



PROJECT PERIODIC REPORT

Grant Agreement number: 283496

Project acronym: transPLANT

Project title: Trans-national Infrastructure for Plant Genomic Science

Funding Scheme: Combination of CP & CSA

Date of latest version of Annex I against which the assessment will be made: 01.08.2011

Periodic report: 1st ☒ 2nd ☐ 3rd ☐ 4th ☐

Period covered: from 1.9.2011 to 31.08.2012

Name, title and organisation of the scientific representative of the project's coordinator:
Paul Kersey, Dr., EMBL-European Bioinformatics Institute

Tel: +44-(0)1223-494601

Fax: +44-(0)1223-494468

E-mail: pkersey@ebi.ac.uk

Project website address: <http://www.transplantdb.eu>

Declaration by the scientific representative of the project coordinator

I, as scientific representative of the coordinator of this project and in line with the obligations as stated in Article II.2.3 of the Grant Agreement declare that:

. The attached periodic report represents an accurate description of the work carried out in this project for this reporting period.

. The project (tick as appropriate):

- ☐ has fully achieved its objectives and technical goals for the period;
- ☐ has achieved most of its objectives and technical goals for the period with relatively minor deviations;
- ☐ has failed to achieve critical objectives and/or is not at all on schedule.

. The public website is up to date.

. To my best knowledge, the financial statements which are being submitted as part of this report are in line with the actual work carried out and are consistent with the report on the resources used for the project (section 2 of the core of the report) and if applicable with the certificate on financial statement.

. All beneficiaries, in particular non-profit public bodies, secondary and higher education establishments, research organisations and SMEs, have declared to have verified their legal status. Any changes have been reported in the project management report in accordance with Article II.3.f of the Grant Agreement.

Name of scientific representative of the Coordinator: Paul Kersey

Date://

Signature of scientific representative of the Coordinator:



PROJECT PERIODIC REPORT

Publishable summary

Grant Agreement number: 283496

Project acronym: transPLANT

Project title: Trans-national Infrastructure for Plant Genomic Science

Funding Scheme: Combination of CP & CSA

Date of latest version of Annex I against which the assessment will be made: 01.08.2011

Periodic report: 1st ☒ 2nd ☐ 3rd ☐ 4th ☐

Period covered: from 1.9.2011 to 31.08.2012

Name, title and organisation of the scientific representative of the project's coordinator:
Paul Kersey, Dr., EMBL-European Bioinformatics Institute

Tel: +44-(0)1223-494601

Fax: +44-(0)1223-494468

E-mail: pkersey@ebi.ac.uk

Project website address: <http://www.transplantdb.eu>

1. Publishable summary

1. Summary description of the project context and objectives

transPLANT aims to establish a scalable, pan-European research infrastructure to support genomic science in plants through the organisation and interpretation of molecular data, from relatively unprocessed, experimental sequence data through to reference annotation and interpreted models. Through a combination of networking, RTD, and service activities, transPLANT will establish a new, open-access database for plant genomics, a virtual resource built from data (and expertise) distributed throughout Europe. 28 well-defined, verifiable milestones will mark transPLANT's progress towards the following goals:

- i. The establishment of a set of reference data for plant genomes, and a single point of access to plant genomic data (work packages 6, 7; milestones MS15-MS20).
- ii. A new repository to archive genomic variation data, which is at present crucially lacking for plant scientists (work package 9; milestone MS23).
- iii. The development of efficient, accurate tools for sequence description, assembly and alignment, and of metrics for assessing their efficacy and accuracy (work package 12; milestones MS27, MS28).
- iv. The development of new tools and algorithms for exploring and exploiting this wealth of genomic information through its association with phenotype (work package 10; milestone MS24).
- v. The development of controlled vocabularies (ontologies) and data structures for the description and exchange of data and meta-data, and the development of a new meta-data driven search area for data identification (work package 3; milestones MS7, MS8).
- vi. The provision of a compute environment for analysing large data sets remotely, in the cloud and on high-performance compute environments, based on standard, open e-science protocols that support the full interoperability of data (work package 5; milestones MS12-MS14).
- vii. Interact with national and international plant (and related) science research communities to ensure that developments are closely correlated with their needs, and to provide training in the emergent resources (work package 2; milestone MS6).
- viii. The provision of advanced training to the user community (work package 4; milestones MS9-MS11).
- ix. The residual project milestones MS1-MS5 (work package 1) relate to the internal management of the project.

2. Work performed since the beginning of the project and results achieved so far

During the first 12 months, activity has commenced in all work packages, and in most of them, an initial milestone has been reached.

An internal project website was established within the first month of the project's start to support the internal operation of the project: A wiki-style tool (Atlassian Confluence) has been used by consortium members as their primary tool for the storage of documents and the

exchange of information. An issue-tracking tool (Atlassian JIRA) has been deployed for the reporting of bugs and the assignment of tasks (Work package 1, **milestone MS1**).

In work package 2, the partners are coordinating the development of the project with other national and international plant genomics projects, and with parallel projects in other domains. The first milestone (**milestone MS5**, “Report on the ELIXIR preparatory phase”) has been produced as an internal document to help guide these activities (ELIXIR, which brings together governmental funders for the purpose of coordination of informatics infrastructure for all the life sciences, provides a broader context within which the transPLANT activities, and their subsequent evolution, necessarily sit). The organization of the first of a series of stakeholder meetings to explore the needs and opportunities of the plant science community, to be entitled “Genomes To Germplasm”, is currently underway. The meeting will be co-organized with the plant science working group of the EU-US Task Force on Biotechnology Research to maximise the breadth of the input and its potential implications, and will be hosted by transPLANT partner INRA early in 2013.

In work package 3, transPLANT partners have contributed to (as co-organizers and participants) international meetings on standards for phenotype descriptions in crop ontologies, and a set of mandatory or optional fields associated with ontology formalisms for descriptors have been defined (**milestone MS7**).

In work package 4, preparatory work has begun on a transPLANT training programme. The first activity has been scheduled for November 2013 (to be hosted by partner INRA), will be focused on working with partial genome sequences for cereal plants (e.g. wheat, barley), has been advertised and is currently open for registration.

In work package 5, DAS servers have been made available for a total of ten plant genomes, providing a reference framework against which users can integrate positional and non-positional feature data (**milestone MS12**). We have also launched the public website for the transPLANT project (**milestone MS15**), and made significant progress towards the provision of a fully integrated search functionality over all the partner resources. The development of programmatic and interactive tools for accessing genomic data will continue throughout the project, providing progressively deeper integration between increasing numbers of distributed resources.

In the RTD work packages (7-12), new services and analysis tools are in development that will ultimately be included in the transPLANT services. In work package 7, we have developed a registry of available genomic resources, which will be maintained in a distributed fashion by the transPLANT partners (**deliverable D7.1**). We have also incorporated data from 10 new reference genomes into Ensembl Plants (which functions as a central hub resource within the transPLANT infrastructure) and subjected these to comparative analysis (**milestone MS18**).

In work package 8, a genome browser has been implemented that links genomic sequences with sequence-associated data such as genetic. Four species have been identified to serve as priority species for this integration: rice, maize, soybean and *Brassica rapa* (**deliverable D8.1**).

A prototype of **web interface to map genotype-phenotype** has been made available in April 2012, with a focus on *Arabidopsis thaliana* (Work package 10: **milestone MS24**).

In work package 11, several partners, led by IPK, have been working towards the development of the LAILAPS search interface.

Finally, in work package 12, a general method for the analysis of transcriptome variants has been implemented. A transcriptome assembly workflow has been evaluated with a transcriptome sample from two parental *Miscanthus* ecotypes (**milestone MS27**).

3. Expected final results and their potential impact and use

The transPLANT project will coordinate national and international genomics programs, and unify presently distinct activities into a network of interconnected tools, data and resources, creating unified points of access to European plant genomics data (and the association of genomic information with phenotypic characteristics). Specifically:

- transPLANT will foster the development of standard representations for genome scale data from plants, especially (but not limited to) the description of phenotypes (WP3). The project will assess and develop methodologies for the analysis of data (WP8, 10, 12); and will develop a set of community-accepted, reference genomic data (WP7) for use in the plant sciences. These activities will be supported through substantial community engagement (WP2 and 4) and will result in the delivery of services (WP5 and 6) for the sharing of data throughout the plant science research community.
- transPLANT will develop new tools for data visualization (WPs 7 and 8), data mining (WP10) and data discovery (WP11), which will be integrated into the transPLANT services. Moreover, transPLANT will develop services designed for use in a “cloud computing” environment (WP5), a model of growing importance for the provision of data access as data volumes increase. This will have profound effect not only for the plant sciences, but also for other related scientific communities, in which the use of cloud computing approaches is still in its infancy. transPLANT will engage with these related communities to share experience and develop sound, universal approaches for efficient exploitation of compute resources, especially in the context of problems involving large data.
- transPLANT will additionally develop a new repository for plant variation data (WP9), a crucial resource underpinning the potential to translate genomic science into improved crops and societal impact.

Potential Impact

The overall goal of the project is to help address the massive problem of feeding the world in the next 40 years. Humans are completely dependent on a relatively small group of crop plants for food, feed and important industrial materials. Securing a stable, sustainable and affordable supply of crop products has always been, and remains today, the single most important long-term requirement for human progress. The agricultural sector in Europe is the third largest business sector. Thus the social and economic impact of research that facilitates crop improvement directly is therefore exceptionally high.

Understanding how genetic variation translates into phenotypic variation, and how this translation depends on the environment is fundamental to our understanding of evolution, and

has enormous practical implications for human health as well as for plant and animal breeding. It is essential to the goal of feeding the world in a sustainable manner. Thanks do the rapidly decreasing costs of sequencing, we are facing a future where we will have complete genome information for large populations of individuals, for which we also have phenotypic data, e.g., in the form of yield, drought resistance, or metabolome measurements, often in several environmental conditions. The challenge will be integrating these data to elucidate the genotype-phenotype map, allowing us to predict phenotype from genotype, as is essential for genomic selection. The impact of genomics is predicted to reduce the wheat breeding cycle, for example, from 15 to 5-7 years. In this way genomics can make a major contribution both to accelerating the rate of improvement and expanding the scope of new characteristics that can be bred into plants. The transPLANT project is focused on realising this goal.

The volume of data, the extent of necessary analyses, and the need to standardise and distribute data to users for application is an essential part of modern plant science and crop improvement. The genome sequences managed in this project will directly facilitate high density genic marker development, identify genes underlying important traits, and provide cost effective ways of accessing genetic diversity in genes of diverse wheat lines and their progenitors. Bioinformatics access to a reference genome sequence will facilitate a step-change in the way wheat breeding and engineering, trait analysis and gene isolation is performed. Genome sequence also provides a computational framework linking the extensive biological knowledge obtained from model plants and non-plant systems into wheat biology and trait analysis, and has the potential to lower barriers in crop research and draw more scientists into wheat research and crop improvement. Furthermore, genomics facilitates the rapid development of transgenic lines that have the potential to benefit seed companies through protected elite germplasm that commands a market premium for seed sales and which can be used for pyramiding other traits. The development of infrastructure components under open source licences, and the release of data without restriction will particularly aid small and medium enterprises in undertaking genomic research and breeding programmes, and the studies of “orphan crops”, not closely related to the most important crop species but still vitally important sources of nutrition in some parts of the world, which currently suffer from limited financial investment.

The project website has the address <http://transplantdb.eu>



PROJECT PERIODIC REPORT

Core of the report for the period

Grant Agreement number: 283496

Project acronym: transPLANT

Project title: Trans-national Infrastructure for Plant Genomic Science

Funding Scheme: Combination of CP & CSA

Date of latest version of Annex I against which the assessment will be made: 01.08.2011

Periodic report: 1st ☒ 2nd ☐ 3rd ☐ 4th ☐

Period covered: from 1.9.2011 to 31.08.2012

Name, title and organisation of the scientific representative of the project's coordinator:
Paul Kersey, Dr., EMBL-European Bioinformatics Institute

Tel: +44-(0)1223-494601

Fax: +44-(0)1223-494468

E-mail: pkersey@ebi.ac.uk

Project website address: <http://www.transplantdb.eu>

TABLE OF CONTENTS

1. Project objectives for the period
2. Work progress and achievements during the period
3. Deliverables and milestones tables

1. Project objectives for the period

The primary objectives for the reporting period were the achievement of 2 deliverables and 9 initial milestones, establishing the ground for further development of the project.

MS1: Internal project website (EMBL) due 28.2.2012 (work package 1).

An internal project website will support the management of the consortium, the update of information between partners, the archiving of records of internal project minutes, and the tracking of issues with services and requests for new features. These functionalities will be developed using a wiki system for flexible discussion and quick documentation; JIRA for issue tracking).

MS5: Report on ELIXIR preparatory phase (EMBL) due 31.8.2012 (work package 2).

ELIXIR will support a model for long-term sustainability whereby centres of excellence in particular fields interact with a central hub (EMBL-EBI) to integrate biological information. We will interact with the evolving ELIXIR process to integrate plant genomics infrastructure.

MS7: A set of mandatory or optional fields associated with ontology formalism for descriptors (INRA) due 31.8.2012 (work package 3).

The community gathered around the INRA Ephesis project has shown that existing ontologies and formalisms need to be extended to fit the needs of real phenotype data. We will capitalize on international works on ontologies like those led by the Plant Ontology Consortium, the Generation Challenge Program, Xembl, and OBOE.

MS9: 1st transPLANT training workshop (HMGU) due 31.8.2012 (work package 4).

The workshops will train users in understanding the data present in the transPLANT database and use of the interactive and programmatic interfaces offering access to it. Workshop material will typically last between 1 and 3 days, and will consist of a series of lectures, demonstrations, and hands-on practical sessions.

MS12: DAS servers provided for sequence and annotation for 10 reference genomes (EMBL) due 31.8.2012 (work package 5).

To enable distributed computing, web services implementations will be provided over European plant genomics resources maintained by the partners. For genome “features”, transPLANT will provide DAS servers for resources for plant-centric data.

MS15: Initial public launch of transPLANT integrative portal (EMBL) due 31.8.2012 (work package 6).

The portal will be run by EMBL-EBI and will maintain a high-availability service in which the key transPLANT data will be integrated, either directly or remotely (DAS protocol).

MS18: 10 reference genomes incorporated in transplant hub and submitted to comparative analysis (EMBL) due 31.8.2012 (work package 7).

This work package aims at building a repository for reference genome and annotation. The first task will be to establish a registry of plant genomic information (**deliverable D7.1**, due M9). The subsequent step will consist in progressively incorporating reference genomes in order to make them available for comparative analysis.

In work package 8 (An infrastructure for handling plant genomic complexity), there are no milestones due during the first 12 months, but one **deliverable D8.1**: Datasets with associations available and integrated into visualization interfaces (due M12). This will consist of a collection of molecular markers and their visualization on genomic sequence.

MS24: Basic GWAS GUI available (GMI) due 31.8.2012 (work package 10).

Using *Arabidopsis thaliana* as a model, we will develop tools that connect genotype and

phenotype databases and visualize the results. The first interface will be available before the first 12 months.

MS27: Implementation of reference-free methods for transcriptome variation analysis (TGAC) due 31.8.2012 (work package 12).

We will explore and develop methods for genome annotation. This work package will aim at building an analysis pipeline from component modules to finally provide virtual plant breeding.

2. Work progress and achievements during the period

The report covers the period from the project's start to month 12. The project contains 12 work packages, and activity has commenced in each of these.

The project has been progressing well, in line with the planned objectives. Two due deliverables have been submitted on time to the project officer. Eight out of nine milestones have been reached as expected, and we are on course to achieve the outstanding milestone (MS9) shortly (the first transPLANT training meeting has been planned, advertised, and registration is open, but for practical reasons (availability of speakers and the desire to maximise the size of the audience), the meeting has been scheduled for November 2012, not August as initially planned).

At the date of the report, 3 research papers have been published acknowledging transPLANT:

1. Paul J. Kersey, Daniel M. Staines, Daniel Lawson, Eugene Kulesha, Paul Derwent, Jay C. Humphrey, Daniel S. T. Hughes, Stephan Keenan, Arnaud Kerhornou, Gautier Koscielny, Nicholas Langridge, Mark D. McDowall, Karine Megy, Uma Maheswari, Michael Nuhn, Michael Paulini, Helder Pedro, Iliana Toneva, Derek Wilson, Andrew Yates, and Ewan Birney (2012) *Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species*. *Nucleic Acids Res.* 2012 January; 40(D1): D91–D97.
2. Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., & Nordborg, M. (2012). *An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations*. *Nature Genetics*, --. doi:10.1038/ng.2314.
3. H. Mehlhon, M. Lange, U. Scholz, and F. Schreiber (2012) *IDPredictor: Predict Database Links in Biomedical Database*. *Journal of Integrative Bioinformatics*, 9(2):190, 2012.

The following reports describe the activities undertaken in the 12 work packages during the reporting period.

Work package number	2		Start date or starting event:			M1
Work package title	Interaction with national and trans-national genomics and informatics activities					
Activity Type	COORD					
Participant number	1	2	4	5	6	10
Participant short name	EMBL-EBI	HMGU	IPK	INRA	IGR PAN	DLO
Person-months per participant	5	5	3	6	14	2

Objectives

Drawing on the partners' existing collaborative networks, we will organise two workshops for interacting with key external project stakeholders. In these workshops, we will present our results, encourage external stakeholders to present their work, and explore avenues by which the external stakeholders may take advantage of the transPLANT project. Similarly, transPLANT will take advantage of developments in adjacent fields. The output of these meetings will be the production of reports written in collaboration with the stakeholder communities to inform the development of the project, and which will be disseminated to funders, policy makers and collaborating institutions. We will also develop other media for information exchange and collaboration with national initiatives in plant genomics.

Lead Beneficiary: INRA

Description of work

Task 1: Interactions with national plant research initiatives

Objective: Potential overlap and duplications with other efforts in the field of plant science will be identified and discussed to determine whether ongoing efforts need to be adjusted or pursued to maintain the objectives of the project. Contacts will be established with leaders of the plant genome sequence initiatives as well as with leaders of other international plant genomics databases. The task will be to establish and maintain collaborations between these projects, and maximize opportunities for synergistic development.

Description: Workshops will be organized between transPLANT and invited representatives of concurrent projects during either international meetings or transPLANT meetings.

Most partners are involved in international plant genomics initiatives where they managed the bioinformatics tasks of these projects. Projects cover both species data management and computational infrastructure developments. Several initiatives for wheat, barley and grapevine genomics have been funded in the frame of national projects (e.g. 3BSeq, POLAPGEN-BD, Muscares, Barlex) and European projects (e.g. *Triticeae* Genome, GrapeReSeq, EU-SOL, GLIP). In addition, most partners have been mandated by their respective governments or international genome consortia to maintain repositories for data from particular plant species (e.g. *Sorghum*,

Brachypodium, *Oryza glaberrima*, cotton, rye, maize, grapevine, *Arabidopsis*). Finally, they have already established strong connections and collaborations with other similar non-European plant bioinformatics infrastructure initiatives (e.g. Gramene, TAIR). Consequently, they are in a strong position to interact with these communities, to exchange information on features under development.

A part of the transPLANT web site and a mailing list will be established to support information exchange with the other initiatives. For the initiatives in which transPLANT partners have been or are playing leading roles (e.g. grapevine, *Brachypodium*, wheat, barley), the partners will serve as contact persons to ensure interaction and rapid integration of data into the transPLANT framework (see work package 7).

The first of the transplant external stakeholder meeting will be held in the context of these efforts, aimed at identifying the infrastructure requirements for translating basic plant science to agronomical application.

Task 2: Interaction with ESFRI research infrastructure programs

Objective: This task is to maintain interactions with supra-national initiatives in plant sciences or bioinformatics infrastructure.

Description: ESFRI, the European Strategy Forum on Research Infrastructures, is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach. There are 35 ESFRI projects currently in progress, including 10 Biological and Medical Science Research Infrastructures (BMSRIs). Among these is ELIXIR, coordinated by EMBL-EBI, which aims at constructing and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society. In partnership with national funders, ELIXIR is developing new models for coordinated funding and technological development. Securing a stable food supply is one of the primary goals that ELIXIR is being developed to address. Specifically, ELIXIR will support a model whereby nodes – centres of excellence in particular fields – interact with a central hub (EMBL-EBI) to integrate biological information. The development of the ELIXIR framework will have implications on the operations of other infrastructures (such as transPLANT) within the domain of biological information: on the model for long-term sustainability, and on the technical and organizational solutions best deployed in the project.

Several transPLANT partners are participating in the ELIXIR process and will report on developments to the consortium as a whole to ensure that the development of the project fits into the emerging ELIXIR framework. The ELIXIR preparatory phase is scheduled to end at M6 of transPLANT; followed by the establishment of a scientific advisory board and the signature of an international consortium agreement by national funders during 2012-2013. In this task, we will interact with the evolving ELIXIR process to establish a sustainable model for the ongoing development of plant genomics infrastructure.

The Partnership for Advanced Computing in Europe, PRACE, is a unique persistent pan-European Research Infrastructure for High Performance Computing (HPC). transPLANT partner BSC is a participant in this infrastructure and we will also seek to align the development of relevant transPLANT infrastructure with developments in this initiative.

The second of the transPLANT external stakeholder workshops will be focused on this theme.

Progress towards objectives and details for each tasks

The global objective of task 1 is to establish and to maintain collaborations between national plant research initiatives and to maximize opportunities for synergistic development.

The global objective of task 2 is to maintain interactions with supra-national initiatives in plant science or bioinformatics infrastructures.

The deliverables of the two tasks are two reports: one for task 1, provisionally entitled 'Transnational research for agronomical application' and one for Task 2, provisionally entitled 'Future developments in IT infrastructure for plant science; leveraging synergies'.

Task 1: Interactions with national plant research initiatives.

Actions:

To establish collaborations with plant scientists, several steps were planned:

It was first necessary to build a network of the user communities working in plant research fields and to identify the key people and the key projects involved in those fields. The second is to communicate with these communities to give them information about transPLANT objectives, its tools and resources. The last step will be to interact more precisely with these communities in several ways, their own project meetings, training meeting or special stakeholder meetings dedicated to some topics of interest shared by these communities and also by transplant partners. These meetings will provide user communities, knowledge on transPLANT tools and resource but also encourage discussions and promote new collaborations for the future.

Results of year 1

Survey

As a first step to obtain this user communities network and to help also the selection of the invitations to those important meetings that we plan to organize in the coming year project, we built a survey that was filled by all the Transplants partners.

This survey is a excel file composed of several sheets, that gives for example information on the main projects in plant genomics in which transPLANT partners are in partnership, the name of their project coordinator, the name of the bioinformatics contact in transPLANT for this project, the species involved, the funding.

Another sheet lists which bioinformatics teams are also involved and have repository mission (short-term, long-term), who are involved in managing the data, what kind of data are loaded into these databases such as genetic resources, genomic annotation, markers (microsatellites, SNPs, ...), phenotypes, where these data are available and if they are available to query and how.

Another sheet lists also which transPLANT partners are involved in international bioinformatics initiatives such as the Bioinformatics Task Force. Networks and collaborations are indeed very important tools to set up collaboration for the future at international level.

This file was updated several times according to new projects accepted at national level, for example also in spring 2012, taking into account the results of the French call 'investment' for the future. It will be available on the transPLANT web site.

Communication events

Communication on Transplant project was already initiated for some projects, i) during the european *Triticeae* closure meeting in 2012 in Versailles and also ii) in the annual workshop of POLAPGEN-BD project in Poland. In particular, the problems covered by WP3 were presented to

geneticists. New communications events are planned in autumn 2012 for other projects such as Breedwheat (wheat genomics French community), Amaizing (maize genomics French community), Rapesodyn (for Rape) and PeaMust (for Pea).

Stakeholder meetings

It is intended to hold a number of key stakeholder meetings in the course of the project to determine the needs of the plant genomics research community, to access information about relevant technologies and to identify the key scientific challenges expected to drive research and applications in future years. These meetings will lead to the production of reports, which will serve as records of the current state of the art, and potentially the basis for awareness-raising publications or even funding applications. To maximise the effectiveness of these meetings, we have decided to co-organise these with the Plant Biology Working Group of the EU-US Task Force on Biotechnology Research. The Task Force is an organisational structure that accommodates funding agencies and scientists from Europe and North America with the goal of coordinating research and collaboration and the funding programmes that support it. The Task Force has previously organised meetings in the plant genomics area e.g. at Hinxton, United Kingdom, in December 2009, and the joint organisation brings the opportunity to bring increased numbers of U.S. participants to any meetings (at no cost to the transPLANT project), to ensure that any resulting report reflects the overall view of the global research community, and to bring conclusions to the attention of funding agencies.

We have identified two areas in which we think it would be beneficial to hold stakeholder meetings within the next year: one entitled « Genomes to Germplasm », which will be hosted by the transPLANT project, and one focused on the capture, coordination and analysis of phenotype data, which will be hosted by a United States-based participant (Doreen Ware, Cold Spring Harbor Laboratory). transPLANT partners (and other relevant scientists) will contribute to the scientific organisation of both meetings, Each of which will feature the participation of 30-40 scientists from both academia and industry. Invitations for the first meeting will be sent out in October 2012 for a meeting date in 2013.

Task 2: Interaction with ESFRI research infrastructure programs.

Since the initiation of transPLANT, the ELIXIR project has now completed its preparatory phase and has entered its construction phase. 13 countries (as well as the coordinating international organisation, EMBL) have now signed a memorandum of understanding and will move towards the formal establishment of the ELIXIR organisation. With the completion of the preparatory phase, a number of documents have been produced describing the needs of the life sciences for informatics infrastructure and providing a blueprint for its delivery through ELIXIR. The appointment process for a founding director for ELIXIR is underway and many national funding agencies are developing concepts of ELIXIR nodes, which will interact with the coordinating hub to provide the integrated European infrastructure.

The impact of ELIXIR on plant sciences will depend crucially on the partners' choice of nodes. The establishment of national centres of expertise, working in relevant domains, working in partnership with EBI provides a model that allows scalable yet integrated development against the backdrop of expected growing importance of information infrastructure in the life sciences. A report (transPLANT **milestone MS5**) has been prepared describing the current state of the ELIXIR project and its potential impact for genomic science. Under the leadership of EMBL-EBI, which coordinates both projects, transPLANT partners will work with the relevant ELIXIR nodes, as they emerge, to ensure all developments are complementary.

If applicable, explain the reasons for deviations from Annex I and their impact on other tasks as well as on available resources and planning

No deviation

If applicable, explain the reasons for failing to achieve critical objectives and/or not being on schedule and explain the impact on other tasks as well as on available resources and planning *(the explanations should be coherent with the declaration by the project coordinator)*

No deviation

Use of resources *(highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1)*

EBI: 0.5 person months (towards D2.1)

INRA: 1 person months

IGR-PAN: 1 person months

IPK: 0,25 person months

Work package number	3		Start date or starting event:			M1
Work package title	Community standards for the interoperability of data resources					
Activity Type	COORD					
Participant number	1	4	5	6	7	10
Participant short name	EMBL-EBI	IPK	INRA	IGR PAN	BIOGEM	DLO
Person-months planned per participant	4	4	12	24	6	6

Objectives

Develop community-accepted standards for data description and submission, covering format, and content and policy.

Lead Beneficiary: IGR PAN

Description of work

Task 1: Standards for phenotype description

Objective: Phenotype is a concept used in many domains of biology, such as transcriptomic, association genetic, the interaction of genotype and environment, and experimentation. The aim of this task is to determine the minimum set of data necessary to describe a phenotype.

Description: For plant genetic resources, the concept of a passport has been established: a minimum set of data needed to describe the resource. In a similar way, we need to determine the minimum set of data necessary to describe a phenotype. In different domains, however, the precise data corresponding to the concept of a phenotype varies greatly, from a simple descriptor attached to a genotype to the whole data set of an experimental trial. Furthermore, a phenotype is always the result of the action of an environment applied on a genotype. The way environmental parameters are recorded differs also depending on the scientific domain considered. Finally, the appropriate level of elaboration of a phenotype condition must be decided. These data will necessarily include genotype traceability, and phenotype and environmental descriptors. The statistical descriptors under development in WP10 also need to be captured in these representations.

Previous experience has shown that to ensure comparability of the data, descriptors must be organized in ontologies. We will capitalize on international works on ontologies like those led by the Plant Ontology Consortium (<http://www.plantontology.org>), the Generation Challenge Program (<http://www.generationcp.org>), Xeml (<http://xeml.codeplex.com>), and OBOE (<http://marinemetadadata.org/references/oboeontology>). The community gathered around the INRA Ephesis project (<http://urgi.versailles.inra.fr/index.php/urgi/Projects/URGI-sofwarewares/Ephesis>) has begun the analysis of these ontologies and has shown that existing

ontologies and formalisms need to be extended to fit

The needs of real phenotype data. The Plant Ontology Consortium has been contacted to discuss the possibility to extend the ontologies they maintain. Furthermore, INRA is working with Bioversity International (<http://www.bioversityinternational.org>) to develop extended ontology formalism from the OBOE and the EQV models. transPLANT partners (e.g. Biogemma, INRA, IPG PAS) will be able to enrich ontologies by providing the phenotypic variables they use.

One aim of transPLANT is to develop services serving as a central point for all data access, and to enable systems biology approaches to complex assemblies of diverse data. Standards for the representation of phenotypic data must be comprehensive, including all types of “-omic” data. Care will be taken that the existing and applied or newly developed descriptors and ontologies will be appropriate for data integration on a wider scale and allow the establishment of inter-relationships between different ontologies. This is necessary in order to support queries over different “-omics” features (e.g., between transcriptome and proteome, between enzyme levels and metabolic concentrations, etc.) in any query system that accesses these data. It is an essential requirement if users need the capacity to analyse the data and plan experiments by proposing a priori hypotheses and specific assays to test them.

Progress towards objectives and details for each tasks

The work in progress is described according to the deliverable:

D3.1. Recommended ontology set for use in phenotype description and epigenetic variability

- P. Kersey, P. Krajewski and C. Pommier participated in the workshop Crop Ontologies for Agronomic Traits organized by EBI with support of BBSRC (Hinxton, 8-9.12.2011), in which researchers, breeders and ontologists from the United States and the European Union were brought together to discuss the potential application of plant (trait, phenotype) ontologies to describe agronomic traits in crop plants. In the workshop the current state of several ontologies was presented, and comparison was made between controlled vocabularies in use in academic environments (often ontologically structured, and describing phenotypes at the level of the individual plant), and the vocabularies in actual use by breeders (which often are country-specific, have a flat structure, and which may describe phenotype at the level of an entire crop, where both values and concepts may be different). Real phenotypic datasets were considered as practical examples of cases requiring standardization. The work consisted of, e.g., taking decisions if a particular data set can be standardized according to the existing ontologies, or if the ontologies should be extended or reconstructed in such a way that the proper standardization is possible. transPLANT is continuing to coordinate with international partners in this area, and has contributed to the organization of a follow-up meeting (with a particular focus on the active development of particular ontologies) to be held in Oregon (local organizers Laurel Cooper, Pankaj Jaiswal, Oregon State University) in September 2012, together with Plant Ontology Consortium, building a Reference Plant Trait Ontology. It will serve as a foundation to a collection of specific ontologies, which will cross-reference the reference ontology to ensure consistency. The specific ontologies will therefore be much easier to maintain by their community, for instance by a species group.

- In collaboration with partners a report on describing current position of ontology formalism (**milestone MS7**) was discussed in a teleconference, prepared and sent to the coordinator. We chose to capitalize on existing work at international level, conducted within the PhenotypeRCN coordination network. This RCN gathers reference of international plant trait ontologies and particularly collaborates with the Plant Ontology Consortium (coordinating Plant Ontology and Gramene Trait Ontology) and with the Generation Challenge Program (coordinating the Crop

Ontology).

D3.2. Format specifications for data exchange by flat file and web services

- The groups from INRA and IPG PAS collaborated on evaluating the data exchange format used by Ephesis database for input of phenotypic data. The format provides slots for inputting metadata concerning description of the experiment, observed variables, observed genotypes, design of the experiment, and for the data themselves. Propositions were developed for representation of data obtained in the experiments conducted in several environments (with factorial structure) and at several time points. Exemplary data from IPG PAS were prepared according to the format and proposed solutions and submitted to Ephesis. A portion of real data obtained in POLAPGEN-BD was also prepared according to the format and successfully submitted to Ephesis for public accessibility.

D3.3. Report on standardisation activities accomplished during transPLANT

- IPG PAS: P. Krajewski and H. Ćwiek worked on standardization of phenotypic observations in POLAPGEN-BD. In this project a broad spectrum of traits is observed on genotypes of barley subjected to drought and other treatments. At first, the morphological traits (such as spike and stem dimensions), productivity traits (grain yield, grain number, number of tillers) and developmental traits (stages) were considered. The observational protocols and units were standardized among partners running phenotyping experiments. Then the trait names were standardized according to unified language requirements and in relation to existing Trait Ontology terms. The work proved to be essential for production of statistical reports concerning all experiments run within POLAPGEN-BD. The effect of this work was presented at the annual workshop of POLAPGEN-BD project held at Institute of Soil Science and Plant Cultivation in Puławy.

Use of resources *(highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1)*

EMBL-EBI 0.5 person-months

IPK 0 person-months

INRA 3 person-months

IPG PAS 9.75 person-months

BIOGEMMA 0 person-months

DLO 0 person-months

Work package number		4	Start date or starting event:						M1	
Work package title		User Training								
Activity Type		COORD								
Participant number	1	2	3	4	5	6	8	9	10	11
Participant short name	EMBL-EBI	HMGU	GFMPG	IPK	INRA	IGR PAN	TGAC	BSC	DLO	KN
Person-months per participant	6	10	2	2	2	8	2	2	2	2

Objectives

Organize a series of training workshops for the transPLANT user community.

Lead Beneficiary: HMGU

Description of work

Task: Organization of training workshops

Objective: Organize a series of training workshops for the transPLANT user community.

Description: We will organize a series of training workshops, held across Europe each focusing on a defined area of the transPLANT project. The workshops will train users in understanding the data present in the transPLANT database and use of the interactive and programmatic interfaces offering access to it. Workshop material will typically last between 1 and 3 days, and will be presented by representatives of the project partners and invited guests, and will consist of a series of re-usable modules including lectures, demonstrations, and hands-on practical sessions. Access to the course materials will be provided to participants after the conclusion of the course. Each workshop will be focused on the needs of a defined user community, e.g. introductory courses aimed at experimental researchers and advanced courses aimed at experienced bioinformaticians. All workshops will be aimed at both academic and commercial participants. Each course will concentrate on resources developed by transPLANT, but will also cover related tools developed by the project partners and by others. Potential foci of individual courses include: Analysis of next generation sequencing data Genome sequence and annotation Integrating “-omics” data: genomics, transcriptomics and proteomics Resources for exploring genotype-phenotype interactions; Programmatic access to molecular biology databases; Cereal genomics; Genomics of dicotyledons. Courses will be hosted by project partners and by other interested organizations. Access to the courses will be offered free of charge, although attendees will be required to pay their own costs for travel and accommodation. Workshops will be promoted through the transPLANT website and through the websites

of the project partners.

Progress towards objectives and details for each tasks

The 1st transplant user-training workshop is scheduled for 12-13 November 2012 and will involve participations from several transplant partners. The reason for the delayed date is mainly availability of both teaching personnel and workshop venue. After surveying both workshop teachers and potential participants it became clear that a date in autumn, e.g. November would allow significantly more users to participate and to hold a workshop with teachers from all relevant transplant partners.

The 1st transplant training workshop will be focused on *Triticeae* (wheat, barley) data resources at transPLANT partner sites as highly relevant and interesting but complex data was generated there recently (several high-impact papers in preparation). This workshop will be dedicated to both data end users such as experimental biologists and data analysts/bioinformaticians.

The training workshop program was set up to teach at least 25 users (both from transPLANT partner institutions but also from all other interested parties worldwide) over 2 full days. Individual transplant partner participations and sessions involve INRA Versailles, EMBL-EBI, IPK Gatersleben, Helmholtz Center Munich and INRA Clermont.

The workshop is announced over transPLANT partner websites, several mailing lists (including transPLANT and *Triticeae*Genome mailing lists), cooperation partners and within connected communities (specifically *triticeae* communities in Europe and US).

The (preliminary) 1st transPLANT user workshop program is attached:

Nov 12th – Day1 (Monday)

10:30 Welcome, computer setup and introduction of workshop objectives and agenda (Delphine Steinbach, Hadi Quesneville, Manuel Spannagl)

11:00 Introduction: about the transPLANT project (Paul Kersey, Klaus Mayer, Hadi Quesneville)

11:30 Introduction to the public transPLANT web hub at EBI (Dan Bolser, Paul Kersey) including a short introduction on search engines developed in transPLANT (IPK Gatersleben: Uwe Scholz, Jinbo Chen) <http://transplantdb.eu/>

12:30 Lunch

13:30 *Triticeae* data@ENSEMBLplants: introduction+data access (Dan Bolser, Paul Kersey)

<http://plants.ensembl.org/>

15:30 Coffee break

16:00 *Triticeae* data@MIPS (Manuel Spannagl, Klaus Mayer):

Concept of and interactive data access to the barley and wheat genome zippers

The barley genome: integration of physical and genetic map, data access and use cases

UK wheat 5x WGS+analysis results: concepts to use this new data resource

Comparative genomics – from models to crops: exploring synteny, visualization tools (CrowsNest)

<http://mips.helmholtz-muenchen.de/plant/triticeae/index.jsp>

18:00 Close of day 1

Nov 13th – Day2 (Tuesday)

09:30 GnpIS tool training session

Quicksearch tool, Advanced search tool (Biomart) and links between Biomart and Galaxy tool

Main focus on Wheat data@URGI Versailles (Delphine Steinbach, Aminah-Olivia Keliet, Nacer Mohellibi, Michael Alaux): introduction, data access, tools, use cases to define: (search by marker, by gene, qtl, snp), graphical viewers: all data centered on genome browser, links to genetic map viewers.

<http://urgi.versailles.inra.fr/gnpis>

12:30 Lunch

13:30 Annotating *triticeae* sequences: the triANNOT pipeline (Phillipe Leroy – external speaker-, INRA Clermont-Ferrand): introduction, use cases

<http://clermont.inra.fr/triannot>

15:00 Coffee break

15:30 Wrap-up, time for questions and discussion (also after and during the individual sessions)

17:00 End of workshop

Required user skills: none

Required user hardware/software: PC or laptop with any operating system and web browser.

We will follow up this 1st transplant user-training workshop with a transPLANT computer demonstration session at the next Plant and Animal Genome Conference in San Diego in January 2013 (<http://www.intlpag.org/2013/index.php/abstracts/workshop-speakers-a-demos>). This conference and computer demonstration will allow us to reach a huge number of –both European and international- potential transPLANT service users. We will both introduce the scope of transPLANT and give very short tutorials on selected transPLANT services and resources. A poster will be presented as well to facilitate discussion and help in case of questions.

If applicable, explain the reasons for deviations from Annex I and their impact on other tasks as well as on available resources and planning

Workshop slightly delayed due to availability and date preferences of venue, workshop teachers and targeted audience. No (negative) impact on other tasks and no impact on available resources and planning.

If applicable, explain the reasons for failing to achieve critical objectives and/or not being on schedule and explain the impact on other tasks as well as on available resources and planning (the explanations should be coherent with the declaration by the

project coordinator)

Workshop date slightly delayed due to availability and date preferences of venue, workshop teachers and targeted audience. Workshop will be held in November. No (negative) impact on other tasks and no impact on available resources and planning.

Use of resources *(highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1)*

HMGU: 1 PM

EBI: 0.5 PM

BSC: 0.31 PM

TGAC: 0.03 PM

Work package number	5		Start date or starting event:						M1
Work package title	Programmatic services for genome-scale data								
Activity Type	OTHER								
Participant number	1	2	4	5	7	8	9	10	11
Participant short name	EMBL-EBI	HMGU	IPK	INRA	BIOGEM	TGAC	BSC	DLO	KN
Person-months per participant	4	4	2	10	1	2	21	4	7

Objectives

Develop programmatic services, and environments for running programmatic analyses, for plant genomes.

Lead Beneficiary: BSC

Description of work

Task 1: Provision of a cloud compute environment

Objective: Provide a cloud compute environment for parallel, portable processing of large data

Description: The aim of this task is to provide a cloud compute programming environment for the project. The environment will be composed, first by a cloud computing infrastructure. For this infrastructure several options can be considered, from European proposals like OpenNebula toolkit or the BSC solution, EMOTIVE cloud; or other international proposals like Eucaliptus or commercial solutions like MS Azure. Based on this middlewares, the project cloud infrastructure will be extended and adapted to meet the purposes of the project.

On top of this cloud computing infrastructure, the COMPSs framework will be contributed from BSC to offer an easy porting and development framework of the applications on top of the cloud computing platform. COMPSs is an innovative programming framework for distributed computing environments that enables unskilled programmers to develop applications that can be run in a distributed infrastructure. The COMPSs runtime has the ability of parallelizing the applications at task level, distributing the execution of parallel tasks in different resources of the underlying infrastructure. While COMPSs was initially designed to run in grids and clusters, the current version has already been enabled to run in the cloud and further developments in this direction are ongoing in the framework of the projects VENUS-C and OPTIMIS. Besides enabling the programming model for cloud computing environments, the extensions in the project

OPTIMIS will consider the inclusion of WS as part of the COMPSs applications. This work is well aligned with the other WP tasks, and will also provide an environment for the algorithms developed in WP12.

A COMPSs-enabled version of Hammer has already been used by EBI for long runs in the MareNostrum supercomputer (using more than 100.000 cpu hours) demonstrating its suitability (Tejedor, E., Badia R.M., Royo R., Gelpi, J.L. Enabling HMMER for the Grid with COMP Superscalar. (2010) Proc. Comp. Sci. 1(1), 2629-2638). In the framework of the tasks, BSC will contribute by setting the cloud computing infrastructure, to offering the COMPSs programming and adapting it for the project needs and by supporting the different project applications to port suitable applications to this environment.

Task 2: Implementation of HPC web services

Objectives: Provide an environment for compute-intensive data analysis

Description: Analysis pipelines above would eventually include operations that require a large amount of computer power, or specific HPC facilities like large shared-memory servers. Those operations will constitute the bottlenecks of the analysis process unless being processed in high throughput servers.

Following the experience gained with the development of a complete set of Biomoby based web services (<http://inb.bsc.es>. S. Pettifer et al. The EMBRACE web service collection. Nucleic Acids Research. (2010) 38. W683-W688), BSC has developed the necessary technology to interface highly demanding applications running on large clusters or shared-memory servers. Interfaces available include standard SOAP and REST access, and programmatic access via Perl and Java APIs. The developed interfaces are compliant with the authentication and security issues that are required to access HPC facilities. Interface backends are powered by COMPSuperscalar and other technologies, to gain the maximum benefit of the specific underlying computer architectures (openMP, MPI, CUDA, etc.).

From this background, operations from tasks 3 identified as computationally demanding will be implemented on BSC HPC facilities, and made available through the appropriate interfaces to be integrated on the selected cloud compute environment.

Task 3: Provision of web services for computational analysis

Objective: To enable distributed computing, web services implementations will be provided over important European plant genomics resources maintained by the partners, according to the standards established in WP3.

Description: Web services will be provided over a range of resources already maintained or newly developed by the transPLANT partners, including the following:

Ensembl Plants: For genome “features” (annotation located to a particular range in a sequence coordinate system), the DAS protocol (Jenkinson AM et al. BMC Bioinformatics. 2008 9 Suppl 8:S3) is a lightweight REST-ful web service already in wide use by genomics resources. transPLANT will provide DAS servers for important resources for plant-centric data; other data types may require the use of alternative technological approaches.

Ensembl Plants (Kersey, P.J. et al. Nucleic Acids Res. 2010 38:D563-9) is a portal for genome scale data for plant genomes. Coding and non-coding gene annotation, variation and alignment data is available in the context of the coordinates of reference genome sequence. In the DAS framework, Ensembl software can function both as a sequence server (providing reference

sequence which can be combined with annotation from other sources) and as an annotation server. transPLANT will use Ensembl to serve reference genomic data, and associated features and gene summary information, through the use of the DAS protocol.

GnpIS

INRA URGI maintains and develops an information system called GnpIS (Samson D. et al. Nucleic Acids Res. 2003 31:179-82) developed first in 2000 to collect and integrate all the data of the French federative program Genoplante. This information system built on different databases connected together. Interfaces allow users to query the data according the type of data they want to access. For example genetic maps are available through GnpMap database, SNPs data are available through GnpSNP database, gene and annotation are available through GnpGenome database (a database based on the Chado (Zhou P. et al. Curr Protoc Bioinformatics. 2006 Chapter 9:Unit 9.6) and Gbrowse (Donlin M.J. Curr Protoc Bioinformatics. 2009 Chapter 9:Unit 9.9.) tools developed in the framework of the GMOD project). This Information system is used and is guided by user needs of INRA Wheat, Grape, Maize and bioagressor (mainly fungi) communities. Increasingly, there are also requirements for the storage of forest genomics data (poplar, pine, oak). GnpIS offers also interfaces for cross querying of resources, one tool based on Lucene technologies to rapidly search into an indexed data warehouse, the other one a advanced search tool based on the BioMart system developed to answer to specific biological questions (Gene, QTL or SNP oriented marts) by supporting queries on diverse datasets. Web service interfaces to these queries will be provided. Web services for phenotypic data (based on the schema developed) in WP3 will also be provided.

MIPSPplantsDB

MIPSPplantsDB (Spannagl M. et al. Nucleic Acids Res. 2007 35: D834-40) is a portal for plant genomes and genome associated data and harbours gene sequences structural and functional annotation data, syntenic information and cooperative data. It serves as community portal for a range of national (barley, rye, maize resequencing), European (tomato, Medicago, Arabidopsis thaliana) and international genome initiatives (e.g. barley, Oryza glaberrima, Brachypodium, Sorghum). The resources will be integrated into the transPLANT DAS framework as both sequence and annotation servers.

Data cart web service for the meta-data driven information retrieval systems

The Data Cart is a concept for collection, transformation and distribution of data in the information retrieval environment developed in WP11. The IR environment will loosely link the partner resources by a reverse lookup from public repositories for gene and protein functions to genomics data. Doing so, a search engine, including recommended system and user specific relevance ranking, is offered. The results of search queries are hits in protein databases or other relevant genome annotation resources and linked genomic data from transPLANT partner. This data will be persisted for later analysis by WP5 web service infrastructure in the Data Cart. Here, the user may individually maintain a stock of transplanted data, which are relevant for his/her planned analysis. Furthermore, there will be the option to collect all results from runs of those analysis pipelines as citable data records. Here, we provide globally uniquely identifier that will be resolved as web service endpoint.

Tools for functional genomics

An infrastructure for genome annotation that significantly outperforms other related methods for protein function has been developed at StDLO (Kourmpetis et al., 2011). We will develop our current, stand-alone software implementation for protein function prediction into a platform that is web-based; employs web-services technology and can deploy the power of cloud-

computing for protein function prediction on a genome-wide scale. Through integration of this platform with the database and analytical resources developed for collections of complete genome sequences, we aim to significantly increase the coverage and specificity of functional annotation of plant genomes.

Task 4: Provision of interoperable data warehousing technologies

Objective: Offer optimized data mining for common queries through the deployment of data warehousing technology

Description: BioMart (Smedley, D. et al. BMC Genomics. 2009 10:22) is a query-oriented data management system widely used in bioinformatics. 34 sites are currently referenced at the BioMart central server (<http://www.biomart.org>), and third party software with BioMart plugins implemented include Bioclipse, biomaRt-BioConductor, Cytoscape, Galaxy, Gitoools, Ruby, Taverna, and WebLab. BioMart provides a set of tools for the easy construction of denormalised databases and (programmatic and interactive) interfaces focused on common queries. Access to BioMarts is available via both RESTful web services and a programmatic API. Crucially, BioMart supports data federation, allowing the performance of distributed queries between different Marts. BioMart already in deployed by EMBL-EBI and INRA to provide access to plant genomic data.

The current version of the BioMart software in use is v0.7. A new version (v0.8) is expected by the end of 2010, supporting new a new (generic) user interface and improved tools for constructing instances of this customised for specific data. We will implement updated data warehouses using the BioMart v0.8 technology and provide public access to these.

Progress towards objectives and details for each tasks

. Task 1: Provision of a cloud compute environment

1. Programming environment

The proposed programming environment consists of a local infrastructure managed by a Cloud middleware (OpenNebula <http://www.opennebula.org> in this case), where COMP Superscalar (COMPSs <http://www.bsc.es/compss>) applications can be executed as well as common web applications. A diagram of this architecture is depicted in Figure 1 and a description of each component can be found in the following sections.

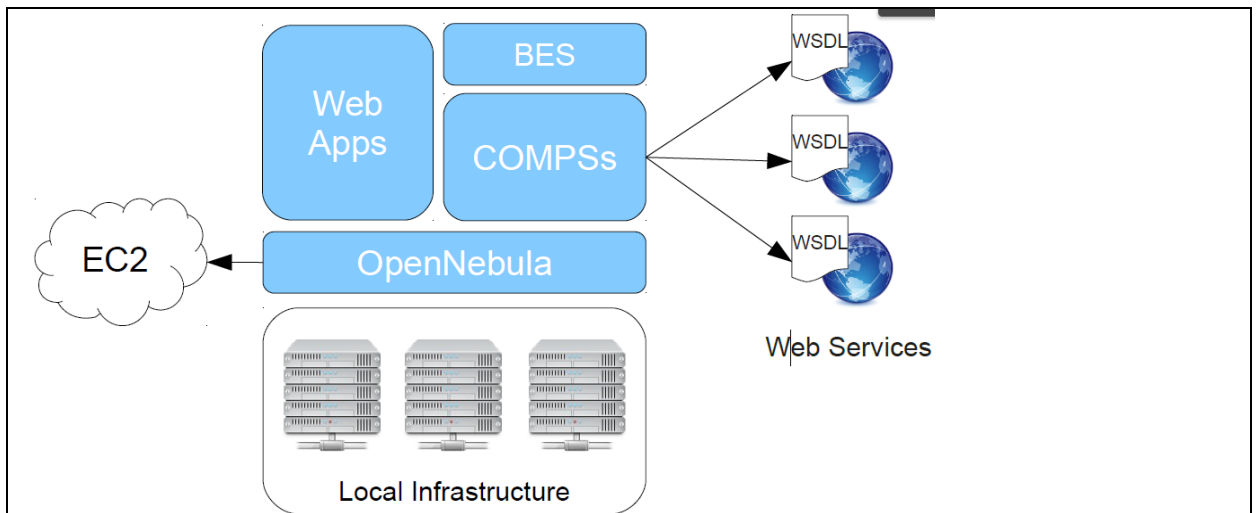


Figure 1 – Execution Environment

1.1. COMPS Superscalar

An application developed using COMPSs' programming model can contain both computation parts and Web Service requests. Once the implementation of the application has finished, it can be packaged and deployed in the system enabling its execution through a Web Service.

When a request arrives to the system, COMPSs' runtime asks for resources (virtual machines) to the underlying infrastructure and distributes the computational work and Web Service calls of the application amongst them ensuring that data dependencies are maintained.

COMPSs does that by following a master-worker architecture where the master launches tasks in the available resources, transmitting them the input files they may need and collecting the results afterwards. Since the input and output data of each task is specified in the source code of the application by using annotations, the runtime is aware of the precise moment the dependencies of a task are satisfied and thus it can be executed. It is worth to mention that the definition of tasks can include a specification of the needs of each task in terms of resources, and that COMPSs will create or destroy virtual machines as the computational load grows or decreases.

1.2. OpenNebula

OpenNebula is a middleware that enables an easy management of a private Cloud infrastructure. It is composed by a front-end that offers different interfaces, such as REST, XML-RPC and Web based, and that is able to manage virtual machines over multiple remote hosts running different hypervisors (e.g. Xen, <http://www.xen.org/>; KVM, http://www.linux-kvm.org/page/Main_Page; or VMWare, <http://www.vmware.com/>). Moreover, using OpenNebula as infrastructure provider also enables the use of Amazon EC2 (<http://aws.amazon.com/en/ec2/>) resources if the local infrastructure is not enough for running a specific application.

1.3. Basic Execution Service (BES)

The BES component is an implementation of a Basic Execution Service (<http://www.ogf.org/documents/GFD.108.pdf>), which has been designed to submit job execution requests to remote servers. It takes a Job Submission Description Language (JSDL) (<http://www.gridforum.org/documents/GFD.56.pdf>) document describing, amongst other information, an application's name and its input parameters, and starts the COMPSs' runtime

(master) in order to execute it. The runtime is also started in a virtual machine, so the BES needs to be able to interact with OpenNebula as well.

1.4. Web Applications

The environment also enables the deployment of already existing web applications in virtual machines instead of developing them with COMPSs.

2. Implementation and Testing

BES and COMPSs' runtime are based in connectors in order to interact with different Grid and Cloud platforms, so two new connectors have been developed to enable them to manage virtual machines in OpenNebula. The correct behaviour of these connectors has been tested in a closed environment where OpenNebula has been installed and configured to manage a Xen host. Also, a virtual machine image has been created, consisting in a plain Debian installation with COMPSs and some contextualization scripts to configure the network and remote access. This image can act both as master and worker, so it is valid for the BES and the runtime.

Some applications has been successfully executed in this testing scenario with a few number of virtual machines and the installation of the fully functional programming environment in a publicly available cluster is currently in process.

. Task 2: Implementation of HPC web services

Development if this Task will wait until the Cloud environment prepared in Task 1 has been fully tested. At that point, the necessary interfaces to include HPC facilities as part of the local environment will be prepared. It should be noted that interfaces of COMPS's to major BSC's computer architectures (large clusters or shared-memory systems) are already available. Applications to be implemented will come for those developed in Task 3 and also from the sequence assemblers being tested in project's WP12

. Task 3: Provision of web services for computational analysis

In the context of Tools for functional Genomics, StDLO has under development a webserver for sequence- and network based function prediction at genome-wide scale, implementing their previously described method (BMRF) and adding to its scope already sequenced crop species.

EMBL-EBI has provided DAS servers, providing access to 10 additional plant genomes through the Ensembl Plants interface during the first year of the transPLANT project:

<i>Brassica rapa</i> (turnip)	<i>Physcomitrella patens</i> (a moss)
<i>Chlamydomonas reinhardtii</i> (a red alga)	<i>Selaginella moellendorffii</i> (spikemosse)
<i>Cyanidioschyzon merolae</i> (a green alga)	<i>Setaria italica</i> (foxtail millet)
<i>Glycine max</i> (soybean)	<i>Solanum lycopersicum</i> (tomato)
<i>Oryza glaberrima</i> (African rice)	<i>Zea mays</i> (maize)

The availability of these DAS sources, that represents the achievement of the first milestone on transPLANT WP5 (MS12), has been published at <http://plants.ensembl.org/das/sources> and in the DAS registry (<http://www.dasregistry.org/listServers.jsp>).

. Task 4: Provision of interoperable data warehousing technologies

The implementation of updated data warehousing technologies has been scheduled for year 2 of the project. However, as initial task in this respect, preparatory to D5.1, INRA updated GnpIS data into their already existing Biomart 0.7 system. Data is available at: <http://urgi.versailles.inra.fr/biomart/martview/>.

Use of resources (*highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1*)

BSC 8.24 PM

EBI 0.5 PM

StDLO 1 PM

INRA 2 PM

TGAC: 0.07 PM

Work package number	6		Start date or starting event:			M1
Work package title	A virtual European Plant Genomics Database					
Activity Type	OTHER					
Participant number	1	2	3	6	10	11
Participant short name	EMBL-EBI	HMGU	GFMPG	IGR PAN	DLO	KN
Person-months per participant	15	12	2	8	4	2

Objectives

Maintain and provide public access to a virtual European Plant Genomics Database, offering a single point of entry to a distributed resource encompassing genomics, variation and phenotype.

Lead Beneficiary: EMBL-EBI

Description of work

Task 1: Provision of services for plant science researchers through a unified portal for plant genomics data

Objective: Provide an integrating portal for plant genomics data to the plant research community

Description: We will set up and run an integrated portal to offer trans-national access to the transPLANT data. The portal will be run by EMBL-EBI. We will maintain public access a registry of plant genomics services and data (developed in WP7) based on the activities of the project participants and other important resources. Access to genomic sequence will be offered through the well-established Ensembl platform that provides visualization and data mining services for genome scale data. Ensembl supports integration of genome-scale data such as variation, regulatory information, comparative genomics and the annotation of coding and non-coding genes. Ensembl also supports direct upload of user data and as a client for the integration of remotely held data served using the RESTful web service protocol DAS. Additional features for data visualization relevant to plant genomics will be developed in WPs 7 and 8 and included in the portal. We will maintain a high-availability service in which the key transPLANT data will be integrated, either directly or remotely, exploiting the services for data interoperability developed in work package 5 and elsewhere. An annual report will be delivered to the European Commission indicating usage, service availability, and the range of data and services integrated at each stage.

Task 2: Integrated Search Services

Objective: Provide trans-national access to new data search and information retrieval systems within the transPLANT portal.

Description: The transPLANT portal will be enhanced through the development of new integrating search services for genotypic and phenotypic information (based on the developments in WP10), and meta-data driven information retrieval (based on the developments in WP11), to enable seamless integration of the entire transPLANT data through a single point of entry.

Task 3: Trans-national access to a phenotypic data repository

Objective: Provide trans-national access to an international repository for phenotype data

Description: Like genomic or expression data, phenotypic data must be stored and kept available on the long term. Therefore, a phenotype data repository must be designed and built. The Ephesis project has been initiated three years ago to store phenotype and environment data in a dedicated module of the URGI information system, GnpIS. This information system is designed to handle multispecies data, from annual crop to trees. It is highly generic and its data model has been inspired by other generic systems such as the Chado database developed by the GMOD consortium, the Genomic Diversity and Phenotype Data Model, and the International Crop Information System (ICIS). INRA scientists are involved in many collaborative European projects, which use Ephesis to serve as a repository for the data of the INRA and its partners. Under transPLANT, access to Ephesis will be made available to the international community to serve as a repository for phenotypic data.

Task 4: Services for the enablement of virtual plant breeding

Objective: Provide trans-national access to a problem-solving environment for plant breeders and translational science

Description: Tools and workflows that will enable the structured use of plant genome and phenotypic data in the design of breeding programs will be developed in WP12. In the current task we will work on the embedding of these computational modules in a e-science infrastructure for virtual plant breeding (IVPB) in accordance with standards and technologies established in WP3. The aim is to build problem-solving environments in which all required modules are available and reusable in a coherent manner for the design and execution of specific experiments that exploit data on genomes, phenotypes, variation and markers for breeding purposes.

Progress towards objectives and details for each tasks

Task 1: Provision of services for plant science researchers through a unified portal for plant genomics data

EBI, as project lead, has developed a website for the transPLANT project (<http://www.transplantdb.eu>). This website provides access to information about the project, but also about the project partners and the other relevant activities (e.g. other nationally or internationally coordinated plant genomics projects, training activities, etc.) with which the project partners are engaged. The website will also host the integrated search services being developed as task 2 (development version accessible at <http://test.transplantdb.eu/ext/search>), which in turn provide access to information about searched-for biological entities in all of the interactive services maintained and developed by the transPLANT project partners, for example, Ensembl Plants at EBI, PlantsDB at HGMU, etc. The website will additionally contain a mirror of the registry of plant genomic information being developed in work package 7 (development

version accessible at <http://test.transplantdb.eu/resource-instances>). Over the course of the project, the website will further develop to function as an interactive hub for all resources (data and analysis tools) in development by the transPLANT partners.

The website has been implemented using Drupal, an industry-standard content management system based on a PHP framework and a MySQL relational database, and deployed at the EBI's London Data Centres, which provide a robust, high availability, replicated environment from which many resources run by the EBI are exposed to external users.

Some pages from the transPLANT site are shown below for illustration: the project homepage (<http://transplantdb.eu>); and the list of plant-related services (<http://transplantdb.eu/resources>).


trans-National Infrastructure for Plant Genomic Science


[Home](#)
[About](#)
[Resources](#)
[Publications](#)
[Meetings](#)
[Partners](#)



Resources

Find descriptions of the various plant-based [databases](#), [services](#), and [software](#) maintained by the transplant partners.



The falling cost of nucleotide sequencing is opening up significant opportunities for crop improvement through plant breeding and increased understanding of plant biology.

Many plant genomes are large and have complex evolutionary histories, making their analysis theoretically challenging and highly demanding of computational resources. Issues include genome size, polyploidy, and the quantity, diversity and dispersed nature of data in need of integration.

transPLANT is a consortium of 11 European partners gathered to address these challenges and to develop a trans-national infrastructure for plant genomic science. Bringing together groups with strengths in data analysis, plant science, computer science, and from the academic and commercial sectors, transPLANT will develop integrated standards and services and undertake new research and development needed to capitalise on the sequencing revolution, across the spectrum of agricultural and model plant species.

transPLANT is committed to establishing the broadest international collaborations for data and standards. Explore the project's aims in more detail on this website, or contact us on transplant_help@ebi.ac.uk.

Meetings

→ Event added Thursday, September 6, 2012 - 16:25

transPLANT user training workshop Monday, November 12, 2012 to Tuesday, November 13, 2012. INRA URGI campus, Versailles, France

→ Event added Tuesday, July 3, 2012 - 08:31

Training workshop in plant pathogenic genomics Wednesday, September 19, 2012 to Thursday, September 20, 2012. EMBL-EBI, Hinxton, Cambridge, CB10 1SD, UK

News

→ Article added Thursday, September 6, 2012 - 16:40

Ensembl Plants, Release 15

- We have added a set of homoeologous SNPs between wheat A, B and D genomes using wheat contigs aligned to *Brachypodium distachyon* as a reference framework, a structural variation dataset for *Sorghum bicolor* has been imported from dGVA, and a variety of small improvements to our assembly, annotation and variation datasets have been incorporated. See the [Ensembl Plants homepage](#) for details.

→ Publication added Thursday, July 5, 2012 - 14:53

IDPredictor: predict database links in biomedical database Mehlhorn H, *Journal of Integrative Bioinformatics*, 2012

[PubMed](#) | [DOI](#)

Ensembl Plants

<http://plants.ensembl.org>


The Ensembl Genomes project produces genome browsers for important species from across the taxonomic range, using the Ensembl software system. Five sites are now available: Ensembl Bacteria, Ensembl Fungi, Ensembl Metazoa, Ensembl Plants, and Ensembl Protists.

Ensembl Plants includes reference genome assemblies and gene builds for 15 plant species, including functional, comparative, and variation data for important crop species.

Genetic and Genomic Information System

<http://urgi.versailles.inra.fr/gnpis/>


GnpIS is a powerful multispecies centralized database information system dedicated to plant, trees and their bioagressors (fungi). It is composed of a set of relational databases, each optimized according to one data domain and connected together to be search as a whole or by theme. Its originality is to bridge genetic and genomic data, allowing researchers and breeders to cross genetic information (i.e Genetic maps, QTL, markers, SNPs, Germplasms, Genotypes) with genomic data (i.e. physical maps, genome annotation, expression data) for species of agronomical interest. One key feature of the system, is the availability of two tools allowing to query through all these databases simultaneously, a quick search tool based on Lucene technologies and an advanced search tool based on Biomart software (GMOD). The system is generic for any kind of species.

IPK WebBlast

<http://webblast.ipk-gatersleben.de/barley/>


This BLAST-service offer accessing to the Hordeum vulgare physical map and genomic sequence data that where released in The international Barley Sequencing Consortium.

Integrated GBrowse for four crops: Maize, Brassica rapa, Rice, and Soybean

<http://www.keygene.com/>


Later this year we will produce a GBrowse with integrated sequence based data for four crops:

- * Maize
- * Brassica rapa
- * Rice
- * Soybean

This is deliverable D8.1 with due date end of August 2012.

MIPS Webservices

<http://mips.helmholtz-muenchen.de/plant/static/gen/newArchitecture.html>


MIPS Website

<http://mips.helmholtz-muenchen.de/plant/genomes.jsp>


Task 2: Integrated Search Services

The integrated search functionality of the transPLANT web-portal allows users to search for resources provided by all partners. Resources include information on genes, transcripts, markers, phenotypes, and metabolic reactions (Table 1). Search results from all partner databases are returned in a consistent, tabular format, ranked by relevance to the search term used. Each result is accompanied by a URL linking the user back to the appropriate source information. The list of results matching a search term can be further filtered by several search facets covering the data source, data type, and species. Multiple facets can be added or removed dynamically, allowing the user to explore the available information before selecting one or more links to retrieve further details.

Case study to highlight the above functionality: A search for "carbamoyl synthase" currently returns 14,217 results. The first 10 most relevant results are displayed to the user, with paging functionality above and below the results table. The user can select a resource of interest from the data source facet. After selecting Ensembl Plants, for example, the user can see the number of results per species update in the species facet, and can select a species of interest. Finally the user can review and select the different data types matching the search, given the current filters, can select one and, in this case, can click on the link to view the information in the Ensembl

browser.

Technical description: Search was implemented using the Solr (v4.0) text search framework. A simple schema was devised to capture information from the range of partner resources available, with most resource specific information going into a free-text description field. Additional fields were defined to capture the associated resource identifier, URL, and species information. One meta-data field records the partner database from where the associated information was obtained.

Data was collected from partners in a simple, tabular format, and was loaded into the Solr search index using the schema described. The free text field was processed using tokenization and filtering rules optimized for English, allowing stemming and wild card search. Faceted fields were indexed separately to facilitate faceted searching. All fields, with the exception of URL, were copied to the description field index so that by default, searches for text in these fields would return the expected results.

The total set of data was indexed by the Solr server in under 10 minutes.

Table of partner resources indexed:

Partner	Database	Data types	No. data points	No. species
EBI	Ensembl Plants	Gene-centric	778614	19
MIPS	PlantsDB	Transcript-centric	447568	6
IPK	CR-EST	ESTs	258697	6
	GEBIS	passport data	201672	50
	MetaCrop	Biochemical reaction	348	50
PAS	PolapgenDB	Phenotype-centric	69	<i>Vitis vinifera</i>
URGI	GnpGenome	Genes, variations, and markers	641525	<i>Hordeum vulgare</i>
	Siregal and GnpMap	germplasm and markers	43624	To be added



Drupal integration: The Solr search index is queried and the results presented within the Drupal powered transPLANT web-portal. Drupal has a modular plug-in system, and an existing module for linking Drupal with Solr was adapted for these purposes. The module depends on several APIs that are themselves provided by other modules, as described below. The module we developed functions to:

- 1) Create a search page using the search API that uses the appropriate connection details of the Solr server,
- 2) Register facets with the facet API, providing the facet blocks of the search page, and
- 3) Process and display search results on the search page, adding paging and formatting elements.

On-going integration strategy: Two different strategies can be employed to keep the search up to date. We can continue to accept static dumps of database information from partners, periodically updating the central index with this information. Alternatively, partners could establish their own Solr instances, using the common schema but indexing specific data. In this system the Drupal 'client' would issue a distributed search across several servers, integrating the results in the presentation layer.

Current status: The search functionality has been implemented and exists in the transPLANT test site (<http://transplantdb.eu/ext/search>). Roll-out to the production site is expected within

the next two months. A screenshot from the development version of the search site is shown below.


trans-National Infrastructure for Plant Genomic Science


Home

About

Resources

Publications

Meetings

Partners

transPLANT search

Enter terms

Provider	Type	ID	Species	Description
PlantsDB	transcript	AT2G39730.1	Arabidopsis thaliana	rubisco activase - Rubisco activase, a nuclear-encoded chloroplast protein that consists of two isoforms arising from alternative splicing in most plants. Required for the light activation of rubisco.
PlantsDB	transcript	AT2G39730.2	Arabidopsis thaliana	rubisco activase - Rubisco activase, a nuclear-encoded chloroplast protein that consists of two isoforms arising from alternative splicing in most plants. Required for the light activation of rubisco.
PlantsDB	transcript	AT2G39730.3	Arabidopsis thaliana	rubisco activase - Rubisco activase, a nuclear-encoded chloroplast protein that consists of two isoforms arising from alternative splicing in most plants. Required for the light activation of rubisco.
PlantsDB	transcript	Sb03g031190.1	Sorghum bicolor	similar to Rubisco subunit binding-protein beta subunit-like - ID=Sb03g031190;Description="similar to Rubisco subunit binding-protein beta subunit-like"
CR-EST	expressed sequence tags	HY01E07u	Hordeum vulgare	gi 1185390 gb AAA87731.1 alphacpn60 precursor [Pisum sativum] RuBisCO subunit binding-protein alpha; gi 1345582 emb CAA30699.1 unnamed protein product [Triticum aestivum] RuBisCO subunit binding-prote; gi 1351030 sp P21239 RUB1_BRANA RuBisCO subunit binding-protein alpha subunit, chloroplast precursor; gi 21554572 gb AAM63618.1 putative rubisco subunit binding-protein alpha subunit [Arabidopsis thali; gi 31711734 gb AAP68223.1 At2g28000 [Arabidopsis thaliana] putative rubisco subunit binding-protein; gi 3790441 gb AAC68501.1 chaperonin 60 alpha subunit [Canavalia lineata];

Current search

Search found 946 items

- rubisco

Filter by provider:

- CR-EST (864)
- PlantsDB (45)
- Ensembl Plants (37)

Filter by resource category:

- expressed sequence tags (864)
- transcript (45)
- protein_coding (37)

Filter by species:

- Hordeum vulgare (835)
- Arabidopsis thaliana (27)
- Oryza sativa (17)
- Brassica rapa (13)
- Solanum tuberosum (12)
- Pisum sativum (9)
- Nicotiana tabacum (8)
- Zea mays (7)
- Sorghum bicolor (5)
- Chlamydomonas reinhardtii (4)

[Show more](#)

. Task 3: Trans-national access to a phenotypic data repository

Access to the phenotypic data repository will be provided by INRA in year 2 of the project, and comprises part of project **milestone MS16**.

. Task 4: Services for the enablement of virtual plant breeding

This work will be done by DLO in year 4 of the project, following on from the work developing the underlying technology in WP12.

Use of resources (highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1)

EMBL-EBI 3.5 PM

Work package number	7	Start date or starting event:	M1
Work package title	A repository for reference genome and annotation		
Activity Type	RTD		
Participant number	1	2	5
Participant short name	EMBL-EBI	HMGU	INRA
Person-months per participant	15	32	9

Objectives

Develop a repository of reference sequence and annotation for the genomes of important plant species.

Lead Beneficiary: HMGU

Description of work

Task 1: A registry of plant genome sequences and resources

Objective: Develop a registry of plant genome sequences and resources.

Description: We will develop a registry of plant genome sequences and resources linking to these. The number of plant genome sequencing and (re-sequencing) projects is increasing; and for many of the more complex cereal genomes, progress towards the complete assembly of a reference genome is incremental, with a variety of projects producing genome, transcriptome and marker-based data from a range of technologies. Drawing on the broad expertise of the consortium, we will maintain a registry of important sequence-based resources for species of agricultural and economic importance as well as model systems. More specifically model genomes such as *Arabidopsis thaliana*, *A. lyrata*, *Medicago* and *Brachypodium* will be included as well as the genomes of tomato, potato, wine, cucumber, maize and apple as well as the extremely challenging and complex grass and *Triticeae* genomes of wheat, barley and rye.

Progress towards objectives and details for each tasks

Task 1: A registry of plant genome sequences and resources

The Registry

We collected repository data for publicly available plant genome database systems making use of both existing compilations and a de-novo search of relevant websites and

systems. For instance, we imported all relevant data from the plant genome website collection at http://www.phytozome.net/Phytozome_resources.php and performed extensive web searches for registering both species-specific and multi-species plant genome resources maintained by both transPLANT and non-transPLANT partners.

A total of 187 distinct plant genome resources is registered at the transPLANT data registry at this time. We collected and structured the following information entities from the identified plant genome resources:

ID: this will be the internal database ID, incremental
Provider_Shortname: short name of the data providers institution (such as research center, university...)

Provider_Details: details (such as full name) of the data providers institution

Resource_Shortname: short name of the particular resource provided...this will be a system name such as *Phytozome*

Resource_Details: details (such as full name, scope) of the data resource

Instance_Name: this will describe a particular instance of the resource, such as the database for a single species within the system (such as the *Arab. thaliana* instance within *Phytozome*)

Instance_Description: details (such as full name) of the instance

Species_Name: scientific species name...this can contain more than one entry

Species_Commonname: common species name

URL: primary URL of the particular instance (should point to the entry page of the instance such as to the *Ara. thaliana* entry page in *Phytozome*)

Version_release_name: this field can hold a release or version tag for the data within an instance...often this will be an annotation or assembly version, such as *TAIR10*

source_name_URL: this field can hold a source URL for the data within an instance...often the data within an instance are derivative and were obtained from another primary resource...this URL can then point towards the original data resource

last_updated_or_release: this field can hold a release version or last update data for the instance OR/AND resource...many resources are built in releases (such as *TAIR* or *Phytozome* in version 8.0) or updated regularly

data_type: this field describes the type of data that is stored or provided from an instance...such as "genomic", "variation" or "expression"

tools: this field can hold the names of useful tools provided within the instance or resources such as *GBrowse*

keywords: keywords to be indexed for search

All registry data was stored in a relational database system and provided to the transPLANT partners in Excel format for proof-reading and adding data. The registry is hosted, maintained and updated at HMGU (<http://mips.helmholtz-muenchen.de/plant/transplant/index.jsp>) but is also fully accessible for search, query and linking (e.g. using cross-references from other data entities produced or integrated within transPLANT) from the official transPLANT website hosted by EBI: <http://transplantdb.eu/>.

To ensure both registries are synchronous we exchange updates and changes to the master registry monthly.

In the near future, changes and updates to the registry can be performed by database providers themselves, lowering the maintaining cost and ensuring expert-curated and -driven information. For that, a web interface hosted by HMGU will be provided with secure access to the transPLANT genomic resources registry.

Ensembl Plants

EMBL-EBI is providing access to many of the genomes described in the registry available for interactive and programmatic analysis through the Ensembl Plants (<http://plants.ensembl.org>) site. Ensembl, originally developed in the course of the

Human Genome Project but subsequently applied to other domains, is a powerful tool suite for the analysis and display of genome scale data, and Ensembl Plants is the EBI's primary user interface for accessing plant data. We have used transPLANT funding to increase our capacity to include additional reference genomes incorporated in Ensembl Plants. In the first year of the grant, we have made 5 releases of Ensembl Plants, and incorporated the following additional genomes: *Brassica rapa* (turnip), *Chlamydomonas reinhardtii* (a red alga), *Glycine max* (soybean), *Cyanidioschyzon merolae* (a green alga), *Physcomitrella patens* (a moss), *Selaginella moellendorffii* (spikemoss), *Setaria italica* (foxtail millet), *Solanum lycopersicum* (tomato), and *Zea mays* (maize). These genomes have been analysed comparatively using the Ensembl Compara functional genomics pipeline, which has 2 elements: a protein-based analysis, which infers evolutionary relationships after clustering and alignment (and which are performed over the domain of all plants) and a pairwise DNA-based analysis, performed using the alignment tools blastZ and lastZ. The inclusion of plants *sensu lato* – i.e. including both the green and the red algae – within Ensembl Plants has been undertaken specifically to support the comparative aspects. The available analyses are given in the table below.

Genome 1	Genome 2	Method
<i>Arabidopsis thaliana</i>	<i>Arabidopsis lyrata</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Brachypodium distachyon</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Brassica rapa</i>	lastz
<i>Arabidopsis thaliana</i>	<i>Chlamydomonas reinhardtii</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Cyanidioschyzon merolae</i>	lastz
<i>Arabidopsis thaliana</i>	<i>Glycine max</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Oryza brachyantha</i>	lastz
<i>Arabidopsis thaliana</i>	<i>Oryza glaberrima</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Oryza indica</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Oryza sativa</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Physcomitrella patens</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Populus trichocarpa</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Selaginella moellendorffii</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Setaria italica</i>	lastz
<i>Arabidopsis thaliana</i>	<i>Solanum lycopersicum</i>	lastz
<i>Arabidopsis thaliana</i>	<i>Sorghum bicolor</i>	blastz
<i>Arabidopsis thaliana</i>	<i>Vitis vinifera</i>	blastz
<i>Oryza sativa</i>	<i>Arabidopsis lyrata</i>	blastz
<i>Oryza sativa</i>	<i>Arabidopsis thaliana</i>	blastz
<i>Oryza sativa</i>	<i>Brachypodium distachyon</i>	blastz
<i>Oryza sativa</i>	<i>Brassica rapa</i>	lastz
<i>Oryza sativa</i>	<i>Chlamydomonas reinhardtii</i>	blastz
<i>Oryza sativa</i>	<i>Cyanidioschyzon merolae</i>	lastz
<i>Oryza sativa</i>	<i>Glycine max</i>	blastz
<i>Oryza sativa</i>	<i>Oryza brachyantha</i>	lastz
<i>Oryza sativa</i>	<i>Oryza glaberrima</i>	blastz
<i>Oryza sativa</i>	<i>Oryza indica</i>	blastz
<i>Oryza sativa</i>	<i>Physcomitrella patens</i>	blastz
<i>Oryza sativa</i>	<i>Populus trichocarpa</i>	blastz
<i>Oryza sativa</i>	<i>Selaginella moellendorffii</i>	blastz
<i>Oryza sativa</i>	<i>Setaria italica</i>	lastz
<i>Oryza sativa</i>	<i>Solanum lycopersicum</i>	lastz
<i>Oryza sativa</i>	<i>Sorghum bicolor</i>	blastz
<i>Oryza sativa</i>	<i>Vitis vinifera</i>	blastz
<i>Oryza sativa</i>	<i>Zea mays</i>	blastz
<i>Oryza indica</i>	<i>Brachypodium distachyon</i>	blastz

Zea mays

Sorghum bicolor

blastz

The data are available through the Ensembl Plants user interface, and will be searchable through the integrated search facility in development for the transPLANT website (see work package 6 report).

Future Developments

Several transPLANT partners also maintain genome-centric resources, each with their own strengths in terms of data types and visualization interfaces, reflecting each organization's own domains of expertise. Over the course of the project, a primary objective is to move forward from the initial concept of the registry to keep a clear track of the versions of sequence assemblies in use for various resources, to support the mapping of positional feature data between successive versions of genomes, and to thereby standardize the use of reference data among the project partners – i.e. the different transPLANT partners will use the same versions of reference genome sequences where possible; and where not, the versions will be well defined (enabling users to identify where data is directly interoperable, and where not); and where not, tools will support the propagation of features to facilitate interoperability as far as possible. We will commence work towards these goals in year 2 of the project.

Use of resources (*highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1*)

HMGU: 4.5 PM

EBI: 3.5 PM

Work package number	8		Start date or starting event: M1			
Work package title	An infrastructure for handling plant genomic complexity					
Activity Type	RTD					
Participant number	1	2	5	7	8	11
Participant short name	EMBL-EBI	HMGU	INRA	BIOGEM	TGAC	KN
Person-months per participant	2	21	12	6	10	12

Objectives

This WP is devoted to the development of tools for the identification and the handling of specific features of plant genomic data of importance for its use in experimental applications. We will consider both inherent biological features, such as duplications and polyploidies, and also the requirements of experimental strategies for integration of genetic markers, sequence associated data, and sequence variation data, on a framework of genome sequence.

Lead Beneficiary: INRA

Description of work

Task 1 Integration of genetic markers and sequence associated data on genomic sequences

Objective: We will integrate and visualize on genomic sequences sequence associated data such as sequence based genetic markers and/or physical contig data (BACs anchored by individual sequence anchors).

Description: These data are notoriously detached from genome backbones. We aim to overcome this by collecting and associating the diverse datasets with the genomes and make the associations electronically available and integrated into visualization interfaces. This action will ask for close collaboration with the respective national and trans-national sequencing consortia to obtain key data for each species.

Task 2 Comparative genomics of plant, visualization of ancient duplications and polyploidy

Objective: Plant genomes are inherently complex and in contrast to vertebrate genomes in almost all cases have undergone polyploidisation and genome rearrangements during their recent evolutionary past. We aim to analyse for syntenic relationships and intragenomic duplications and rearrangements using established analysis software. The results will be integrated and populated through database platforms to make this information usable to the broader user community.

Description: Ancient and recent polyploidisation events are scientifically interesting features of

plant genomes. Numerous evolutionary and functional open questions are associated with these features. To fully understand the evolutionary history of individual plant genomes and to exploit homologous and paralogous relationships between genes, a full understanding of evolutionary relationships among closely and more distantly related plant genomes, is necessary. We will analyse syntenic relationships, intragenomic duplications, and other rearrangements using established analysis software (e.g. blastz/multiz, Mavid, Mauve, DAG chainer). Information gained on genome scale analysis will be made usable on the level of smaller genomic segments or even individual genes. To assist in the analysis of these features we will analyse the dynamics of retention of duplicated genes and genome segments and, upon availability, for sub-/neofunctionalisation of duplicated genes using transcriptional data as a proxy. The results will be integrated into the transPLANT services and databases to make this information available to the broader user community.

Task 3 Resolving the conceptual and practical implications of pan-genomics

Objective: Recent advances in sequencing technologies allow today to sequence at a reasonable cost several plant genomes of the same species in order to describe their pan-genome. The “Pan-genome” is a concept describing the full complement of sequences in a species i.e. a superset of all the sequences in all the strains of a species. The pan-genome can be subdivided into the "core genome" containing sequences present in all strains, a "dispensable genome" containing those present in two or more strains, and finally "unique sequences" specific to single strains. This task will be focused on the identification of sequence variations, their storage in databases and their display in genome browser.

Description: The re-sequencing of several individuals belonging to the same species has revealed a high level of sequence variations. These sequence variations are supposed to be at the origin of phenotypic variations among them. Plants of agronomical interest are organized around strains. These strains, also called accessions, have these polymorphisms fixed such as almost no difference is observed between individuals of the same strain at the genetic and phenotypic level. They are the starting material of most crop improvement programs. Important genomic programs aim at inventorying these sequence variations in order to link them to the observed phenotypes. Single nucleotide polymorphisms (SNPs), presence/absence variations (PAVs), and copy number variations (CNVs) identification, is today the primary goal of study trying to understand strains agronomical performances.

In order to identify these sequence variations, we will need, first, to assemble genomes from short reads using reference sequences when available. Then, the resulting sequences have to be multi-aligned, and all observed sequence variations extracted from these alignments. As multiple tools and strategies are possible when considering these problems, sharing the experiences among the partners of the project will be an important goal of this task. The main deliverables of this task will be specifications of important data to collect and optimal strategies to obtain them. Collaboration between platforms for data exchanges will be also strongly encouraged.

Task 4: Implementation of a pan-genome browser

Objective: Provide database schemas for sequence variation storage and paradigms for their visualization.

Description: We will search for a smart use of the already existing solution able to display sequence variations among several strains. We will pragmatically test various solutions based on tools such as Ensembl, GBrowse (over Chado or Bioseq::feature), or other genome browsers. We will explore how we can configure these tools to display this information. Performance and readability of the solution will be an important goal. The use of DAS will support the integration

of species-targeted solutions developed at different sites.

Progress towards objectives and details for each tasks

. Task 1 Integration of genetic markers and sequence associated data on genomic sequences

The goal of this task is to link genome sequences with all types of sequence-associated data such as genetic markers and other sequence based data. After some discussion, four species were identified to serve as role models to implement this integration: Rice, Maize, Soybean and *Brassica rapa*. In a next step, an inventory of public resources was made and data was collected for these four species. These include existing genome browsers for the four species, at multiple locations, the sequence databases (EMBL, NCBI) and other types of public data, e.g. for genetic markers and maps.

In the next step, data was combined and checked on nomenclature. If possible, corresponding entries were identified and discrepancies were eliminated.

Data are visualized in GBrowse and consist of genomic sequence data, mostly on a pseudo-chromosome level, gene information (orientation, splicing, function), annotated information such as transposons and restriction sites, clone information (BACs) and genetic markers and maps. The GBrowises are hosted on a KeyGene server and a link to this server is provided on the transPLANT website.

HMGU was integrating genetic markers and physical contig data for the complex *triticeae* organism *Hordeum vulgare* (barley). Different anchoring strategies were used and combined to obtain a high-resolution ordered gene map. All data can be downloaded from <http://mips.helmholtz-muenchen.de/plant/barley/index.jsp> and maps are visualized in GBrowse and CrowsNest (see Task 2).

. Task 2 Comparative genomics of plant, visualization of ancient duplications and polyploidy

This task will analyse syntenic relationships, intragenomic duplications, and other rearrangements.

INRA has improved the code of an existing pipeline able to detect intragenomic duplication (Fiston *et al.* Genome Res. 2007 Oct;17(10):1458-70. Epub 2007 Aug 28). The new prototype is currently under test.

To analyze syntenic relationships, rearrangements and intragenomic duplications HMGU enhances the CrowsNest tool (<http://mips.helmholtz-muenchen.de/plant/crowsNest/index.jsp>).

CrowsNest is a whole genome interactive comparative mapping and visualization tool comparing genetic, physical and hierarchical (fingerprinted contigs) maps in the plant kingdom. CrowsNest is specifically designed to visualize synteny at macro and micro levels. It allows to intuitively exploring rearrangements, inversions, and deletions at different resolutions, to transfer knowledge about function and conservation between several plant species and to derive evolutionary information.

CrowsNest consists of two main integrated parts: 1) the web-based user interface with the integrated comparative visualization tool, and 2) the comparative analysis pipeline.

The pipeline was designed to perform a variety of different tasks: i) anchoring unfinished genome to a model reference genome, ii) determining orthologs and paralogs, iii) calculating conserved synteny, and iv) calculating features such as syntenic quality index and dS/dN ratios.

CrowsNest currently harbours data from the model grass organisms *Brachypodium distachyon*,

Sorghum bicolor and *Oryza sativa* (rice) as well as from the crop plant *Hordeum vulgare* (barley). We are working on the integration of new model and crop plant species as well as adding functionality to the analysis pipeline and improving the code, evaluating the results and defining ways to share these results with transplant partners.

. Task 3 Resolving the conceptual and practical implications of pan-genomics

This task is focused on the identification of sequence variations and their storage in databases. The main deliverables will be specifications of important data to collect and optimal strategies to obtain them.

These objectives overlap with activities in work package 12, in particular the deliverable 12.1 that refers to the “Development of sophisticated statistical methods to model variation in large plant genomes”. We organized a joint WP8/WP12 meeting to obtain a work plan for deliverable 12.1 and to treat the overlaps with the activities in this task. From the discussions, it comes out that we will consider three kind of variation according to their size. Small-size variations such as SNPs or small indels (*e.g.* 1-30bp), medium-size variations such as small inserts or deletions of 30 to 100bp, and large-size variations when larger than 100bp. For each of these sequence variation classes, we will adopt different strategies combining different tools in a pipeline. The tools used will be tested for their performances in WP12.1. The size of the variation that these tools are able to efficiently detect will be an important outcome, and will allow refining our three range size classes. The pipelines will be implemented as a part of WP12 task 1.

Several tools already exist for identifying small-size variations. In particular, INRA partner has developed previously a novel “Mapping Analysis Pipeline for High-Throughput Sequences” (MAPHiTS: <http://urgi.versailles.inra.fr/Tools/MAPHITS>). The pipeline allows the detection of SNPs and small indels by comparing high-throughput Illumina short-reads (GAIIx or HighSeq) with a reference sequence from the same or a different species. This pipeline is based on public softwares (BWA, Bowtie, SAMtools, VarScan and Tablet) and homemade tools. It can filter out the called SNPs according to genome coverage, allele frequency, pValue, and SNP positions in the read, and run in parallel on a computer cluster.

Other software such as Pindel (Ye *et al.* Bioinformatics. 2009 Nov 1;25(21):2865-71. Epub 2009 Jun 26.), BreakDancer (Chen *et al.* Nat Methods. 2009 Sep;6(9):677-81. Epub 2009 Aug 9.) SVDetect (Zeitouni *et al.*, Bioinformatics. 2010 26 (15): 1895-1896.) or SECluster (Wong *et al.*, *Genome Biology* 2010, **11**:R128) are already available to detect medium and large-size sequence variations. They are currently under evaluation by Biogemma partner on maize sequences.

Another approach for large sequence variations, such as large inserts, consists in assembling the re-sequenced genomes from their short reads using reference sequences when available. The Columbus module from the Velvet package is available for this task. The resulting assembled sequences have then to be aligned on the references, and all observed sequence variations extracted from these alignments. We tested BLAT, Lastz and Mummer3 as alignment software and choose Mummer3 as it is the fastest, gives good results, and has tools to detect variations in its alignments.

. Task 4: Implementation of a pan-genome browser

This task will search for a smart use of the already existing solution able to display sequence variations among several strains.

INRA tested GBrowse over Bioseq::feature to display small-size sequence variation obtained on grapevine accessions. The result appears satisfactory for that class of sequence variation. They are displayed on the URGI genome browser at:

http://urgi.versailles.inra.fr/gb2/gbrowse/vitis_12x_pub/

EBI have been working on extending the Ensembl schema and API to support the representation of Pan-Genomes. This requires the ability to support multiple assemblies in one database, and the ability to re-use certain sequences in multiple assemblies. A set of visual user interfaces will allow users to see the structural organization of the Pan Genome, the presence/absence of particular genes/genomic segments in individual lines, and sequence variation within conserved segments, while still allowing users to browse through individual genomes. Initial work has been focused on schema and API design, and testing the system with bacterial genomes. Future work will involve interface development, and optimization for improved performance necessary to support large(r) plant genomes.

TGAC is also exploring strategies to deploy a genome browser with extended client capabilities. The effort has been into proving an intuitive interface and a client engine that can implement fast data transfer. The first prototype for this browser is available at:

<http://tgac-browser.tgac.ac.uk>

If applicable, explain the reasons for deviations from Annex I and their impact on other tasks as well as on available resources and planning

No deviations from Annex I on available resources and planning.

Use of resources (*highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1*)

HMGU: 2PM

INRA: 3PM

BIOGEMMA: 2.94PM

KEYGENE: 4.22PM

EBI: 0.5PM

TGAC: 0.57PM

Work package number	9		Start date or starting event:	M1
Work package title	An archive of plant genomic variation			
Activity Type	RTD			
Participant number	1	3	7	8
Participant short name	EMBL-EBI	GFMPG	BIOGEM	TGAC
Person-months per participant	42	3	6	10

Objectives

Develop an archive of plant genomic variation.

Lead Beneficiary: EMBL-EBI

Description of work

Objective: Develop a distributed infrastructure for handling of plant genomic variation, including software for submission, archiving, exchange and update

Description: transPLANT will develop an distributed archive of plant genomic variation (single nucleotide polymorphisms (SNPs), insertion/deletion events (indels), copy number variants (CNVs), plugging a critical gap in the infrastructure of the plant science community. This new resource will supplement but not overlap with existing repositories (such as the U.S. National Center for Biotechnology Information's resource dbSNP, which is currently the leading resource for archiving of SNP data but which has a clear medical focus, and other resources for CNVs and other structural variants). An infrastructure will be developed whereby domain-specific repositories can assemble and gather information related to particular projects and broker submission of mature data to a central archive for subsequent perpetual archiving.

Owing to the large size of these data, we will implement a distributed solution that allows sharing of localized information between remote nodes based on the use of common technology and accepted reference genomes. This model will also support collaboration with dbSNP and other international collaborators. In outline, the model proposed is as follows. A central hub interacts with a number of distributed nodes, using agreed reference sequence as the currency for two-way data exchange. Local repositories accept submissions through a submissions interface, and communicate with the hub using through a client - server model. Assignment of non-redundant identifiers, and projection of variations between assembly versions, is performed within the hub. A pre-existing repository can communicate with the hub through the implementation of the client interface (B) as a layer on top of its own existing interface. Code will also be implemented to support data exchange with key international collaborators via a flat-file format (F).

Progress towards objectives and details for each tasks

Task 1 Development of an archive of plant genomic variation.

The goal of the system in development is to provide an infrastructure for managing the onslaught of variation data now being produced for many plant species. The key challenges in managing this data relate to the fact that its meaning is provided through its positioning on a reference sequence; but as reference sequences are updated, so existing data needs to be migrated forward to be seen in the context of the latest reference and annotation. The solution in development rests on the idea that variations - single nucleotide polymorphisms (SNPs) and insertion/deletion events ("indels") - are positional features, and can be projected from one genome assembly to another provided the genome assemblies as a whole have been mapped against each other. As mapping a few chromosomes is a simpler task than individually aligning millions of variant together with their associated flanking sequence, which is the classical method of finding the location of variants in a new sequence, a much more computationally effective approach becomes possible.

The development of such a system requires the development of a number of components:

1. An agreed set of meta data to describe relevant parameters of an experiment.
2. An agreed data exchange format for the submission and release of data.
3. Submission (and data verification) system, to capture data and (appropriate) meta data.
4. An archiving system, to provide persistent storage and document-level retrieval of both data and meta data.
5. A system to map between locations in different versions of the same genome sequence, enabling.
6. A system to project positional features from one version of a genome sequence to another.
7. A local data store to hold the data needed during the processing, and to store derived data resulting from the projection of originally submitted data onto future assemblies.
8. A system for merging the results of various submissions on the same reference assembly, and for assigning identifiers to variation loci.
9. A persistent store of the mappings between sequences, for purposes of data authentication and allowing users to update features from outside the system.
10. A tool for exporting data into the Ensembl Plants variation schema, which will be used as the primary point of access for this data.
11. A tool for exchanging variation managed in the infrastructure with the main potential international collaborators, dbSNP at NCBI.

Before commencing development, a set of specifications was drawn up, based on consultations with U.S. collaborators at the Gramene resource, and other resources outside the plant genomics domain with which EMBL-EBI is involved, which face similar needs (e.g. WormBase, VectorBase, etc.).

To date, work has focused mainly on points 1-8. Work on points 8 is ongoing, work on points 9-10 is still to be done. The relationship of EBI's services with those of NCBI in the area of variation will be taken forwards by a new variation team leader recently appointed at EBI.

1. The most important meta data for inclusion in any submission is the identity of the genome sequence and version on which variant calls have originally been made (but the system also stores the identity of the reads, to allow for potential re-calling; and of the wider set of sequences against which variants have been called, not just the molecule on which the variant was located). Fortunately, after a lengthy period in which there has been no effective identifier for genome assembly versions, a new system for identification has recently been implemented as an extension to the International Nucleotide Sequence Database Collaboration. These new "genome collections" identifiers will be the fundamental denominator of sequence identity in the system, and any alternative descriptors/identifiers will be mapped to this.
2. Variant Call Format (VCF) is a text file format (most likely stored in a compressed manner) that has developed as a standard for the representation of variant information in the context of the 1000 genomes project (<http://www.1000genomes.org>). It contains meta-information lines, a header line, and then a variable number of data lines, each of which describes a position in the genome. The infrastructure will accept submissions in VCF format and allow users to retrieve submitted files in this format.
3. A provisional (web-based) user interface has been developed, to allow users to submit VCF files and appropriate meta data, and to verify certain elements of the meta data against reference values before deposition in the persistent archive (step 4). A screenshot of the tool is shown in figure 1, below.
4. The archive layer of the infrastructure will be implemented as an extension to the European Nucleotide Archive (<http://www.ebi.ac.uk/ena>), which already archives the raw sequence data from which both reference sequence and variant calls are generated. ENA already has tools for storage, back-up and fast, indexed retrieval of individual documents, and prior experience of storing huge data volumes (variation call data is large, but a variant call is effectively a reduced form of the sequence reads themselves). The ability to utilize ENA infrastructure has significantly increased the speed of progress on this work package, and allowed a focus on the missing pieces of the infrastructure, i.e. QC tools on data submission, feature-mapping pipelines, etc. To allow the use of the ENA infrastructure, tools have been written to support the automated submission of quality-checked data into the ENA, and for retrieval using the ENA's data access libraries. Both ENA browser and submission service offer programmatic access via a REST API (<http://www.ebi.ac.uk/ena/about/browser>), using XML and FASTA.
5. If features are not to be unnecessarily lost or otherwise mis-propagated, it is essential to utilize a fast, accurate genome aligner to provide the basis for the genome-genome mapping. We conducted a test of various alignment methods, to see (i) how quickly they ran (ii) what proportion of a given genome assembly they

were able to map to another version of the same assembly and (iii) what effect this had on the ability to propagate variants correctly (which was assessed by comparing the results with *de novo* calls utilizing the underlying reads on the new assembly). Additionally, the results were compared with the results of propagation on a per-variant basis using the traditional method of flanking sequence alignments. The following methods were used: ATAC (<http://kmer.sourceforge.net>), NUCmer (<http://mummer.sourceforge.net>), BLASTZ (http://www.bx.psu.edu/miller_lab). For the eleven genomes of the five species tested, ATAC consistently delivered the highest fraction of sequenced mapped (sometimes reaching 99% where other tools only reach 80%) and consistently required the least execution time (~50x faster than next best for dissimilar genomes). NUCmer, despite employing similar techniques to ATAC at it's best only reach mapping coverage matching ATAC, whilst requiring significantly more time to complete. The general-purpose aligner BLASTZ was unable to keep up with both ATAC and NUCmer both in terms of performance and execution time, as it is not optimized to work on near identical sequences. Lastly, the flanking sequence alignment approach was able to reach a slightly worse propagation precision compared to ATAC, with an execution time highly dependent on the number of features - with our sample size of 500k being 5-60x slower. On this basis, the gapped ATAC method was chosen for use. The method is usable quick and is able to map more sequence, and to propagate positional features, more accurately than the alternative approaches.

6. The test of the alignment methods was obviously dependent on the existing of a method for feature propagation, although this is in itself straightforward and therefore not effectively under test. All that is required is to process the mappings produced by the genome aligner, and for each mapping to update the position of all features that fall within it. Using the *de novo* calls described in point 5 as a gold standard, precision (or positive predictive value) and recall (or true positive rate/sensitivity) of feature propagation can be measured. When projecting from version 6.1 of the rice genome to version 7, feature propagation based on ATAC mapping reached a precision of 99.41%, and recall of 99.47%; with flanking sequence alignments reaching 99.14% and 98.60% respectively. When projecting from version 2 to version 7, ATAC reached precision and recall of 98.12% and 96.69%; with flanking sequence alignments reaching 97.60% and 95.81%. For closely as well as distantly related pairs of sequences the flanking sequence alignment approach has a higher rate of false negatives, losing more features between versions; and a higher rate of false positives, placing features incorrectly.
7. The operational data store has been implemented in MongoDB (<http://www.mongodb.org>), a document-store database system. As such, MongoDB has some advantages applied to this project compared with a classical Relational Database Management System, namely data conceptually belonging together - describing the samples, genotypes and metadata of a position in the genome does not have to split and later rejoined across tables of a relational schema. Furthermore it is better suited to handle very large datasets and allows horizontal partitioning/sharding. The underlying document format consists of indexed JSON files. A schema has been derived to provide access to this data – see

figure 3 below.

8. A merging procedure has been developed. Data that has been sharded with MongoDB across multiple physical servers can be processed efficiently in parallel using the MapReduce programming model (<http://research.google.com/archive/mapreduce.html>). Each document in the database is processed independently and in parallel during the "map" step and combined in some way to form the output during the "reduce" step. Selecting an appropriate reduce step guarantees that all features sharing the same position are grouped together and are merged. Work is now proceeding on a system for the assignment of unique identifiers.

Figure 1 The provisional variation submission interface

submission data

Alias: (e.g. Maize HapMap II)

Center name: (e.g. CSHL)

analysis data

Title: (e.g. Capturing Extant Variation from a Genome in Flux: Maize HapMap II)

Description: (e.g. A comprehensive characterization of genetic variation across 103 inbred lines...)

study data

Provide an existing sample accession (e.g. SRP011907)

Title Resequencing of 50 rice individuals

Abstract

Rice is a staple crop that has undergone substantial phenotypic and physiological changes during domestication. Here we resequenced the genomes of 40 cultivated accessions selected from the major groups of rice and 10 accessions of their wild progenitors (*Oryza rufipogon* and *Oryza nivara*) to >15 * raw data coverage. We investigated genome-wide variation patterns in rice and obtained 6.5 million high-quality single nucleotide polymorphisms (SNPs) after excluding sites with missing data in any accession. Using these population SNP data, we identified thousands of genes with significantly lower diversity in cultivated but not wild rice, which represent candidate regions selected during domestication. Some of these variants are associated with important biological features, whereas others have yet to be functionally characterized. The molecular markers we have identified should be valuable for breeding and for identifying agronomically important genes in rice.

Description

assembly data

Provide an existing genome assembly accession (e.g. GCA_000005005)

sample accessions

Upload a file

rice_v7_1k.vcf 0.1MB

sample name in VCF file Sample accession in ENA

SRR063669.bam

SRS086325

Oryza sativa

Figure 2 The overall workflow

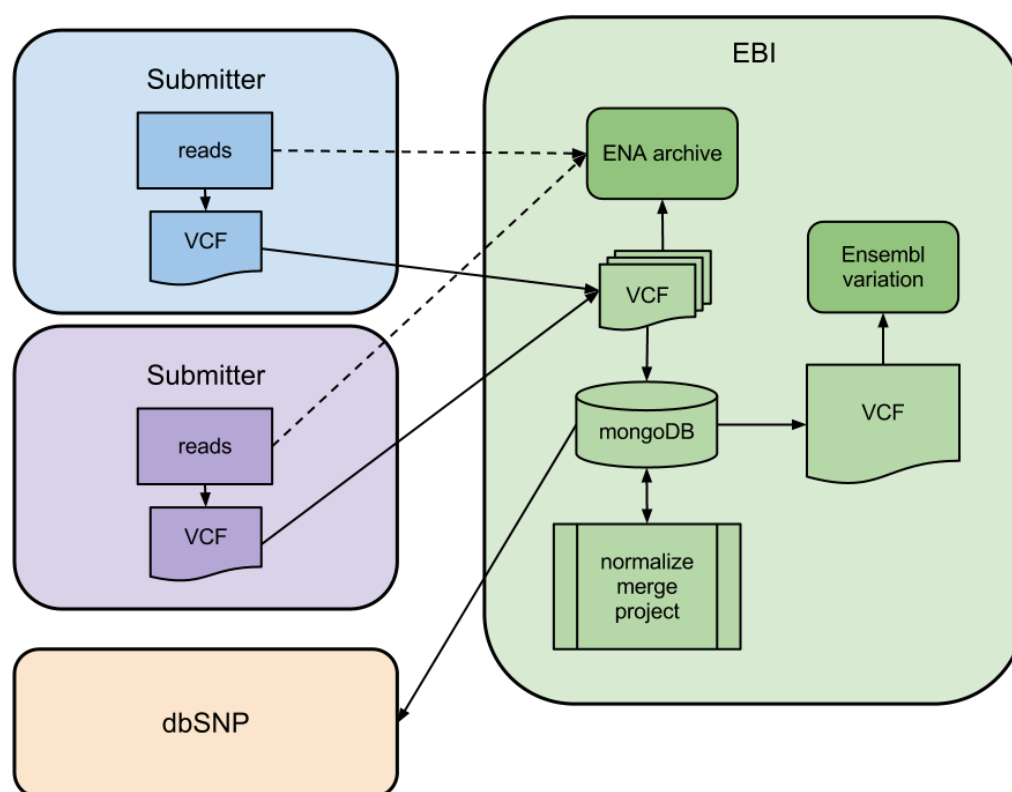
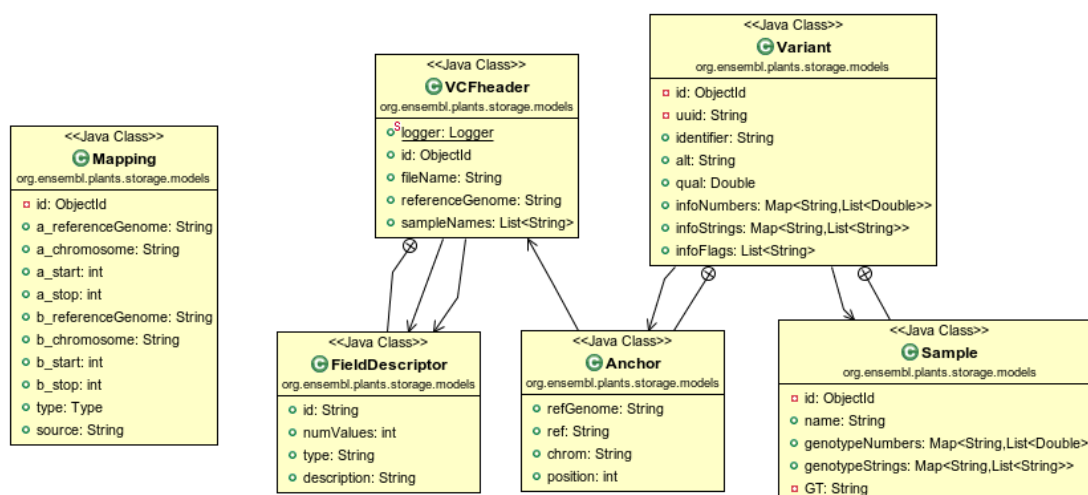


Figure 3 The data representation UML class diagram



Next steps

The document DB needs to be extended to provide a persistent store for the mapping data. Export into the Ensembl schema then needs to be implemented. We will then commence the upload of data provided, in the first instance, by selected partners and close collaborators. This process is expected to begin in the last quarter of 2012 and the infrastructure should be open for more general submission by the end of 2013, ahead of the initially planned schedule. Much of the work in year 2 of the project will focus on testing the pipeline with actual use cases for data (examples from *Arabidopsis thaliana* and maize are prepared), and for increasing the robustness of the infrastructure to

support development from prototype to a production system.

Perspectives and future directions

An alternative approach is not to archive variation calls at all, but merely to archive sequence reads used to infer the existence of variants; and then to re-use existing reads, alongside any new reads that may be available, to re-call variants against each new reference sequence. The advantages of this approach are: (i) it would provide consistency; and (ii) as archives for sequence reads already exist, the need for (much of) the infrastructure being developed in this work package would be obliterated. However, the approach (i) is computationally expensive and inefficient (especially when a new assembly consists mainly of rearrangement of previously existing local sequence) and (ii) fails to archive original calls, which may have been referenced in publications etc. (iii) requires transPLANT to effectively take responsibility for declaring the “best” variation calling methods, as the project would be preferring to use its own calls over user submissions. For this reason, we have followed the approach outline above.

However, the two approaches are not incompatible; indeed, were transPLANT to re-call existing variations, the transPLANT-generated results could be submitted into the new infrastructure alongside the previous results. Over time, even the best-annotated data set may become outdated. We reserve the right, when need is greatest, to re-analyse genomes and make new variant calls as appropriate, but will continue to archive any user-generated calls submitted to us.

For handling larger scale variations, we will take our cues from the conclusions emerging in work package 8, as to the right way to handle these, and implement appropriate mechanisms into this infrastructure as appropriate.

Use of resources (*highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1*)

EMBL-EBI, 3.5 person months

Work package number	10	Start date or starting event:			M1
Work package title	Tools for elucidating the genotype-phenotype map				
Activity Type	RTD				
Participant number	3	5	6	7	11
Participant short name	GFMPG	INRA	IGR PAN	BIOGEM	KN
Person-months per participant	33	2	48	1	18

Objectives

Develop tools for flexible, easy-to-use tools for genome-wide association mapping, and statistically optimal data descriptors.

Lead Beneficiary: GFMPG

Description of work

Task 1: A web-interface that allows real-time GWAS in *A. thaliana*

Objective: Develop a web-interface that allows real-time GWAS in *A. thaliana*.

Description: The power of GWAS in organisms for which inbred lines are available has recently been demonstrated (Atwell et al., Nature 2010; Huang et al., Nature Genetics 2010). Genotyped inbred lines can be distributed throughout the plant genetic community, making it possible for anyone capable of growing and phenotyping plants to carry out GWAS (Atwell et al., 2010). Analysing the data remains a stumbling block for many groups, however. The problem is two-fold: first, there is the statistical problem of carrying out association analyses that involve millions of polymorphisms; second, there is the bioinformatics problem of visualizing and interpreting the results in terms of genome annotation. We will develop a web application that lets individual users upload their phenotypic data, and analyse online, in real time. The results will be visualized as Manhattan plots, with individual points annotated and hyperlinked to the genome annotation.

Task 2: Meta-analysis of pleiotropy

Objective: Develop tools for the meta-analysis of pleiotropy

Description: The primary rationale for the tools described under Task 1 is to enable individual groups to analyse their data quickly and painlessly. However, an important secondary goal is to make it possible to compare the results of individual studies with other published results, for systems-level insights into pleiotropy. The current Arabidopsis database at GMI contains several hundred different phenotypes, and we have already made tantalizing observations, such as connections between seed dormancy and flowering time, and between different kinds of

resistance (Atwell et al., 2010; Todesco et al., 2010). There will be an option to upload data permanently (with password protection until public release) in order to build an ever more complete description of phenotypic variation, ultimately making phenotypic associations part of the genome annotation. Types of interfaces envisioned range from simple listings of phenotypes with which a particular polymorphism appears to be associated, to network displays of correlations between phenotypes.

Task 3: Multi-factor GWAS models

Objective: Construct an interface for analysis of traits in a polygenic background.

Description: Most GWAS to date analyse polymorphisms one at a time even though most traits have a polygenic background. This is obviously suboptimal from a statistical point of view (Platt et al., Genetics 2010). We will extend the interface described above to allow more sophisticated model that use particular polymorphisms as co-factors in the analysis. For example, we could envision gradually building a more refined genotype-phenotype map by including experimentally verified loci.

Task 4: Modelling correlated phenotypes

Objective: Construct an interface for modelling correlated phenotypes

Description: While pleiotropy refers to unexpected phenotypic correlations, many phenotypes are obviously correlated, e.g., because they measure the same thing under different environmental conditions, or because they are components of a known biochemical network. GWAS in these cases should be carried out with the benefit of prior knowledge. For example, if we measure flowering time under several different light and temperature regimes, these should be co-factors in the model. We will extend the basic tools above to allow this kind of analysis.

Task 5: Development of statistical descriptors

Objective: Develop statistical descriptors for use in data repositories.

Description: Given the large amount of phenotypic data to be stored in the proposed databases, the information which is to be kept must be chosen carefully. Neither the raw data nor the final summaries that appear in papers are suitable for integration of different studies. We will carry out research designed to find appropriate statistical descriptors for the different types of data covered by the project. If possible, the descriptors will form sufficient statistics, that is, they will contain all information necessary to estimate parameters of interest to biologists. As some random variables (traits) observed in the genetic experiments have complicated distributions, conditioned by several nuisance variables and classifiers (e.g. distribution of binding score for genomic locations of different annotation), the number of parameters of interest, and consequently of descriptors, may be large for some experiments. Note also that if the raw data are not directly available, the set of statistics must contain variance measures appropriate for estimation of errors made in data integration and map construction. The descriptors will be found for situations in which the biologists deal with pleiotropy and correlated phenotypes. Special attention will be paid to traits that are observed in the form of profiles (in protein-DNA interaction studies, metabolomic assays, etc.), for which complicated data processing methods exist, in order to choose the best representation. The Task will consist of methodological development based on generalized mixed models, multivariate data analysis and functional data analysis methods that will serve as the theoretical basis for descriptors. Here, a special role is envisioned for functional data analysis methods that provide analogues of usual analyses done for uni- or multivariate data (analysis of regression, principal component analysis), but applicable for data obtained in the form of profiles (functional regression, functional PCA) and

appropriate for binding signals over chromosome, chromatograms, spectrograms etc. The Task will also involve development in the area of integration of all types of descriptors for “data to knowledge” transformation and genotype-phenotype map construction (feature extraction and feature selection methods using multivariate and machine learning approaches). This task will be tightly linked with D3.4 to use the same descriptor format in data repository and analysis tools. INRA scientists and data providers will also contribute to the setup of common descriptor ontology usable with all species of agronomical interest. Defining that ontology will also ensure the right level of data precision in the phenotype data repositories.

Task 6: Computational aspects of sufficient data descriptors

Objective: Optimize the computing of statistical descriptors

Description: Taking into account the data architecture and data processing models considered in the project, it is necessary to optimize the way in which the descriptors described in Task 5 are computed and used in the databases. For some of them, pre-computation and permanent storage may be proper (cached sufficient statistics). For some, provision of proper functions and calculation “on the fly” will be more convenient (SQL queries development, user defined functions). For some, the optimal trade-off between precision and computation time will be looked for. Known algorithms will be scaled for large data sets. In case of excessive computational cost approximate versions of descriptors will be defined. Correction for experiment-specific design will be taken into account. We will also optimize the set of descriptors provided by the database with respect to the possible queries. The Task will consist of studies on real and simulated data. The main tools will be mathematical statistics methods, decision trees, machine learning approaches (neural networks) used for training data sets and queries.

Progress towards objectives and details for each tasks

. Task 1: A web-interface that allows real-time GWAS in *A. thaliana*

A stand-alone application has been developed by the GMI group, and a paper describing is under review for *Plant Cell*. A more comprehensive version was demonstrated at the International Arabidopsis meeting in Vienna in July 2012, and should go online by the end of the year.

. Task 2: Meta-analysis of pleiotropy

Tools for accomplishing this task are part of the web-interface discussed above. Research papers describing the application are under preparation.

. Task 3: Multi-factor GWAS models

A paper describing such a model/method was published this summer (Segura *et al.*, *Nature Genetics*, 2012). Research continues.

. Task 4: Modelling correlated phenotypes

A paper describing such a model/method was published this summer (Korte *et al.*, *Nature Genetics*, 2012). Research continues.

. Task 5: Development of statistical descriptors

Work is in progress at IGR PAN on: (a) extending results on existence of sufficient statistics in linear models to models with more than one random effect and characterized by orthogonal block structure, and; (b) describing conditions under which the sufficient statistics for fixed parameters in a linear model can be used for computation of sufficient statistics in a corresponding mixed model.

Discussions with KeyGene were held to further define and plan research activities.

. Task 6: Computational aspects of sufficient data descriptors

Data from KeyGene's automated phenotyping device were gathered for 90 lines of the Arabidopsis Nordborg population. These data are being further analysed and tested for suitability as a test set for WP10. An inventory is being performed for suitable tools for data reduction and statistical descriptors.

Use of resources *(highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1)*

GMI: 11.88 person-months

IPG PAS: 2.79 person-months

KeyGene: 0.6 person-months

Work package number	11	Start date or starting event:	M1
Work package title	Meta-data driven information retrieval systems		
Activity Type	RTD		
Participant number	1	4	
Participant short name	EMBL-EBI	IPK	
Person-months per participant	3	42	

Objectives

Development of a cross database information retrieval infrastructure using search engine technology

Lead Beneficiary: IPK

Description of work

Task 1: Development of the information retrieval infrastructure

Objective: Develop an infrastructure for meta-data aware information retrieval

Description: This task is organized in relation to the particular system modules (search engine web frontend, search engine backend, text index system, relevance ranking logic, data format converters, training of relevance, system installation and maintenance).

Search engine web frontend: In addition to result browsing and collection in the data cart, the frontend will support the feedback of user relevance ratings and the tracking of user frontend interaction for an automatic estimation of data record relevance. The exploration of search results will be supported by a detail browser and feedback system. The original data is displayed and the user might rank the relevance of the hit for later training of the user ranking profile.

Search engine backend: At the technical level the reengineering of the storage backend and text index framework is necessary to meet requirements of scalable and well performing query execution. Doing so, the storage backend will switch away from classical structured storage in relational database to NOSQL systems, also known as triple-stores or attribute value databases, which are optimized for high-traffic web sites and a read only access. Furthermore, we will use cloud computing based on distributed Apache LUCENE, which is the state of the art open source text index system.

Text index system using existing data access interface for transPLANT databases (use of WP5) and IPK ex-situ Genebank: The proposed information retrieval system will utilizes the LAILAPS search engine for transPLANT databases (use of WP5) and the IPK ex-situ Genebank (a collection of

agricultural and horticultural plants which aims to conservation and distribution of plant genetic resources; the IPK Genebank holds one of the most comprehensive collections worldwide and provides a major contribution to the prevention of genetic erosion; currently over 146 thousand accessions from 2,649 plant species and 779 genera are available). Here we will combine a feature model for relevance ranking, a machine learning approach to model user relevance profiles, ranking improvement by user feedback tracking and an intuitive and slim web user interface. Queries are formulated as simple keyword lists and are expanded by synonyms. Furthermore a full data export as a RFC 4180 standard compliant flat comma separated file has to be provided by WP7 and IPK ex-situ Genebank will be implemented.

transPLANT specific ranking features: Benchmarking and the investigation of user criteria for relevance rating show the need for additional features. Consequently, will extend the scoring functions. Promising effects are expected from the consideration of link degrees between data records and the use of “Statistically Improbable Phrases” to predict relevance influencing keywords.

Data cart: The Data Cart is a concept for collection, transformation and distribution of data in the information retrieval environment. The search engine fills references to data items that result from a search query in this container. The data cart will offer function for filtering, versioning, data download and data format transformation. Furthermore, data privacy will get special focus and will be ensured by user authentication and data encryption. To ensure platform independence and well scaling in respect to high voluminous data the data cart API will be implemented as Representational State Transfer (REST) architecture. This style of software architecture is optimized for distributed hypermedia systems such as the data access URL’s used in life science databases. Access to the data cart will be provided as part of the suite of web services developed in WP5.

Data format converters: The format transformation will support in the initial version the export to basic bioinformatics data formats (FASTA, JMOL, SBML, CSV, attribute value pairs, text, and binary). By a plug-in mechanism, the list of supported data formats is expandable on individual user needs. Doing so, data converters can be implemented on demand and dynamically registered.

Relevance ranking training for phenotype queries: The crucial step for the search engine training is a set of true positive relevance rankings for query results of user cases. In order to express the relevance of an database entry, our experience motivates manual curated relevance reference lists, which will be separated into three confidence classes: high, medium and low. This enables a use case related and end-user specific training of neural networks for a customized relevance ranking. The first option to get these ranked reference lists is manual rating of delivered search results by the user. In this way we have the possibility to link user personal background with user profiles and profile specific relevance criteria. To combine end user and domain expert knowledge, use cases, which have a common general interest for daily use, will be identified in this deliverable. Initial use cases for the manual curation are queries in protein or gene functional annotations and NCBI literature databases and retrieval of gene/marker data relevant for important traits.

System installation and maintenance tool: In order to maintain the search engine infrastructure, an installation and maintenance tool will be implemented. This software package is the final deliverable and will be the central dashboard and control automated update of installed search engine instances. This include the update of database indexes, the maintenance of ranking parameter like keyword and synonym list, as well as the update of the relevance ranking model. Beside maintenance, the installation and set-up of new search servers is the second core function. The idea is to provide an installer, which can be used to set-up individually customized

search engine installations.

Task 2 An interface for the information retrieval system in the transPLANT portal

Objective: Develop an interface to enable the integration of the information retrieval system within the transPLANT portal

Description: An interface will be developed within the transPLANT portal to provide integrated access to the information retrieval system developed in WP11.

Progress towards objectives and details for each tasks

Task 1: Development of the information retrieval infrastructure

Actions

The transPLANT consortium will provide an information infrastructure for genomics resources. Those are distributed among the partners and will be integrated at several levels. One level is the information retrieval (IR) over genomics meta data, like functional annotation for genes or other genomics regions. The transplant-IR environment should meet the existing database infrastructure, which is a network of distributed, but interlinked information systems. In consequence we compiled a list of resources the partner use to annotate their genomic data. Those repositories, i.e. UniProt, Gene Ontology, PFAM, are the points of intersection, and on the other hand well-known data hubs for information search.

Results of year 1

The result of the review of partners data bases, the approach of combining search engine technology with an annotation reference network (Mehlhorn H, Lange M, et al.: IDPredictor: predict database links in biomedical database. J. Integr. Bioinform. 2012, 9(2):190). The approach is to search for keywords in the mentioned public repositories and map relevant hits back to the partner databases. This concept of search and reverse identifier lookup is implemented as prototype (<http://lailaps.ipk-gatersleben.de>) using the LAILAPS search engine (Lange, M. et al. J Integr Bioinform. 2010, 7(2):110), which combines a keyword based search, recommender system and an AI-based user specific relevance ranking.

In order to prepare the implementation, we collected references from those partners, who annotate their genome data using vocabulary, ontologies or textual description of gene function from public databases. The resulting compilation comprises a list of 8 major databases (Trait Ontology, Pfam, Gramene, Plant Ontology, SwissProt, TrEMBL, Gene Ontology, PDB). At the first step we indexed those databases in LAILAPS and already cross-linked IPK data bases GBIS, MetaCrop, Cr-EST and the EBI Ensembl database. The cross-linking of GNpIS resource from INRA is in progress.

The results of search queries are relevance ordered links to genomic data from transPLANT partner. This data must be collected, persisted and shared for later analysis using the web service infrastructure of WP5. Furthermore, there will be the option to collect all results from runs of those analysis pipelines as citable data records. The result of this concept is the e!DAL system (Arend D, Lange M, et al.: The e!DAL JAVA-API: Store, Share and Cite Primary Data in Life Sciences. IEEE Internl. Conf. on Bioinform. and Biomed., Philadelphia, U.S.A., 2012). e!DAL is a comprehensive storage backend for primary data management. It stands for (electronical Data Archive Library) and implements an enhanced and file system like storage system. Main features are version and meta data management, data citations, support for information retrieval,

persistent identifiers and its easy and modular integration into existing data frontends and information systems, i.e. the LAILAPS search engine.

Task 2 An interface for the information retrieval system in the transPLANT portal

The transPLANT portal provides a central point of entry to the partner resources. IPK provided here links and summaries to the major databases, software and information systems. Next step is the integration of the meta-data search engine, which is under developing in this WP 11, as information retrieval tool for the transPLANT portal.

If applicable, explain the reasons for deviations from Annex I and their impact on other tasks as well as on available resources and planning

No deviation

If applicable, explain the reasons for failing to achieve critical objectives and/or not being on schedule and explain the impact on other tasks as well as on available resources and planning *(the explanations should be coherent with the declaration by the project coordinator)*

No deviation

Use of resources *(highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1)*

EBI: 0.75 person month

IPK: 10 person month

Work package number	12	Start date or starting event:		M1
Work package title	Implementation of resource-intensive algorithms for plant genomics data			
Activity Type	RTD			
Participant number	5	8	9	10
Participant short name	INRA	TGAC	BSC	DLO
Person-months per participant	12	30	20	18

Objectives

Distributed implementations of resource-intensive algorithms in a high-performance compute environment.

Lead Beneficiary: EMBL-EBI

Description of work

Task 1: Strategies for genome sequencing and assembly

Objectives: Evaluation and development of strategies for genome sequencing and assembly.

Description: This task will focus on the process of data generated by the latest technology in sequencing including the next generation sequencing (NGS) platforms but also looking into single-molecule technologies. The objective of this work is to develop a dynamic matrix of, on the one hand, sequencing technology and assembly characteristics and, on the other hand, biological questions (e.g. on SNP discovery, the discovery of structural variation haplotype composition in heterozygous and polyploid genomes, the epigenome, etc.) addressed through large-scale sequencing. Re-sequencing algorithms are traditionally organised around alignment tools with extensions for calling variants, in general single nucleotide polymorphisms (SNPs), but other approaches have been recently introduced around graph-based frameworks. De novo assembly algorithms are very demanding on memory resources and, in some plant species these tools will need to cope with heterozygosity and allopolyploidy (where the challenge is to distinguish between different homeologous haplotypes). We will compile benchmark datasets for a variety of sequencing project objectives, assess the range and types of sequencing data minimally required to address these objectives and evaluate and compare the performance of tools and algorithms available for analysis on these datasets. The matrix should provide a dynamic decision support system that can be consulted by the plant research community in the design and execution of large-scale genome sequencing and assembly projects. Particular areas of focus will include the following:

NGS assembly algorithms for large polyploidy plant genomes. The current assembly algorithms for NGS data have been design for vertebrate diploid genomes (typically ~50% repeat content with heterozygous diploid genomes). Although in plants is relatively simple to generate fully

homozygous individuals, the challenge is in the ability to distinguish between homeologous sequences in polyploidy species. The evaluation of assembly algorithms will be focused on the traditional quality metrics for consensus sequences (N50, number of contigs) as well as the resources required such as RAM memory and performance. This task will also cover the challenges around transcriptome sequence assembly.

Scaffolding algorithms to use read-pairs and st One of the challenges in large plant genomes is the repeat content that in some cases can be more than 80% of the genome. Highly repetitive genomes are difficult to assemble resulting in large number of contigs. Although most of the current algorithms make use of read-pair information sometimes the data is not fully used. Alternative approaches that use the pair end data once the bulk of the genome is assembled could be better suited. This scaffolding stage should also prepare for dealing with strobe reads. Effectively strobe reads are a generalisation of pairs into n-tuples. A good and robust approach for scaffolding and more general genome refinement and finishing will help to implement better biology.

Re-sequencing and population genetics for (Allo)polyploidy. In the near future we will have the genome sequence for several crops. These genomes are characterized by complex architectures including high repeat content and polyploidy. These features challenge some of the well-established concepts in population genetics that will need to be rethought in the context of plant genomes. One example is the modelling of polymorphisms in allopolyploidy species where it is required to distinguish homeologous from heterozygous events.

Reference-free / metagenomics analysis algorithms. For some other species the availability of genome sequences will be more distant. In the recent years we have seen new emerging techniques based on assembly/alignment hybrids approaches designed to work in the presence of highly heterogeneous samples, or in situations where there is not available reference sequence.

Task 2 Data structures for algorithm optimisation

Objective: Evaluation and development of Data structures for algorithm optimisation

Description: The algorithms for the analysis of the datasets generated by the NGS platforms are characterised by high demand on resources. In particular assembly algorithms are based on approaches that rely on the access to the whole datasets to be able to for example remove noise or low quality data in the sequence reads. This is particularly challenging with large plant genomes such as wheat. The aim of this task is to develop optimisation strategies that will help algorithm developers to write software that can efficiently use the available hardware. This task will first analyse the selected algorithms by means of profile tools and the tracing and visualization tools developed at BSC. These analyses will be performed both on supercomputers at BSC as well as resources made available by other consortium partners. The results of these analyses will provide recommendations about the most appropriate computer architecture and programming model. Conclusions will be then used to the optimization of selected algorithms (with particular focus on genome assembly, the most demanding group). Possible actions can include redesigning the structure for algorithms and data organizations, specifically improving data layouts for locality and concurrency, and efficient use of memory, and support for communication and computation overlap. Prototypes of optimized versions will be benchmarked using test genome data. The aims in task 1 focus on the NGS algorithms from a user perspective, whereas the aims in task 2 emphasise the aspect around software from the perspective of the developers.

Task 3 Gene annotation and functional genomics

Objective: Exploit synteny between plant species to improve genome annotation

Description: Plant genome annotation is a particularly challenging task because of their large size, polyploidy and repeat content. Despite immense progress made in the past decade on development of large-scale experimental methodology, functional gene annotation in plants continues to greatly lag behind in the deciphering of new gene and genome sequences. Even for the widely used model species *Arabidopsis thaliana*, one third of the proteins still lack a functional annotation. For lineage-specific or highly divergent proteins the probability of identifying a functionally characterized homolog is small, and traditional homology-based tools cannot deal with sub- and neo-functionalization of recent paralogs. One approach is to integrate experimental data to maximize the accuracy and coverage of function prediction. To this end, computational methods have been developed that can accurately predict protein functions from experimental data on a large-scale or provide leads for hypotheses of function and the design of targeted experiments. The need for wide and user-friendly availability of such methods becomes even more critical as the number of genome sequences and experimental datasets for non-model crops is soaring. An aim in this task is to explore the comparative genomics tools to take advantage of the conserved synteny between some of the crops species (for example grasses) to annotate large and complex genomes.

In this task we will also test methods, tools and pipelines for gene and repeat annotation. Their performances on complex plant genomes will be assessed. Pipelines components will be developed in such a way that they can be added to improve existing pipelines, or assembled in new pipelines. These pipeline building blocks will be shared among partners and freely distributed to the scientific community. These component will be designed to be used in pipelines implemented over computer grids or clouds

Task 4 Development of tools for the enablement of virtual plant breeding

Objective: Develop a pipelining infrastructure for the support of multi-step analyses for the enablement of virtual plant breeding.

Description: Two factors are essential for continued successful improvement of crop species by classical means of plant breeding. First, adequate genetic variation needs to be available. Second, the technological route to the exploitation of this variation needs to be optimized. Material from wild relatives, ancestors, and landraces held in germplasm collections of crop species often contains a wealth of genetic variation. Most importantly, this will offer a useful gene pool, providing many new, but also old and better alleles that were lost during domestication and selection targeted at only a narrow range of desirable agricultural traits. Exploiting this resource in modern breeding in particular has the potential to genetically enrich extant crops with alleles that can improve traits that have recently become important in the face of new challenges and requirements regarding climate change, sustainable production and a growing demand for more and better food. The challenge in efficient exploitation of germplasm material lies, firstly, in the ability to identify adequate alleles for a desired trait directly at the DNA sequence-level and, secondly, in the immediate availability of DNA markers associated with or causal to such a trait. Targeted, high-resolution panels of DNA markers can be designed to monitor the exclusive introgression of the genomic region from the germplasm accession that carries the desired trait. From a technological point of view, the challenges in exploiting multiple genome sequences for breeding purposes lies in the nature and scale of the computational management that these data require.

The aim of this task is to initiate the development of an infrastructure for virtual plant breeding (IVPB). The core system should deliver proof-of-principle in the form of an in silico-designed breeding experiment, using genome sequences from a germplasm collection for one selected crop and one selected trait. The development part of the IVPB covers the automation and

standardization of the core computational processing and analysis of the sequence data, including raw sequence processing, quality control assessment, de novo assemblies of novel insertions, mapping of variants and multi-genome alignment. All data produced will be stored and managed in a genome database adapted for or developed specifically for multi-genome comparison and display. This database and the genome view will be both browsable and searchable using both text-based and sequence-based querying.

Progress towards objectives and details for each tasks

Task 1: Strategies for genome sequencing and assembly

The aim of this task is to develop strategies for the implementation of bioinformatics analysis that are intrinsically demanding in computational resources. In the first year of the project we have focused on a number of activities:

Evaluation of assembly algorithms. TGAC has access to a variety of whole-genome datasets that we have used to learn about some of the general features of the assembly tools with an emphasis on the concrete challenges emerging from the use of algorithms that have been designed to work with human and other mammalian genomes to perform with plants. We have focused on the evaluation of these tools when working with data generated by the most popular next-generation sequencing platforms. The main limitation of these technologies is that sequences are short; this is compensated by the high-throughput at a low cost that these instruments can achieve but this is in itself a challenge for the software tools, as they need to deal with massive datasets.

In order to evaluate the different tools we have prepared a number datasets that range from a simple simulated genome (based on the Assemblathon project) to a more complex and highly-heterozygous real dataset (from the red clover genome project). An intermediate datasets is based on a collection *Saccharomyces cerevisiae* samples from the UK National cultures of yeast collections.

The assembly tools that will be evaluated in a first round of assessment are:

Assembler	Approach	Read tracking
ABYSS	de Bruijn	mapping
SOAPde novo	de Bruijn	mapping
Velvet	de Bruijn	mapping
SGA	String graph	?
ALLPATHS	de Bruijn + OLC	?
Newbler	OLC	complete tracking
WGS (Celera)	OLC	complete tracking
Fermi	String graph	?

TGAC and BSC have initiated the work and we expect to start testing the tools on real sequence data within the next three months.

Milestone MS27 – Implementation of reference-free methods for transcriptome variation analysis. TGAC has led the development of this milestone within Work Package 12. The full report of the

milestone explains in detail the activities carried out over the first year of this project relevant to this part of the project. In summary we built a bioinformatics pipeline that works in three stages starting from the sequencing reads for the different samples as input:

1. Generation of a “transcriptome reference” to be used in the mapping downstream
2. Mapping of the sequencing reads using tools developed to cope with splice boundaries.
3. Analysis of the alignments to generate a list of variants (with associated quality scores).

Deliverable 12.1 Development and test of sophisticated statistical methods to model variation in large plant genomes. Aspects of this activity are complementary to one of the objectives in Work Package 8. As TGAC and INRA are both responsible for these activities in Work Packages 8 and 12, we organised a one-day workshop to develop a work plan for this deliverable. Please refer to Work Package 8 for a detailed explanation of the strategy. As a direct outcome of this meeting we have devised a pipeline that will work on the different types of genetic variations and available technologies to generate an output in VCF format with all the variants discovered in the input samples and with quality scores for each entry.

Task 2 Data structures for algorithm optimisation

Evaluation of the performance of selected algorithms. Task 2 seeks to propose solutions for the improvement of the use of resources by assembly algorithms. Task 1 (see above) has selected a number of algorithms in basis of their extended use. At BSC, the analysis of the performance has just started. Software to be tested has been initially installed in a shared-memory high performance computer (Altix II, ultraviolet, architecture cc-NUMA, 12 Intel Xeon 8-core CPUs E7-8837 at a 2.67GHz, 1.5Tb of shared memory). This kind of architecture is the most commonly used for assembly algorithms, even though the main reason is the large amount of RAM memory required, without taking necessarily benefit of the available parallelism. In a later phase, for those algorithms where the need of memory is not so important, a second computer (8 node cluster provided by 2 6-core Intel xeon CPU's, and 96Gb RAM memory per node) will be used for testing. It should be noted that the prototype of cloud environment developed in Work Package 5 is being installed and tested in this second computer. In this way, the suitable assemblers could be also tested using the cloud.

In this early phase, programs listed in the table above have been installed and installations have been checked using the test provided by the applications themselves. In the next months, we will proceed to the assembly of the different datasets prepared in task 1. Parameters to be measured include among others:

- Validity of the results obtained
- Use of Memory (Peak values and average)
- Use of scratch space
- Influence of communication network to access data disks
- Type of parallelism if any

- Basic profiling (for open-source codes) both in serial and parallel versions and efficiency in CPU usage

Task 3 Gene annotation and functional genomics

This task focuses on genome annotation of complex plant genomes because of their large size, polyploidy and repeat content. Objectives are to improve structural gene annotation by exploiting synteny between plant species, to scale gene and repeat annotation pipelines for large polyploid genomes, and to improve functional gene annotation by integrating experimental data to maximize the accuracy and coverage of function prediction.

INRA worked on the REPET package (Flutre *et al.* 2011, Plos One), a set of pipelines for repeat analysis. The pipelines were improved for speed and accuracy including a new transposable element classifier called PASTEC (Pseudo Agent System for Transposable Element Classification), and a new method of transposable element detection based on LTR transposable elements structural features, using LTR_Harvest (Ellinghaus *et al.* 2008, BMC Bioinformatics). New strategies to scale-up the pipelines for annotating large genomes are under development. A new prototype pipeline for long satellite detection (>500bp) has been developed and is under testing.

DLO worked on the sequence- and network based function prediction tool BMRF (Kourmpetis *et al.* 2011, Plant Phys). Its performance was assessed for several species in the context of the CAFA community wide assessment of protein function annotation (<http://biofunctionprediction.org/>). Furthermore, the method was extended for application to crop species. To do so, inter-species transfer of annotations was added, co-expression networks for various crop species were obtained and their use as input data was tested. Currently, a prototype webtool to allow access to the resulting sets of predicted gene function annotations is under development.

Task 4 Development of tools for the enablement of virtual plant breeding

A major challenge for successful plant breeding in the framework of virtual breeding is the transition from genetic maps and markers intervals of a quantitative trait-of-interest (QTL) to the actual genes responsible (at least in part) for that trait. Given a physical map and/or genome sequence, the translation of genetic map data to the physical map generally results in large lists of candidate genes from which to choose. In human genomics, large-scale text-mining methods based on the assessment of abstracts of scientific papers combined with extensive thesauri in so-called 'nanopublications' using 'triple stores' have recently been successfully applied to predict gene functions and interactions between proteins as well as helped gene prioritization in QTL regions. Activities therefore also relate to Task 3 in WP12. Such methods would be an attractive and essential element of the future IVPB imagined. It is suggested that such approaches could be complementary to ontology-based analyses, but such suggestions should be tested before offered to the wider plant community as infrastructural tool.

DLO has started up collaborative talks with relevant experts (group Mons in Leiden/ Rotterdam) to see what such methods require and if such methods can be successfully applied to plant data. The various data sources and tools to use such approaches should be identified and mastered. They need to be integrated and quality of their performance on plant species-specific datasets needs to be assessed. We have defined several use cases to be explored and compiled a first list of potentially relevant plant literature resources, databases and plant ontologies, including taxonomy and chemical compounds. We are discussing with Wageningen University library to get access to their extensive plant thesaurus. This will allow assessing the value of such

approaches for annotation and prioritization of plant genes in breeding as step in the creation of IVPB.

Use of resources (*highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex 1*)

INRA: 5.4 PM

TGAC: 5.61 PM

BSC: 2.78 PM

DLO: 5.6 PM

3. Deliverables and milestones tables

Deliverables

Del. no. ¹	Deliverable name	WP no.	Lead beneficiary	Nature ²	Dissemination level ³	Delivery date (proj. month) ⁴	Actual delivery date	Comments
D7.1	A registry of plant genomic information	WP7	HMGU	P	PU	Month 9	31.5.2012	
D8.1	Datasets with associations available and integrated into visualisation interfaces	WP8	KN	R	PU	Month 12	7.9.2012	

Milestones

Milestone no	Milestone name	WP no	Lead beneficiary	Delivery date from Annex I	Achieved (Yes/No)	Actual / Forecast achievement date	Comments
MS1	Internal project website	WP1	EMBL-EBI	Month 6	Yes	7.9.2011	
MS5	Report on ELIXIR preparatory phase	WP2	EMBL-EBI	Month 12		31.8.2012	
MS7	A set of mandatory or optional fields associated with ontology formalism for descriptors	WP3	INRA	Month 12	Yes	28.7.2012	
MS9	1st transPLANT training workshop	WP4	HMGU	Month 12	No	13.11.2012	
MS12	DAS servers provided for sequence and annotation for 10	WP5	EMBL-EBI	Month 12	Yes	31.8.2012	

¹ Deliverable numbers in order of delivery dates. Please use the numbering convention <WP number>.<number of deliverable within that WP>. For example, deliverable 4.2 would be the second deliverable from work package 4.

² Please indicate the nature of the deliverable using one of the following codes:

R = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other

³ Please indicate the dissemination level using one of the following codes:

PU = Public

PP = Restricted to other programme participants (including the Commission Services).

RE = Restricted to a group specified by the consortium (including the Commission Services).

CO = Confidential, only for members of the consortium (including the Commission Services).

⁴ Measured in months from the project start date (month 1). Even though they should be available upon request at the indicated date, deliverables will be submitted to the Commission (and approved) at the time of the next following periodic report.

	reference genomes						
MS15	Initial public launch of transPLANT integrative portal	WP6	EMBL-EBI	Month 12	Yes	15.3.2012	
MS18	10 reference genomes incorporated in transPLANT hub and submitted to comparative analysis	WP7	EMBL-EBI	Month 12	Yes	31.8.2012	
MS24	Basic GWAS GUI available	WP10	GMI	Month 12	Yes	10.4.2012	
MS27	Implementation of reference-free methods for transcriptome variation analysis	WP12	TGAC	Month 12	Yes	31.8.2012	



PROJECT PERIODIC REPORT

Project management during the period

Grant Agreement number: 283496

Project acronym: transPLANT

Project title: Trans-national Infrastructure for Plant Genomic Science

Funding Scheme: Combination of CP & CSA

Date of latest version of Annex I against which the assessment will be made: 01.08.2011

Periodic report: 1st ☒ 2nd ☐ 3rd ☐ 4th ☐

Period covered: from 1.9.2011 to 31.08.2012

Name, title and organisation of the scientific representative of the project's coordinator:
Paul Kersey, Dr., EMBL-European Bioinformatics Institute

Tel: +44-(0)1223-494601

Fax: +44-(0)1223-494468

E-mail: pkersey@ebi.ac.uk

Project website address: <http://www.transplantdb.eu>

This management report covers the period from M1 to M12 (from 1.9.2011 to 31.8.2012).

1. Consortium management tasks and their achievement

The objectives of the management effort are:

- To coordinate the work programme, and the provision of public services.
- To manage strategic direction of the project
- To ensure good coordination with the European Commission

Management activities are a constant part of the implementation of the project.

a. Coordination of the Work Programme and Public Services.

Because of the distributed structure of the project, which is a combination of CP & CSA actions with 11 Partner institutions distributed across seven countries, the maintenance of good levels of **communication** is an important task. The project manager coordinates actions.

The main rhythm of the Project is set by **monthly phone teleconferences** chaired by the project manager. These are attended by at least one person from each group. The teleconferences are used to monitor the progress of the partners towards project milestones and the submission of the deliverables; the discussion of new projects and events; and to spread information (e.g. pertaining to workshops, administrative questions or scientific developments of interest.). The minutes of the teleconferences are written by the project manager and are available to all members on the internal website.

A general **mailing list** eu_plants@ebi.ac.uk has been established, to which all partners are subscribed.

The first **Annual General Meeting** (AGM) was organized at the European Bioinformatics Institute in Hinxton (UK) on November 17-18, 2011 and all consortium members attended. Each partner presented an overview of their future contribution with regard to the different work packages and the strategies for the first year of the project were discussed. The second AGM is currently in planning (for a scheduled date early in 2013). It is proposed inviting key collaborators from outside the consortium (including collaborators from outside the European Union) to attend (part of) the meeting to increase coordination with other international efforts.

Teleconferences and videoconferences (either organized by the coordinator or by the WP leaders using conventional teleconferencing or internet-based telephony) are used to coordinate the work within and between the work packages.

Helping collaboration across the Project. Given the large number of work packages (12), interactions between partners involved in different work packages are an important part of the project's management. Communication and collaborations are followed up and supported by the manager, in particular during the monthly teleconferences.

An access-restricted **internal website** is also maintained for the exchange of information restricted to partners. These include information relevant to the internal management of the project, and the exchange of data and documents among collaborating project partners. The latter is supported through the use of an easy (wiki-like) editing tool, Atlassian Confluence, of which extensive use is made within the project. A **bug tracker** (Atlassian JIRA) is used to report and track specific issues raised by partners.

Telephone conferences opened by the project manager on specific topics within transPLANT also support collaborations within the work packages.

Co-operation with other projects/programmes

- The work package 2 is dedicated to « Interaction with national and trans-national genomics and informatics activities». Actions are reported in the corresponding WP report: Communication events allowed introducing the activity of transplant to the scientific community; Interaction with ESFRI research infrastructure programs, in particular the interaction with the project ELIXIR.
- We have approached the “data infrastructure for agriculture” AGINFRA project (<http://aginfra.eu/>) for discussions about potential collaborations as both projects mature.
- Internationally, we are contributing to the plant-working group of the EU-US task force of Biotechnology Research, which is co-chaired by project partner Klaus Mayer of HMGU. EMBL-EBI is additionally collaborating closely with the NSF-funded Gramene project in the US to ensure the development of interoperable resources on both sides of the Atlantic.
- In the context of the work package 3, transPLANT partners are participating in the efforts of the Plant Ontology Consortium (coordinating Plant Ontology and Gramene Trait Ontology).

In general, transPLANT partners are trying to help coordinate European- and US-based initiatives in our domain.

b. Managing the strategic direction of the project

The coordinator is assisted in the strategic direction of the project, through *ad hoc* meetings of a **Strategy Committee**, which exists to consider strategic and high-level management decisions and to make recommendations to the full consortium.

The committee is composed of the project coordinator and three project participants: Klaus Mayer (HMGU), Hadi Quesneville (INRA), Pawel Krajewski (PAS), each of whom lead both a Coordination and an RTD work package, and who thus have direct involvement across the full spectrum of project activities (all partners are involved in the service (OTHER) activities of the project. In 2011, the Strategy Committee met in person at the Annual General Meeting (November 17-18, 2011 at Hinxton, UK) and communicated via teleconference on May 9, 2012. Minutes of strategy committee meetings are taken by the project manager and are communicated to the full consortium via the internal project website.

c. Reporting to the EU

During the period M1 – M12, two deliverables have been submitted to the Project Officer. The achievement of milestones is recorded on the consortium’s internal website.

2. Changes in the consortium

No contract amendment has been requested so far.

3. List of internal project meetings

The first **Annual General Meeting** (AGM) was organized at the European Bioinformatics Institute in Hinxton (UK) on November 17-18, 2011 (see above).

4. Development of the project website

A public website has been developed and hosted by the partner EBI, coordinator of the project. The URL is www.dbtransplant.eu. The release of the public portal was achieved on March 2012, and reported as the milestone MS15.

5. Project planning and status

As of month 12, the two deliverables that were due for submission have been duly submitted. 8 project's milestones due within the first 12 months have been reached. An additional milestone was planned for the first period: the organisation of the first training course (MS9). The training meeting has been organised, advertised, but for practical reasons (to maximise the availability of trainers and potential attendees) it has been scheduled for November 2012 (instead of August 2012 as originally planned). No adverse effects on the future development of the project are expected.

The Gantt chart presented below summarises the work performed since the project's start.

