better access to the intersection of genotype-variation data and the user's own variants. Most notably, VEP predicts the effect of all variants, including structural variants - a feature not available from other comparable tools. For broader utility, VEP can run independently of an Ensembl core database using GTF and FASTA files, and has a simple option to filter out common variants, which is popular in the clinic. All Ensembl variants are available in GVF format dumps, including clinical significance information and global minor allele frequencies. To facilitate clinical use of data from the 1000 Genomes Project, UNIMAN reformatted 1000 Genomes data for clinical use. Parallel, broad-scope and powerful pathogenicity inference support has also been provided by release of a range of 'UMD-Predictor' tools, now as stand-alone applications (UMD-HTS system).

To improve query and display options generally, FIMIM developed two client components: One is a simple table grid for variation data. The other is NCBI's sequence viewer, which was modified so that the viewer (javascript code) can be installed on a local server. The sequence viewer is used for visualising locations of variants selected on the table grid. Components are available on Github. Also, we have used SPARQL technologies for accessing semantic knowledge bases.

Additional integration approaches concentrated on improving the innovative 'COEUS' Semantic Web Application Framework and the DiseaseCard portal (with a new navigation tree), both now with SPARQL endpoint engines. Work on 'GRID Technologies' has proceeded, for example by developing webservice and AJAX components for demonstrating use of VarioML/JSON, and a MongoDB implementation was made for storing Vario-ML objects and to provide a database backend for REST-based web services. Related to this we automated linkage of genotype to phenotype associated findings with genotype to phenotype databanks and repositories: The work extends previous work on G2P scientific discovery workflows towards seamless access and acquisition of these data from relevant public repositories via standard Web-Services and reusable workflows. Discovered discriminant SNPs are automatically hyper-linked to relevant public G2P

repositories (e.g., dbSNP, GWAS Central, DiseaseCard, Ensembl, Pubmed etc) to help biomedical researchers locate and interpret established relevant information.
Hence, WP6 has been highly active and worked across the range of individual tools, federated approaches and centralised approaches to data integration – which we conclude is the best way to increasingly bring the totality of G2P information together for holistic analysis, in research and clinical settings.
Scheduled Deliverables from WP6 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D6.1	State-of-the-Art Document for GRID Technologies	M18	PU
D6.2	Successful Initial Integration	M18	PU
D6.3	Routine Calculation of Consequence Data	M24	PU
D6.4	Successful Integration Using Standards	M36	PU
D6.5	Successful Integration of G2P Data Into DiseaseCard	M36	PU
D6.6	State-of-the-Art Document on New Search Approaches for G2P Data	M48	PU
D6.7	Prototype For New Search Approaches	M66	PU

Scientific papers that were primarily the result of work done in WP6 are as follows:

- Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM. Ensembl 2013. Nucleic Acids Res 2013 Jan;41(Database issue):D48-55.
- Arrais JP, Rosa N, Melo J, Coelho ED, Amaral D, Correia MJ, Barros M, Oliveira JL. OralCard: A bioinformatic tool for the study of oral proteome. Arch Oral Biol 2013 Jul;58(7):762-72.
- Gaspar P, Moura G, Santos MA, Oliveira JL. mRNA secondary structure optimization using a correlated stem-loop prediction. Nucleic Acids Res 2013 Apr 1;41(6):e73.
- Lopes P, Oliveira JL.An innovative portal for rare genetic diseases research: The semantic Diseasecard.J Biomed Inform. 2013 Aug 21. [Epub ahead of print]
- Lopes P, Oliveira JL. COEUS: "semantic web in a box" for biomedical applications. J Biomed Semantics 2012 Dec 17;3(1):11.
- Gaspar P, Oliveira JL, Frommlet J, Santos MA, Moura G. EuGene: maximizing synthetic gene design for heterologous expression. Bioinformatics 2012 Oct 15;28(20):2683-4.
- Lopes P, Mendonça R, Rocha H, Oliveira J, Vilarinho L, Santos R, Oliveira JL. A Rare Disease Patient Manager. In 6th International Conference on Practical Applications of Computational Biology & Bioinformatics 2012;154:173-80.
- Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM. Ensembl 2012.Nucleic Acids Res. 2012 Jan;40(Database issue):D84-90.

- Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Vogel J, Searle SM. Ensembl 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D800-6.
- Lopes P, Dalgleish R, Oliveira JL. WAVe: web analysis of the variome. Hum Mutat 2011 Jul;32(7):729-34.
- 4th International Semantic Web Applications & Tools for Life Sciences Workshop (2011). COEUS: A Semantic Web Application Framework. Pedro Lopes and José Luís Oliveira.
- Tsiliki G, Zervakis M, Ioannou M, Sanidas E, Stathopoulos E, Potamias G, Tsiknakis M, Kafetzopoulos D. Multi-platform data integration in microarray analysis. IEEE Trans Inf Technol Biomed. 2011 Nov;15(6):806-12.
- Lopes P, Oliveira JL. An Extensible Platform for Variome Data Integration. 10th IEEE International Conference on Information Technology and Applications in Biomedicine. 2010 Nov; Corfu, Greece.
- McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010;26(16):2069-70.
- Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Beroud C, Dobson G, Lehvaslaiho H, Taschner PEM, den Dunnen JT, Devereau A, Birney E, Brookes AJ, Maglott DR. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med 2010 Apr 15;2(4):24.
- Kanterakis A, Potamias G, Zacharioudakis G, Koumakis L, Sfakianakis S, Tsiknakis M. Scientific discovery workflows in bioinformatics: A scenario for the coupling of molecular regulatory pathways and gene-expression profiles. Stud Health Technol Inform 2010;160(Pt 2):1304-8.
- 3rd International Workshop on Practical Applications of Computational Biology and Bioinformatics (2009). DynamicFlow: A Client-side Workflow Management System. Pedro Lopes, Joel Arrais and José Luís Oliveira.
- Koumakis l, Moustakis V, Tsiknakis M, Kafetzopoulos D, Potamias G. Supporting Genotype-to-Phenotype Association Studies with Grid-enabled Knowledge Discovery Workflows. 31st Annual International Conference of the IEEE Engineering in Medicine and Biological Society 2009; p 6958-6962.
- Koumakis L, Sfakianakis S, Moustakis V, Potamias G. Discovery of Genotype-to-Phenotype Associations: A Grid-enabled Scientific Workflow Setting. BMIINT: Biomedical Informatics & Intelligent Methods in the Support of Genomic Medicine, AIAI 2009 workshop, CEUR Proceedings (2009) 475:45-56.

WP7: DATA FLOWS

Work Package 7 (WP7) was responsible for collecting G2P data to populate the databases constructed by the GEN2PHEN project, and exchanging datasets between these and other databases. The work of populating G2P databases with comprehensive and/or the latest available, high quality data represents a core challenge for the G2P field. Typical steps in this process include efficiently collecting the information, thoroughly curating it (for governance, quality and semantics), and suitably transforming the data (for syntax), in order for it to be possible to enter it into the relevant

database. WP7 enabled us to undertake and optimise approaches for tackling these and other related challenges.
Activities were therefore directed towards; i) developing an LSDB Federation populated with public domain G2P datasets, ii) populating genomics databases with G2P data and with core data elements extracted from dbSNP via a tool called 'dbSNP-lite', iii) curation of population-specific frequency datasets, and iv) developing quality control guidance for data submitters and receiving databases.
To accomplish these goals, WP7 needed to utilise tools and systems built by WP3, WP4, and WP5 earlier in the project. Most WP7 activities were, therefore, scheduled to begin after the first year. However, given the rapid progress made in GEN2PHEN from the outset, some WP7 work was initiated ahead of that schedule.
Specific achievements were as follows:

- We established many relationships with others regarding data exchange, such as the Pennsylvania State University 'PhenCode' project, the InSiGHT and ENIGMA cancer consortia, the Nordic Centre of Excellence in Disease Genetics, the FinDis.org database, various national databases (e.g. the TREAT-NMD FP6 project), the NHGRI GWAS Catalog, the international Serious Adverse Events Consortium (iSAEC).
- Using a prototype webservice on the LOVD software and the Vario-ML exchange format, we created 'live' federated views of LSDB record citations and/or content within the NGRL Universal Browser, WAVe (see WP6), OneSearch (see WP8), and Café Variome (see below). In the NGRL Universal Browser, 21 LSDBs were thereby federated and an analysis of 187 others highlighted issues that will challenge broad federation. Nevertheless, this enabled dynamic views of BRCA1 and BRCA2 data from two new LOVD databases alongside existing dbSNP, Ensembl and BIC datasets, and relevant disease genes.
- We defined quality control criteria for both the genotyping process and for the datasets themselves (verification of sample relationships and ethnic origins) based on general practices in use for GWAS. A system which can perform the core battery of tests was generated.
- To archive LSDB data in danger of loss in situations where financial support ceases or curators stop maintaining the resource, we made contact with such LSDB curators and offered to revive their stagnant LSDBs and/or help them change to one of our standard database platforms.
- The FINDbase database of disease causing and pharmacogenomic variants was enlarged with population frequency data for genetic variants, making it the richest database of this nature in the world. During GEN2PHEN its content almost doubled, and it has become the world's first 'database-journal' by affiliating with a new open access scientific journal.
- Exonic data for 2000 genes were converted from the 1000 genomes project into a format suitable for clinical use. The process included identifying and processing 1000 genome data, generating VarioML terms, and submitting these data for presentation within Café Variome Central (www.cafevariome.org), from where they can be optionally uploaded into LSDBs.
- Ensembl Variation was expanded from below 50M to >180M variants, covering 16 species and germline and somatic short variants (SNPs, SNVs and InDels)
- We produced and applied the 'dbSNP-lite' tool, to extract and validate core dbSNP content in a form useful for producing foundational marker layers in genomics G2P databases
- The content of GWAS Central database was gathered and processed, such that the next imminent release will hold nearly 30 million p-values for over 1,600 studies, making it the largest open online collection of such data
- The complete record set from the 'Professional content' of Mendelian mutations in the Human Gene Mutation Database (HGMD: www.hgmd.cf.ac.uk/ac/index.php) were made available through the GWAS Central browser.
- GWAS Central India was launched and first round data populated into the system

Additionally, via WP7 we became acutely aware of the many social, legal and practical obstacles that obstruct data sharing and exchange. This prompted us to think beyond standard approaches, and to realize that instead of just data *sharing*, data *discovery* also needs to be addressed. Therefore, we devised a novel software platform, called 'Café Variome', by which the existence of, rather than the substance of, data are made openly accessible via suitable search interfaces. Its operational principles are as follows: a) data providers and users need to use only one data format; b) labs can submit data either *en masse* or one at a time as they are processed (e.g., via a dedicated 'submit button' now on diagnostic lab software from PhenoSystems and Interactive Biosoftware); and c) final data access (post discovery) can be open or controlled or blocked, as per the preference of the submitting lab for each particular user. The system has been piloted by discovering and moving datasets full circle between the COL1A1 LOVD database, Café Variome, and DMuDB. To test the software at scale, we deployed it as a universal website 'clearing house' for variant data discovery, including all public and various private datasets (currently 1,489,687 variants from 11 major sources). Several disease consortia, national diagnostic lab networks, and DMuDB are now in the process of setting up Café Variome to openly publish all their records for discovery purposes.
Scheduled Deliverables from WP7 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D7.1	dbSNP-lite Established	M24	PU
D7.2	Archives Established from Federated LSDBs	M30	PU
D7.3	Pre-submission Quality Control Guidelines	M36	PU
D7.4	Documented Successful Federation of LSDBs	M45	PU
D7.5	List of Populated Foundational LSDBs	M54	PU
D7.6	Summary of Success in Gathering Population Frequency Data	M60	PU
D7.7	Summary of Success in Populating Genomics G2P Databases	M60	PU

Scientific papers that were primarily the result of work done in WP7 are as follows:

- Smedley D, Schofield P, Chen CK, Aidinis V, Ainali C, Bard J, Balling R, Birney E, Blake A, Bongcam-Rudloff E, Brookes AJ, Cesareni G, Eppig J, Flicek P, Gkoutos G, Greenaway S, Gruenberger M, Hériché JK, Lyall A, Mallon AM, Muddyman D, Reisinger F, Ringwald M, Rosenthal N, Schughart K, Swertz M, Thorisson GA, Zouberakis M, Hancock JM. Finding and sharing; New Approaches to Registries of Resources and Services for the Biomedical Sciences.Database (2010) 2010, baq014.

WP8: GEN2PHEN KNOWLEDGE CENTER

Work Package 8 (WP8) was concerned primarily with internet-based exchange of information to and from the community - including G2P data and knowledge about GEN2PHEN and G2P databasing in general. This was centered on a sophisticated web portal, or 'Knowledge Center' (KC). Critically, we intended for this information exchange to be bidirectional - enabling users to find and access what they need at the KC, and also letting them upload data/knowledge/opinion to the KC. A related activity in this Work Package concerns providing training and training information – to provide support to GEN2PHEN Partners and training for the general community. As part of this, GEN2PHEN participated in several summer schools on bioethics and health law, in Toulouse, that extended beyond the G2P community.
A first version KC was constructed during the first year of the project (www.gen2phen.org) and content management systems plus a design path for future versions of the KC were decided upon. Version 2.0 of the KC was created in years two and three of the project, enabling more sophisticated access control, content categorization, tools for community feedback, discussion and contribution.

This was then enhanced with extensive pertinent information, such as news items and event listings, catalogues of G2P resources, and the provision of GEN2PHEN generated resources and training tools to Partners and the general public (via private and public zones of the website). The forum, news and file sections of the KC were particularly useful in helping Partners run open debates and share code and ideas. It also helped drive forward and develop the Bioresource Research Impact Factor (BRIF) topic, leading to a pilot system for biobank ID issuance being launched in collaboration with P3G and a system of harmonization of bioresource citation in scientific papers that has been accepted for presentation (poster) to the international meeting of medical journal editors (Peer review congress, Chicago, September 2013).
Building on this wide layer of project related information and the discussion support system, we next added in several components that furnished G2P data with a holistic focus. These included a complete list of all LSDBs, with powerful search options and rich annotations per database. Plus, most notably, we launched the 'OneSearch' system to enable users to search across the complete G2P public data domain. This was achieved by leveraging webservices, such as those created in LOVD, and the Café Variome technology (see WP7).
In the last two years of the project, a major upgrade to the design and technology behind the KC was executed, leading to the website attaining a high and stable level of system usage: >500 visitors/week, each for an average duration of 3 minutes.
The training provision in WP8 took on many forms, within and external to the Consortium. It began with a survey on this matter, reported in early 2009. In addition to running face-to-face training events, and offering support services, we provided extensive virtual training materials, instructional videos, and User Manuals, etc (via the KC). These covered LSDB curation, ethics, GWAS software, and SME Partner services, and were conducted online and physically in Europe, North America, and India. In many cases this exploiting materials developed in other WPs, not least WP9. Further details and long-term access to these materials are provided at the KC itself.
Scheduled Deliverables from WP8 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D8.1	Procedures to Establish Internal Training Needs	M12	CO
D8.2	Launch of the GEN2PHEN Knowledge Centre	M18	PU
D8.3	Interim Report on Training Activities	M30	PU
D8.4	Update of the GEN2PHEN Knowledge Centre	M48	PU
D8.5	Final Report on Knowledge Management and Training Activities	M66	PU

Scientific papers that were primarily the result of work done in WP8 are as follows:

- Webb AJ, Thorisson GA, Brookes AJ; GEN2PHEN Consortium. An Informatics Project and Online ''Knowledge Centre'' Supporting Modern Genotype-to-Phenotype Research. Hum Mutat 2011 May;32(5):543-50.

WP9: DISSEMINATION, USE AND FUTURE SUSTAINABILITY

Work Package 9 (WP9) was tasked with effectively spreading information about the project to relevant stakeholders. These communication activities are key to GEN2PHEN, as the project needed to interact with, spread knowledge to, and build solid partnerships with, many other G2P data management initiatives. Work was planned to encompass activities related to project communication, making proposals for rewarding and incentivizing database contributions, and exploring business models that guarantee the sustainability of useful resources.
Actual dissemination activities were initiated during the first year, about one year earlier than planned, to exploit many opportunities that presented themselves. Thus, in less than 12 months, >50 dissemination actions were carried out, from paper submissions to oral presentations at congresses.

Moreover, a GEN2PHEN Communication Plan had been finalized and put into action, and the needed communication tools had been developed.
Regarding incentives and rewards issues, the work was mainly focused on two initiatives: the 'Bioresource Research Impact Factor' (BRIF) and the 'Researcher Identity' (see WP1 and WP5). On these issues, an article on tracing biological collections was quickly published at the Journal of the American Medical Association. Soon after this, a white paper was published entitled "Identifying Users And Contributors On The Biomedical Internet", an article on BRIF was published in Nature Genetics in 2011, and an abstract was presented at the 2009 European Society of Human Genetics Conference and published in the European Journal of Human Genetics. An international workshop in Toronto, Canada was then collaboratively organized, entitled "Identifying Researchers on the Biomedical Web (IRBW2009)". A Final Report on Incentives and Rewards in the Field of Biomedical Research Databases was finally produced.
The rapid pace of dissemination work only increased as the project moved onwards and more results became available. This led to the production and use of a vast array of dissemination tools (diptych project brochure, scientific posters, and flyers that showcase specific project results), the organization of further major public conferences and workshops (e.g., on 'Data Sharing and Sustainability', an 'I4Health workshop', and two BRIF workshops Jan 2011 and Oct 2012), and regularly hosting GEN2PHEN booths at the annual ESHG Conferences. All of this was tracked by a regularly updated inventory/repository of project results contributed to by the overwhelming majority of Partners.
The final set of advertising and information flyers individually covered BRIF, Café Variome, Wave, DiseaseCard, Coeus, LSDB curation, and HGMD, and these will all remain available long term at the project KC. Ultimately 787 dissemination activities were carried out during the project, presentations were made at many hundred meetings around the world, and we published 73 scientific papers.
Regarding the critical matter of long-term sustainability and business models, the broad range and diversity of project results were all documented and compiled in the form of 'result fiches'. The resulting GEN2PHEN inventory of results was then used as the basis for organising sustainability and use plans. Specifically, results owners were asked to share their plans for promoting the use of the results beyond the GEN2PHEN project. More than 19 project fiches and sustainability strategies were thereby gathered (see D9.6). Most of the results owners declared their intention to make their results and solutions available in the public domain, while others (a minority) may be commercially exploited.
Scheduled Deliverables from WP9 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D9.1	GEN2PHEN Communication Plan	M12	CO
D9.2	Report on the GEN2PHEN Deployment of Communication Tools	M24	PU
D9.3	Draft Report on Incentives and Rewards in the Field of Biomedical Research Databases	M30	PU
D9.4	Final Report on GEN2PHEN Communication Activities	M66	PU
D9.5	Final Report on Incentives and Rewards in the Field of Biomedical Research Databases	M66	PU
D9.6	Report on Sustainability Models for G2P Data Systems	M66	PU

Scientific papers that were primarily the result of work done in WP9 are as follows:

- Dalgleish R, Molero E, Kidd R, Jansen M, Past D, Robl A, Mons B, Diaz C, Mons A, Brookes AJ. Solving bottlenecks in data sharing in the life sciences. Hum Mutat 2012 Oct;33(10):1494–6.

- Cambon-Thomsen A, Thorisson GA, Mabile L on behalf of the BRIF workshop group. The role of a bioresource research impact factor as an incentive to share human bioresources. Nat Genet 2011; 43:503-4.
- Bravo E, Cambon-Thomse A, De Castro P, Mabile L, Napolitani F, Napolitano M, Rossi AM. Citation of bioresources in biomedical journals: moving towards standardization for an impact evaluation. European Science Editing 2013;39(2):36-8.
- Mabile L, Dalgleish R, Thorisson GA, Deschenes M, Hewitt R, Carpeter j, Bravo E, Filocamo M, Gourraud PA, Harris JR, Hofman P, Kauffmann F, Muñoz-Fernandes MA, Pasterk M, Cambon-Thomsen A; on behalf of The BRIF working group. Quantifying the use of bioresources for promoting their sharing in scientific research. Giga Science 2013 May 1;2(1):7 (open access).

WP10: PROJECT MANAGEMENT

In close collaboration with WP1, WP10 was responsible for all the project management tasks needed for correct implementation of the project, including administrative, financial and legal procedures.
Throughout the project, activities were focussed on setting up the management structure, establishing and clarifying procedures with Partners, and completing the financial tasks needed for appropriate distribution of funds and correct justification of costs incurred in. Recurring tasks included enabling appropriate progress reporting (monthly and official periodic reports), maintaining communication within the Consortium and with the Commission, ensuring Consortium awareness on ethics (including devising operational policies for the project), providing support for appropriate reporting and distribution of funds, and ensuring the correct justification of costs incurred - building on experience acquired as the project proceeded. Doing this well required regularly getting acquainted with new tools developed and used by the Commission for online reporting (including SESAM, NEF, FORCE, ECAS), which were progressively implemented.
In total, four amendments to the Grant Agreement were prepared over the years, involving changes in the partnership and the workplan. Support for meeting organisation was provided, and work of the different management bodies was coordinated. Formal review of all deliverables was arranged, and timely submission of each deliverables enforced. Training of partners in existing and new FP7 procedures was a particularly important activity.
However, perhaps the most important task in WP10 was the development of complete synergy with WP1 in the leadership tasks of the project, which was achieved via very frequent and open contact between the project manager and the scientific co-ordinator. This enabled fully aligned decision making and global steering of the project with maximal efficiency. An optimal relationship between the project manager and the scientific co-ordinator was well established by the time of the mid-term review, which we jointly orchestrated with great success.
Scheduled Deliverables from WP10 were all produced successfully, and most are publically available at the project's Knowledge Center, as follows:

D10.1	Project Handbook	M3	CO
D10.2	Technical and Financial Annual Reports #1	M12	CO
D10.3	Technical and Financial Annual Reports #2	M24	CO
D10.4	Technical and Financial Annual Reports #3	M42	CO
D10.5	Technical and Financial Annual Reports #4	M66	CO

SUMMARY

In summary, the GEN2PHEN project was extremely successful - it met or exceeded all its initial goals, and as such had a large impact on the field of G2P data management and exploitation.

GEN2PHEN worked positively with any other projects and teams, and created over 2000 new inter-connected genotype-phenotype (G2P) databases, based on a partly centralised and partly federated approach, plus an extensive range of support components including curational software, curation tools, advanced search systems, semantic and syntactic standards, operational and ELSI policies, and data collection regimes.

One of the more substantial challenges that remains for the field is that of maximising data use while protecting the privacy of individuals – which is now the focus of the recently announced Global Alliance initiative. The new models for data sharing, knowledge access, and open data discovery that we have explored and validated should therefore be extremely valuable.

The comprehensive ambition we applied to integrating and optimising the G2P field was both necessary and effective. It involved multiple external collaborations to ensure synergistic progress between us and others, encompassing clinical diagnostics, gene specific databasing, whole genome databasing, and human plus model organism datasets – to the benefit of all these stakeholders.

GEN2PHEN has clearly helped to bring about the holistic G2P database ecosystem it originally envisaged. This has already furnished considerable benefit to both research and healthcare, and moved the field a long way forward on its path towards truly personalised medicine based upon the knowledge engineering and 'I4Health' concepts we formally enunciated. The project should therefore provide a solid foundation for future projects aimed at specialized sub-domains, e.g., rare disease or systems biology.

1.4 The potential impact and the main dissemination activities and exploitation of results

The GEN2PHEN project was successful in its goal of having a major impact on the Genotype-To-Phenotype (G2P) data domain. In many fundamental ways, the project catalysed and realised a quantum leap in G2P data organisation and knowledge compilation, including new holistic data querying capabilities and sophisticated interfaces for data visualisation and analysis. This has facilitated far more effective research into the genetic basis of disease, with parallel benefits for clinical diagnostics in support of genomic medicine, stratified medicine, and management and prevention of genetic disease.

At the start of the project, G2P databasing and knowledge management was a very immature field - both in technical terms, in data quality terms, and in ethico-legal terms. Thus it was ineffectual in many ways, meaning that G2P data was barely being shared or exploited by the community as a whole. This implied a great scope for GEN2PHEN to have a big impact, but the highly complex and challenging nature of the field actually presented many significant obstacles to achieving that potential impact. Indeed, this is why little progress had been made prior to the GEN2PHEN project, other than in a few large scale centralised approaches with a very particular mission (e.g., Ensembl, HGMD) where certain datasets could more easily be brought under control, and as part of simpler ethico-legal settings where the governance of sensitive patient data was tightly controlled (e.g., hospital datasets).

The success of G2P data management in previous centralised and straightforward settings had created the impression that progress in the field would require ever more centralised strategies. But that ignores the many benefits that would be gained by involving the wide community in the tasks of G2P databasing and data integration (e.g., expertise data curation, engagement and incentivisation, flexible and tailored approaches to patient data protection, full understanding of data provenance and context which is key for astute data usage). Arguably, one of the main impacts of GEN2PHEN was that it resoundingly broke the mould of how G2P databasing should be done. It demonstrated that it is not only powerful but also quite possible (if approached appropriately) to involve many different stakeholders and empower them to undertake their own G2P databasing and data integration activities. This really had not be tried on a large scale before GEN2PHEN, but as a consequence of our efforts it is now widely accepted that a combined federated and centralised approach is the most effective strategy.

The main reason GEN2PHEN was able to succeed with multi-nodal federated G2P databasing alongside centralised approaches was that the project had been skilfully designed to achieve this from the start. This allowed many diverse areas of development to proceed hand in hand, with maximum synergism and cross-fertilisation between groups and tasks. This relates directly to the design of the 10 workpackages of the project, and so the full list of project impacts will be presented in the frame of the different WPs.

WP1: SCIENTIFIC COORDINATION

This WP had several major impacts. Its core mission was to monitor project progress and encourage a positive and collaborative atmosphere. This 'social' dimension directly mirrors the 'technical' objective of bringing different resources together in a federated manner. Our success in this regard was substantial, convincing doubters within and beyond the project that this new networked approach to G2P databasing was both a viable and an effective way to proceed. Indeed, in a final survey of the consortium members regarding the most important impact of the project, we received answers such as:

- The most important impact was "Enhancing the community interaction" [Helen Parkinson, EBI]

- The most important impact was that it managed to "bring together all the different groups involved in the area, and provide a forum for them to exchange ideas. This was very useful to help direct efforts" [Frank Schacherer, BIOBASE]
- The most important impact was the joint development of systems that "will echo in the years for the community to harvest from, and I expect much of this will still be further developed in future collaborations with gen2phen and new partners in new configurations" [Morris Swertz, University Medical Center Groningen]
- The most important impact was "that the project won the argument about the role of federated approaches" [anon, EBI postdoc]
- The most important impact was the creation of "collaborations beyond Gen2Phen" [David Atlan, PhenoSystems]
- The most important impact was that "the project has brought together groups of people to address important issues in this field" [Andrew Devereau, University of Manchester]
- The most important impact was its ability "to bring together different groups involved in the area, to construct a community and provide a forum to exchange ideas including the legal and ethical part" [Anne Cambon-Thomsen, INSERM]

This important new strategic approach and philosophy was explained to the global community in a key, high-profile publication (Nature Reviews Genetics (2009) 10, 9-18. Genotype-Phenotype Databases: Challenges and Solutions For The Post-Genomic Era. G.A.Thorisson, J.Muilu and A.J.Brookes).
A second major impact from WP1 relates to ethics of G2P databasing and data use. Here we explored Consortium and external views and awareness of issues relevant to the systems being developed. Our growing understanding and delineation of the relevant issues was notable, and frequently shared with external groups by multiple modes of dissemination. Not least, this entailed a deep analysis of, and involvement in, the issue of Data Protection in the context of EU guidance and emerging regulation in this important area.
Subsequent WPs developed a lot of impactful software that has facilitated better G2P research. But the challenge of directly impacting clinical utility turned out to be far greater than anticipated - primarily because of the lack of accessibility of the health data, the very different needs and expectations of clinical versus research users, and the fact that G2P data generally describe weak or uncertain effects of questionable use in the clinic. But rather than give up on this topic, we refocused our efforts to fully understand the nature of the problem and suggest solutions. This was done by leading extensive community discussion (international workshops, surveys, publications) on the question of how to bridge the IT gap between research and healthcare, and by beginning to develop a completely new databasing platform for pathogenicity inference of DNA mutations ('PathoKB'). Thereby we worked out, and catalysed a broad recognition of the fact, that there needs to be a major push (in terms of funding, training, innovation, commercialisation) to develop a new discipline that emphasizes 'Knowledge Engineering' as its core paradigm for Integration and Interpretation of Information for Individualised Healthcare ('I4Health'). This enlightenment was widely disseminated, and is now apparent in many recently funded projects and in the general vision of the new round of Horizon 2020 and IMI funding calls, representing a significant area of impact.
A particularly strong example of this is in the developing focus of the global 'IRDiRC' initiative, for Rare Disease research, diagnostics, and therapy. A Rare Disease ecoSYsteM for PATHogenicity Inference (RD-SymPathI) workshop in the frame of I4Health was organised for and with IRDiRC, leading to abig impact on the structure and focus of their current working groups and tactical priorities. As such, several members of GEN2PHEN have been invited into that project, as well as being full and associate partners of the FP7 RD-CONNECT project. These things are therefore direct descendants and beneficiaries of the work done by GEN2PHEN.

Finally, for WP1, an enduring impact is apparent in the durable and productive community of teams and relationships we forged, reflected by the Consortium agreeing to continue working together on the goals of GEN2PHEN and their evolution, as part of a coalition called the 'GEN2PHEN-Alliance'.

WP2: DOMAIN ANALYSIS AND COMMUNITY RELATIONS

This WP2 was tasked with establishing the needs of the field, both strategic and technical. As such, it was not intended to have a direct impact itself, but provided the all important guidance to the other WPs for them to translate this into final impactful products. However, it is interesting that GEN2PHEN partners did comment that:

- The most important impact was "to map what were the situation and the needs in this field in various communities, based on empirical data" [Anne Cambon-Thomsen, INSERM]

Since all the findings of WP2 were carefully documented and disseminated, then to the extent that other groups read these deliverables and learn from them, the WP2 could be said to have a direct impact itself. Indeed, the work of WP2 was undertaken in partnership with many external groups, experts and G2P field stakeholders, including G2P data creators, database technologists, biobank teams, G2P data end-users, LSDB curators, and members of several initiatives (e.g. HVP, GAIN, WTCCC, P3G consortium, human genetics societies and genetic journals, etc). Therefore, it is more than reasonable to argue that these many external teams and initiatives were positively impacted by the work of WP2.

WP3: STANDARD DATA MODELS AND TERMINOLOGIES

This WP was tasked with developing standards to underlie the G2P database development and data exchange within GEN2PHEN. This includes data models, data formats and terminologies by which the data are represented in repositories, formatted for exchange and annotated for analysis. Core objectives related to use case development, iterative data modelling and format development to support WP4, WP5 and WP6. Such an ambitious standards development program simply had to be (and was) done in close collaboration with all other major teams internationally having a shared interest in these standards, so that the systems built by each of us would be able to interoperate effectively.

In terms of major impact, the particular standards work of WP3 that deserve to be mentioned would include:

- The generic 'PaGE-OM' reference model for G2P data, later enhanced significantly as basic and advanced versions of the 'Observ-OM' data model. This brings an unprecedented, immense flexibility of scope, achieved by a completely new approach to data modelling for the field whereby all entities are abstracted into just 4 concepts. These models are now being further developed or assimilated in new EU and national projects, and have already dramatically lowered the barriers in data integration and software interoperation.
- The novel G2P data exchange formats MAGETAB and Vario-ML, derived from high-level object modelling activities for each of the sub-domains: LSDBs, high-throughput data, and phenotypic descriptions, with metadata and supporting evidence. The formats have already been tested, implemented and positively adopted by several others, e.g. EGA and the ENGAGE community.
- The Locus Reference Genomic (LRG) standard for the reporting of sequence variant positions. This standard is unique and powerful in that it is the only non-versioning reference structure for key genomic regions of interest. This feature brings long-term utility and clarity for specifying variant coordinates (nucleotide and amino acid residue positions) in genes, and hence a level of certainty and useability of G2P data that simply did not exist before GEN2PHEN. External adoption of LRGs has therefore been considerable, such that ~1,000 LRGs have so far been adopted for routine use in diagnostic and research applications.

WP4: GENETICS G2P DATABASES

This WP was concerned with creating solutions for single gene/disease databasing and data analysis (typically for Mendelian disease). At the start of GEN2PHEN this sub-domain was plagued with litle support, no system integration, and many siloed and diverse approaches. Via GEN2PHEN we brought all the main players together under one roof, identified the best features of all the systems, devised a common data model, and build a dramatically larger, more standardised and fully federated ecosystem of such data resources. Basically, we converted a messy and small field (comprising a few tens of disconnected operational databases) to a fully unified network of many thousand resources with hundreds of new database curators.

Our efforts have thus radically increased the amount, quality, uniformity, and completeness of gene-related G2P information available online. This significantly improved scale and accessibility of available gene/disease specific G2P datasets has benefitted not only the research community but also the majority of diagnostic laboratories for whom access to such data is essential in order to guide their interpretation of mutations observed upon diagnostic testing of patients.

Undoubtedly, this realm still faces many more challenges before universal, consistent and precise pathogenicity inference can be said to be possible, but GEN2PHEN has at least now delivered a very robust basal set of resources upon which other components and integration structures can be assembled.

WP5: GENOMICS G2P DATABASES

This WP created solutions for large scale G2P databasing and data analysis, ranging up to whole human genome data perspectives. Many academic and commercial components were connected in this work, which significantly benefitted the SMEs involved. But the main impact of WP5 was the creation of GWAS Central - the world's largest, most comprehensive and most powerful open public database of genome wide association study data. With its content of >1600 studies, orders of magnitude more markers and p-values (>0 million) than any alternative resource, and its user community accessing >7,000 page views per week, the impact of this database on G2P research is substantial.

But beyond just being a widely used and data rich database, this resource also explored novel and important new ideas that are essential for the next stage of evolution of the G2P databasing domain. Examples would be: multi-database collaboration towards semantic standardization of all phenotype content using MESH and HPO terms to enable intelligent data searching; highly sophisticated data mining (search and display) options; software to facilitate data submission and curation; semantic Web and 'Linked Data' support; canned queries; the use of digital IDs to unambiguously identify data, databases, and researchers submitting and consuming data; incentive and reward issues to promote data sharing; ‘nano-publication’ and 'micro-attribution' strategies to recognise and reward data sharing (collaboratively with several external groups, not least Thomson Reuters). These exemplars and experiences are today influencing the direction of development of other major projects, in this and related areas of G2P databasing.

WP5 also allowed us to reveal some real-world (rather than theoretical) insights into the culture of G2P data sharing. Specifically, we conducted a ‘Publicity Project’ that involved sending emails to corresponding authors of all GWAS publications, requesting they check and enhance their published content in GWAS Central. Response rates and patterns revealed a lot about when groups do and do not want to share data, and this will soon be published whereupon we expect it to significantly impact policy and strategy of major initiatives such as the 'Global Alliance'.

Finally for WP5, the full GWAS Central system has been assembled into a virtual machine, and others are now adopting this (e.g., GWAS Central India). As part of this we have been able to convince the iSAEC consortium of major pharmaceutics companies to release their GWAS data on adverse reactions to several major drugs. This will positively impact the field of phrmacogenomics.

WP6: INTEGRATION AND DATA ACCESS TECHNOLOGIES

This WP explored the integration of G2P information in a variety of ways. One aspect of this concentrated on centralised systems for bringing data together and enhancing its utility, using the UniProt and especially the Ensembl platforms. In essence this involved using GEN2PHEN resources to extend and improve G2P aspects of genome data annotation in these systems. Allied to this was work to practically implement and use the LRG standard described above (WP3),and development of the Variant Effect Predictor (VEP) to help inference of the effect of all variants, including structural alterations. These activities add to the major impact that these centralised resources at EBI and Swissprot have on the biomedical field.

Other aspects of the integration work concerned technology development (e.g., SNP-DAS, GRID-based solutions), improvements in the reporting and representation of phenotype (development and systematic use of ontologies), the incorporation of more specialized resources (such as 'DiseaseCards'), reformatting 1000 Genomes data for clinical use, pathogenicity inference support by a range of 'UMD-Predictor' tools (now as stand-alone applications) and creating the innovative 'COEUS' Semantic Web Application Framework. Such advances will each have had a positive impact on particular users and audiences they were designed to serve.

WP7: DATA FLOWS

This WP was concerned with how data flows to and from G2P databases, and around the emerging G2P data ecosystem. Naturally, a lot of effort here was devoted to populating quality data into the databases emerging from the GEN2PHEN project, so that these resources could have a greater impact.

But perhaps more important in terms of impact, WP7 looked at key issues and novel ideas relating to how to maximise the rate of data flow. In particular, the project brought out the real blocking issues and challenges, which were not technological at all. The main obstacles to data sharing and flow were the lack of clarity and regulation about who owns genomic data, and what rights or obligations do people have to share this information. This is further complicated due to ethical questions of consent, of the special nature of genomic data which even if shared only in aggregated form can never be really made anonymous. By surfacing these issues, and disseminating and discussing our findings widely, we have positively impacted and advanced the topic of data sharing.

Furthermore, given the many social, legal and practical obstacles that obstruct data sharing and exchange, we took a step beyond traditional approaches, and realized that instead of just data sharing, knowledge sharing and data discovery must needs to be emphasized. Therefore, we devised a novel software platform, called 'Café Variome', by which the existence of rather than the substance of data are made openly accessible via suitable search interfaces. This approach can be fully open in all contexts, risks nothing for the data owner, and yet shares knowledge and brings 90% of the user benefits of full data sharing. Consequently it has received an extremely positive reception amongst myriad groups struggling with their own data sharing. This includes diagnostics laboratories (where national deployment of Cafe Variome is underway in several countries, and various rare disease consortia are evaluating the tool), and possibly also pharmaceutical and publishing companies as well (initial discussions underway). The principle of Open Data Discovery as a parallel track to data sharing is therefore a major breakthrough created by GEN2PHEN, the impact of which will be enormous.

WP8: GEN2PHEN KNOWLEDGE CENTER

This WP had to facilitate the exchange of information to and from the community, primarily by creating and running the project's 'Knowledge Center' (KC). This portal emphasized bidirectional information exchange, discussion, and training. As such, it was designed to have a major impact by being one of the key dissemination and field deveopment tools of the project. Since the system attained a high and stable level of system usage (>500 visitors/week, each for an average duration of

3 minutes), it is fair to say we were very succesful in achieving the intended level of impact for this aspect of the project. The KC also directly furnished G2P data, with a holistic focus. Specifically, it enabled users to use one interface to search across the vast majority of extant G2P databases - further increasing the impact of the KC.
WP8 also managed the training provisioned by GEN2PHEN. This involved many forms and modes of training within and external to the Consortium, covering LSDB curation, ethics, GWAS software, and SME Partner services, conducted online and physically in Europe, North America, and India. GEN2PHEN also participated in several summer schools on bioethics and health law that extended beyond the G2P community. The large range of high-level training activities we ran will have had a substantial and fundamental impact on improving activities in G2P knowledge management and use.

WP9: DISSEMINATION, USE AND FUTURE SUSTAINABILITY
This WP was tasked with effectively spreading information about the project to relevant stakeholders, and exploring business models that guaranteed the sustainability of project resources. It did this by developing and using any communication tools, and by various modes of interaction and consultation within and beyond the project (surveys, interviews, workshops, etc). A Final Report on Incentives and Rewards in the Field of Biomedical Research Databases was also produced, covering relating issues such as the Bioresource Research Impact Factor (BRIF). This workpackage was therefore important as it facilitated the delivery of the impact described in the other WPs detailed above, and contributed to ensuring the longer term viability of these resources so that their impact will be maintained.

WP10: PROJECT MANAGEMENT
This WP10 dealt professionally with project management issues, and similar to WP9 this contributed to the project's impact by enabling the goals of the project to be achieved. Key facets of this would be the work done by WP10 to ensure the timely and quality production of the many deliverables of the project, adapting budgets and priorities according to match evolving needs and levels of progress made, and resolution of disputes and disagreement.

SUMMARY:
The GEN2PHEN project was an extremely successful program of work. It met or exceeded all its initial goals by a long margin, and as such had a large impact on the field of G2P data management and exploitation. It worked positively with many other projects and teams, and created over 2000 new inter-connected genotype-phenotype (G2P) databases, based on a partly centralised and partly federated approach. This has been transformative in enabling effective user and data interactions for researchers and medical professionals who need to explore and interpret all the existing and new G2P data, which has dramatically grown in size and depth throughout the project's lifetime.
The project has also been very useful in revealing and detailing some of the more substantial challenges that remain for the field, such as how to maximise data use while protecting the privacy of individuals. The new models for knowledge access and open data discovery that we have explored and validated will be extremely valuable in addressing this area of challenge. Additionally, we identified and explored the core remaining challenge of bridging the gap between research and healthcare in terms of data management and use, and we showed how to address this by Knowledge Engineering strategies, especially in the context of Rare Disease. Thus, GEN2PHEN not ony achieved a dramatic impact by virtue of the many useful tools and systems it developed, but its deeper evaluation and probing of the field provides an expertly informed springboard for guiding future development and prioritisation in the field. The whole consortium is rightly very proud of the impact of all these achievements.

.

1.5 Address of the project public website and relevant contact details.

Project coordinator: Prof. Anthony J. Brookes, University of Leicester

Project manager: Carlos Díaz, Synapse Research Management Partners

Contact details:
www.gen2phen.org

List of Partners

- University of Leicester (ULEIC), UK. Anthony J Brookes
- European Molecular Biology Laboratory - The European Bioinformatics Institute (EBI - EMBL), Germany. Paul Flicek; Helen Parkinson
- Fundació IMIM (FIMIM), Spain. Carlos Díaz (Until February 2012)
- Leiden University Medical Center (LUMC), Netherlands. Johan den Dunnen
- Institut National de la Santé et de la Recherche Médicale (INSERM), France. Anne Cambon-Thomsen; Christophe Béroud
- Karolinska Institutet (KI), Sweden. Jan-Eric Litton
- Foundation for Research and Technology (FORTH), Greece. Giorgos Potamias
- Commissariat à l'Energie Atomique (CEA), France. Mark Lathrop
- Erasmus University Medical Center (EMC), Netherlands. (Until June 2009)
- Institute for Molecular Medicine Finland, University of Helsinki (UHFCG), Finland. Juha Muilu
- University of Aveiro – IEETA (UAVR), Portugal. José Luis Oliveira
- University of Western Cape (UWC), South Africa. (Until June 2009)
- Council of Scientific and Industrial Research (CSIR), India. Samir K Brahmachari
- Institute of Genomics and Integrative Biology (IGIB). Debasis Dash
- Swiss Institute of Bioinformatics (SIB), Switzerland. Livia Famiglietti
- University of Manchester (UNIMAN), UK. Andrew Devereau
- BioBase GmbH (BIOBASE), Germany. Edgar Wingender
- deCODE genetics ehf (deCODE), Iceland. Hakon Gudbjartsson
- PhenoSystems SA (PHENO), Belgium. David Atlan
- Biocomputing Platforms Ltd Oy (BCP), Finland. Timo Kanninen
- University of Patras, Greece. George Patrinos (From July 2009)
- University Medical Center Groningen (UMCG), Netherlands, Morris Swertz (From March 2012)
- University of Lund (ULUND), Sweden, Mauno Vihinen (From March 2012)
- Synapse Research Management Partners, SL (SYNAPSE), Spain. Carlos Díaz (From March 2012)