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: 01.08.2012

Periodic report: **1st** □ **2nd** ⛀ **3rd** □ **4th** □

Period covered: **from** 1.9.2012 **to** 31.08.2013

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: 01.08.2012

Periodic report: **1st** □ **2nd** ⛀ **3rd** □ **4th** □

Period covered: **from** 1.9.2012 **to** 31.08.2013

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

Between month 13 and 24, all 12 work-packages continued their activities. All, 8 planned milestones were reached, and all 10 deliverables due within the reporting period were submitted to the European Commission,.

The second milestone (**MS2)** of Work package 1 "Management" was reached with the submission of the first annual report on time (month 14).

In work package 2 "Interaction with national and trans-national genomics and informatics activities", the partners are coordinating the development of the project with other national

and international plant genomics projects, and with parallel projects in other domains. We have undertaken a survey of potential stakeholders in plant genomics infrastructure, and are currently in the process of evaluation of the results. We have additionally held a stakeholder meeting, , co-organized with the plant science working group of the EU-US Task Force on Biotechnology Research, and the Gramene project, bringing together 40 invited participants from 10 nations to discuss themes around the topic "Genomes to Germplasm". We are currently in the progress, in partnership with the participants at the meeting and the wider community, in preparing a paper for publication reporting on the findings of the meeting and the perception of future needs, which will be used for subsequent engagement with funding bodies. A report on these activities (**D2.1**) has been delivered to the European Commission.

In work package 3 "Community standards for the interoperability of data resources", we have agreed the use of a set of ontology standards and a data format for the exchange of data within the consortium, and a draft "Minimal Information About a Phenotypic Experiment" Standard has been drawn up. Our goal is to work around the development and application of existing standards where possible; and we have been in communication with other standardisation and ontology development initiatives in Europe and the United States to this effect. We have commenced the process of wider community consultation about the proposed standards for phenotyping with other interested groups. Two deliverables: **D3.1** - Recommended ontology set for use in phenotype description and epigenetic variability and **D3.2** – Format specifications for data exchange by flat file and web services – have been produced.

In work package 4 "User Training", a training course was hosted in November 2012 by the partner INRA in Versailles (France), (**milestone MS9)**. The training focused on data resources and tools for Triticeae species (primarily wheat and barley), and included tutorials on data access and use from INRA, EBI-EMBL, IPK and HGMU, and was targeted at (experimental) biologists, breeders, and bio-informaticians. A second training course was hosted by the partner PAS in Poznań, Poland, 27-28 June 2013 (**milestone MS10**). The workshop focused on plant genomics resources with two foci: (i) Triticeae (again, due to high demand for the first workshop) (ii) standardisation and annotation of plant phenotypic data. The target audience was experimental biologists and plant breeders without prior informatics knowledge. A total of 65 people have attended the workshops held to date. Further workshops are in planning for next year.

In work package 5 "Programmatic services for genome-scale data", we have provided DAS servers provided DAS servers for sequence and annotation of reference plant genomes (**milestone MS13**), allowing this data to be shared with other resources (inside and outside the consortium).

We have delivered BioMart data warehouses for data from EBI and INRA, using both the existing version 0.7 software and the recently released version 0.8 software, and have explored alternative data warehousing technologies that may offer better longer term prospects as data volumes continue to increase (**D5.1**). In addition, we have made substantial progress in the development of high-performance and cloud computing platforms, with a specific focus on critical tools utilised for genomic analysis, including those developed/analysed in WP12 (which focuses on tool development).

In work package 6 "A virtual European Plant Genomics Database", we have reworked the project website, separating out community-facing features from internal project features. New tools for the submission of variation and analysis of variation data have been embedded in the portal. 9 resources from 5 partners have been incorporated in the transPLANT portal through an integrated search mechanism, which has been improved through the deployment of faceting over an agreed common data model (**milestone MS16**).

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 "A repository for reference genome and annotation", we have established a registry of plant genomics databases and services. The Ensembl Plants functions as a central hub resource within the transPLANT infrastructure and we have accommodated data from an additional 6 genomes into the resource and submitted these to comparative analysis, taking the number of genomes added since the start of the project to 16 (milestone **MS19**) and the total number available to 25.

In work package 8, transPLANT partners have integrated genetic marker data from wheat, barley, maize and oak; and have been working on developing analytic approaches for identifying structural variants from sequence read data. Progress for milestones and deliverables due later in the project is on course.

In work package 9 "An archive of plant genomic variation", a first implementation of the variation repository has been completed and publicly advertised on the transPLANT website, and the first data has been accessioned. (**D9.1 and milestone MS23**). We will now work, at first primarily with collaborators but increasingly with the wider community, to archive, integrate and provide access to plant variation over the remainder of the project.

Work package 10 deals with "Tools for elucidating the genotype-phenotype map". In this context, we have continued to develop new methods and tools for data analysis. Several papers have been published, some tools have been released, and further tool development is in progress. Two deliverables Statistical descriptors for genotype-phenotype map construction (**D10.1**) and Software for analysis of genome-wide association data (**D10.2**) have been completed.

In work package 11 "Metadata-driven information retrieval systems", the LAILAPS search interface has been released (**milestone MS25**) and trained (**D11.1**). Work will now continue on integrating this interface within the transPLANT portal, based on a new portable Drupal module currently in development.

Finally, in work package 12 "Implementation of resource-intensive algorithms for plant genomics data", statistical methods to model variation were developed and tested by the partner INRA (**D12.1**). INRA also delivered software for the analysis of repeats in plant genomes (**milestone MS28**). Assessment of assembly algorithms has been performed and potential optimisations have been explored, to identify programs particularly suitable for deployment within a high-performance compute environment in development in work package 5.

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 (WP7 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 to 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: 01.08.2012

Periodic report: 1st □ 2nd ⛀ 3rd □ 4th □

Period covered: from 1.9.2012 **to** 31.08.2013

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

The primary objectives for the reporting period were the achievement of 10 deliverables and 8 milestones, pursuing the development of the project.

MS2: 1st annual report submitted (EMBL) due 31.10.2012 (work package 1).
The first annual report submitted to the Commission and to the panel of external reviewers in preparation to the first Commission's review in Brussels.

In work package 2 (Interaction with national and trans-national genomics and informatics activities), there are no milestones due in the second period, but one deliverable D2.1: A report entitled "Translational research for agronomical application" (due date shifted from M18 to M24 with the agreement of the project officer). This will report on a EU-US communities' discussion to identify infrastructure needs for translational application of plant science research, and on the results of a community survey in this thematic.

In work package 3 (Community standards for the interoperability of data resources), no milestones were planned for the second period, but two deliverables due on 31.8.2013: D3.1 - Recommended ontology set for use in phenotype description and epigenetic variability; D3.2 - Format specifications for data exchange by flat file and web services.

MS9: 1st transPLANT training workshop (HMGU) due 31.8.2012 (work package 4).
MS10: 2nd transPLANT training workshop (EMBL) due 31.8.2013 (work package 4).
MS9 is reported here because its delivery was shifted one month after the end of the first period (November 2012) due to organisational reasons. The workshops 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.

MS13: DAS servers provided for sequence and annotation for 15 reference genomes (EMBL) due 31.8.2013 (work package 5).
We continue the implementation of web services to enable distributed computing. For genome "features", transPLANT will provide DAS servers for resources for plant-centric data. In this work package, deliverable D5.1 will provide updated data warehouses developed for genomic annotation and variation data.

MS16: Ensembl Plants, MIPS Plants DB and GnpIS integrated in transPLANT portal (EMBL) due 31.8.2013 (work package 6).
The transPLANT portal is 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).

MS19: 15 reference genomes incorporated in transplant hub and submitted to comparative analysis (EMBL) due 31.8.2013 (work package 7).
This work package aims at building a repository for reference genome and annotation. We progressively incorporate reference genomes in order to make them available for comparative analysis. Work package 7 will deliver D7.2, a DAS-based server/client interface for integrating omics data.

MS23: Initial public launch of variation repository (EMBL) due 31.8.2013 (work package 9).
This work includes developing a procedure for whole genome-alignment between releases and projection of variant features between releases. The repository will be described in the deliverable D9.1.

In work package 10 (Tools for elucidating the genotype-phenotype map), no milestones were planned for the second period, but two deliverables due on 31.8.2013: D10.1 – Statistical descriptors for genotype-phenotype map construction led by the partner PAS; D10.2 - Software for analysis of genome-wide association data led by the partner GMI.

MS25: Software core released (IPK) due 28.2.2013 (work package 11).
The focus of this milestone is to provide a query system for genomics meta-data, like functional annotation of genes or other genomics regions. The release and training of the search engine software will be described in the deliverable D11.1.

MS28: Software for the analysis of repeats (INRA) due 31.8.2013 (work package 12).
The first software developed in this work package will focus on the analysis of repeats in plant genomes. This will be complemented by the development and testing of statistical methods to model variation (D12.1).

2. Work progress and achievements during the period

The report covers the period from month 13 to month 24. The project contains 12 work packages, and activity has continued in each of these.

The work programme has been progressing in line with the intended plans. Ten due deliverables have been submitted on time to the project officer. Eight milestones have been reached as expected. The work package reports (section 3, below) describe in detail the progress made (see also the publishable summary).

At the first review, it was requested that the partners propose indicators of impact against which the progression of the project could be measured. We propose that the project's impact should be assessed in the following areas:

- Coordination/support:
  - Community engagement
    - Awareness of transPLANT as a vehicle for addressing (longer-term) infrastructural needs
  - Standards
    - Phenotyping standards paper published with global collaborators
  - Training
    - Over 100 scientists trained in transPLANT partner resources
- Services
  - Standardised data made available over transPLANT partner and community-generated services
  - Improved, interoperable programmatic access for partner resources
  - Increased usage of transPLANT-funded services
- Research and Technical Development
  - New variation archive online
    - Archive receiving submissions from outside the consortium
    - International collaboration initiated (e.g. with NCBI)
  - High impact research papers published

Summary of progress in the context of impact indicators.

- Ontologies and file formats to describe the results of phenotypic experiments have been drawn up, and we are in discussion with other initiatives to ensure compatibility of efforts. Work on a joint publication with the European Plant Phenotyping Network is aimed at maximising the community both contributing to standards development, and made aware of current recommendations.
- The training programme is underway. 66 scientists have been trained at the first two (of five planned) transPLANT-organised training events (additionally, transPLANT has been presented, and transPLANT funded resources taught, at other, externally-organised events).
- transPLANT partner resources have been improved, and made more inter-operable, and more accessible to programmatic access (see reports of work packages 5, 7).
- The first version of the transPLANT variation archive has been developed and made public.

A number activities conduced in the second reporting period have helped build awareness of transPLANT. These include (i) the training activities (ii) the stakeholder meeting (iii) the user survey (iv) the establishment of public mailing list, transPLANT announce (v) engagement with database groups with regards to maintenance of content in the transPLANT data resources registry (vi) First connection with smaller communities including those working on oak and rye. Other developments prepare the way for new lines of engagement,

including (i) the establishment of cross-domain search using an open architecture, permitting the inclusion of external resources alongside the transPLANT partner services, (ii) the reorganization of the website to separate the public-facing component from internal project description (iii) the opening of the transPLANT variation archive.
Over the course of year 3, project partners will (i) analyse and respond to survey results (ii) seek to publish the report from the stakeholder meeting in an appropriate journal (iii) advertise transPLANT services (e.g. the variation archive) and possibility for direct engagement (e.g. through incorporation in search engine) widely in the scientific community (iv) further our engagement with smaller communities, extending beyond the transPLANT collaborative network (v) continue our programme of training activities. Through these activities, we aim to grow the transPLANT user community and maximize the impact of the project.

During the second period, 11 research papers have been published acknowledging transPLANT:

1. M. Lange, J. Chen, and U. Scholz (2012) Information Retrieval in Life Sciences: The LAILAPS Search Engine. In U. Goltz, M. Magnor, H.-J. Appelrath, H. K. Matthies, W.-T. Balke, and L. Wolf, editors, INFORMATIK 2012, 42. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 16.–21.09, Braunschweig, Germany, Lecture Notes in Informatics (LNI), volume P-208: 1552–1558.
2. Korte, A. et al. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44, 1066–1071.
3. Arend, D. and Lange, M. and Colmsee, C. and Flemming, S. and Chen, J. and Scholz, U. Information Retrieval in Life Sciences: The e!DAL JAVA-API: Store, Share and Cite Primary Data in Life Sciences. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 4-7 October 2012, Philadelphia, U.S.A., pages 511-515, 2012. DOI: 10.5447/IPK/2012/13.
4. Brenchley, R. et al. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491, 705–710 (29 November 2012) doi:10.1038/nature11650.
5. Thomas Nussbaumer; Mihaela M. Martis; Stephan K. Roessner; Matthias Pfeifer; Kai C. Bader; Sapna Sharma; Heidrun Gundlach; Manuel Spannagl (2012) MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Research 2012; doi: 10.1093/nar/gks1153.
6. Seren U, Vilhjálmsson BJ, Horton MW, Meng D, Forai P, Huang YS, Long Q, Segura V, Nordborg M. (2012) GWAPP: A Web Application for Genome-Wide Association Mapping in Arabidopsis. Plant Cell 24:4793-4805.
7. Radivojac P. et al. (2013) A large-scale evaluation of computational protein function prediction. Nat Methods. 2013 Jan 27. doi: 10.1038/nmeth.2340.
8. H. Mehlhorn, M. Lange, U. Scholz, and F. Schreiber. (2013) Extraction and prediction of biomedical database identifier using neural networks towards data network construction. In P. Ordez de Pablos, M. D. Lytras, and R. Tennyson, editors, Cases on Open-Linked Data and Semantic Web Applications, pages 58–83. Information Science Reference (an imprint of IGI Global), 2013. doi = {10.4018/978-1-4666-2827-4}, isbn = {781466628274}.
9. Steinbach D, Alaux M, Amselem J, Choisne N, Durand S, Flores R, Keliet AO, Kimmel E, Lapalu N, Luyten I, Michotey C, Mohellibi N, Pommier C, Reboux S, Valdenaire D, Verdelet D, Quesneville H. (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database (Oxford). 2013 Aug 19;2013:bat058. doi: 10.1093/database/bat058.
10. Long, Q. *et al.* (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* **45,** 884–890 (2013).

11. Lange M, Henkel R, Müller W, Waltemath D, Weise S (2013) *Information Retrieval in Life Sciences: a programmatic survey*. In: Chen M, Hofestädt R (Eds.): Approaches in Integrative Bioinformatics – Towards Virtual Cell. Springer, in press.

A total of 14 publications have now been reported since the project start.

–

The following reports describe the activities undertaken in the 12 work packages during the reporting period in more detail (and in the context of plan outlined in the description of work).

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 organize 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 on-going 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. TriticeaeGenome, 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 glaberima*, 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 meetings 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 ELXIR 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 and bioinformatics infrastructures.

The deliverables of the two tasks are three reports: one for task 1 (D2.1), provisionally entitled 'Transnational research for agronomical application', lead by INRA, another one for Task 2 (D2.2), lead by EBI, provisionally entitled 'Future developments in IT infrastructure for plant science; levering synergies' and a final report (D2.3) on diffusion activities on concurrent projects, lead by IPG.

Task 1: Interactions with national plant research initiatives.

Actions:

Stakeholder meetings

It was 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 future funding applications. Each meeting is planned to 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.

The planned focus of the first stakeholder meeting was translating basic plant science to agronomical application. Specifically, an agenda was drawn up around the theme of "Genomes to germplasm", following discussion within the project and with external collaborators. The meeting was co-organized these with the Plant Biology Working Group of the EU-US Task Force on Biotechnology Research. The Task Force is an organizational structure that accommodates funding agencies and scientists from Europe and North America with the goal of coordinating research and collaboration and the funding programmes supporting it. The Task Force has previously organized meetings in the plant genomics area e.g. at Hinxton, United Kingdom, in December 2009, and the joint organisation provided the opportunity to bring increased numbers of U.S. participants (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 direct attention of funding agencies. Further funding for US participation, and co-funding of the costs of professional meeting organization, were provided through additionally co-organizing together with the Gramene project, which maintains plant informatics resources in the United States. Together with the same co-organizers, we are currently in the early stages of preparing a related meeting, whose immediate focus will be the capture, coordination and analysis of phenotype data, to be held in the United States during 2014.

Results of year 2

Task 1

<u>Stakeholder meetings</u>

The 1st transPLANT external stakeholder meeting called "Genomes to Germplasm" was co-organized by the Plant Bioinformatics Working Group of the EC-US Task Force on Biotechnology Research, the transPLANT project, and the Gramene project.

This meeting brought together 40 research scientists, informaticians and crop breeders from Europe

(23), USA (16) and 1 from Philippines. They address the biological and informatics challenges associated with our attempts to catalogue genomic variation and apply it to increase our understanding of plant biology to improve crop plants.

Specific points of discussion included (i) how natural variation is being sampled, and the likely future applications of this data (ii) informatics needs and solutions: what infrastructure and data standards are available, and what components are missing or underdeveloped, particularly in the context of globally distributed activities (iii) connections between germplasm resources and genomic databases and (iv) tools needed to practically apply these data for the purposes of plant breeding and crop development.

Methods:

The meeting held at Versailles in the INRA campus from February 28th to March 2nd 2013. It has been professionally facilitated and designed to ensure that the discussions are shaped by the ideas of the participants.

We shared ideas and thoughts in advance and during the meeting on a web site. We used a social networking platform called Ning. It is a closed network and has been customized to suit our needs. It also allows to share documents and to engage discussions on a forum restricted to people of the meeting.

At the beginning, the organizers (Paul Kersey, Klaus Mayer, Hadi Quesneville, and Doreen Ware) have sketched some ideas about likely areas of interest and have recommended some papers that might be read in advance of the meeting.

During the meeting, small groups were formed to discuss specific topics. The results of their discussions where then presented to the whole audience for further discussions. New groups were formed according to the topics rising from these discussions. At the end of this process, ideas and thoughts were shared on the Ning web site. We drafted this report from the material found there.

Meeting results:

The results of the discussion during the meeting were summarized on a social networking site, and can be visualized here (http://genomes2germplasm.ning.com/). A summary of this information has been presented below.

1 Introduction

Plant species play a critical role in life on earth, transforming the sun's energy into biological materials that provide humans with food, fuel, and bio-active compounds. The explosive growth in the human population experienced over recent decades has only been possible because of advances in agricultural technique but also in crop breeding, which has delivered new varieties with significantly higher yields and improved resistance to biotic and abiotic stress. But the world's population is continuing to grow, with the present population of about 7 billion, predicted to rise to between 8 and 16 billion by 2100 according United Nations estimates (http://esa.un.org/unpd/wpp/ExcelData/DB01_Period_Indicators/WPP2010_DB1_F04_BIRTHS_BOTH_SEXES.XLS). This will occur in the context of accelerating environmental changes that will alter the suitability of land for agricultural purposes and reduce the total number of cultivable areas, and increased competition for the land that remains. Plant-based sources are increasingly used as replacements for fuels and chemicals for declining mineral resources. If these challenges are to be surmounted without massive human misery, there is a strong need for the accelerated development of better crops, with higher yields, better suited to their environments, and capable of being deployed rapidly to meet with shifting patterns of environmental stress.

More prosaically, crop breeding must deliver more quickly new varieties that are adapted to a changing world. Critical to this, is the better characterization of the germplasm, not only of existing elite varieties, but also wider genetic resources and wild relatives of crop plants, which contain the genetic

material from which new varieties will be developed. Fortunately, the ongoing development of new technologies for high throughput genotyping, and high throughput (and high precision) phenotyping, make this a feasible goal. But, much is still to be done in fundamental researchers, particularly in regard of the modelling of genetic and environmental interactions. Reductionist analyses will need to be combined in systems-wide approaches if we are to understand specific ecophysiologies and determine the best (actual and possible) crops for specific geographical locations.

2 Basic science challenges

The basic science challenges that need to be addressed within a 5 to 10 years framework are:

- Functional classification of plant genes. Today we have still an incomplete realization of this goal even after a decade of work on the model species *Arabidopsis thaliana*. High throughput genotyping and phenotyping technologies should be able to finally enable this.
- Predictive plant biology should focus on how to predict which germplasm will perform "best" for a given environment. This would request (i) to generate a complete inventory of plant genetic and phenotypic diversity, (ii) to characterize the plant microbiome, and (iii) to understanding a plant in terms of its local ecosystem.
- We should be able to engineer plant biology to fulfill specific goals. Predictive modeling should have a positive impact on understanding plant phenotypes. Hence plant improvements should result of genetic improvement, but also modification of abiotic environments or biotic conditions, such as modifying the microbial communities to affect specific outcomes for a genotype.

We need to increase the possibilities for direct work in crops, but plant models are still important for cost-effective development of basic science (e.g. for synthetic engineering, long term development potential).

3 Translational biology

Translational biology will benefit from basic sciences translation. Challenges for a more "applied" science in a 5 to 10 years future will be:

- Improve plant varieties quicker, improving breeding methods but also by direct genetic editing (i.e. GMO). Traditionally, molecular biology has been too expensive for breeders to be interested in, but costs are falling. Long-term crop improvement is dependent on a complete science/application stack, from basic biological research to field phenotyping. For plant improvement, there is a need for understanding at the cellular, organismal and population level yield but also quality. A better characterization of the germplasm (elite varieties, genetic resources, wild relatives…) through high throughput / high precision genotyping and phenotyping capacities to model GxE interactions, is a mean to deliver more quickly new varieties that are adapted to a changing world.
- Synthetic biology should represent new markets for plants (secondary metabolites, biotechnological applications such as phytoremediation). Plant metabolites are already enormously important for drug development. A more ambitious goal is to re-engineer plant physiology (e.g. new organs for the storage of products closer to consumption products).
- Maintaining biodiversity of existing and future crop species to facilitate introgression, new domestication, and a diversification of agriculture (vegetables, energy, forestry, …). We should be able to leverage the benefits developed first for high-value crops to niche crops, making new technologies commercially viable across more crops, more markets and more breeders.

4. Vision of the future: An Information-Enabled Environment for primary research and translation in plant biology

Biology has become, in the last two decades, an information science. High-throughput technologies have been increasingly used to catalogue the natures of living systems and to assay for their occurrences and behaviours. The result has been an enormous growth in the amount of biological data available for the twin purposes of understanding life and applying this knowledge for human benefit. Yet the yields of these developments lag behind the rate of narrow technological development. In part, this is because researchers are still exploring the potential of larger and deeper data sets than were hitherto available. But the huge size and great complexity of the data itself poses challenges for organization, analysis and insight. These challenges are made more acute by the relatively low costs of the experimental apparatus, which has led to a more dispersed approach to data generation than seen in other data-intensive fields (e.g. high energy physics). While this democratic approach is in it highly welcome, putting new tools in the hands of the entire scientific community, there are associated difficulties. Although costs are continually falling, this has opened up possibilities of more extensive sampling, while the challenges of the custodianship and sharing of biological materials may be reducible but are unlikely to disappear. Effective integration of the data produced through scientific and fieldwork is likely to remain essential for the foreseeable future.

How do we facilitate the application of new technologies to support the faster development of crop plants? Our vision is that this depends on the dynamic, large-scale integration of relevant data, transformed into information and delivered through usable tools into the hands of basic scientists, systems modellers and ultimately plant breeders, for whom this new knowledge will become the central building blocks of their craft. The main problems to solve are scale and variety of data.

4.1 Enabling technology (infrastructure)

Through a better integration of genomic data, breeding should be more efficient as well as all biotechnologies such as molecular breeding, transgenesis, and mutagenesis. Possible solutions and associated challenges covers data integration, data sharing and other less-specific topics.

In the age of big data, classical “libraries” on knowledge need to be replaced by query able, digital archives. A dictionary of plant life should appear as data warehouses for genomic information. We should move then to data-driven science where every biologist has tools and skills to work as a bioinformatician. Some questions are plant specific but most are shared questions with other domains. We should then re-use solutions found by other communities (e.g. environmental researchers, human health researchers, and so on).

Scale has changed and mode of operation has changed as well. Existing data models and standards for their use exist in some form but are currently insufficient. There is a need for decoupling knowledge stewardship from service provision. Core data objects (universal, stable) need to be centrally archived and managed by dedicated entities with informatics skills adapted to big data management. Experiment-specific data (e.g. phenotypic, epigenetic) might be more transitory and local.

In that context, data discovery should be promoted allowing searching heterogeneous data varying in their standardization (from raw information to ontology structured information). Queries in natural languages should be possible. In parallel, mechanism for showing and improving data quality should be developed using user feedbacks, but automation of data capture is essential for scalability. Hence, comprehensive access to phenotypes under different environments should be possible to gain a comprehensive understanding of these experiments (GxExP) and to enable breeding by design and/or predictive biology.

Technologies to support distributed development need to be widely adopted. There is a need for shared API for federated databases such as different resource providers implement the same API. Toolkits will be available for implementation of services, but test suite for API must be also available to make

sure the API works. Ontologies must be developed as they represent important keys to query databases. Dynamic data exchange between databases must be also facilitated and encouraged to see the same data integrated in various context.

Challenges and obstacles are data ownership, data access, IP restrictions, data versioning of large data collections (very difficult in a centralized database), the need for sustainable funding, cultural and commercial barriers to share data, sustainability of tool development. The role of private companies in the schema is central as they are potentially big data providers and consumers. What public services will, commercial entities, need and expect has to be determined as well as the appropriate interfaces between public and private research, should be clarified to solve most of the obstacle.

4.2 Tools

Tools must improve their ability to model agricultural and environment management practices. Tools implementing methods that will optimize integration with biological and environmental data would improve prediction of phenotypes. Integration of systems biology information will help to make breeding decisions.

New computational methods should be available to a broad user community more quickly and reliably. We should develop reliable and tested software. Testing needs to include evaluation of scalability (usability for very large data sets). Software development and distribution support for new methods, should possibly be centralized and community supported.

5 Implementation in specific areas

A cyber infrastructure should support the general challenges described in “Science” and “Translation”. They are here translated in specific scientific areas.

5.1 Germplasm: Seeding the data, development of resources for germplasm

We should move toward a full genetic and increasing phenotypic characterization of all well-defined plant strains. Ideally we should have genome sequences of all crops and wild relatives, and access to haplotype structures for breeding by design. We need then collaborative characterization of large panels of genetic resources (passport or research-derived), exchange deregulation of genetic resources for research purposes, and common sets of materials that can be broadly used for crossing and evaluation across sites.

5.2 Genomes and Epigenomes

Current models developed for humans are not powerful enough for plant genome complexity. In essence, a genotype is fixed for a stock and the set of genotypic information is finite and universal. A model should be to develop a universal catalogue of genomic variation, with distributed custodianship (and possibly implementation). For breeding and research purpose, the genomes from same population will need to be projected to the same physical reference and/or genetic maps. Some solution exists, but it is time to build a community standard.

Epigenomes can be seen more as phenotypes, as they are not finite and universal, because depending on development stage, cell type, and environments.

5.3 Phenotyping

A new face of phenotyping is appearing. A phenotype could be both an attribute and a measurement (e.g. metabolomics is one form of molecular phenotyping). Automated phenotyping and field phenotyping are different things. Field phenotyping is geographically distributed by definition. Unlike genotype data, a stock may be phenotyped any number of times, and the results may be specific to local conditions. Reference phenotyping (standard assays and conditions), might be undertaken by stock centers or large scientific projects, but there will be a persistent need for additional phenotyping activities thereafter. A lot of phenotyping will need to be done locally as it is easier to transfer knowledge than germplasm because of regulatory barriers.

We need to have a distributed, dynamically accessible data sharing model with components connected by common interfaces. Open APIs would allow interactivity to, for example, genomic repositories, GIS system for climate monitoring, metagenomic (inc. microbial) data, epigenetic, or phenotypic archives (standard reference phenotypes). This requires the use of controlled vocabularies.

5.4 Systems modelling and Predictive Biology

To take the full advantage of the predictive power to predict phenotype from genotype and environment, we need to study the genetics, physiology, biochemistry, …, of response to biotic and abiotic environments. The strategy is to decipher the biology of traits elaboration by crossing diverse levels of information on reference populations.
To achieve this goal, we need to be able perform complex molecular phenotyping of core set of plants with diverse germplasms. Measurements should include expression profiling, metabolomics, methyl-seq, etc ... These data should be integrated with genotypes to develop predictive models.

5.5 Crop Breeding

Genomic selection, genetic and epigenetics information need to be integrated in the information systems. Fast pre-screening system for genetic resources will speed up generation cycles. Methods must be high throughput and cost-efficient. Indeed, sequencing costs must still drop for applying genomic selection to minor crops. Identify genes controlling recombination should allow to increase recombination to access genes in low recombination regions.

6 Teaching and Training

As the need develops for all biologists to become (at-least part-time) bioinformaticians, the acute shortage of training capacity is becoming a serious problem. Increasing numbers of scientists and breeders are needed with a broad-based skill-base encompassing molecular and field biology, statistics and computer science. Yet even within the relevant domains, there is sometimes resistance to, or ignorance of the potential of new technology, as well as rational skepticism/insight based on deep domain knowledge. Deeper and wider channels of communication are required, strengthening ties between technologists, molecular biologists, germplasm collections, and breeders. E-learning methods should be developed to face the challenge.

7 Connecting to the wider Community

A major challenge faced by all scientists, including the plant genomics community, is popular suspicion and distrust of their work. While this may result from genuine dislike of the purposes to which scientific knowledge is applied or from rational skepticism about the scientific claims, it can also result from ignorance, miss-information and preference for the superficially “natural”, even when that term cannot be defined in any meaningful way. These problems are clearly highlighted in the context of the debate over genetically modified foods: while risk assessment is innately hard even for professionals, it may be that public ignorance has contributed to the intensity of some of the opposition to the development of this technology.

New way to communicate with practicing professional, policy makers, and the public has to be invented to improve the positive impact of plant research on the public and private policy makers, managers, but also change the acceptance of scientific results by broadening awareness and receiving more public support

Perspectives:
D2.1 (A report entitled “Translational research for agronomical application”) has been delivered. The expected outcome of this report is a journal paper, and a document intended to form a potential basis for future coordinated funding calls between the European Union and the United States. This is currently being drafted by the meeting organisers, and will be further developed in consultation with the meeting attendees, and additional members of the global community, before submission for publication.

A new user survey

Following the recommendation of the 2012 project review, a new, larger survey on "transPLANT User Needs" was initiated to collect the bioinformatics stakeholders' needs in the field of agronomical research. The goal of this survey is to identify potential needs that are not already covered by the TransPLANT project and to help drawing the landscape of possible overlaps with other projects, in order to better coordinate developments and avoid redundancies.

This survey was made accessible on the transPLANT web site: http://www.transplantdb.eu/survey and on the URGI web site. It was sent for dissemination to transPLANT partners, transPLANT projects collaborators, and transPLANT partners' networks. It is addressed to both scientists from academic and private sectors, working on wheat, barley, maize, pea, sunflower, rapeseed genomics and genetics.

The survey contains 41 questions, grouped by sections. The first section (Q1 to Q7) gets information on the person answering the survey. The second (Q8-Q20) gets information on the data that the user is manipulating and analyzing, the storage needed, the submission process to database repositories, the data types, the data to be shared, and the required queries. The third section (Q21-Q33) concerns the tools used to visualize the data: what is used, what is missing, what are the difficulties, what are the needs in terms of tools and computing resources. The last section (Q34-Q41) asks questions about existing projects in which the user is involved and his expectations about the outcomes of the transPLANT project.

The survey has been online since the 6th June 2013 and distributed to different user networks. Seventy persons have already responded. The survey will remain open until late autumn 2013.

Perspectives:
The results of the survey will be analyzed, and a report prepared on this analysis.

Specific interaction with national genomics projects

The group at IPG PAS coordinates data processing for two Polish genomic projects: POLAPGEN-BD investigating biotechnological tools for obtaining cereal genotypes tolerant to drought and GENSEK investigating genetic and genomic determination of rye resistance to pathogens. In both projects we actively promote application of generally accepted standards, and the ones developed within transPLANT, by presentation at project meetings and proposing data collection and storage solutions. As an example, we have supervised dataset submissions to databases (Metabolights, in ISA-TAB format).

Task 2: Interaction with ESFRI research infrastructure programs.

The ESFRI programme ELIXIR will establish a pan-European research infrastructure for all biological information, with implications for the long-term development of resources for the plant sciences as in other fields. 15 countries (as well as the coordinating international organisation, EMBL) have signed a memorandum of understanding and will move towards the formal establishment of ELIXIR as an independent legal entity. The ELIXIR Consortium Agreement (ECA), which provides the legal basis for ELIXIR, has been approved and will come into effect once ratified by EMBL and 5 countries. This is expected in the coming months. To date, EMBL, the UK and Sweden have already ratified. ELIXIR will be launched as a separate legal entity in Brussels on 18 December.

ELIXIR's five-year programme is currently being developed and this will be presented to the Board for first discussion in November. ELIXIR's founding director, Niklas Blomberg, has been appointed and the ELIXIR Hub staff now numbers 8 members of staff. Recruitment is currently underway for a Chief Technical Officer. Five ELIXIR Pilot Actions (http://www.elixir-europe.org/system/files/documents/elixir_pilot_points_trifold_without_bleed_0.pdf) have taken place

and Nodes have submitted proposals for the second wave of Pilots, which will commence in 2014.

The 15 Members have all submitted Node applications and these have been reviewed by ELIXIR's independent SAB. EMBL-EBI is the largest ELIXIR Node. Other ELIXIR Nodes also have, or are developing, bioinformatics services relating to agriculture, for example; the Portugal ELIXIR Node has a focus on woody plants; the Netherlands ELIXIR Node has interest in agriculture; and the ELIXIR Italy Node also provides services of relevance to agriculture. National Nodes will formally become ELIXIR Nodes once the ECA has been ratified and after signature of the Collaboration Agreement between the Node and the ELIXIR Hub. This is expected in 2014.

EMBL-EBI, as coordinator of both ELIXIR and transPLANT, is continuing to keep transPLANT partners informed on developments within ELIXIR; and is opening discussions with the ELIXIR nodes with a plant science scope aimed at ensuring developments are complimentary to the framework established in transPLANT, and that transPLANT data is made available to all users of ELIXIR services. These dialogues will intensify as the nodes become operational.

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

EMBL-EBI: 2.5 person months (reaching 50% of the total planned)

HMGU: 1 person month (reaching 20% of the total planned)

IPK: 0.25 person months (reaching 17% of the total planned)

INRA: 8.02 person months (reaching 150% of the total planned)

IGR-PAN: 3 person months (reaching 29% of the total planned)

DLO: 0.02 person months (reaching 1% of the total planned)

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 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://marinemetadata.org/references/oboeontology). The community gathered around the INRA Ephesis project (http://urgi.versailles.inra.fr/index.php/urgi/Projects/URGI-softwares/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 an 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 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. This is an essential requirement if users are to have the capacity to analyze 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 deliverables:

D3.1. Recommended ontology set for use in phenotype description and epigenetic variability (completed, details submitted as report on D3.1)

The process of standardisation of the annotation of biological information must cover practices, formats and vocabularies. Use of controlled, semantically structured vocabularies (ontologies) can support quality control, and enable data exchange, interoperability between information systems, and knowledge discovery through advanced data mining. The activities of transPLANT is this area are carried out in collaboration with groups of experts working together in related initiatives like the U.S National Science Foundation's Phenotype Research Coordination Network (http://www.phenotypercn.org), Plant Ontology (www.**plantontology**.org/), and the Crop Ontology (http://www.cropontology.org). The goal is to agree on a single set of vocabularies (with identities and relationships between different vocabularies) made explicit that can be promoted worldwide, and to then seek to share labour on the ongoing development of this vocabulary set.

The recommendations of the transPLANT project have been informed by these collaborations and concern:

- plant description (Plant Ontology),

- trait description (Crop Ontology; additions to this ontology should include semantic linking using the Plant Ontology and PATO when possible),

- chemical phenotyping (Brenda, Golm Metabolom Database, ChEBI, KEGG, others),

- environment ontologies (EO),

- experimental design and investigation (OBI),

- epigenetic traits (Medical Subject Headings Thesaurus OWL version, Ontology for Genetic Interval, Gene Regulation Ontology, Gene Ontology, Subcellular Anatomy Ontology).

The deliverable associated with these recommendations (D3.1) has been prepared and submitted.

D3.2. Format specifications for data exchange by flat file and web services (completed, details submitted as report on D3.2)

The goal of third activity is the development of recommendations concerning the standardisation of

data sets containing phenotypic observations and meta data, to support integrative analysis of phenotypes and "-omics" data measured at different levels of plant organization, and their further integration with existing knowledge. Laboratory protocols for measurements of different plant phenotypic traits (internal, external, molecular) require very different data collection schemes and device-specific preprocessing algorithms, although there are also unifying themes; for example, all approaches involve the collection of data from clearly identified "samples" or "experimental units". Standardization achievements have already been undertaken for various types of high-throughput molecular data (such as that derived from transcriptomics, proteomics, and metabolomics experiments), but these are domain-specific and generally do not allow for inclusion of morphological, yield-related, quality or resistance traits.

We have therefore proposed the "Minimum Information about Plant Phenotypic Experiments" (MIAPPE) checklist, which is based on existing "Minimum Information" recommendations (e.g. those created for transcriptomics, metabolomics, proteomics or genomic sequence analysis). Furthermore, we have developed an implementation of this standard in the form of a file format for exchange of phenotypic information between databases, web services and data analysis tools. The format is based on the ISA-TAB structure, which has already been applied successfully in other biological domains. Our ISA-TAB implementation is compatible with the structures used in other plant science applications, but differs due to definition of two additional files, the "trait definition file" and "sufficient data file", specific for the domain of plant phenotyping.

After the successful development, testing and implementation work, the consortium has agreed on application of the developed standard for collection of the phenotypic data. The presented standards and recommendations will be used by the transPLANT consortium to develop tools for data storage, exchange, information retrieval and integrated data analysis of phenotypic data. The deliverable associated with these recommendations (D3.2) has been prepared and submitted.

In addition to this, we are also discussing these standards with a wider community, comprising our collaborators and other groups active in this rapidly evolving area (e.g. plant phenotyping centres, crop breeders involved in field trials, etc.). We are working in conjunction with other projects also interested in this area with the goal of establishing a single common standard, through the combination of our own efforts with those of other groups working with this data. For example, we are already in discussion with the developers of the Trait Ontology; with the European Plant Phenotyping Network (participation in the EPPN workshop and teleconference planned for September 2013); with Bioversity International; and with various European consortia interested in this domain. We will push these activities forward over the remaining period of the grant.

D3.3. Report on standardisation activities accomplished during transPLANT

In collaboration with other groups also working in this area, further efforts are ongoing to ensure the continued development of formats and vocabularies for data representation and their adoption as global standards, as follows:

- P. Krajewski (IPG PAS), C. Pommier (INRA) and D. Bolser (EMBL-EBI) participated in the Crop Plant Trait Ontology Workshop at Oregon State University 2012, Corvallis, Oregon, USA, 13-15.09.2012. The workshop gathered plant breeders, biologists and bioinformaticians from ten countries, seven US states and two plant agribusinesses. It was hosted by the Plant Ontology and the Trait Ontology, and co-organized by transPLANT, European Bioinformatics Institute, GARNet, Generation Challenge Program, Sol Genomics Network, and SoyBase; and followed from a previous meeting, co-organized by transPLANT and hosted at the EBI, in December 2011, also focused on the application of ontologies for the annotation of crops. The goal of the workshop was to engage researchers associated with major cultivated crops worldwide, widen their awareness and showcase the latest developments in ontologies for plants. P. Krajewski and C. Pommier presented the work on standardisation in transPLANT to that community. The main outcome of the workshop was the conclusion that there is a need for a broad, coordinated effort to create a semantic framework for meaningful cross-species queries using a Common Reference Ontology for Plants. This Reference

Ontology will encompass all green plants and will facilitate queries for related gene expression and phenotype data from plant genomics, genetics experiments from the various species- and clade-specific databases and describe accessions in the various international crop germplasm collections. The other outcome of the meeting was an agreement of joint work between transPLANT partners and the Crop Ontology group led by Elizabeth Arnaud (Bioversity International) on species-specific ontologies; the current state of this collaboration is described in the report on D3.1.

- P. Krajewski and H. Ćwiek participated in the 4th Scientific Workshop of POLAPGEN-BD project in Kraków, Poland, 25-26.03.2013. They described to the scientists working on a systems biology project the need for standardisation of the experimental data and metadata. They presented the methods of standardisation developed in transPLANT, also in collaboration with Crop Ontology group. Following the workshop, in collaboration with the POLAPGEN-BD WP leaders, the phenotypic trait names and annotations were standardized within the project and in relation to existing ontologies. This work concerned yield-related traits, physiological traits, metabolomic traits, and physical characteristics of the plant. The standardized names are used in the project's database.

- H. Ćwiek participated in 10th Extended Semantic Web Conference & Semantics for Biodiversity Workshop in Montpellier, France, 26-30.05.2013. She presented to the Crop Ontology group the work done in transPLANT and discussed the course of collaboration.

- D. Bolser participated in the PRO-PO-GO Meeting, Buffalo University, 15-16.05.2013. The meeting was proposed to promote the coordination of the Gene, Protein, and Plant Ontologies and of other reference ontologies used in plant biology. The ontology annotation data in Ensembl Plants was presented, including coverage, display and sources of manual annotation, as well as the pipelines for automatic annotation. The workshop highlighted the links between existing orthogonal ontologies, and the need to coordinate development among them.

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

Use of resources

EMBL-EBI: 1.5 person-months (reaching 50% of the total planned)

IPK: 2 person-months (reaching 50% of the total planned)

INRA: 3.6 person-months (reaching 55% of the total planned)

IGR PAN: 12 person-months (reaching 91% of the total planned)

DLO: 0.06 person-months (reaching 1% of the total planned)

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

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

Lead Beneficiary: HMGU

Description of work

Task: Organisation 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 bioinfomaticians. 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

Two user training workshops have been held during the reporting period. The first was held in Versailles (France), 12th-13th November 2012, and focused on current developments in plant data resources at transPLANT partner sites, with a special reference to *Triticeae*, and emerging (but complex) data (barley, bread wheat, rye, Aegilops; Brenchley et al., 2013, Nature; IBSC, 2013, Nature) recently generated for this tribe. The 2nd transplant workshop was held in Poznan, Poland, from 27-28 June 2013 at the Adam Mickiewicz University campus, Dept. of Biology. Local organization was carried out by the Institute of Plant Genetics, Polish Academy of Sciences (Pawel Krajewski and colleagues).

Workshop 1

A total of 44 participants from 20 different institutions attended the workshop.

Nov 12th – Day1 (Monday)
10:30 Welcome, computer setup and introduction of workshop objectives and agenda **(Hadi Quesneville, Manuel Spannagl)**
11:00 Introduction: about the transPLANT project **(Paul Kersey)**
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/
12:15 Introduction into triticeae (data) analysis concepts and generation **(Klaus Mayer)**
13:00 Lunch
14:00 Introduction into triticeae (data) analysis concepts and generation **(Klaus Mayer)**, evtl. Follow-up/questions
14:30 Triticeae data@ENSEMBLplants**: introduction + data access (Dan Bolser, Paul Kersey) http://plants.ensembl.org/**
15:30 Coffee break
17:00 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
In between this session: Coffee break
13:00 Lunch
14:00 Annotating triticeae sequences: the triANNOT pipeline **(Phillipe Leroy – external speaker-,**
INRA Clermont-Ferrand): introduction, use cases
http://clermont.inra.fr/triannot
15:30 Coffee break

16:00 Wrap-up, time for questions and discussion (also after and during the individual sessions)

17.00 End of workshop.

Workshop 2

Topics related to Tricitiate species (due to the high demand for training on this topic indicated by the first training meeting) and standardisation and annotation of plant phenotypic data were also introduced and discussed during the workshop. The workshop was targeted at (experimental) biologists and breeders who have needs to use these resources in everyday work to interpret own observations.

To ensure appropriate workshop conditions and resources and to prevent misunderstandings of workshop objectives and contents, user registration required an application explaining the researchers background and motivation on the workshop.

A total of 21 participants attended the course, coming mainly from eastern-European countries such as Poland. Teaching sessions were contributed by the transPLANT partners INRA Versailles, EMBL-EBI, Helmholtz Center Munich, IPG PAS and the Adam Mickiewicz University, Poznań, as an invited guest talk. The workshop was announced over transPLANT partner websites, several mailing lists (including transPLANT and TriticeaeGenome mailing lists), cooperation partners and within connected communities (specifically *triticeae* communities in Europe and the US).

The 2nd transPLANT user workshop program is outlined below:

Day 1

9:00 Opening

9:05 - 13:00 Plant data resources at HMGU: PlantsDB (HMGU)

MIPS PlantsDB (http://mips.helmholtz-muenchen.de/plant/genomes.jsp) is a database framework for integrative and comparative plant genome research and provides data and information resources for individual plant species (including Medicago, *Arabidopsis*, *Brachypodium*, *Sorghum*, maize, rice, barley and wheat). Building up on that, state-of-the-art comparative genomics tools such as CrowsNest are integrated to visualize and investigate syntenic relationships between monocot genomes. Results from novel genome analysis strategies targeting the complex and repetitive genomes of *triticeae* species (wheat and barley) are provided and cross-linked with model species. The MIPS Repeat Element Database (mips-REdat) and Catalog (mips-REcat) as well as tight connections to other databases, e.g. via web services, are further important components of PlantsDB.

Specific topics:

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

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

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

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

14:00 – 14:45 Application of new web technologies in biological research and databases (Adam Mickiewicz University, Poznań)

As biologists venture into bioinformatics, they often have to trade their favorite graphical computer desktop environments for a cumbersome command line versions of very useful scientific programs. Most recently the Web has rapidly evolved into a platform suitable for user-friendly applications that exhibit a level of richness and interaction that could barely be envisioned several years ago. The new web technologies are closing the gap to native applications and fill an important need for the development of data analysis software that provides bioinformatics functionalists to biologists without requiring prior knowledge of programming and scripting languages. We will present design and key technologies underlying some of the recent biological databases and web applications. We show their roles in integration, management and visualization of biological data in mirEX, our latest database of expression levels of Arabidopsis pre-miRNAs. We also show the use of different common programming languages to build an Ajax-driven biologist-friendly web applications. Finally, we demonstrate application of our recently-developed web service for annotation of highly-divergent tryptophan-containing Argonaute-binding proteins in grass genomes.

15:00 - 18:00 Genomic, genetic and phenomic plant data at the INRA URGI: GnpIS.

GnpIS is an integrative information system for plant and pest genomics hosted and developed at the URGI, INRA. It stores and allows data mining of genomic, genetic and phenotype data for several species such as grape, wheat, maize, Arabidopsis, rice, poplar (apple, in progress) but also fungi. During this training, we will focus on some specific use cases of the URGI information system. We will explore wheat portal and genomic resources, then genetic and phenotype data for Vitis and Hordeum and finally explore integrated data set through BioMarts.

Day 2

9:00 - 12:00 Plant data resources at EMBL-EBI: Ensembl Plants (EMBL-EBI)

Ensembl Genomes (http://www.ensemblgenomes.org) is an integrative resource for genome-scale data from non-vertebrate species, including plants and plant pathogens. The project exploits and extends the vertebrate-focused Ensembl technology, originally developed for the human genome. Ensembl Plants (http://plants.ensembl.org) provides a complementary set of resources for plant species, providing a consistent set of interactive (web) and programmatic (API and MySQL) interfaces. We currently host 22 plant genomes and provide access to data including: reference sequence, gene models, transcriptional data, polymorphisms and comparative analysis. Since its launch in 2009, Ensembl Genomes has undergone rapid expansion, with the goal of providing coverage of all major experimental organisms, and additionally including taxonomic reference points to provide the evolutionary context in which genes can be understood. Against the backdrop of a continuing increase in genome sequencing activities in all parts of the tree of life, we seek to work, wherever possible, with the communities actively generating and using data, and are participants in a growing range of collaborations involved in the annotation and analysis of genomes.

12:00 – 14:00 Methods of standardisation and annotation for plant phenotypic data (IPG PAN)

One of the transPLANT aims is to develop community-accepted standards for data description and submission, covering format, and content and policy. In this part of the workshop we will

concentrate on the topics of biological data formatting and annotation possibilities. Currently available tools and Internet resources helpful for these tasks will be presented.

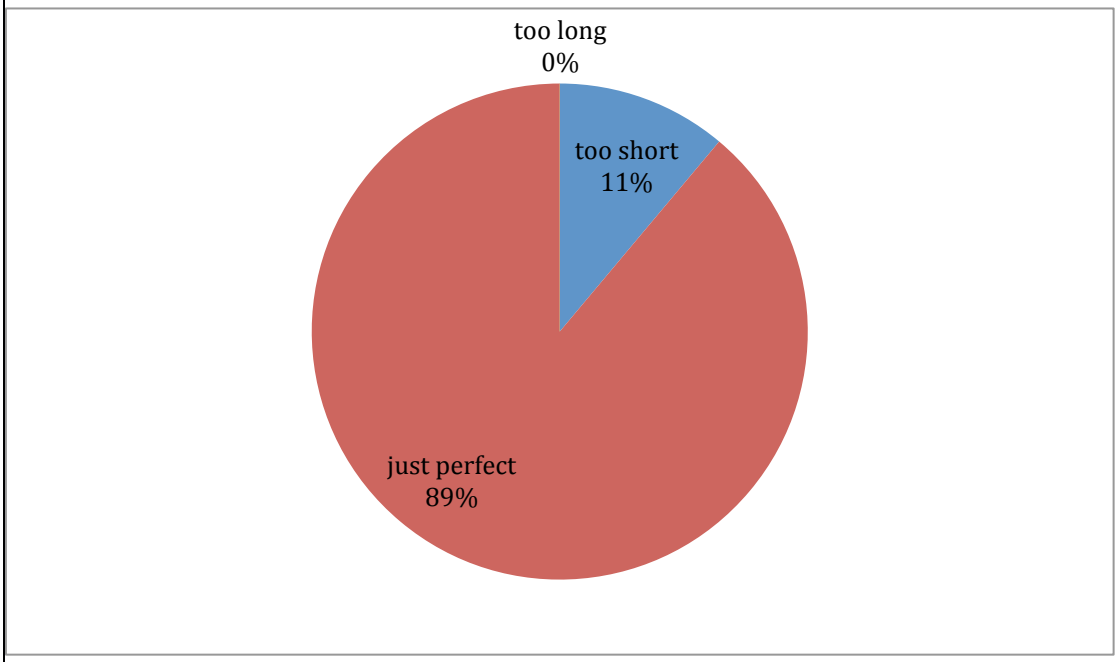
14:00 End of workshop

All sessions consisted of short database/resource introductions followed by extensive "hands-on" training workshops with a number of exercises to solve for the participants. All workshop presentations as well as training materials were uploaded to a dropbox server where participants were able to access it in a convenient way.

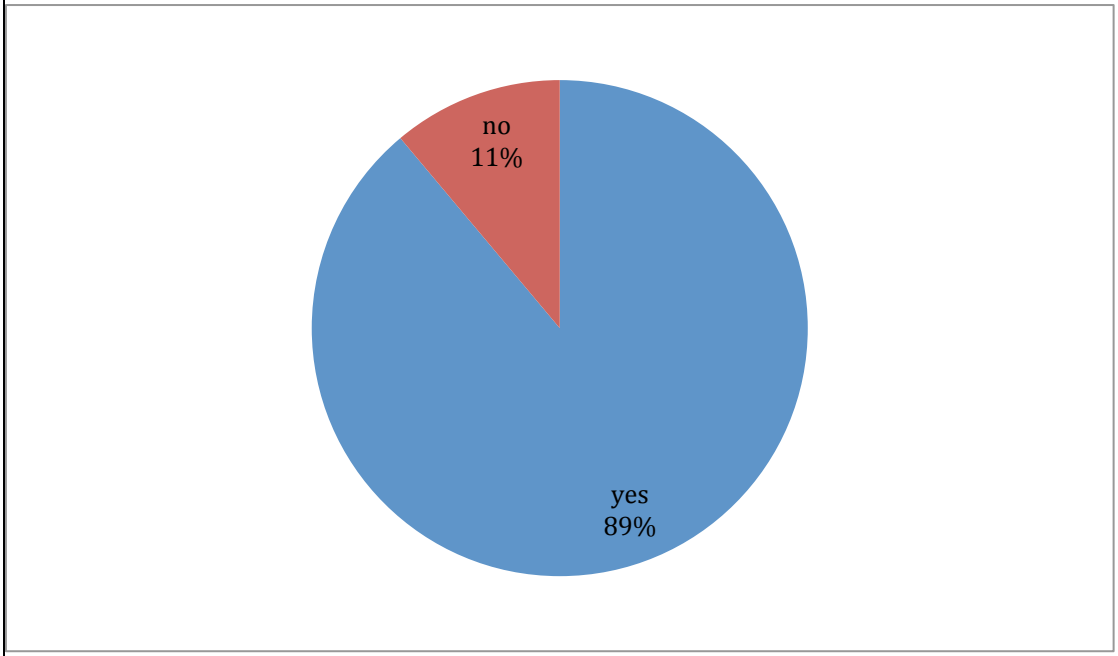
User Feedback

All participants were asked to evaluate the 2nd transPLANT user training workshop and to provide feedback on how to improve organization, content and maximize impact on their daily work. We received back a total of 18 filled evaluation forms, a detailed assessment is provided in the following:

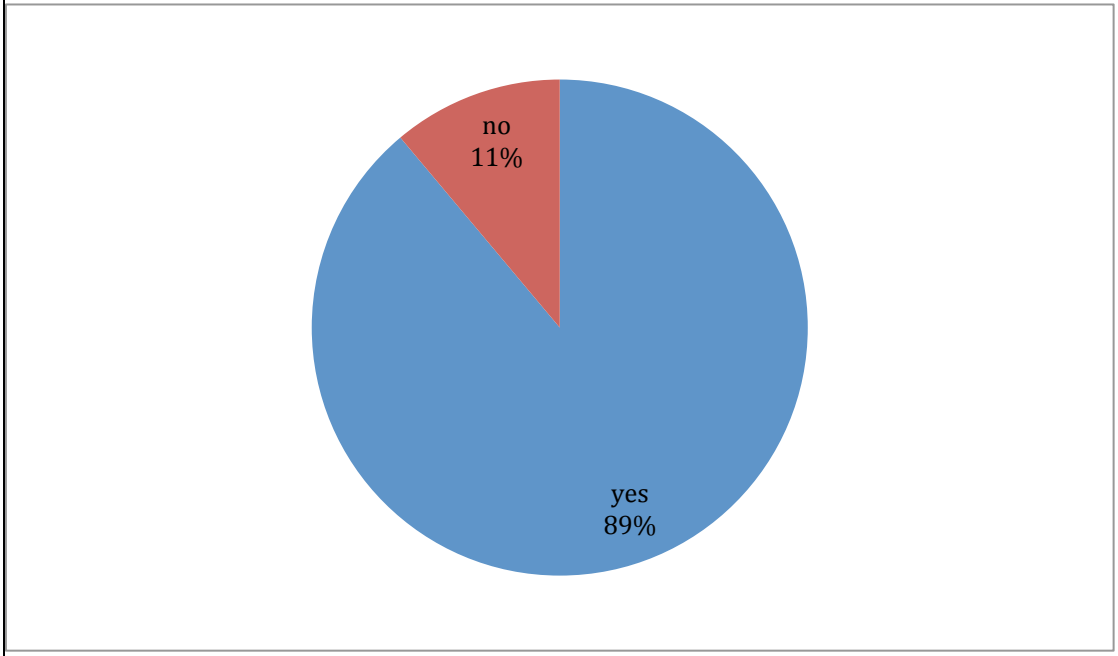
Do you think the workshop duration of 2 days was...?



Do you think the daily timing was appropriate?



Did the workshop meet your expectations?

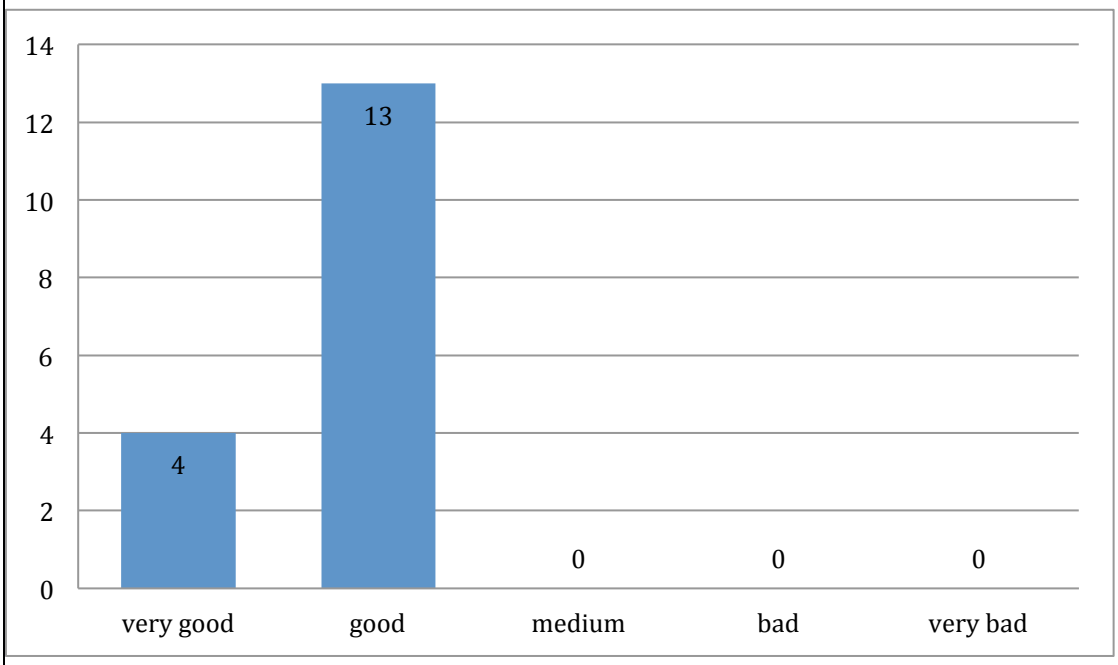


If no, some remarks:

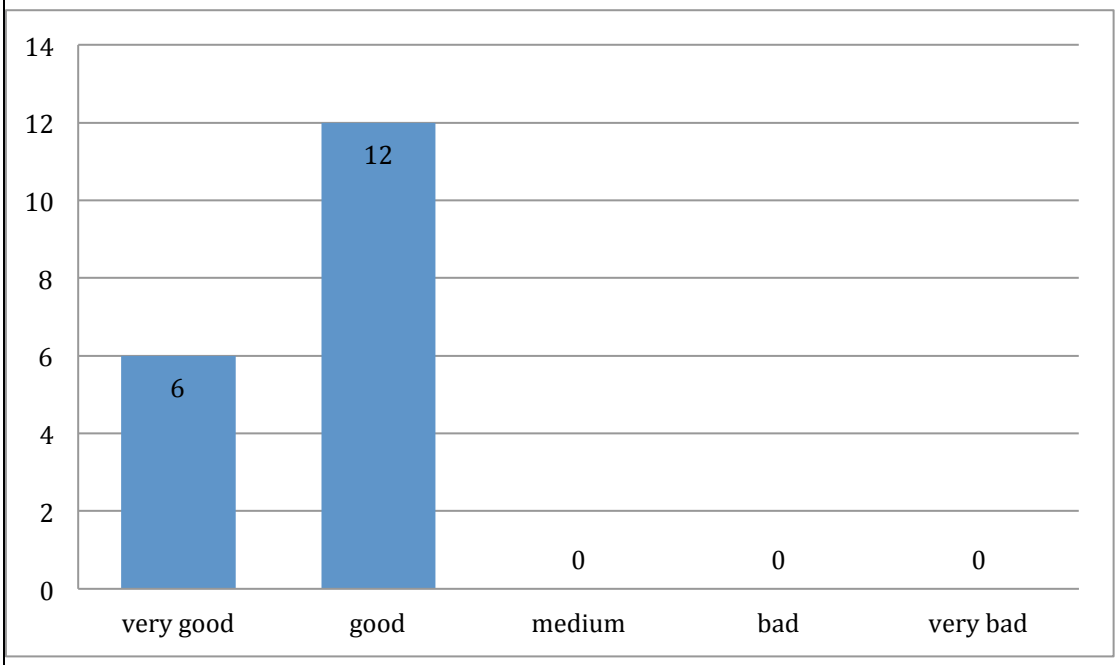
- Expected to work more with sequences, markers, data than only operate on ready databases
- Workshop focused on the triticeae which are not the subject of my studies

Grades for:

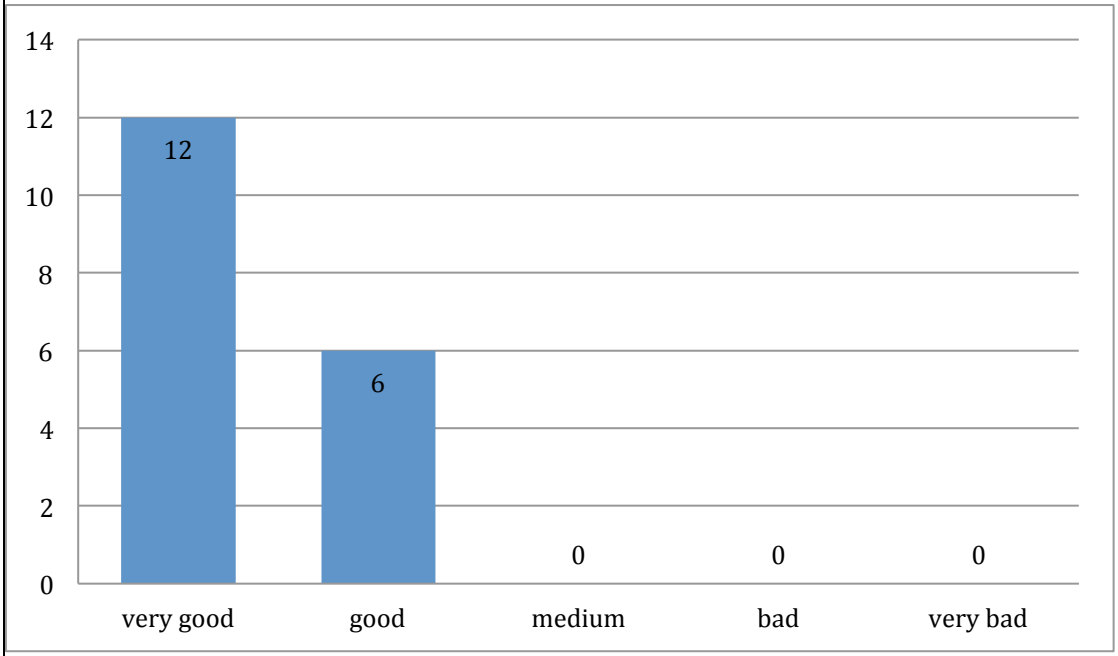
Workshop tutorials overall:



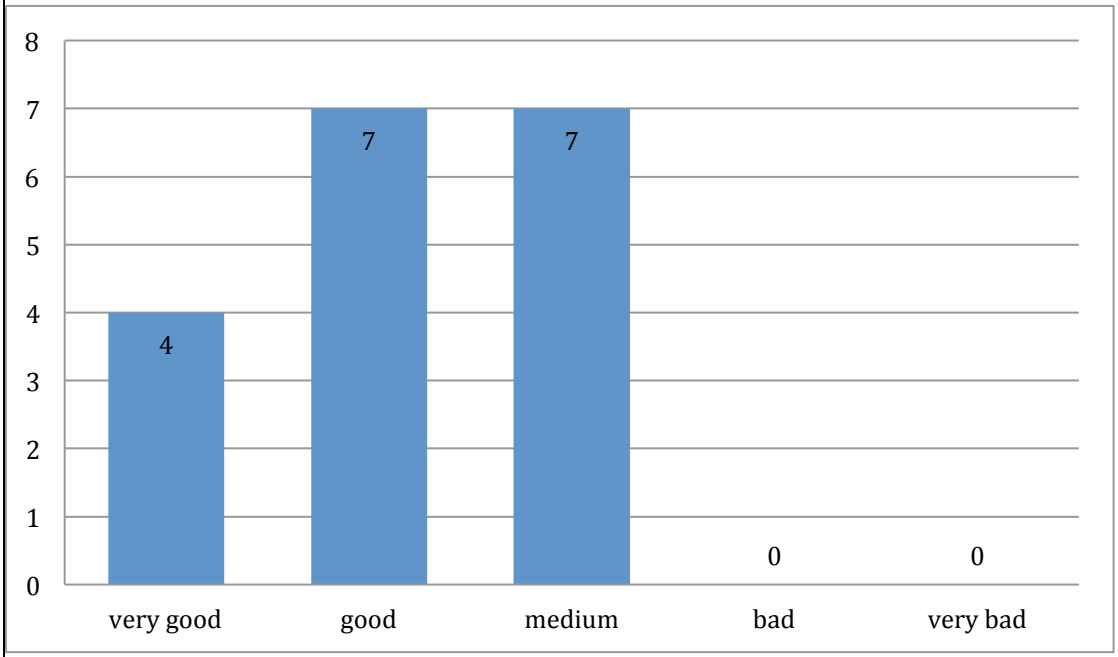
Workshop teachers:



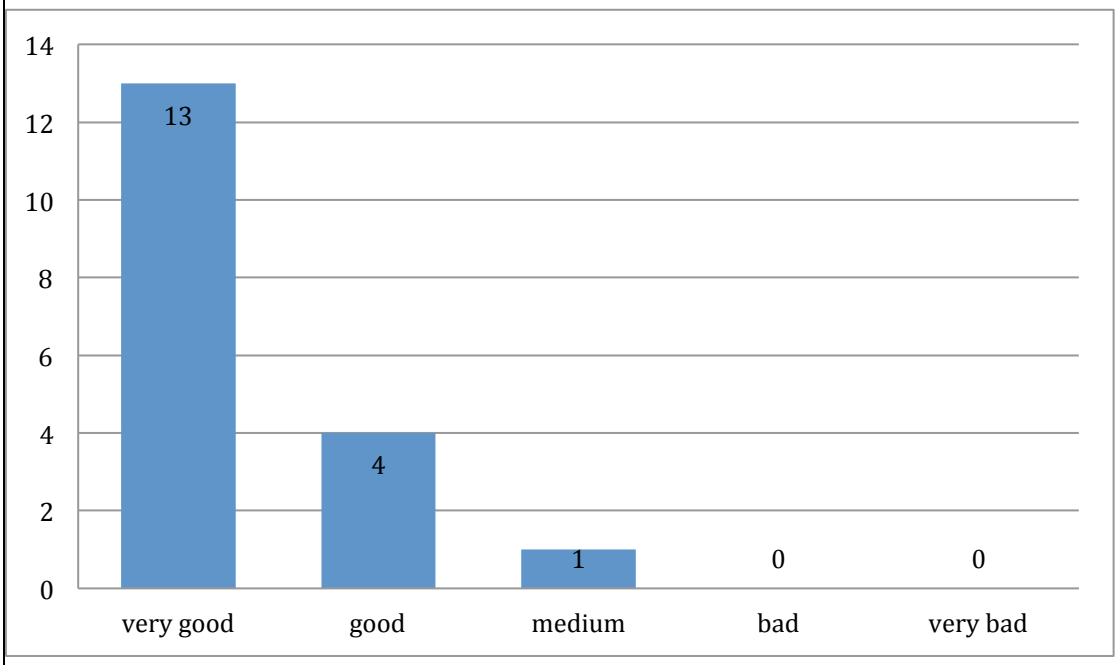
Opportunity to ask questions and discuss your problems/use cases:



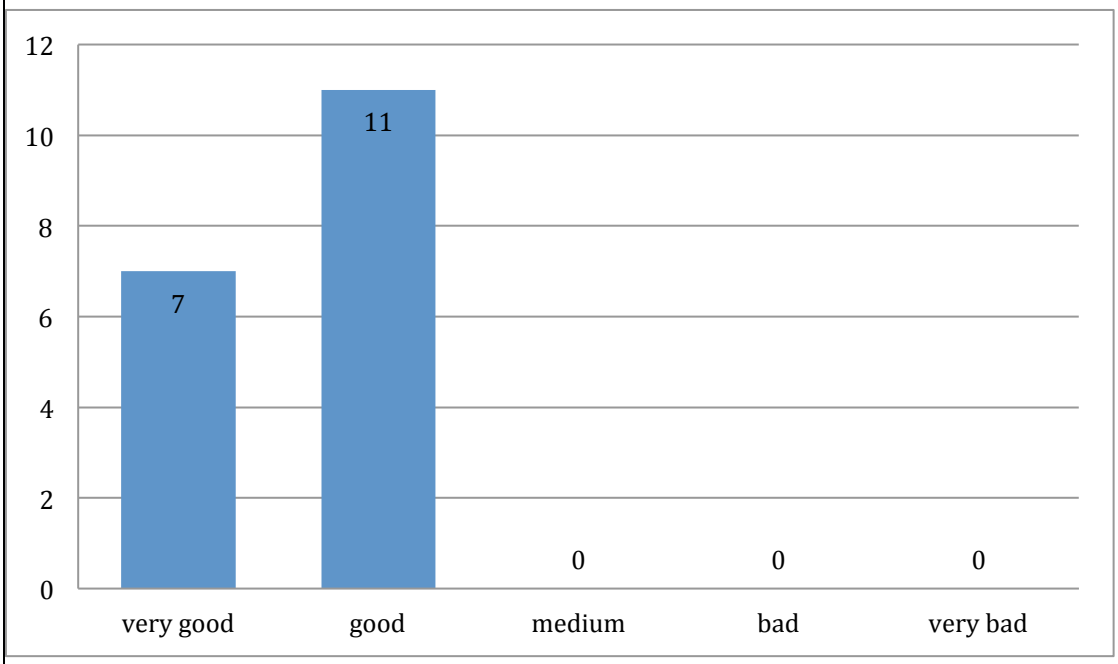
Relevance of tutorial contents for your daily research work:



Workshop location in Poznan:



Overall grade:



Selected remarks by users:

- Encourage participants to bring to workshop their research questions that teachers can help with
- Highlight the unique-ness of each web interface so users can decide which functionalities best fit their questions.

To access a number of different user communities and to present transPLANT services and data resources to a large crowd of international researches we chose the Plant and Animal Genome Conference (PAG XXI) in San Diego in January 2013 (http://www.intlpag.org/2013/index.php/abstracts/workshop-speakers-a-demos) as a venue for a transPLANT computer demonstration. In this 20-minutes session we introduced the scope of transPLANT and gave short tutorials on selected transPLANT services and resources. These included the public transPLANT portal (with search interfaces) at EBI as well as the transPLANT data resources and services at Helmholtz Center Munich (MIPS PlantsDB), EMBL-EBI (Ensembl Plants) and INRA Versailles (GnpIS). A poster was also presented to facilitate discussion and assist in case of questions and problems.
A similar transPLANT computer demonstration and dissemination event is planned for the next Plant and Animal Genome Conference (PAG XXII) in San Diego in January 2014.
transPLANT resources and services were also introduced during the CropLife training course "Research Commercialization in a Large Enterprise" held at the Carlsberg laboratory (Copenhagen, Denmark) at April 9th, 2013 . This course involved a full day training on transPLANT-facilitated barley genome resources including MIPS PlantsDB (Helmholtz Center Munich) and Ensembl Plants (EMBL-EBI) as well as an introduction into the transPLANT project and web portal. A total of ~20 international participants from both academia and industry attended this workshop.

The 3rd transPLANT user training workshop has been scheduled for October 2014 and is to be held in Wageningen, the Netherlands.

Internal Training

COMPSs tutorial workshop
In February 2013, in combination with the AGM meeting in Hinxton (UK), a 2-day training workshop was organized directed to transPLANT partners for the development of parallel applications using BSC's COMPSs, selected technology for powering project's cloud infrastructure.
7 participants attended coming from EBI, INRA URGI, IPK Gatersleben, IPG PAN, BSC and HMGU/MIPS, received a broad view on the internal functioning of COMPSs runtime and its programming methodology, as well as instructions on how to make use of the cloud compute environment, and on how to deploy it in their own infrastructure. The programme is described here:

February 13th

11:00 – 12:15 Introduction to COMP Superscalar (COMPSs)
1.Overview
2.Programming Model
2.1.Steps, 2.2.Data Types, 2.3.Task Types, 2.4.Application Example
3.Resource Configuration
12:15 – 13:00 transPLANT Infrastructure

1.Architecture
2.Programming Model Execution Service (PMES)

13:00 – 14:00 Lunch Break

14:00 – 18:00 Hands-on (Breaks On Demand)

1.Virtual Machine Setup
2.Application Description: Hmmer
3.Hands-on: Hmmer
4.Configuration, compilation and execution
5.Monitoring & Debugging
6.Demo: Gene Detection Application
7.Hands-on: BLAST

February 14th

9:00 – 12:00 Audience Provided Applications & Feedback

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.

Use of resources

EMBL-EBI: 1 person-months (reaching 25% of the total planned)

HMGU: 3.1 person-months (reaching 41% of the total planned)

IPK: 0.5 person-months (reaching 25% of the total planned)

INRA: 0.43 person-months (reaching 22% of the total planned)

IGR PAN: 8 person-months (reaching 100% of the total planned)

TGAC: 0.36 person-months (reaching 20% of the total planned)

BSC: 0.6 person-months (reaching 46% of the total planned)

DLO: 0.02 person-months (reaching 1% of the total planned)

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 an 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.

MIPSPlantsDB

MIPSPlantsDB (Spannagl M. et al. Nucleic Acids Res. 2007 35: D834-40) is a portal for plant genomes and genome associated data and harbors gene sequences structural and functional annotation data, synteny 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 use may individually maintain a stock of transplant data, which are relevant for his/her planed 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 DLO (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

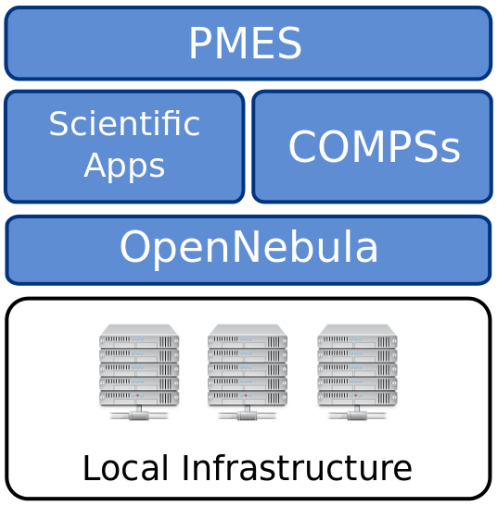
Description: BioMart (Smedley, D. et al. BMC Genomics. 2009 10:22) is a query-oriented data management system widely used in bioinformatics. 34 sites are currently referenced at the BioMart central server(http://www.bioomart.org), and third party software with BioMart plugins implemented include Bioclipse, biomaRt-BioConductor, Cytoscape, Galaxy, Gitools, 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

During the second year of the project, we have continued the development of the project's cloud environment. After analysis and testing of the initial prototype, the architecture of the cloud platform has suffered some modifications; the current one can be seen in the diagram presented below.



Scheme of the Cloud environment logical architecture

The physical resources are virtualized by means of the OpenNebula middleware, and the Programming Model Execution Service (PMES), which was formerly named BES, acts as a single entry point for the platform in the form of a SOAP Web Service.

The PMES enables the execution in the virtualized resources of two different kinds of applications: COMPSs and also, in the present version, stand-alone applications. This functionality has been added mainly to cover the use of preexisting applications or workflows, which now can be run on the cloud

platform without modification. Although the use of COMPSs as workflow manager is expected to improve the performance on workflow executions, especially in an HPC environment, the necessary modifications required to adapt preexisting software will unnecessarily delay their implementation. Allowing a direct execution is expected to increase the short-term usability of the system. COMPSs based applications are executed by launching the COMPSs runtime in a virtual machine; whereas to execute standalone applications, the PMES is also able to run any specific command with certain inputs, either directly (for low demanding applications) or launching the necessary, possibly pre-installed, VM's. In both cases, after the execution, the PMES transfers back the results to a place specified by the user. A graphic web interface is under development to facilitate the user interaction with the platform. The testbed for the infrastructure is being moved to a pre-production structure, including an externally accessible front-end, and increasing the computational resource to 4 cluster 12-core, 96 Gb RAM, nodes with access to a several Tb common storage system. In parallel (see task 3) a series of pre-packed virtual machines accessible though PMES are being installed and tested.

Task 2: Implementation of HPC web services

Development of this Task is delayed until the Cloud environment prepared in Task 1 has been fully implemented. 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. In the mean time, the testbed (see Task 1 above) is being enlarged to include 48 cores and a total of 384Gb of distributed memory to be able to test the applications at a larger scale.

Task 3: Provision of web services for computational analysis

Software available through transPLANT Cloud environment

Access to the transplant cloud environment will be made through a web interface that will activate PMES (see task 1, above). Users will be offered a choice of virtual machines (VM's) including prepacked sets of tools. Ideally, each VM will provide a well-defined operation. Task 3 until the end of the project will be populating the VM offer including both general bioinformatics tools and applications and software produced within transPLANT project. Table below shows a list of the currently available VM's and those being tested.

Application	Type	Status
Biomoby registry	Internal use	VM Available
General Bioinformatics Biomoby based WS	COMPSs / Stand-alone	VM Available
Homology based gene detection WF based on Blast and Genewise	COMPSs	VM Available
Blast	COMPSs	VM Available
MAKER (Gene identification suite)	Stand-alone	VM available
REPET (INRA)	COMPS / Stand-alone	In progress
AbySS (Genome Assembler)	Stand-alone	In progress
SOAPdenovo2 (Genome Assembler)	Stand-alone	In progress
MaSuRCA (Genome Assembler)	COMPSs / Stand-alone	In progress

DAS Genome Servers

EMBL-EBI has provided access to 6 additional plant genomes through the Ensembl Plants interface during the second year of the transPLANT project:

- The bread wheat D-genome progenitor (*Aegilops tauschii*)
- Barley (*Hordeum vulgare*)
- Banana (*Musa acuminata*)
- Barrel clover (*Medicago truncatula*)
- Potato (*Solanum tuberosum*)
- The bread wheat A-genome progenitor (*Triticum urartu*)

The availability of these DAS sources, that represents the achievement of the second milestone on transPLANT's WP5 (MS13), has been published at http://plants.ensembl.org/das/sources and in the DAS registry (http://www.dasregistry.org/listServers.jsp).

These new servers take the total number of species for which DAS servers are available through Ensembl Plants to 25, of which 16 have been made available through transPLANT funding. Both the genomic sequence (allowing for its use as a reference by other annotation servers), and the annotation (genes, transcripts, translations) are made available within Ensembl Plants (as data on that reference) as DAS servers.

Servers are also maintained for older versions of the genome, so that users can continue to visualise older annotation. Publishing annotation via DAS is based upon a common system for identification of reference sequence versions; allows for data sharing among consortium members and other sites and for visualisation in most commonly used genome browsers (for example, the Ensembl Genome Browser and Gbrowse are both DAS clients).

Ensembl

DAS sources available

#	URI	Title	Species	Assembly	Maintainer	Taxon ID	Test range	Capabilities	Description
1	ENSEMBL PLANTS_1_GCA_000347335.1	Aegilops_tauschii.GCA_000347335.1.reference	*Aegilops tauschii*	GCA_000347335.1	helpdesk@ensemblgenomes.org	37682		das1:entry_points das1:sequence das1:features	Aegilops_tauschii Reference server based on GCA_000347335.1 assembly. Contains 429891 top level entries.
2	ENSEMBL PLANTS_3_GCA_000347335.1	Aegilops_tauschii.GCA_000347335.1.transcript	*Aegilops tauschii*	GCA_000347335.1	helpdesk@ensemblgenomes.org	37682		das1:stylesheet das1:features	Annotation source for Aegilops_tauschii transcript
3	ENSEMBL PLANTS_4_GCA_000347335.1	Aegilops_tauschii.GCA_000347335.1.translation	*Aegilops tauschii*	GCA_000347335.1	helpdesk@ensemblgenomes.org	37682		das1:stylesheet das1:features	Annotation source for Aegilops_tauschii translation
4	ENSEMBL PLANTS_1_v.1.0	Arabidopsis_lyrata.v.1.0.reference	*Arabidopsis lyrata*	v.1.0	helpdesk@ensemblgenomes.org	81972		das1:entry_points das1:sequence das1:features	Arabidopsis_lyrata Reference server based on v.1.0 assembly. Contains 695 top level entries.
5	ENSEMBL PLANTS_3_v.1.0	Arabidopsis_lyrata.v.1.0.transcript	*Arabidopsis lyrata*	v.1.0	helpdesk@ensemblgenomes.org	81972		das1:stylesheet das1:features	Annotation source for Arabidopsis_lyrata transcript
6	ENSEMBL PLANTS_4_v.1.0	Arabidopsis_lyrata.v.1.0.translation	*Arabidopsis lyrata*	v.1.0	helpdesk@ensemblgenomes.org	81972		das1:stylesheet das1:features	Annotation source for Arabidopsis_lyrata translation
7	ENSEMBL PLANTS_7_v.1.0	Arabidopsis_lyrata.v.1.0.prediction_transcript	*Arabidopsis lyrata*	v.1.0	helpdesk@ensemblgenomes.org	81972		das1:stylesheet das1:features	Annotation source for Arabidopsis_lyrata prediction_transcript
8	ENSEMBL PLANTS_1_TAIR10	Arabidopsis_thaliana.TAIR10.reference	*Arabidopsis thaliana*	TAIR10	helpdesk@ensemblgenomes.org	3702		das1:entry_points das1:sequence das1:features	Arabidopsis_thaliana Reference server based on TAIR10 assembly. Contains 7 top level entries.
9	ENSEMBL PLANTS_3_TAIR10	Arabidopsis_thaliana.TAIR10.transcript	*Arabidopsis thaliana*	TAIR10	helpdesk@ensemblgenomes.org	3702		das1:stylesheet das1:features	Annotation source for Arabidopsis_thaliana transcript
10	ENSEMBL PLANTS_4_TAIR10	Arabidopsis_thaliana.TAIR10.translation	*Arabidopsis thaliana*	TAIR10	helpdesk@ensemblgenomes.org	3702		das1:stylesheet das1:features	Annotation source for Arabidopsis_thaliana translation
11	ENSEMBL PLANTS_7_TAIR10	Arabidopsis_thaliana.TAIR10.prediction_transcript	*Arabidopsis thaliana*	TAIR10	helpdesk@ensemblgenomes.org	3702		das1:stylesheet das1:features	Annotation source for Arabidopsis_thaliana prediction_transcript
12	ENSEMBL PLANTS_1_v1.0	Brachypodium_distachyon.v1.0.reference	*Brachypodium distachyon*	v1.0	helpdesk@ensemblgenomes.org	15368		das1:entry_points das1:sequence das1:features	Brachypodium_distachyon Reference server based on v1.0 assembly. Contains 83 top level entries.
13	ENSEMBL PLANTS_3_v1.0	Brachypodium_distachyon.v1.0.transcript	*Brachypodium distachyon*	v1.0	helpdesk@ensemblgenomes.org	15368		das1:stylesheet das1:features	Annotation source for Brachypodium_distachyon transcript
14	ENSEMBL PLANTS_4_v1.0	Brachypodium_distachyon.v1.0.translation	*Brachypodium*	v1.0	helpdesk@ensemblgenomes.org	15368		das1:stylesheet	Annotation source for

Figure 1 Screenshot of DAS sources available in Ensembl Plants

Released software

INRA improved its REPET package, now at its v2.2 release adding a new pipeline based on Tallymer,

called TallymerPipe, as a pre-processing tool for fast repeated region detection. Using the REPET pipelines, INRA tested a new strategy, to cope with very large genomes such as the wheat. This strategy is an iterative approach and that includes 1) Detection of the most easy to find transposable elements (TEs); 2) TE annotation and splicing of the corresponding sequences from the initial contigs; 3) Detection of the other TEs with sensitive parameters to build a second TE library; and 4) Annotation of the original contigs with the concatenation of the two TE libraries. REPET is available at http://urgi.versailles.inra.fr/Tools/REPET

DLO adjusted and improved the network-based biological process prediction tool BMRF. Currently, a prototype web tool to allow access to the resulting sets of predicted gene function annotations is available (www.ab.wur.nl/bmrf). Prediction power was further improved by combining BMRF with Argot2, an orthogonal sequence-based. The combination of BMRF/Argot2 using co-expression networks outperformed the individual methods significantly.

A more detailed information for both packages is available at the WP12 report.

IPK Gatersleben, in close project collaboration with European/German Plant Phenotype Network (EPPN/DPPN) released the e!DAL system (Arend D et al.: IEEE Internl. Conf. on Bioinform. and Biomed., Philadelphia, 2012), that will be applied as Data Card for the collection, transformation and distribution of data in the information retrieval environment in WP11. It stands for (electronical Data Archive Library) and implements an enhanced and file system like storage system. It can be used as embedded, local or client-server based data repository. Main features are version tracking, Dublin Core meta data, ISO standard data citations (DOIs), and its easy and modular integration into existing data frontends and information systems.

TGAC has set up a number of instances of the TGAC Browser (implemented using an Ensembl backend) for different plant communities (http://tgac-browser.tgac.ac.uk/) and released the source code into github (https://github.com/TGAC/TGACBrowser). Future releases will include visualisation of polyploid genomes.

Task 4: Provision of interoperable data warehousing technologies

Task regarding data warehousing technologies were moved from first to the second year of the project due to the concerns about the stability and future availability of operative updates of Biomart 0.8. This question has not yet been definitively resolved, so during the second year WP5 has been assaying both the original choices together with other approaches. This has been fully detailed in deliverable 5.1, just released. A summary of the achievements and conclusions follows.

BioMart-based data warehousing.

New data warehouses have been made containing the latest data from Ensembl Plants (9 data releases) and INRA (12.4 data releases) in the BioMart data warehousing system, using both version 0.7 and 0.8.

The latest Ensembl Plants release (release 19, August 2013) contained 25 gene-centric data sets, of which 16 have been newly provided since the initiation of transPLANT funding, and 10 variant-centric data sets, of which 4 are new since the commencement of transPLANT. Data is available through the Ensembl Plants portal http://plants.ensembl.org.

The Ensembl Plants BioMarts can also be accessed in the BioMart v0.8 user interface through the BioMart central portal (http://central.biomart.org/). The interface provides more powerful features to sort data, and provides Java and SPARQL endpoints in addition to the support already offered for Perl, SOAP and REST in BioMart v0.7. Updated databases are supplied to the operators of the BioMart Central portal and made available with each release of Ensembl Genomes.

INRA created a new dataset dedicated to the exploration of genetic resources (germplasm) and another one to explore phenotypic trials. These two new datasets are also linked together and allow users to do queries that explore the two datasets at the same time based on common objects (i.e. accessions in this case). INRA data is available through GnpIS BioMart plants datasets at http://urgi.versailles.inra.fr/biomart/martview/.

INRA collaborated with BioMart central team to make its databases and its corresponding datasets available through the BioMart Central portal. As result, the genome annotation databases and its 10 datasets are now online through BioMart v0.8. The second INRA BioMart database was also added. It is more focused on genetic data and it contains 4 datasets dedicated to genetics maps, one on NGS variant, one on genetic resources and one on phenotypic results. Results are also available in a format exportable in table sheets, in SPARQL or in JAVA.

Testing of local installations of BioMart v0.8

Before a more comprehensive commitment was made to local deployment of the new BioMart software, an analysis of its suitability for further developments was made. The new version of the software was installed at INRA and migration of all existing version 0.7 datasets to version 0.8 was attempted. Unfortunately, several bugs were found. While some of the simplest existing datasets could be migrated, for example, genomic annotation datasets for Arabidopsis and maize, more complex datasets (in particular those making interoperability between genomics and genetics i.e. genetic markers, genomic annotations and genetic resources) failed to migrate. These root causes was diagnosed due to new constraint that appeared in version 0.8 and which had not been present in version 0.7. It was indeed not possible to fully migrate both v0.7 database and the v0.7 query system into v0.8. A possible solution would be to keep both v0.7 and v0.8 databases on line, using a unique query interface in v0.8 that is plugged on the two types of databases. This solution is indeed heavier to maintain. Additionally, the rate of recent BioMart development has been slow.

Testing of an alternative system, InterMine

An alternative system with overall features similar to Biomart, InterMine, has been tested by INRA. A data warehouse for the grapevine has been developed using InterMine and made available to public. INRA succeeded in loading indeed genetic marker data (with their position in cM), QTLs and SNP markers into the same instance of InterMine already holding the structural and functional annotation. INRA continued the integration work by adding also genetic resources data (passport data) and phenotyping data (trial data).

INRA also tested the set-up of links between this mine and external tools.

- Links with the GnpIS URGI portal to have access to more detailed information not contained in the mine and also with Gbrowse/Gmod tool.

- Links to external mines. INRA tested the functionality with for example a link with FlyMine tool and the link with another local mine at INRA was also tested.

- Link to BioMart datasets contained in GnpIS, but also accessible at EBI Ensembl Plants. For that we chose an example present in both information system, based on GrapeVine (12X) data.

To setup the InterMine query interface required to make changes in Java parsers and also XML edition because InterMine was not adapted for the required datatypes. Data is now accessible through the QuickSearch and Query Builders tool and through the use of the Region tab. INRA InterMine portal is available at: http://urgi.versailles.inra.fr/grapemine.

Future Directions

We will continue to develop InterMine and the new system for variation data mining over the remainder of the project. Regular updates of our BioMart v0.7 databases will be produced until replacement systems are in place and fully functional. For some data types, BioMart 0.7 may still prove serviceable for the medium term.

Use of resources

EMBL-EBI: 2 person-months (reaching 63% of the total planned)

INRA: 9.3 person-months (reaching 113% of the total planned)

BIOGEMMA: 2.1 person-months (reaching 210% of the total planned)

TGAC: 0.16 PM person-months (reaching 12% of the total planned)

BSC: 7.1 person-months (reaching 73% of the total planned)

DLO: 1 person-months (reaching 25% of the total planned)

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 visualisation 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 visualisation 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

In year 2 of the project, we have worked to restructure the web portal developed in year 1 of the project (available at http://transplantdb.eu) to reflect the increased range of transPLANT services that are now available. Changes undertaken include (i) removal of project-specific information from the transPLANT homepage, and its relocation to a sub-domain of the site, where information about the project's funding, partners and purpose is available (ii) incorporation of the new, integrated search facility in a prominent position on the transPLANT homepage (iii) use of "mouse over" pop-up boxes to summarise the content available behind certain key links (iv) introduction of new information boxes on the homepage, signposting the major features of the site (and of associated partner resources). One of these boxes contains a newsfeed highlighting the latest relevant development; the others will be updated regularly to showcase information of particular importance or novelty. The overall design is cleaner and the useful content of the site more apparent. A screenshot showing the layout of the revised site is shown in figure 1.

Figure 1. The revised transPLANT web portal.

In addition, a new sub-domain has been created, www.transplantdb.eu/variation, which provides access to the new submission tool for variation data (see work package 9) and the new query tool for variation data (see work package 5), both of which have been developed to run as specific transPLANT services

Task 2: Integrated Search Services

In the report for year one, we described the development of a framework for integrated search (using the open source search engine Apache Solr) over multiple resources provided by the transPLANT partners. In the past year, we have continued this work as follows:

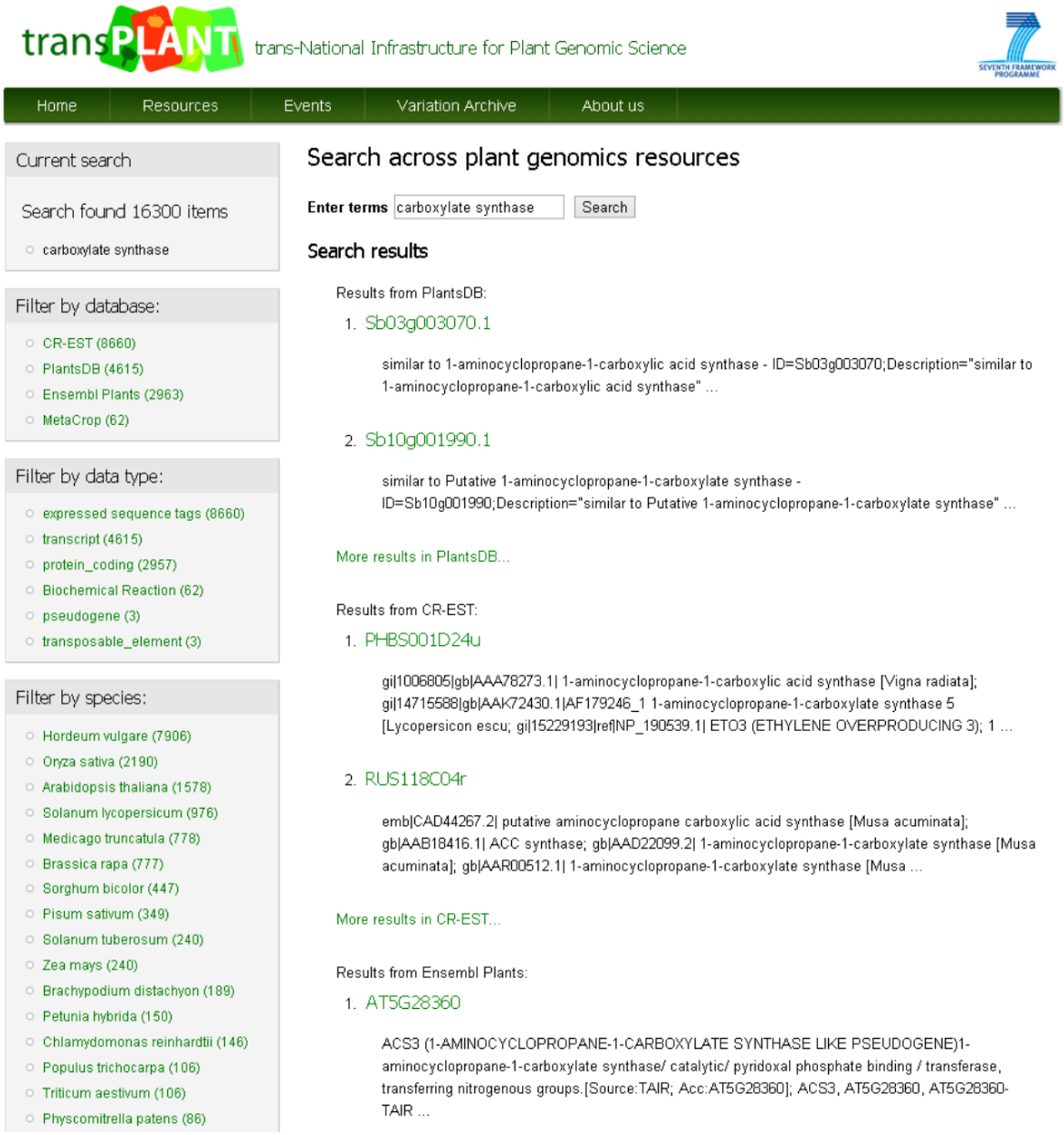
(i) developing a configurable framework which enables data from a particular resource to be collected via FTP and indexed locally, or for the dynamic querying of local Solr servers hosted by individual data providers

(ii) extension of the simple data model for standardized representation of equivalent object types in different resources, supporting certain attributes of certain data types in order to support more sophisticated faceting and visualization logic.

(iii) reworking the visualization of the search results. These were originally displayed in a tabular form that has now been restructured, to make the top hits from more resources immediately visible. Navigation through the search results is provided mainly through the use of faceting, as described above. Examples of the faceting in use are shown in figure 2.

Figure 2. The grouped and faceted display of the transPLANT integrated search.

A search for "carbamoyl synthase" currently returns 14,269 results. The 2 most relevant results from each source database 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.



In addition to supporting search of the three resources originally envisaged, the integrated search also encompasses several additional transPLANT partner resources, namely CR-EST, GEBIS and MetaCrop from IPK and PolapgenDB from PAS. A summary of the complete data currently integrated is shown in Table 1.

Table 1: Summary of partner resources indexed:

Partner	Database	Data types	No. data points	No. species
EBI	Ensembl Plants	Gene-centric	1,072,657	26
MIPS	PlantsDB	Transcript-centric	263,401	6
IPK	CR-EST	ESTs	218,927	6
	GEBIS	Passport data	148,696	5,140
	MetaCrop	Biochemical reactions	585	286
PAS	PolapgenDB	Phenotype-centric	93	*Hordeum vulgare*
UGRI	GnpIS (Vitis)	Variations, markers and genes	168,179	*Vitis vinifera*
	Siregal	Germplasm	16,266	3,278
	Ephesis	Trials, phenotypes, and accessions	334	6

The first implementation of the faceted search was made public on the transPLANT website (http://www.transplantdb.eu) in November 2012, and the improved version was made public (along with a wider revision of the transPLANT portal in August 2013). We will continue to work on enhancing search facilities over the remainder of the project.

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

Access to the phenotypic data repository will be provided by INRA in year 3 of the project, and is represented by project milestone MS16 (due month 36).

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 development of the underlying technologies in work package 12.

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

Use of resources

EMBL-EBI: 3 person months (reaching 43% of the total planned)

DLO: 0.01 person-months (reaching 1% of the total planned)

The EBI has made the primary contribution to establishing the transPLANT site and its search facilities. Other partners are expected to contribute more resources in this area as more powerful integrative services are deployed.

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

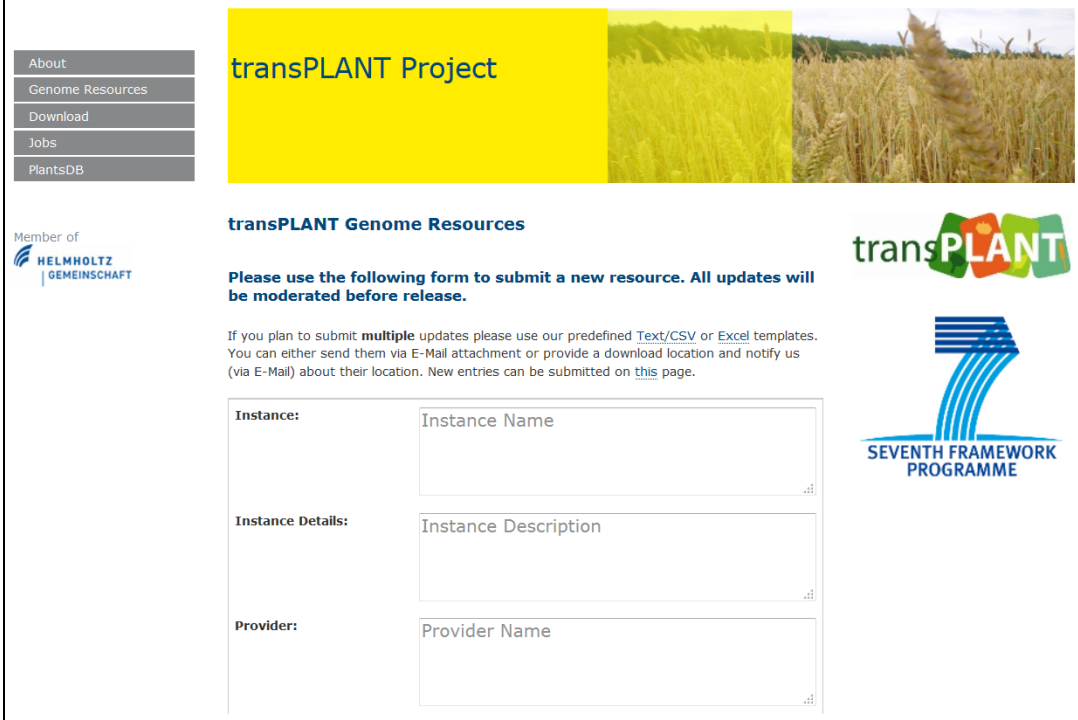
Task 1: A registry of plant genome sequences and resources

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.
214 distinct plant genome resources are registered at the transPLANT data registry at this time and the registry was updated regularly since its initiation. Updates to the registry include new resources for the complex and agricultural important *triticeae* crops bread wheat and barley (HMGU, INRA, EBI) as well as references to databases and resources focusing on sequence variation and re-sequencing. We have 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, resp. http://mips.helmholtz-muenchen.de/plant/transplant/genomeResources.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.
Changes and updates to the registry can now be done by database providers and curators on their own, leveraging the efforts for maintenance and research and ensuring expert-curated and -driven information and up-to-date content. For that, a web interface hosted by HMGU (at http://mips.helmholtz-muenchen.de/plant/transplant/genomeResources.jsp) was developed which enables database curators to upload their database updates and/or changes to the transPLANT genomic resources registry in a convenient way.



All change and update requests are being checked and curated if necessary to avoid spam or false entries inserted into the registry. We also added a bulk upload function to assist database curators who have multiple update or change requests, avoiding filling the form multiple times. This upload functions accepts a very simple tab-delimited format (Excel-compatible) for which we provide text and Excel templates. Database curators wanting to register their resources or services for the first time are also directed to use the bulk upload function. We contacted database curators and providers for all resources currently registered with transPLANT to make them aware of the new update and data change functionalities and invited them to actively help keeping all meta-information regarding their resources up to date.

Outreach

Making potential users aware of this very useful plant genome resources registry, run by transPLANT, is an important task. We took opportunity to present the registry to a broader public at several international conferences, meetings and user trainings including PAG 2013 (Plant and Animal Genome, San Diego USA), all transPLANT trainings, *triticeae* genome training workshops and many more. We are planning to present the transPLANT registry including their new data curation and update functionalities at upcoming PAG 2014 where representatives from many major plant genome databases and resources will be present. Additionally we are investigating the possibility to publish a note or short technical paper in a scientific plant journal describing the transPLANT registry to maximize outreach. Finally, a number of individuals/groups/institutions known to the project partners to be maintaining notable plant informatics resources have been directly emailed and invited to curate updates to their own resources in the registry. We will extend this approach as the project develops, for example, by additionally contacting members of the public who sign up to the transplant_announce mailing list.

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 second year of the grant, we have made four releases of Ensembl Plants, and incorporated the following additional genomes: *Aegilops tauschii* (the bread wheat D-genome progenitor), *Hordeum vulgare* (barley) *Musa acuminata* (banana), *Medicago truncatula* (barrel clover), *Solanum tuberosum* (potato), and ***Triticum urartu*** (the bread wheat A-genome progenitor). The representation of the barley genome specifically resulted from collaboration with the transPLANT partners at HGMU, who provided the primary genome assembly and annotation for representation, and the same gene models are now available through both sites. 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. Pairwise alignments are provided for rice against every other genome, *Arabidopsis thaliana* against every other genome (except barley), and 14 other pairwise comparisons. The precise comparisons available are shown in Figure 1. For several pairs of genomes, these detailed analyses are also used to support assertions of synteny (which can also be visualised in the user interface). A full list of species for which syntenic data is available is provided in Figure 2.

Figure 1 Pairwise genomic comparisons available in Ensembl Plants.

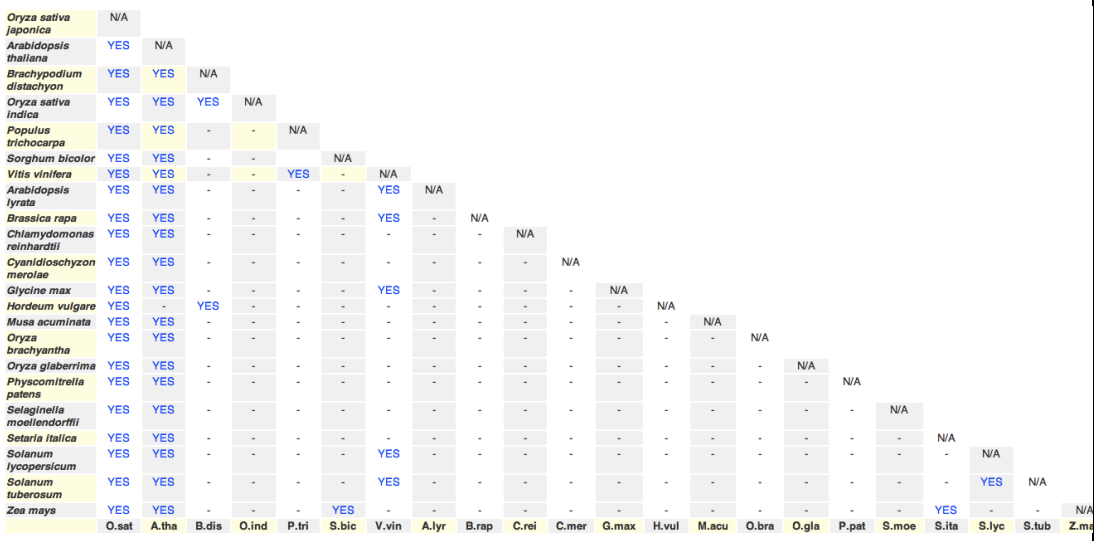
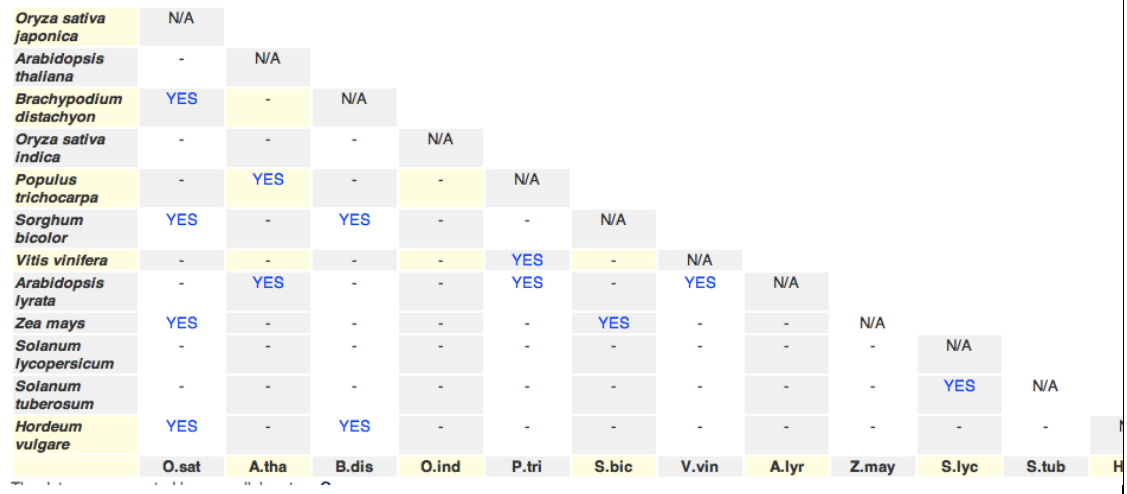


Figure 2 Syntenic comparisons available in Ensembl Plants.



The protein-centric analysis has (as of July 2013) placed 8,919,135 proteins from 23 plant genomes and selected outlying eukaryotic species (*Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans, Ciona intestinalis* and *Saccharomyces cerevisiae*) into 43,771 clusters. The bread wheat precursor genomes, which are the only genomes in Ensembl Plants presently missing from these analyses, will be included in the protein-centric analysis with the release due September 2013. In addition, 7 of these genomes: *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Cyanidioschyzon merolae*, *Physomitrella patens*, *Oryza sativa*, *Solanum lycopersicum* and *Vitis vinifera* have been included in a broad range taxonomic analysis aimed at identification and presentation of large protein families from across the taxonomy.

The data are available through the Ensembl Plants user interface, and are searchable through the integrated search facility available through the transPLANT portal (see work package 6 report).

Ensembl Plants: note on funding.

Ensembl Plants is currently funded by multiple sources. transPLANT, as acknowledged on http://plants.ensembl.org, has paid for a total of 2 PM in work package 5, and 5.75 PM in work package 7 in line with the stated milestones/deliverables. The contribution in work package 5 has paid for the provision of DAS-based services and enhanced data warehousing (see MS12, MS13, MS14, D5.1). The contribution in WP7 has covered the integration of new genomes into the resource and within the comparative analysis pipeline (see MS18, MS19, MS20, D7.3). Protein-based comparisons, and some DNA-based comparisons, are run at EMBL-EBI. Some of the DNA-based comparisons are provided through our collaboration with Gramene (http://www.gramene.org), a U.S.-based resource, with whom we collaborate closely. The Gramene grant additionally pays for a part-time position at EMBL-EBI which is deployed to develop new methods for protein classification in plant species (not linked to any transPLANT deliverable). We additionally receive UK national government funding to work specifically on wheat and barley.

Future Developments

We will continue to integrate new genomes as they become available for public release. One of the major challenges in this respect is the (emerging) wheat genome, which currently exists in a highly fragmented state but which initiatives are currently underway to develop (though

owing to its size and complexity, a complete reference-quality sequence is not likely to be available within the duration of the transPLANT project. We currently make some resources available for wheat (although in a different form than that used for more complete genomes) and transPLANT partners are working with the International Wheat Genome Sequencing Consortium to ensure consistent representation of the evolving best sequence across different sites.
In the previous year's report, we noted the importance of maintaining connections between different versions of a reference genome. We are currently developing a genome mapping service that will automatically convert genome coordinates from one version of a reference sequence to another one. This is expected to be publicly released in September 2013.

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

none

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)*

none

Use of resources

EMBL-EBI: 5.75 person-months (reaching 62% of the total planned)

HMGU: 13.6 person-months (reaching 57% of the total planned)

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 visualisation 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, visualisation of ancient duplications and polyploidy

Objective: Plant genomes are inherently complex and in contrast to vertebrate genomes in almost all cases have undergone polyploidization 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. Plant of agronomical interest are organised 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 visualisation.

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

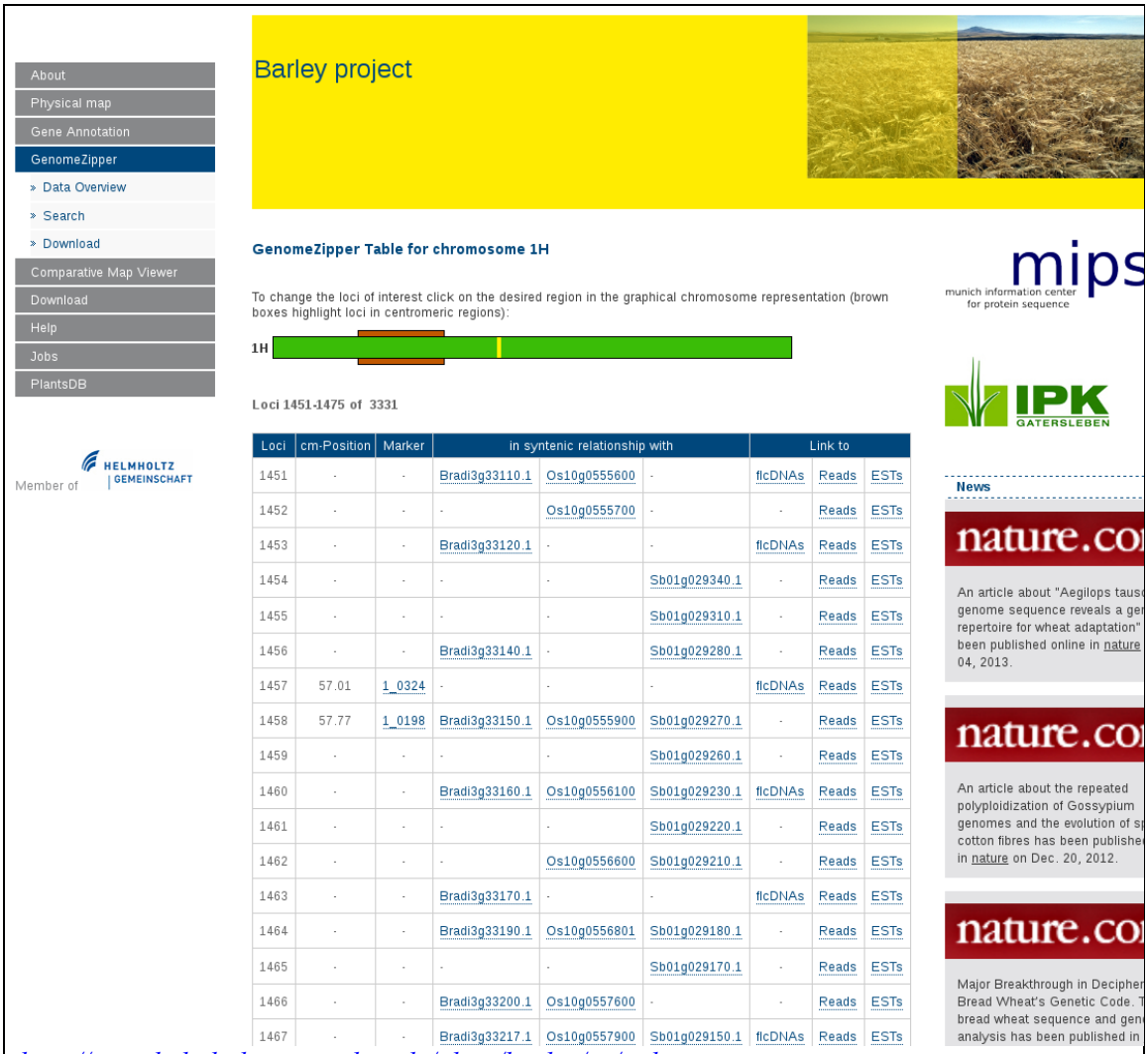
transPLANT partners have integrated genetic marker data from wheat, barley, maize and oak.

Barley

HMGU has integrated genetic markers from three different maps as well as 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 (IBSC, Nature 2012). 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).

In addition, all GenomeZipper (Mayer at al., Plant Cell 2011) data for barley has been integrated into PlantsDB (a platform for integrative and comparative plant genome research maintained at HGMU) and is available for search and browse through various interfaces. The GenomeZipper uses a novel approach that incorporates chromosome sorting, second generation sequencing, array hybridization and systematic exploitation of conserved synteny between crop and model grasses. It has allowed the assignment of 86% of an estimated total of about 32,000 barley genes to individual chromosome arms. A series of bioinformatically constructed 'zippers' integrate gene indices of rice, *sorghum and brachypodium* in a conserved synteny model and assemble 21,766 barley genes in a putative linear order.

The results from the GenomeZipper are available in tabular format (Excel and tab-delimited format) as well but the interactive representation via dedicated web interfaces has some major advantages for the end user: as the GenomeZipper concept integrates and consists of a high number of different elements such as markers, genes, fl-cDNAs, ESTs, raw sequence reads etc. the web representation allows for direct linking and referencing elements from e.g. an overview page. In a flat-file or text-based GenomeZipper representation users are required to retrieve additional information or data such as the sequence individually for every single element, possibly even from different databases or genome resources. For that reason, overview element pages were created in PlantsDB for all elements used in the GenomeZipper, summarizing additional relevant information and data such as the element sequence.



http://mips.helmholtz-muenchen.de/plant/barley/gz/index.jsp

Wheat

A UK-funded wheat consortium has generated a 5x whole genome survey sequence for bread wheat. As the bread wheat genome is very complex due to its large size (~17Gb), high repeat content (80%) and polyploidy (hexaploid) a novel approach was developed to analyse for the gene repertoire of wheat (Brenchely et al., Nature 2012) in the absence of longer continuous genome sequences. Within that analysis, associations of genic wheat sub-assembly sequences with the genes of grass reference organisms such as *Brachypodium distachyon* were established. All data produced in that study have been integrated into PlantsDB and cross-references have been established between the genes from grass reference organisms used in the wheat analysis with the gene reports (giving much additional information about the gene) of the corresponding and already existing PlantsDB database instances. All data in PlantsDB is indexed in the integrative transPLANT search (see work package 5).

Currently HMGU is in the process of collecting and anchoring marker/sequence data produced within the IWGSC bread wheat project (http://www.wheatgenome.org/), which will be visualized within the MIPS PlantsDB GenomeZipper interfaces soon.
http://mips.helmholtz-muenchen.de/plant/wheat/uk454survey/index.jsp

Partner URGI have integrated markers and physical contig data for several *Triticum aestivum* chromosomes. The physical maps have been visualized in a dedicated GBrowse: http://urgi.versailles.inra.fr/gb2/gbrowse/wheat_phys_pub/. Currently URGI is in the process of setting up the 3B reference sequence and their annotations. New markers, QTLs and metaQTL have been integrated into GnpIS (an information system for all genomic data stored at URGI and also indexed in the transPLANT portal). They have been anchored on the physical map used for the wheat reference genome sequencing by the IWGSC. Work is in progress to integrate all these genetic markers on the 3B reference sequence.

Maize

New insertion mutants tracks (Brutnell and Vollbrecht et al. from plantGDB) can be visualized on a URGI genome browser dedicated to the B73 reference maize sequence genome:

http://urgi.versailles.inra.fr/gb2/gbrowse/Zea_mays_ZmB73_pub/

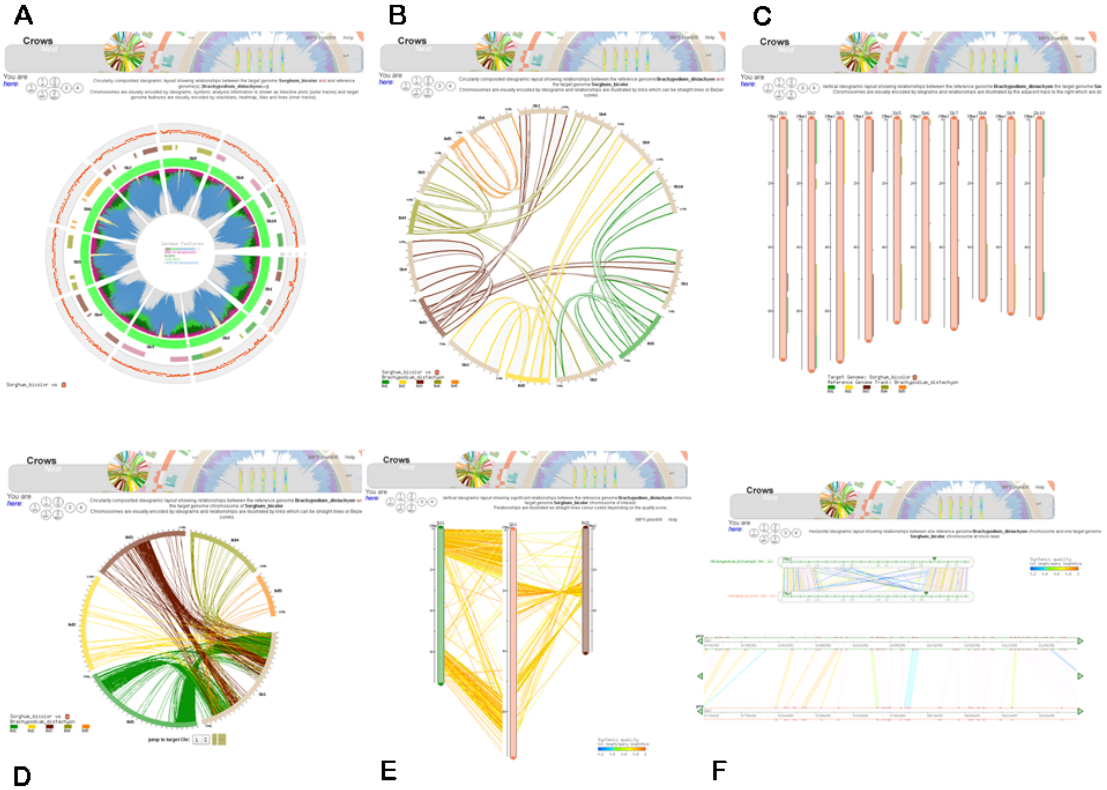
Oak

Nineteen Oak genetic maps, containing more than 14065 markers and 515 QTLs (Submitted by Ehrenmann et al) were integrated in GnpIS. This work was done in preparation to map and to integrate them on the oak reference genome sequence being sequenced.

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

To analyse 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 explore rearrangements, inversions, 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: i) the web-based user interface with the integrated comparative visualization tool, and ii) 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 provides different levels for the visualisation of syntenic data, from genome-wide overviews (A-C) down to micro-syntenic views on a gene-by-gene level (F).



CrowsNest currently harbors data from the model grass organisms *Brachypodium distachyon, Sorghum bicolor* and *Oryza sativa* (rice) as well as from the crop plant *Hordeum vulgare* (barley). This facilitates knowledge transfer from model grasses to crops such as barley, a valuable tool not only for breeders. We are currently working on the integration of new model and crop plant species (bread wheat and wheat relatives) as well as adding functionality (e.g. visualization of gene families) to the analysis pipeline and improving the code and server capacity, evaluating the results and defining ways to share these results with transplant partners.

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

INRA investigated a new method aimed at detecting large indels from short Pair-End (PE) reads. It is based on reference-guided assembly, where all the alignments of re-sequenced reads onto a reference sequence are used to derive consensus draft of contigs. The unmapped reads are *de novo* assembled, and anchored on the obtained contigs using PE information, in order to assemble the newly sequenced genome. This approach has been proven to generate a more accurate assembly than *de novo* approach, using less computing time. Finally, the resulting assembly is aligned onto the reference to predict structural variations (SV) from the unaligned sequences, either on the reference or the guided-assembly. We expected that this method could retrieve larger indels than others. INRA have tested Velvet-Columbus[1] for guided assembly, followed by Nucmer[2] for alignment. For large indels detection, it scuffled contigs using PE information with SSpace[3] beforehand.
INRA also implemented a Depth Of Coverage (DOC) approach to detect deletions based on the fall of read coverage expected when a deletion occurs in the sample studied. To reduce false positive rate due to local falls of coverage, only deletions longer than 500bp were kept.
These approaches have been benchmarked as a part of the deliverable 12.1 from WP12. It shows that, as expected, the reference-guided assembly approach has good results on both large insertions and deletions when compared to other methods. Specificity is good, but sensitivity is still quite low, even with scaffolding and parameters fine-tuning. The DOC approach gives very high sensitivity and specificity in addition to good enough boundaries retrieval, but only allows to detect deletions. It appears nevertheless a promising approach to complement the reference-guided assembly method.
KeyGene has started a study to identify the characteristics and use of Copy Number Variants (CNVs) and the screening of suitable software tools for their identification. Often microarray data are used for detection of CNV, but here we want to focus on sequence based detection. An initial search listed a large number of tools, including CNV-seq, CNVnator, FREEC, readDepth, CNVHitSeq, SegSeq, CNVFinder and EWT. The next step is to test several tools and analyze and compare the results.
Biogemma has tested different tools in order to detect presence/absence variations from Maize MatePair resequencing data and its alignment on B73 reference genome. Based on these tests and on experimental validation results (Sequence Capture), we were able to identify Pindel[4] as the best tool to efficiently call deletions at breakpoint level with low false-positive rates (<17%).

Task 4 Implementation of a pan-genome browser

EMBL-EBI has carried out some experimental work based on the use of the ATAC genome alignment tool to select representative sets of sequences and genes for inclusion as a pan-genome to be made available through the Ensembl Plants browser (part of D8.2, due month 36). A method is being developed for iteratively comparing new genomes against a set of representative sequences and selectively adding novel portions (based on sequence alignment

[1] Velvet: algorithms for de novo short read assembly using de Bruijn graphs. D.R. Zerbino and E. Birney. Genome Research 18:821-829

[2] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: Versatile and open software for comparing large genomes, Genome Biol. 2004;5(2):R12. Epub 2004 Jan 30

[3] Boetzer M, Henkel CV, Jansen HJ, Butler D and Pirovano W. 2010. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics.

[4] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009 Nov 1;25(21):2865-71. Epub 2009 Jun 26.

and gene content). The method has been applied (in test) to *Arabidopsis* and *Oryza* genomes, and has additionally been tested on larger numbers of non-plant genomes, with alternative stringency parameters. Further development will be undertaken next year.

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)*

n/a

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 (reaching 50% of the total planned)

HMGU: 7 person-months (reaching 43% of the total planned)

INRA: 0.4 person-months (reaching 28% of the total planned)

BIOGEMMA: 4.08 person-months (reaching 117% of the total planned)

TGAC: 0.1 person-months (reaching 7% of the total planned)

KN: 0.38 person-months (reaching 38% of the total planned)

If applicable, propose corrective actions

No correction needed

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 task

Task 1 To develop an archive of plant genomic variation.

The first steps in developing an archive for plant genomic variation lie in the establishment of a conceptual framework, followed by the definition of a logical operating procedure for dealing with new data, deleted data, redundant data etc. We choose to define a variant locus through its positioning on a reference sequence; but reference sequences for plant genomes are both approximations and abstractions, which creates the first challenge for managing these data. As reference sequences are updated, existing data needs to be migrated forward to be seen in the context of the latest reference and annotation; a useful archive needs to solve this problem, and not just store the data as submitted.

dbSNP, a variant database maintained by the NCBI, addresses this problem through the alignment of each variant locus (together with its flanking sequence) individually against the new reference. Our alternative solution exploits the nature of variant loci as positional features, which can be projected from one genome assembly to another provided the genome assemblies as a whole have been mapped against each other allowing a transformation from the coordinate system of one to that of the other, which is a more computationally effective approach. Having defined the approach, the core components of the data management system required can then be defined as follows:

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.

In last year's report, we reported the initial implementation of components 1-7, and the commencement of the initial work on component 8 (details of this work are were provided in that report, and are also summarized in the submitted deliverable D9.1). This year we have focused on completing the first implementation of component 8, and the first implementation of components 9 and 10, as follows:

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. A system exists for recording the mapping between descriptive identifiers for locations (based on sequence identifier and version, and the location coordinates) to compact stable identifiers (which are remain the same even when the coordinate system changes).

A format for identifiers for use in the system has been specified as follows: vc[A-Z]$_1$[0-9A-Z]$_n$ with a n initial n of 4 which will grow as demand for identifiers increases. This format is compatible with the accession space provisionally assigned for the non-plant species in the evolving suite of resources for variation data in development at EBI. A system for the maintained and allocation of these identifiers has been implemented, utilizing MongoDB.
An example of am excerpt from a submitted VCF file, and an accessioned version of the same file, are shown in figure 4.

9. Genome-genome mappings have been generated between every version of every plant genome that has been modified in the course of the history of the Ensembl Plants resource (i.e. since September 2009). In total, there have been 8 changes to the sequence of 8 genomes in that time. The historical mappings are stored and can be consulted whenever a new submission on an old sequence is provided, preventing the need for wasteful recalculation. This data has been made available for public use through the Ensembl Plants user interface from September 2013, which provides a feature whereby users can upload their own positional features and automatically map their co-ordinates into the appropriate coordinates on the latest sequence version.

10. Export to Ensembl is done via the intermediate a flat-file dump to VCF (Variant Call Format). Exported VCF is validated with the VCF validation tool from the VCF validation package (http://vcftools.sourceforge.net/docs.html) and loaded into the Ensembl framework using an existing VCF -> Ensembl loader.

Work on component 11 will follow the conclusion of on-going discussions with NCBI about future load-sharing for variation databases.

A further major source of effort this year has been the conversion of these individually tested components into an integrated pipeline, with the goal that chromosomal coordinate mapping is automatically triggered by updates to genome sequences, that updated chromosomal coordinates triggers the re-location of existing variant loci on their new locations, so that new submissions are automatically merged with existing data, and that all data are exported into Ensembl variation databases for access and visualization with each release. However, due to the complexity of the overall pipeline, we anticipate a period of semi-automation, in which more manual quality control will occur and in which modifications may be needed to the pipeline in order to support unforeseen use cases or data complexities. A logging system has also been implemented, recording when updates are generated, and on which data sets and genome assembly versions they have operated. This will allow data to be backed out of the system, if required (e.g. due to erroneous variant analysis or wrongful submission), , in addition to providing a clear data audit trail to users of the system.

The pipeline has been tested extensively and used internally to migrate existing data between successive versions of genomes. Documents have been written to formalise the standard operating procedures intended on submission of data, deletion of data, submission of new assembly versions. The file system on which the MongoDB instance is stored is backed up nightly and the MongoDB instance duplicated weekly to a physically separate file system.

Figure 1 Excerpt from submitted and accessioned VCF files, showing the insertion of new vc identifiers into the record after data merging. The initial data files were provided on a one-per strain basis, all identifying variants against the same reference sequence. The second file represents variation in each strain in a matrix format. Unique accession numbers (e.g. vcZ92WSQ) have been assigned to each variant locus on the reference chromosome. Files are available via FTP at ftp://ftp.ebi.ac.uk/pub/databases/transplant/variation/, and the data has been visualised in Ensembl Plants.

```
##fileformat=VCFv4.1
##qual=10000 if varscan p-value=0
##program: VarscanToVCF.py input_file=Lambrusque_Campmarcel_Final_SNPs.VarScan
#CHROM POS ID REF ALT QUAL FILTER INFO
chr1 305 . C T . . AF=1.0000;DP=5;RBQ=0;ABQ=63
chr1 740 . C T . . AF=1.0000;DP=7;RBQ=0;ABQ=64
chr1 797 . A G . . AF=1.0000;DP=11;RBQ=0;ABQ=67
chr1 799 . C T . . AF=0.3636;DP=11;RBQ=67;ABQ=65


#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Araklinos Cabernet Carignan Castellana
Chouchillon           Colorino           Espadeiro           Heben           Jaen
Lambrusque_Campmarcel Lambrusque_E Lambrusque_PSL2  Listan  Malvasia  Maska  Medouar
Orlovi Savagnin Sultanine Teulere Tsolikoouri Vitis
chr1 64 vcZ92WSQ A T 0.0 . . GT . . . . . . . . . . . . . . . . . . . . . . 1/1
chr1 78 vcZ92WSR A G 6.69006781 . . GT . 1/1 . 1/1 . . 1/1 . 1/1 . . . . . . . . . .
. . .
chr1 80 vcZ92WSS C T 6.69006781 . . GT . 1/1 . 1/1 . . 1/1 . 1/1 . . . . . . . . . .
. . .
chr1 83 vcZ92WST T G 0.0 . . GT . 1/1 . 1/1 . . 1/1 . . . . . . . . . . . . . . . .
chr1 86 vcZ92WSU A G 0.0 . . GT . . . . . 1/1 . . . . . .
```

Current Status and Plans for Wider Public Submissions

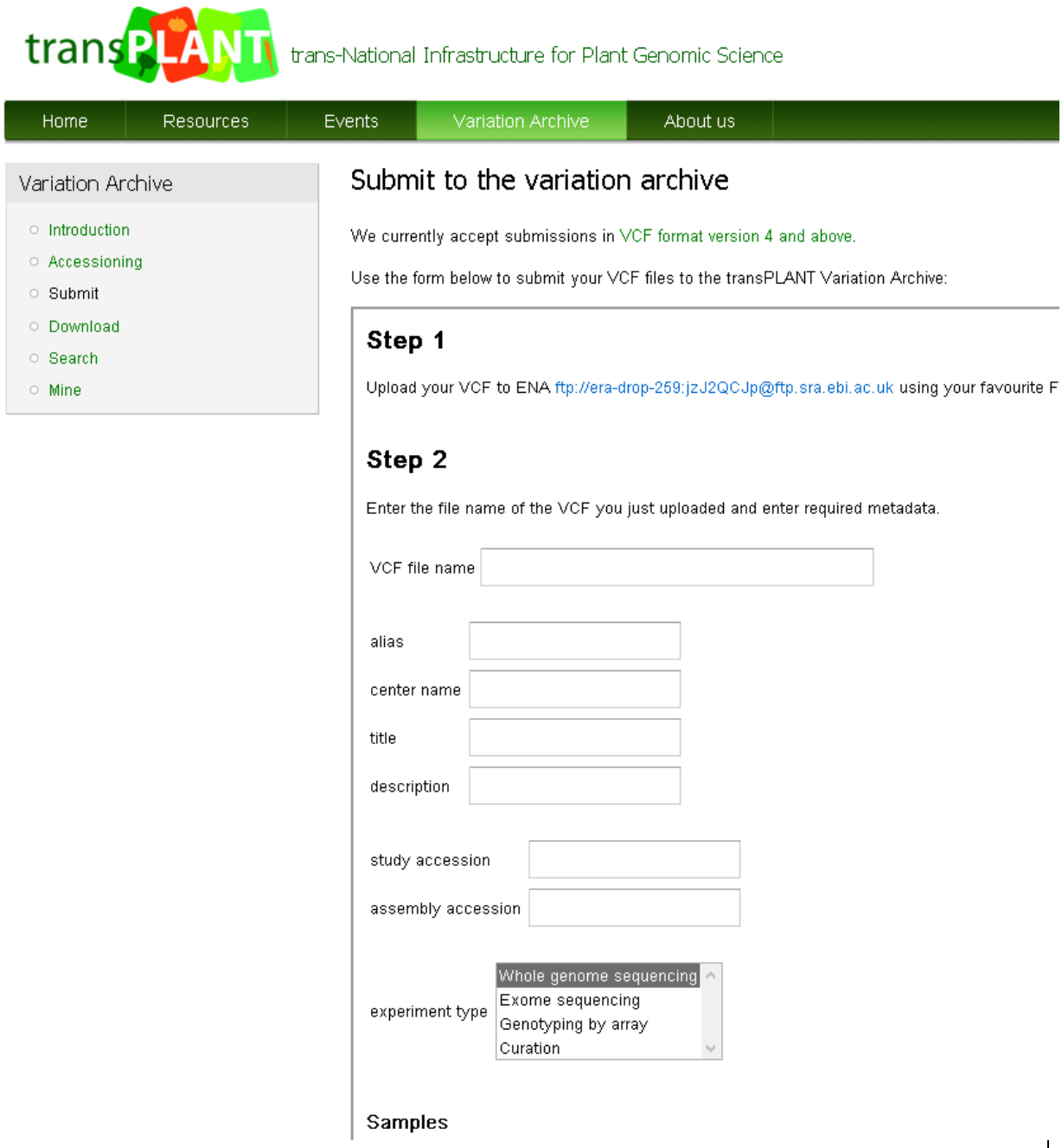
The first variation data has been accessioned in the archive, with the accessioning of data sets from grapevine and barley. Details are given in table 1, below.

Table 1. Initial data sets accessioned in the variation archive

Species	Number of varieties	Number of variants	Number of variant loci on reference genome
Hordeum vulgare	5	24,392,914	15,252,361
Vitis vinifera	23	116,454,085	25,840,400

The data submission tool has been published at http://www.transplantdb.eu/variation/submit and we are now ready to accept submissions from members of the scientific public. A screenshot from the submission portal is shown in Figure 2. Deliverable D9.1 and MS23 have both been achieved on schedule.

Figure 2. The transPLANT user interface for the submission of variation data:



We plan to initially focus mainly on collaborators' data, or data of specific interest to the research of transPLANT partners, while the infrastructure is tested in production. The next priority is fully incorporating the data sets produced by the 1001 Arabidopsis genomes project (D9.2, due month 36). Once we have proven the ability of the system to function well in practice, we will advertise widely to the potential user community.

Two important issues raised by the reviewer's last year are (i) the possibility of variant mis-calling and (ii) the need to provide links to biological material. In terms of point (i), the job of the archive is to store the variant calls that submitters wish to store therein; but individual services might want to only make certain variant data available. We foresee an increasingly important role for meta data (and associated search) allowing downstream users of the archive layer, both end users and intermediate

users (service providers) to select data sets (and data points within individual data sets) that meet certain specified criteria. In terms of point (ii), the archive is again dependent on the description provided by submitters of the sample within which the variant call has been made. A new resource (the Biosamples database; http://www.ebi.ac.uk/biosamples/) has recently been developed by EBI (together with a partner resource at NCBI) to provide identifiers to biological material which can then be cross-referenced by multiple databases holding molecular information, and which could become a standard means of linking between digital and physical bio-resources. We are now using Biosamples as the ultimate store for sample descriptions associated with submitted variant data (including sample-specific metadata as a first step towards this goal.

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.

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

Work is on schedule.

Use of resources

EMBL-EBI: 8 person-months (reaching 27% of the total planned)

TGAC: 0.05 person-months (reaching 1% of the total planned)

If applicable, propose corrective actions

No corrective actions required.

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). Analyzing 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 bioinformatical 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 analyze 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 analyze 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 analyze 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 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 analogs 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 (Seren et al, 2012, *Plant Cell* 24:4793-4805). A more comprehensive version was demonstrated at the International Arabidopsis meeting in Vienna in July 2012, and will be published as part of the 1001 Genomes Project.

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 last summer (Segura *et al.*, *Nature Genetics, 2012*). We are currently working on an alternative method for allelic heterogeneity that estimates the variation attributable to particular loci rather than attempting to pinpoint causative sites.

Task 4: Modeling correlated phenotypes

A paper describing such a model/method was published last 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.

The report on Deliverable D10.1, submitted separately, concerns description of the necessary theoretical background. The numerical implementation is described for the case of factorial

experiments and linear mixed models, and currently depends on functions that could be found in open-source R packages. Extensions of the methodology and numerical procedures to other experimental situations and models are under development, in particular in the area of repeated measurement experiments with application in image phenotyping. In this report we also present a first version of a web tool performing the computations, and describe possible applications of this service.

Task 6: Computational aspects of sufficient data descriptors

For the study on statistical descriptors, phenotype data on 90 *Arabidopsis* lines from an automated phenotyping device (LemnaTec) was used as input for parameter reduction. Growth curves for leaf area and flowering measurements in time were analyzed. Two fitting procedures (Gompertz and Richards models) were used for model reduction, based on non-linear estimation techniques. These will be compared to results from IPG-PAS.

Information Theory, as a method for data compression, can also be used for unraveling the genotype-phenotype map and might be especially useful for complex traits. The Mutual Information principle was used for pairwise relationships between traits and markers and compared to three more mainstream analysis methods: LASSO and stepwise regression and a simple correlation-based test. For an optimal comparison between methods, simulation data was generated, using various genetic models (number of loci, heritability). Data was analyzed by all four methods and the power and false discovery rates were assessed. We concluded that multivariate regression (LASSO and stepwise) outperforms the univariate approaches in most cases. When a limited number of loci is involved Mutual Information outperformed the simple correlation test. Future investigations will focus on the use of Mutual Information in a multivariate approach.

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

none

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)*

none

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

GFMPG: 11.88 person-months (reaching 72% of the total planned)

IGR PAN: 9.13 person-months (reaching 25% of the total planned)

KG: 8.59 person-months (reaching 51% of the total planned)

If applicable, propose corrective actions

none

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 en 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. The underlying databases and information systems are distributed among the partners, but should be discoverable by an integrated search capability. A first iteration of search, implemented using Apace Solr technology, has already been deployed (see WP6 report). In this work package, we are developing a more sophisticated search with embedded capacity for ranking results according to their associated meta data, using the LAILAPS search engine.

Results of year 2

The software core of the metadata search engine (see deliverable 11.1 – search engine software core released and trained) has been developed. For this, the LAILAPS search engine (see http://lailaps.ipk-gatersleben.de/) was extended with enhanced concepts for a distributed search by reverse linked genome annotations and full text indexed metadata repositories. The partners' genome annotations pipelines link to databases collecting ontologies, protein sequence and other information such as passport, characterization and evaluation data for plant subspecies/cultivars. One example is IPK germplasm collection GBIS (http://gbis.ipk-gatersleben.de/gbis_i/). Twelve of the most popular databases are indexed in LAILAPS (see http://lailaps.ipk-gatersleben.de/indexed_databases.html). Those databases were used by the partners databases as source for function descriptions of transcripts, ontology terms, protein families. The "text" of this records are either copied into annotation description or linked by identifiers. From those annotation targets, LAILAPS links back to individual annotations contained within 12 genomics resources of transPLANT partners (see http://lailaps.ipk-gatersleben.de/linked_databases.html).

LAILAPS supports the integration by reverse identifier mapping. The sources of those mappings are sequence annotations that where provided by the partners or downloaded from UniProt ID mapping service. Based on this concept, LAILAPS supports the integrative search over the distributed data sources and links meta data (traits, gene functions, agronomic factors) to transPLANT partner databases.

The LAILAPS search engine in version 1.0 has released. The software is hosted at the IPK infrastructure and is running on a dedicated server, maintained by IPK administrators (permanent staff financed by IPK core funding).

An easy to use web frontend has been developed. In this web frontend the following features are implemented:

- Query suggestion functionality: This feature is also known as the “did you mean” function. During the typing of the query he systems shows dynamically similar terms which are indexed by LAILAPS.

- Special context based relevance ranking: The order of the search hits are computed by an artificial intelligence driven relevance ranking: The user has the possibility to give feedback about the quality of a hit. This feedback and the monitored user behaviour on the result page are used to improve the relevance ranking system automatically.

- Various faceted based filtering options for selection of annotation resources (indexed databases) or for selection of linked partner databases, for types of data links as well as for synonyms of the search terms.

- Detailed presentation of the original data sources for the linked databases in separate browser windows or tabs.

Furthermore, the maintenance of genome annotation is supported by an easy to use import interface. Here, the partner or further collaborators may upload their genome annotation that includes internally used identifiers, accession to mapped annotation, the referenced annotation database, and the annotation evidence.

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

The transPLANT portal acts as the central point of entry to the partner resources, and has been implemented in the content management system Drupal. To provide a seamless integration but modular integration of the metadata search engine LAILAPS an interface and GUI component has to be provided.

Results of year 2

For an easy integration of LAILPAS in different web sites or portals we have implemented a JavaScript portlet which can be used in any HTML-site. To give a proof of concept, this has been integrated in the IPK database OPTIMAS_DW (http://www.optimas-bioernergy.org/optimas_dw). To prepare the integration in the transPLANT portal hosted at the EBI, we have commenced an alternative implementation as a Drupal-Module (Drupal is the open-source content management system used to manage the transPLANT web site). First tests have been successful and the data sources will be provided to the partner EMBL for testing the integration in the transPLANT portal early in year 3 of the project.

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)*

The subtask "training of the LAILAPS relevance ranking system" was not reached until the end of August 2013. For this manual training it is essential to hire a scientist with background in plant biology, experience in text mining, and data curation. The acquisition of a suitable candidate was the reason for delay. Beginning from September 1st , 2013, a scientist was hired, who has long year experience in data curation for plant pathway and metabolic databases. We will finish this subtask at the end of February 2014. The delay of this subtask has no negative influence other tasks in WP 11 and/or other work packages. The results of the training can be included in the LAILPAS system at this later stage without any issues.

Use of resources

EMBL-EBI: 0.75 person-months (reaching 50% of the total planned)

IPK: 9.25 person-months (reaching 46% of the total planned)

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: TGAC

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 focuses 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. Resequencing 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 need to cope with heterozygosity and alloploidy (where the challenge is to distinguish between different homeologous haplotypes). We are compiling benchmark datasets for a variety of sequencing project objectives, to 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 include the following:

NGS assembly algorithms for large polyploidy plant genomes. The current assembly algorithms for NGS data have been designed for vertebrate diploid genomes (typically ~50% repeat content with

heterozygous diploid genomes). Although in some plants it 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 focuses on traditional quality metrics for consensus sequences (N50, number of contigs, completeness of genic regions) as well as the resources required such as RAM memory and performance. This task also covers the challenges around transcriptome sequence assembly.

Scaffolding algorithms to use read-pairs. t 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 long reads coming from single molecule technologies. A good and robust approach for scaffolding and more general genome refinement and finishing will help to implement better biology.

Resequencing 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 alloploidy species where it is required to distinguish homologous 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 no 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 is first focused on the analysis of selected algorithms by means of profile tools and tracing and visualization tools developed at BSC. These analyses are being performed both on supercomputers at BSC as well as resources made available by other consortium partners in the future. The results of these analyses will provide recommendations about the most appropriate computer architecture and programming model. Conclusions will be then used for 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 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 are testing methods, tools and pipelines for gene and repeat annotation. Their performances on complex plant genomes will be assessed. **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.

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 second year of the project we have focused on developing approaches to determine the quality of genome assemblies for species without a reference, and continued work exploring transcriptome assembly approaches.

Assessing assembly quality

There are a number of tools available for assembling genomic sequence, with a range of parameters that can be adjusted and will influence the quality of the resulting assembly. Typically a researcher will try a number of combinations, however there are few tools available to help them determine which of these approaches produced the highest quality assembly. Attempts to validate and systematize the choice of assembly algorithms, software or parameters - such as Assemblathon[3] and GAGE[18] - have been unable to give a definitive answer in all but the specific organisms chosen as test cases.

TGAC has been working to develop tools to address this issue, based on the principle that the genome information content reflected in the sequencing data should not be altered by the assembly process. We can model the accuracy of the assembly by checking how this information is conserved, and its completeness by checking how much of the information is included. We have been initially working with data from Illumina paired end datasets, looking at their kmer spectra (that is, the frequency at which small fixed-size motifs appear in the dataset). We have created a tool called the Kmer Analysis Toolkit (KAT), and used it to analyse correlation between motif frequency distributions. Exploring these distributions gives us information about the expected heterozygosity of samples, the rate of sequencing error and in some cases can identify contamination. This method does not make any assumptions about the underlying sequencing technology and is flexible to model biases and errors in the input set of reads.

We have compared the kmer spectra of the input reads to the appearance of the kmers in the final sequence files produced by well-known genome assemblers like ABySS[19] and SOAPdenovo[9]. Even for low copy number motifs, the results showed situations where the presence of motifs in the output from current assembly heuristics is not coherent with the read spectra (figure 1). We can use this approach to choose between a set of assemblies to identify which contains the most accurate copy numbers of motifs. We have validated KAT's output using sequencing reads generated for the reference genome of *Saccharomyces cerevisiae*. In addition, we can measure the completeness of an assembly, which we have seen is highly correlated to the number of genes present in the assembly as computed by CEGMA[15] (figure 2).

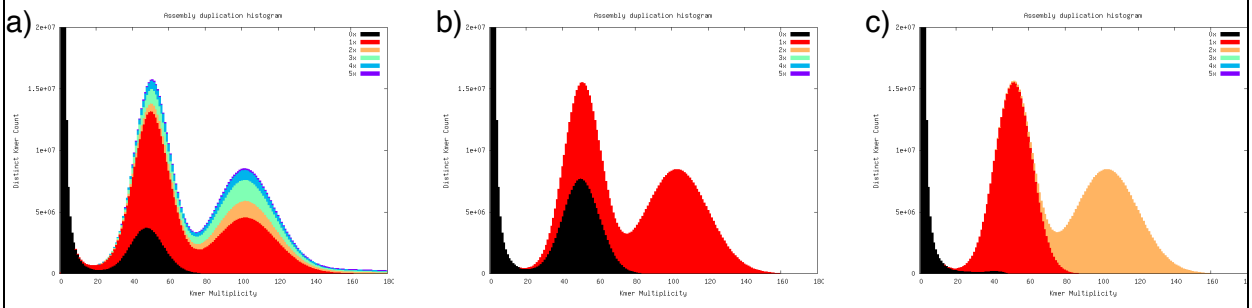


Figure 1: Analysis for the heterozygous *Fraxinus excelsior* sample "Tree 35". Kmer spectra of the reads, where colour indicates copy number on the assembly for (a) current assembly (b) theoretical scenario for haplotype collapsing (c) theoretical scenario for haplotype separation. High frequency kmers were omitted in the theoretical read spectra for simplicity.

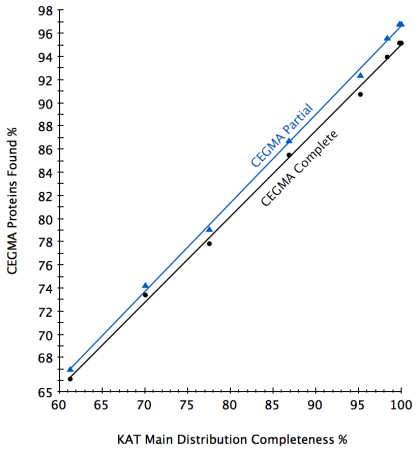


Figure 2: Correlation between percentage of items from the main distribution included in the assembly (x axis), versus percentage of CEGMA proteins found (y axis) in test dataset for *Hymenoscyphus pseudoalbidus*, the ash dieback pathogen.

A manuscript describing KAT is in preparation, and the software and source code will be made publicly available soon.

Transcriptome assembly

Continuing on from the work developed for milestone 27 (delivered in August 2012) TGAC has been comparing reference-free transcriptome assemblies to reference-guided assemblies using a high quality reference from a related species, or a draft reference from the species in question. We have used Trinity[6], a popular reference-free transcriptome assembler, and cufflinks[21], a popular reference-guided assembler to assemble sets of RNASeq data from the grass *Miscanthus*, using the sorghum genome as a related reference.

Trinity produces a much larger number of transcripts compared to both the sorghum-guided, or draft *Miscanthus*-guided assemblies. In addition, comparison of these transcripts to known plant proteins from public databanks (e.g. Uniprot[22]) reveals that many of the Trinity transcripts are truncated, or do not appear to code for proteins, whereas the sorghum reference-guided set has a much greater proportion of full-length transcripts and similarity to known coding sequences. However, performing reference-guided assemblies using a related species genome will result in species-specific transcripts being lost. Similarly using a draft genome from your species may also result in complete transcripts failing to assemble where regions are missing from the draft assembly (see Table 1). As such, we are exploring ways to combine reference-free and reference-guided sets of transcripts to develop a more comprehensive and high quality transcript set, which can be used for variant detection when working with a species which does not have a high quality reference available.

Transcript assembly type	**Number of transcripts**	**Transcripts with homologues**		**Transcripts without homologues**	
		Complete	**Partial**	**Putative ORF**	**No putative ORF**
De novo	450,200	95,399 (21%)	142,024 (32%)	45,367 (10%)	164,731 (37%)
Sorghum guided	45,535	32,796 (72%)	9,619 (21%)	188 (0.4%)	2,932 (6%)

Draft *Miscanthus* guided	91,516	8,465 (9%)	48,512 (53%)	7,019 (8%)	27,510 (30%)

Table1: Comparison of *Miscanthus* transcript models assembled using different approaches. Transcript models were compared to known proteins in Uniprot and were tested for presence of an Open Reading Frame (ORF) to identify putative novel coding transcripts.

INRA and BIOGEM have assessed the ability of different tools to detect SNPs and indels, particularly large indels, using Next Generation Sequencing (NGS) data with a reference genome assembly. The study focused on long indel (insertions/deletions) detection, anchored on a reference sequence, in the aim of having tools able to move from species description with a single reference genome toward pan-genome full description. We tested two different strategies, BIOGEM tested one based on Mate-Pair (MP) and INRA on Pair-End (PE) reads. **Results are presented in deliverable 12.1.** Briefly, we showed that the PE strategy appears to be more efficient than MP, and is also cheaper. The optimal strategy is to combine several tools for detecting different types of structural variants (SVs), MAPHiTS[11] for SNPs and small indels, Pindel[24] for medium ones, DOC[25] for large deletions and guided assembly with the Colombus module of Velvet[26] for large insertions. In addition, when the highest possible specificity is required, predicted SVs cross-validation between tools would be a very efficient strategy.

Task 2 Data structures for algorithm optimisation

Analysis of different strategies for genome assembly.

During this period BSC have carried out a computational analysis of several genome assembly strategies that were selected previously. This analysis consists of performing a detailed execution profiling, a computational performance analysis and the study of the data structures presented in the applications. The analysis is performed on supercomputers at BSC, the list of the assemblers under study are: abySS_1.35[19], SOAPdenovo2[10] Velvet[26], SGA[20], WGS[12], ALLPATHS-LG[1], MaSuRCA[27] and MIRA[2]. For space reasons, we only show some results for abySS_1.35, SOAPdenovo2 and MaSuRCA.

1) Some profiling results:

Figures 3, 4 and 5 show the profiling performed for AbySS_1.35 and SOAPdenovo2 and MaSuRCA assembly applications. The bar on the left of each figure shows how the execution time is spent among the main execution stages for each assembler. The other two bars of each figure show the execution time of the functions executed in the two main stages).

a) AbySS_1.35 is an assembly strategy based on the construction of a distributed de Bruijn graph. Their main stages are: *ABYSS stage:* where all possible substrings of length K (k-mers) are generated from the sequence reads, and then processed to remove read errors and initial contigs are built. On average this stage takes around 60% of the time. *Mapping stage:* where mate-pair information is used to extend contigs by resolving ambiguities in contig overlaps. On average this stage takes around 39% of time.

b) SOAPdenovo2 assembler also uses a de Bruijn graph strategy. It is designed to assemble Illumina short reads of large plant and animal genomes, although it also works on bacteria and fungi genomes. The profiling (figure 4) reveals that the main stages of the execution are: *Pregrap* stage with around 67% of execution time and *Mapping* stage takes around 30% of execution time. Those two stages are analogous to the described stages for AbySS_1.35.

c) MaSuRCA is a novel and more sophisticated assembly application that combines the efficiency of

the de Bruijn graph and Overlap-Layout-Consensus (OLC) approaches. MaSuRCA can assemble data sets containing only short reads from Illumina sequencing or a mixture of short reads and long reads (Sanger, 454). It is aimed to assemble very large-scale genomes of plants. The main stage of MaSuRCA corresponds to an error correction stage (quorum_error_cr steps in the figure 5, taking around 85% of the execution time). By performing this error correction it is possible to create a more simplified and accurate graph in the next stage, therefore, the computational complexity of the whole workflow is reduced.

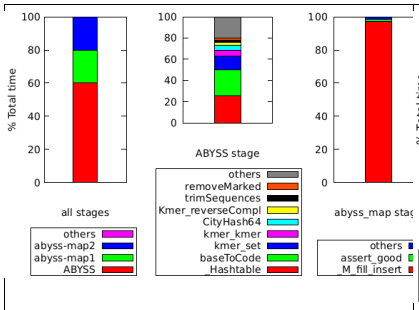


Figure 3

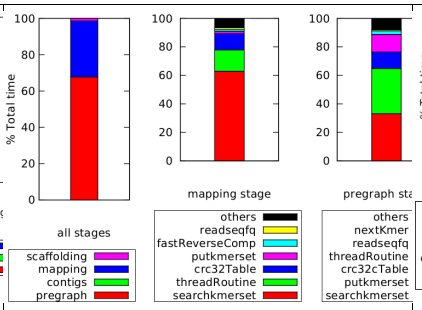


Figure 4

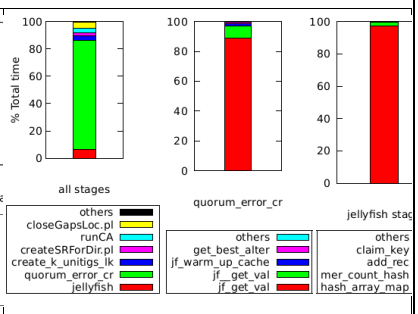


Figure 5

2) Some performance results:

After the profiling of the main stages of the assemblers, we are interested in understanding the parallel capabilities of them. It is important to remark that each application uses different parallel mechanisms in order to improve the performance (including sometime a combination of them: pthreads, MPI, etc). Also, It is common to find that some parts of workflows are parallel implementations and other parts have only sequential implementations. Therefore, we cannot expect a lineal speedup when they run over parallel machines. This issue is important because it shows that there is room for performance improvements if some parallel stages are optimized, or if some sequential parts are parallelized.

Figures 6, 7 and 8 show some performance scalability results for abySS_1.35, SOAPdenovo2 and MaSuRCA assemblers, when several processors are used in the execution. In each figure, the blue lines represent the scalability of the whole work-flow of each application. The other lines are the scalability of the main stages described in section 1 of this task.

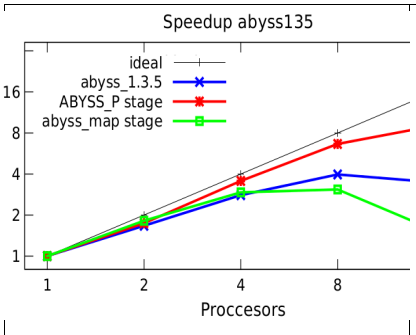


Figure 6

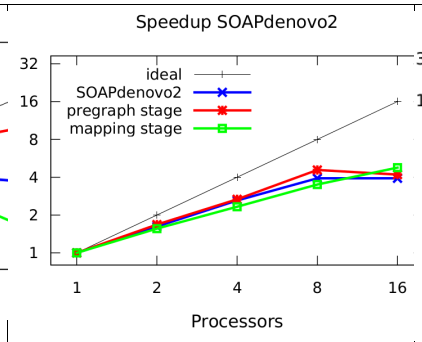


Figure 7

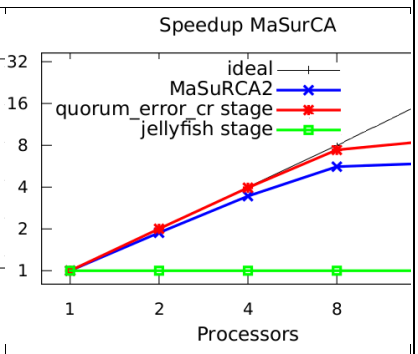


Figure 8

As we can see, the selected assemblers have a modest parallel behavior, globally speaking. However, some stages scale better than others. For example, for abySS_1.35 assembler, the first stage, that is *ABYSS* stage scales better than the *mapping* stage. Similar behavior can be found in MaSuRCA. Although theoretically assembly work-flows are rich in parallelism, it is important to take into account

that the dynamic behavior of the applications depends on several factors like: the input data set, the I/O capabilities and the configuration of the used machine, the way the application code is written, etc.

Currently we are using several computational tools in order to establish which optimization opportunities can be applied to improve the performance of the assembly workflows. It includes the use of ompSs programming model[14], the COMPss programming framework. Also, possibly the redesigning of some data structures that the de Bruijn graph and OLC strategies commonly use. The improvement of the data layout is also an option to exploit more data locality and concurrency, etc.

Additionally, we are installing the applications in the cloud environment developed for the transPLANT project, it is being done by creating a virtual machine where all the applications will be installed. Those installations will be available for all the users in the near future.

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 have improved the REPET package[5], now at its v2.2 release[17] adding a new pipeline based on Tallymer[8], called TallymerPipe, as a pre-processing tool for fast repeated region detection. Using the REPET pipelines, INRA tested a new strategy, to cope with very large genomes such as the wheat. This strategy is an iterative approach and can be summarized as follows:
1) Detection of the most easy to find transposable elements (TEs), with stringent parameters, to build a first TE library. These often correspond to young TEs and less degenerate ones.
2) TE annotation and splicing of the corresponding sequences from the initial contigs. We then obtain a reduced genome sequence.
3) Detection of the other TEs with sensitive parameters on the reduced genome sequence to build a second TE library.
4) Annotation of the original contigs with the concatenation of the two TE libraries.
The rational here is that these large genomes are mostly made of few TE families which are easy to find because they are present in a number of copies. They will be detected in the first step and this will allow us to reduce the genome size by an important factor. Using this approach INRA were able to reduce the wheat chromosome 3B from 986Mbp to ~230Mb, a reasonable size for running the second step, detection of TEs with sensitive parameters. **Milestone MS28** has been reached and described the REPET tools developed so far for repeat detection.

DLO have adjusted and improved the network-based biological process prediction tool BMRF[7]. Participation in the Critical Assessment of Function Annotation (CAFA) community assessment of protein function annotation demonstrated that BMRF performs particularly well in the presence of a (limited) set of existing function annotation[16]. Currently, a prototype web tool to allow access to the resulting sets of predicted gene function annotations is available (www.ab.wur.nl/bmrf). Prediction power was further improved by combining BMRF with Argot2[4], an orthogonal sequence-based method that also performed well in CAFA. The combination of BMRF/Argot2 using co-expression networks outperformed the individual methods significantly. The combined method was applied to the proteomes of several crop species (rice, poplar, soybean, tomato, potato in progress). It allows tentatively annotating the wealth of proteins for which there is no knowledge on function and it offers a means to analyze variability of biological process functions across plant species. The combined BMRF/Argot2 approach is submitted for publication. Results are available online. Furthermore, initial tests indicate that combining BMRF with text-mining approaches[23] also helps to improve performance.

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. Knowledge of the functions of such genes may help selection. 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'[13] 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 being offered to the wider plant community as an infrastructural tool.

Current plant ontologies are yet not sufficiently developed to make good use of the 'nanopublication' concept in plant genomics. More efforts towards such ontologies are necessary. To contribute to candidate gene prediction, the link between prediction of biological processes (with BMRF) and QTL regions was explored in QTL and GWAS datasets from Arabidopsis and rice. Such an approach would enable us to find connections between traits (phenotype) and biological processes and facilitate the integration of trait QTLs with eQTLs. Preliminary results demonstrate the added value of such approaches for candidate gene prediction. The concept of 'virtual plant breeding' was discussed with various stakeholders and will result in defining next activities. A first version of a 'position paper' describing the current status of virtual plant breeding and the activities DLO proposes to perform was written and is being discussed internally. This document will subsequently be distributed among the transPLANT partners.

References

[1] Butler et al., ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008 18(5):810-20

[2] Chevreux et al., Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99. 1999 pp. 45-56

[3] Earl et al., Assemblathon 1: a competitive assessment of de novo short read assembly methods Genome Res. 2011 21(12):2224-41

[4] Falda et al., Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. BMC Bioinformatics. 2012 13(Suppl 4):S14

[5] Flutre et al., Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011 31;6(1):e16526

[6] Grabherr et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 29(7):644-52

[7] Kourmpetis et al., Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. Plant Physiol. 2011 155(1):271-81

[8] Kurtz et al., A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008 9:517

[9] Li et al., De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010 1:265–272

[10] Luo et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo

assembler. Gigascience. 2012 1(1):18

[11] MAPHiTS http://urgi.versailles.inra.fr/Tools/MAPHiTS Cited on 21/09/13

[12] Myers et al., A Whole-Genome Assembly of Drosophila. Science. 2000 287:2196-2204

[13] Nanopublications http://www.nanopub.org/ Cited on 14/09/13

[14] OmpSs http://pm.bsc.es/ompss Cited on 22/09/13

[15] Parra et al., CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007 23(9):1061-7

[16] Radivojac et al., A large-scale evaluation of computational protein function prediction Nature Methods 2013 10:221–227

[17] REPET http://urgi.versailles.inra.fr/Tools/REPET Cited on 21/09/13

[18] Salzberg et al., GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012 22(3):557-67

[19] Simpson et al., ABySS: a parallel assembler for short read sequence data. Genome Res. 2009 19(6):1117-23

[20] Simpson et al., Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012 22(3):549-56

[21] Trapnell et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 28(5):511-5

[22] Uniprot Consortium, Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 2013 41: D43-D47

[23] van Landeghem et al., The potential of text mining in data integration and network biology for plant research: a case study on Arabidopsis. Plant Cell. 2013 25(3):794-807

[24] Ye et al., Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009 25(21):2865-71

[25] Yoon et al., Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009 19(9):1586–1592

[26] Zerbino & Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 18:821-829

[27] Zimin et al., The MaSuRCA genome assembler. Bioinformatics. 2013 Sep 6.

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

n/a

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)

n/a

Use of resources

INRA: 6.93 person-months (reaching 103% of the total planned)

TGAC: 12.26 person-months (reaching 60% of the total planned)

BSC: 5 person-months (reaching 39% of the total planned)

DLO: 6 person-months (reaching 64% of the total planned)

3. Deliverables and milestone tables

Deliverables

Del. no. 4	Deliverable name	WP no.	Lead beneficiary	Nature5	Dissemination level6	Delivery date (proj. month)7	Actual delivery date	Comments
D7.2	Interfaces for integrating omics data within the transPLANT user interface	WP7	EMBL	P	PU	Month 18	27.3.2013	
D2.1	A report entitled "Translational research for agronomical application"	WP2	INRA	R	PU	Month 24	31.8.2013	Delivered on revised schedule agreed with project officer at first review meeting
D3.1	Recommended ontology set for use in phenotype description and epigenetic variability	WP3	INRA	R	PU	Month 24	31.8.2013	
D3.2	Format specifications for data exchange by flat file and web services	WP3	EMBL	R	PU	Month 24	31.8.2013	
D5.1	Updated data warehouses developed for genomic annotation and	WP5	INRA	P	PU	Month 24	31.8.2013	Delivered on revised sched

[4] 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.

[5] Please indicate the nature of the deliverable using one of the following codes:
R = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other

[6] 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).

[7] 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.

	variation data							**ule agreed with project officer at first review meeting**
D9.1	**Variation repository, first release**	**WP9**	**EMBL**	**P**	**PU**	**Month 24**	**31.8.2013**	
D10.1	**Statistical descriptors for genotype-phenotype map construction**	**WP10**	**PAS**	**R**	**PU**	**Month 24**	**31.8.2013**	
D10.2	**Software for analysis of genome-wide association data**	**WP10**	**GMI**	**P**	**PU**	**Month 24**	**31.8.2013**	
D11.1	**Search engine software core released and trained**	**WP11**	**IPK**	**P**	**PU**	**Month 24**	**31.8.2013**	
D12.1	**Development and test of sophisticated statistical methods to model variation in large plant genomes**	**WP12**	**INRA**	**P**	**PU**	**Month 24**	**31.8.2013**	

Milestones

Milestone no	**Milestone name**	**WP no**	**Lead beneficiary**	**Delivery date from Annex I**	**Achieved (Yes/ No)**	**Actual / Forecast achievement date**	**Comments**
MS9	1st transPLANT training workshop	WP4	HMGU	Month 12	Yes	13.11.2012	
MS2	1st annual report submitted	WP1	EMBL-EBI	Month 14	Yes	27.9.2012	
MS25	Software core released	WP11	EMBL-EBI	Month 18	Yes	21.01.2013	
MS10	2nd transPLANT training workshop	WP4	EMBL-EBI	Month 24	Yes	31.8.2013	
MS13	DAS servers provided for sequence and	WP5	EMBL-EBI	Month 24	Yes	31.8.2013	

	annotation for 15 reference genomes						
MS16	Ensembl Plants, MIPS Plants DB and GnpIS integrated in transPLANT portal	WP6	EMBL -EBI	Month 24	Yes	31.8.2013	
MS19	15 reference genomes incorporated in transPLANT hub and submitted to comparative analysis	WP7	EMBL -EBI	Month 24	Yes	22.7.2013	
MS23	Initial public launch of variation repository	WP9	EMBL -EBI	Month 24	Yes	31.8.2013	
MS28	Software for the analysis of repeats	WP12	INRA	Month 24	Yes	31.8.2013	

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 1.9.2012 **to** 31.08.2013

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 M13 to M24 (from 1.9.2012 to 31.8.2013).

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 coordinate with the European Commission on technical aspects of the management of the contract.

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. A project manager is employed, alongside the scientific coordinator, to mange these 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.

We use a general **mailing list** eu_plants@ebi.ac.uk, to which all partners are subscribed.

The second **Annual General Meeting** (AGM) was organized at the European Bioinformatics Institute in Hinxton (UK) on February 14-15, 2013 and was attended by 28 participants, representing all consortium partner institutions. Each work package leader presented the strategy towards the second years' objectives, in particular the preparation of the 10 deliverables and 8 milestones due at the end of year 2, also taking into account the comments received during the first Commission's review in October 2012.

Additional **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 supported by the project manager, in particular by keeping track of agreed action points 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 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. To date 53 technical issues have been recorded and 39 have been resolved, the remaining ones being currently dealt with by the EBI team.

Issues reported by external users (via the links to provide feedback available on the pages of the transPLANT website and transPLANT-supported services) are converted into internal JIRA issues as appropriate after review by developers.

Co-operation with other projects/programmes

Work package 2 is dedicated to "Interaction with national and trans-national genomics and informatics activities". As part of this program, on February 28 – March 2, 2013 the meeting "Genomes to Germplasm" was hosted by the partner INRA in Versailles (France) to better coordinate strategic thinking, coordinated planning and potential joint funding through the joint organisation of transPLANT, Gramene, and the Plant Bioinformatics Working Group of the EC-US Task Force on Biotechnology Research. Representatives of many international plant-genomics-themed projects were invited to participate.
In addition, in work package 2, we have also been connecting with other projects and initiatives through the user survey; and also through the establishment of an open project mailing list. More details are provided in the work package report and the associated deliverables.

In the context of the **work package 3** (standardisation), -a number of meetings have been held with representatives of other projects, in order to support the development of global standards:

P. Krajewski (IPG PAS), C. Pommier (INRA) and D. Bolser (EMBL-EBI) participated in the Crop Plant Trait Ontology Workshop at Oregon State University 2012, Corvallis, Oregon, USA, 13-15.09.2012.

- P. Krajewski and H. Ćwiek participated in the 4th Scientific Workshop of POLAPGEN-BD project in Kraków, Poland, 25-26.03.2013.

- H. Ćwiek participated in 10th Extended Semantic Web Conference & Semantics for Biodiversity Workshop in Montpellier, France, 26-30.05.2013. She presented to the Crop Ontology group the work done in transPLANT and discussed the course of collaboration.

- D. Bolser participated in the PRO-PO-GO Meeting, Buffalo University, 15-16.05.2013. The meeting was proposed to promote the coordination of the Gene, Protein, and Plant Ontologies and of other reference ontologies used in plant biology.

These meetings are described in more detail in the report on activities undertaken in work package 3.

Also in the area of standards, transPLANT has been contacted by the FP7 AgINFRA project to explore involvement in a working group established by AgINFRA to study the interoperability issues of germplasm databases and to publish germplasm descriptions in Linked Data format.. After internal discussions, transPLANT partner IPK, which has been previously involved (together with other gene banks) in similar efforts (for example, in extending the Dublin Core standard for gene banks), has joined the AgINFRA working group).

The training workshops organised in **work package 4** have included guest presentations from participants in other FP7-funded programs, including the TriticeaeGenome and POLAPGEN projects. Participants in the meeting have included individuals involved in numerous other projects and programs. For full details, see the report on work package 4.

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 (IPG PAN), 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). During the second project's period, the Strategy Committee met in person at the Annual General Meeting (February 14-15, 2013 at Hinxton, UK) and communicated via teleconference on December 3, 2012 and June 3, 2013. Minutes of strategy committee meetings are written by the project manager and are communicated to the full consortium via the internal project website.

c. Reporting to the EU

During the period M13 – M24, ten deliverables have been submitted to the Project Officer on time. The achievement of the 8 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

All WPs: The second **Annual General Meeting** (AGM) was organized at the European Bioinformatics Institute in Hinxton (UK) on February 14-15, 2013 (see above).

WP5: COMPSs tutorial at Hinxton (UK) on February 13-14, 2013. Participants: INRA, EBI, BSC, IPG PAN.

WP6: teleconference organized by the Search Working Group March 2013. Participants: EBI, INRA, PAS, IPK, HGMU, GMI.

WP8/WP12: Visioconference "transplant structural variants tools benchmark" on May 21, 2013. Participants: INRA, BIOGEM, KN.

WP10: Visit of P. Krajewski (IPG PAS) to KeyGene on November 5, 2012.

Monthly teleconferences with the involvement of all project participants.

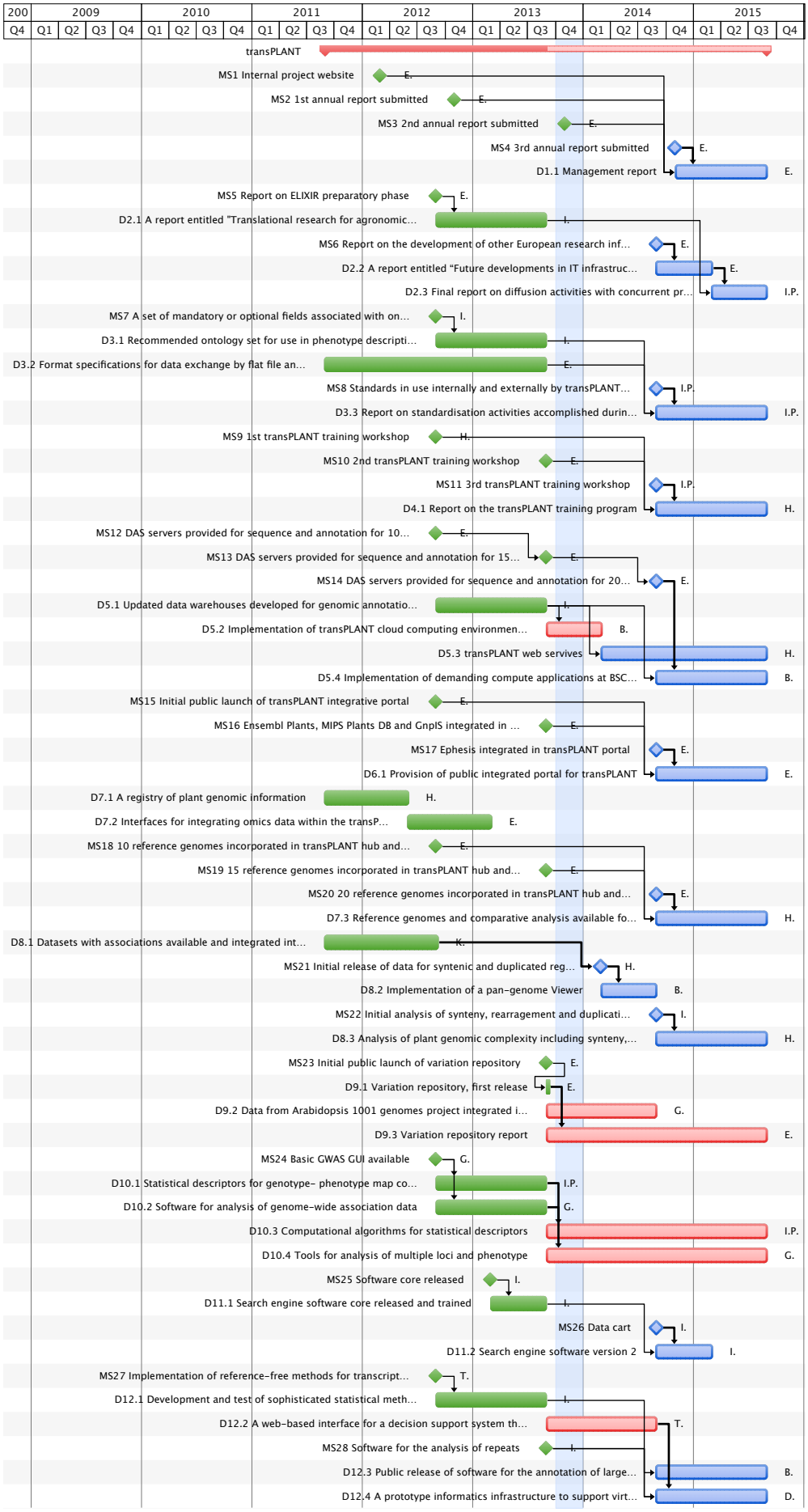
4. Development of the project website

A public website has been developed and is hosted by the partner EBI, coordinator of the project. The URL is http://transplantdb.eu. The release of the public portal was achieved on March 2012, and reported as the milestone MS15. Details of the latest developments to the site are provided in the report on work package 5.

5. Project planning and status

As of month 24, the ten deliverables that were due for submission have been duly submitted. 8 project's milestones due until month 24 have been reached. 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 against the original work plan. Horizontal bars represent deliverables, and diamonds represent milestones. A colour code indicates the status of the work. Green indicates achieved deliverables and milestones. Red indicates ongoing work, and blue indicates planned activities, not yet started. The blue vertical bar paints the current quarter at the time of the report.


200
2009
2010
2011
2012
2013
2014
2015
Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
transPLANT
MS1 Internal project website
E.
MS2 1st annual report submitted
E.
MS3 2nd annual report submitted
E.
MS4 3rd annual report submitted
E.
D1.1 Management report
E.
MS5 Report on ELIXIR preparatory phase
E.
D2.1 A report entitled "Translational research for agronomic...
I.
MS6 Report on the development of other European research inf...
E.
D2.2 A report entitled "Future developments in IT infrastruc...
E.
D2.3 Final report on diffusion activities with concurrent pr...
I.P.
MS7 A set of mandatory or optional fields associated with on...
I.
D3.1 Recommended ontology set for use in phenotype descripti...
I.
D3.2 Format specifications for data exchange by flat file an...
E.
MS8 Standards in use internally and externally by transPLANT...
I.P.
D3.3 Report on standardisation activities accomplished durin...
I.P.
MS9 1st transPLANT training workshop
H.
MS10 2nd transPLANT training workshop
E.
MS11 3rd transPLANT training workshop
I.P.
D4.1 Report on the transPLANT training program
H.
MS12 DAS servers provided for sequence and annotation for 10...
E.
MS13 DAS servers provided for sequence and annotation for 15...
E.
MS14 DAS servers provided for sequence and annotation for 20...
E.
D5.1 Updated data warehouses developed for genomic annotatio...
I.
D5.2 Implementation of transPLANT cloud computing environmen...
B.
D5.3 transPLANT web servives
H.
D5.4 Implementation of demanding compute applications at BSC...
B.
MS15 Initial public launch of transPLANT integrative portal
E.
MS16 Ensembl Plants, MIPS Plants DB and GnpIS integrated in ...
E.
MS17 Ephesis integrated in transPLANT portal
E.
D6.1 Provision of public integrated portal for transPLANT
E.
D7.1 A registry of plant genomic information
H.
D7.2 Interfaces for integrating omics data within the transP...
E.
MS18 10 reference genomes incorporated in transPLANT hub and...
E.
MS19 15 reference genomes incorporated in transPLANT hub and...
E.
MS20 20 reference genomes incorporated in transPLANT hub and...
E.
D7.3 Reference genomes and comparative analysis available fo...
H.
D8.1 Datasets with associations available and integrated int...
K.
MS21 Initial release of data for syntenic and duplicated reg...
H.
D8.2 Implementation of a pan-genome Viewer
B.
MS22 Initial analysis of synteny, rearragement and duplicati...
I.
D8.3 Analysis of plant genomic complexity including synteny,...
H.
MS23 Initial public launch of variation repository
E.
D9.1 Variation repository, first release
E.
D9.2 Data from Arabidopsis 1001 genomes project integrated i...
G.
D9.3 Variation repository report
E.
MS24 Basic GWAS GUI available
G.
D10.1 Statistical descriptors for genotype- phenotype map co...
I.P.
D10.2 Software for analysis of genome-wide association data
G.
D10.3 Computational algorithms for statistical descriptors
I.P.
D10.4 Tools for analysis of multiple loci and phenotype
G.
MS25 Software core released
I.
D11.1 Search engine software core released and trained
I.
MS26 Data cart
I.
D11.2 Search engine software version 2
I.
MS27 Implementation of reference-free methods for transcript...
T.
D12.1 Development and test of sophisticated statistical meth...
I.
D12.2 A web-based interface for a decision support system th...
T.
MS28 Software for the analysis of repeats
I.
D12.3 Public release of software for the annotation of large...
B.
D12.4 A prototype informatics infrastructure to support virt...
D.