Project No. *283496*

**transPLANT**

**Trans-national Infrastructure for Plant Genomic Science**

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

# D2.1

**Report: "Translational research for agronomical application"**

Due date of deliverable: 31.8.2013
Actual submission date: 18.9.2013

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: INRA

| **Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)** | | |
|---|---|---|
| **Dissemination Level** | | |
| **PU** | Public | x |
| **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) | |

| **Contributor** |
| --- |
| EBI, HMGU, INRA |

| **Introduction** |
| --- |
| Deliverable reference number:D2.1<br><br>This community-authored report identifies infrastructure needs for translational application of plant science research. It has been drafted during the 1st transPLANT external stakeholder meeting, and refined remotely thereafter.<br><br>The 1st transPLANT external stakeholder meeting called "Genomes to Germplasm" was co-organised by the Plant Bioinformatics Working Group of the EC-US Task Force on Biotechnology Research, the transPLANT project, and the Gramene project.<br><br>This meeting brought together 40 research scientists, informaticians and crop breeders from Europe (23), USA (16) and 1 from the Philippines. They addressed the biological and informatic challenges associated with current (and likely future) attempts to catalogue genomic variation and apply it to increase our understanding of plant biology and improve crop plants.<br><br>Specific points of discussion included (i) how natural variation is being sampled, and the likely future applications of these 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 was held at Versailles in the INRA campus from February 28th to March 2nd 2013. It was professionally facilitated and designed to ensure that the discussions were shaped by the ideas of the participants.<br><br>We shared ideas and thoughts in advance and during the meeting on a web site, using a social networking platform called Ning, customized to suit our needs. It also allows users to share documents and to engage in early discussion in a closed forum (restricted to meeting attendees and other identified individuals).<br><br>At the beginning, the organisers (Paul Kersey, Klaus Mayer, Hadi Quesneville, and Doreen Ware) sketched some ideas about likely areas of interest and recommended some papers that might be read in advance of the meeting.<br><br>During the meeting, small groups were formed to discuss specific topics. The results of their discussions were then presented to the whole audience for further discussions. New groups were formed according to the topics rising from these discussions. Participants were encouraged to engage with multiple groups on the full range of topics covered in the meeting, before eventually contributing to the later stages of the meeting in their domains of greatest expertise. At the end of this process, ideas and thoughts were shared and summarised on the Ning web site. We drafted this report from the material found there.<br><br>**Next Steps**<br>We are continuing to work, together with participants at the meeting, on a further iteration of the report, in |

order to prepare a document suitable for publication in an appropriate scientific journal. The intended publication will provide a focal point for further community engagement and for a subsequent process of continual refinement based on contributions and feedback received and on the evolving scientific need.

**Perspective: User survey**

To widen the potential field of user input and opinions, and following the 2012 EC review recommendations, a new large 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 analysing, 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 is online since the 6th June 2013 and was distributed to different user networks. Seventy persons have already answered it (it is still open). We intend, to analyses more deeply the results of this survey in autumn 2013, and to provide a report on this analysis.

**Results (if applicable, interactions with other workpackages)**

The discussion results of the meeting can be seen on the Ning web site (http://genomes2germplasm.ning.com/; N.B. this site is password protected) The content is summarized as follows.

# 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/Excel-Data/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

Challenges that need to be addressed by basic science in a 5 to 10 years future include:

- 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 require (i) the generatation of a complete inventory of plant genetic and phenotypic diversity, (ii) to characterization of the plant microbiome, and (iii) a better understanding individual plants in terms of their local ecosystems.
- We should be able to engineer plant biology to fulfill specific goals. Predictive modeling should have a positive impact on understanding plant phenotypes. Hence future plant improvements should result not only from genetic improvement, but also from modification of abiotic environments or biotic conditions, e.g. through modification of 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. long term development potential).

# 3 Translational biology

Society will benefits from the translation of advances in basic sciences. Challenges for a more "applied" science in a 5 to 10 years future will be:

- More rapidy development of new plant varieties, through conventional breeding methods but also by direct genetic editing (i.e. GMO). 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. Traditionally, molecular biology has been too expensive for breeders to be interested in, but costs are falling. 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 means to deliver more quickly new varieties that are adapted to a changing world. The ultimate goal here is breeding by design: the rapid development of new crops matching specifications determined by the relevant environmental or economic niches.
- Synthetic biology should provide new markets for plant products (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. through the growth of new organ types for the storage of products closer to consumption products).
- Maintaining biodiversity of existing and future crop species, facilitating 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 behaviors. 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 itself 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 for 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 field work 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 problem to solve are scale and variety of data necessary to deliver this vision.

## *4.1 Enabling technology (infrastructure)*

Through a better integration of genomic data, breeding should be more efficient as well as molecular breeding, transgenesis, and mutagenesis.

In the age of data-driven science, every biologist needs the tools and skills necessary to work as a bioinformatician – and data analysis becomes a routine part of every biologist's professional life. To support this, new types of knowledge store become necessary: classical "libraries" need to be replaced by queryable, digital archives, providing access to genomic and phenotypic information. Some questions are plant specific but most are shared questions with other domains. There is thus the potential for the re-use of solutions already found by other communities (e.g. environmental researchers, human health researchers, etc…), adapting them to accommodate specific features of plant biology, and their application to crops.

Data models, standards, and infrastructures for their use exist in some form but are currently insufficient. Compared to the recent past, the scale of the data has changed, and mode of operation has changed as well. There is a need for distributed models for data management, decoupling knowledge stewardship (which needs to be decentralized to appropriate experts) from service provision (which depends on certain universal standards, but which is itself a distributable problem). Core (universal, stable) data objects 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, but seemless access needs to be provided making the actual location of data transparent to users.

In that context, data discovery should be promoted allowing to search heterogeneous data at various levels (from raw information to ontology structured information). Queries in natural languages should be possible. Automation of data capture is essential for scalability, but in parallel, mechanisms for showing and improving data quality should be developed using user feedback.

Technologies to support distributed development need to be widely adopted. A common API is needed to bind separate national and domain-focused resources into an effective federated database. Toolkits must be made available for implementation of services, as well as a comprehensive test suite to ensure the reliability of the API over time.. Ontologies, defining controlled lists of terms (usable in annotation) and the relationships between them, must be further developed as they represent important keys to query databases: many existing ontologies (e.g. for plant morphology, experimental meta data, etc., are currently under-developed and, where extent, under-used). Dynamic data exchange between databases must be also facilitated and encouraged, enabling users to see the same data integrated in many contexts.

Challenges and obstacles include 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 data sharing, and the need for a sustainable model 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 commercial entities will need and expect, and the appropriate interfaces between public and private research, need to be determined.

### *4.2 Tools*

We need a better toolset to improve our ability to model agricultural and environment management practices. Tools implementing methods that will optimize integration of biological, environmental and agricultural data will improve prediction of phenotypes and the development/targeting of appropriate crops for particular environments.. Integration of systems biology information will help to make breeding decisions.

New computational methods need to be made available to a broad user community more quickly and reliably, through documented, tested and reliable software. In the developmental phase, software testing needs to include evaluation of scalability (usability for very large data sets) so that resources are invested in projects with a viable future. The right model will combine support for the competitive development of new ideas and algorithms with a longer term view to provide professional and/or community support for software packages proven to be of value to a broad user base.

## 5 Implementation in specific areas

A cyberinfrastructure is needed support the general challenges described in “Science” and “Translation”. It is discussed below in terms of specific components.

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

We are moving towards a full genetic and increasing phenotypic characterization of many plant strains. Ideally , this will include not only the genome sequences of all crops, but also their wild relatives, and provide access to the haplotype structures of each individual strain/stock for the purpose of enabling breeding by design. To achieve this goal, the data needs to be integrated, structured and made available to users. As a first step,we need collaborative characterization of large panels of genetic resources (passport or research-derived), an easier exchange 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*

In essence, a genotype is fixed for a stock and the set of genotypic information is finite and universal. Epigenomes can be seen more as phenotypes, as they are not finite and universal, because depending on development stage, cell type, and environments. We propose the development of 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. Current models developed for humans are not powerful enough for plant genome complexity.

### 5.3 Phenotyping

The nature of phenotyping in plant systems has changed rapidly in recent years, encompassing molecular phenotyping (e.g. metabolomics profiling), large scale automated phenotyping (in specially designed greenhouses), and increasingly sophisticated methods of field measurement.. Yet experimental phenotyping and field phenotyping have different characteristics. 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. Phenotype data is enormously diverse but a basic model can consider a phenotype could be as attribute, a measurement and experimental metadata, describing the conditions under which a phenotype was observed. But any data model will need to be highly flexible and extensible, and capable of connecting with a large number of different data archives each differing in the information they contain.

### 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, etc. of plant response to biotic and abiotic environments. The strategy is to decipher the biology of traits by crossing diverse levels of information on reference populations.
To reach this goal, we first need to be able to 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 genotype data and made available for the development of predictive models.

### 5.5 Crop Breeding

Selection programs using all genomic information as a whole (also called genomic selection) should use also genetic and epigenetic information integrated into information systems. Fast pre-screening system for genetic resources will speed up generation cycles. Methods must be high throughput and cost-efficient. Indeed, further falls in sequencing costs are required before genomic selection can be applied to minor crops. Identification of the genes controlling recombination should allow increased recombination to access to genes in low recombination regions.

## 6 Teaching and Training

The acute shortage of training capacity in the relevant skill set is becoming a serious problem. Increasing

numbers of scientists and breeders are needed with a broad 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. Deeper and wider channels of communication are required, strengthening ties between technologists, molecular biologists, germplasm collections, and breeders. Where possible, scalable e-learning methods should be developed to meet 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. It is appropriate to recognize that this may derive from genuine dislike of the purposes to which scientific knowledge is applied or from rational skepticism about scientific claims. But it may also result from ignorance, mis-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. Without success in this area, the ability of plant science to deliver social benefits on a large scale will be seriously reduced, even as the need for them increases.

| **Publications** |
| --- |
| 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. |