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

WP2: Report on User survey

transPLANT project, October 2014
Delphine Steinbach INRA URGI

Introduction

A user survey was launched in June 2013 (end of year 2) to collect the bioinformatics stakeholders' needs in the field of agronomical research.

The objectives were :

- To identify potential needs that are not covered by transPLANT project.
- To help define the landscape and possible overlaps with other projects.
- To better coordinate development in the field and to avoid redundancies.

The survey was online the 6th June 2013 and is now closed.

It was made accessible via the transPLANT web site here.

It was distributed to different user networks, to both scientists from academic and private sectors, working on wheat, barley, maize, pea, sunflower, rapeseed…, genomics and genetics. More than 200 people were reached personally by email.

It invited participants to answer a mixture of multiple choice and free text questions (41 in total), covering topics about biological data types, databases, analysis tools and infrastructure needs.

- The first section gets information on the person answering the survey.
- The second gets information on data that 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 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 asks questions about existing projects in which the user is involved and his expectations about the outcomes of the transPLANT project.
- All the questions are fully described on transPLANT web site at: http://www.transplantdb.eu/sites/transplantdb.eu/files/UserNeeds_Survey_Questions.pdf

In 2014 we have analysed the results of the user survey to build this report that can be also addressed directly on the new transplant web site at this url: http://www.transplantdb.eu/survey

Summary

Main results:

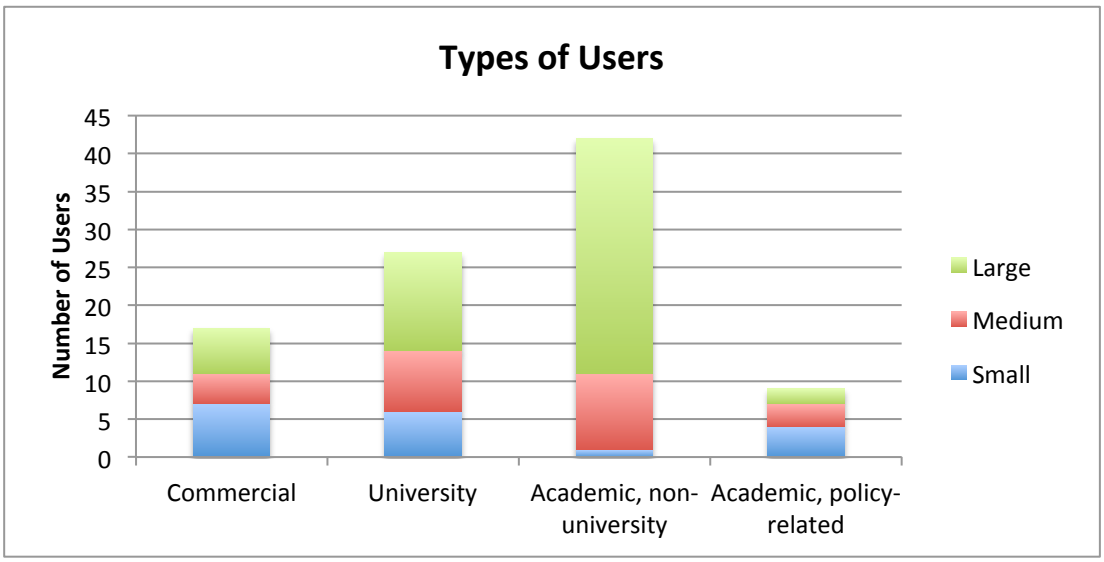
- Of the 74 respondents, 80% have the problem of data integration. They have difficulties because of the scale of the data (56,3%): algorithms and tools are not well adapted. They have limited information about strengths and weaknesses of available software (46,9%).
- However, they agree to share with others their elaborated (compute and expert) data concerning genomic sequencing (55%), resequencing (52,5%), genotyping (50%), and RNA-Seq derived expression data (52,5%).
- Respondents have problems in analysing their data because of species complexity (43,8 %), rapid technological change (40,6%), and because of problems hiring skilled staff.
- To solve these problems, respondents would like to have more user training sessions (85,7%), to have a unique web portal (71,4%) that is easy to query with guidelines to access tools and data resources. They would like to have a forum (45,7%) to exchange with people and to receive newsletters (28,6%) to be able to get summary information. Their preferred way to be informed is on their own (55%), by

mailing (51%) and by newsletter (41%).

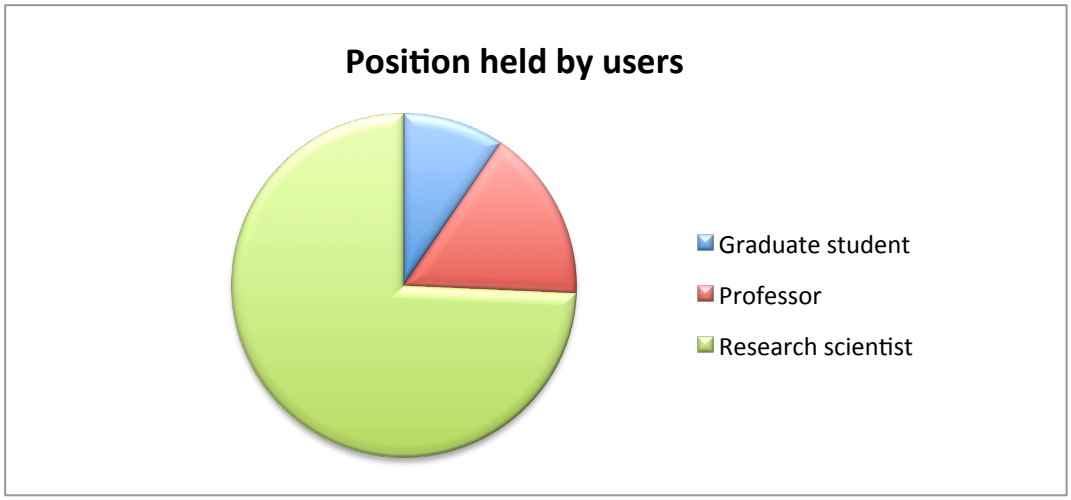
Detailed report:

- There were 74 respondents to the survey, from several different countries, institutes, universities or private compagnies.
    - **13 different institutes**:
        - INRA (28/74 answers), Embrapa, CNRS, Ecole Nationale Superieure Lyon, Agro Bio Institute, GEVES, Welience, CIRAD, CEA/Genoscope (French sequencing center),
        - Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
        - Institute for Agricultural and Fisheries research (ILVO)
        - Institute of Plant Genetics Polish Academy of Sciences
        - John Innes Center
    - **11 universities:**
        - University of Paris-Sud, Agrocampus (France)
        - University of Zurich
        - University of Agriculture in Krakow, Poznań University of Life Sciences, Poznańska Hodowla Roślin Sp. z o. o, Adam Mickiewicz University, Poznan, Poland
        - Maastricht University
        - Wageningen university
        - Aberystwyth University
        - University of Nottingham
    - **6 compagnies:**
        - KWS LOCHOW GMBH, Biogemma, Vilmorin, Florimont-Desprez, Maisadour semences, Syngenta seeds.

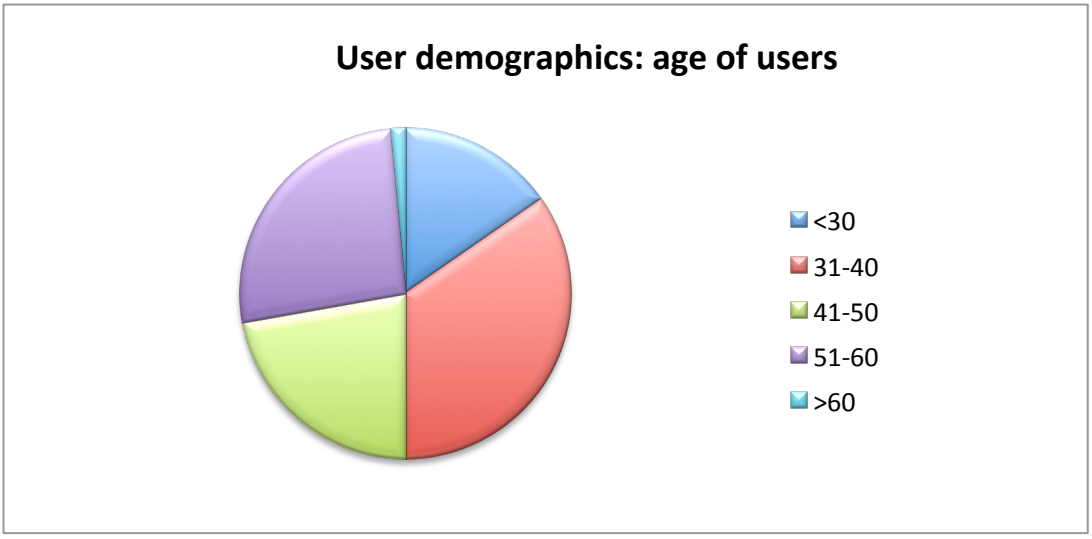
See figure below showing category of provenance according to size of their hosting structure.



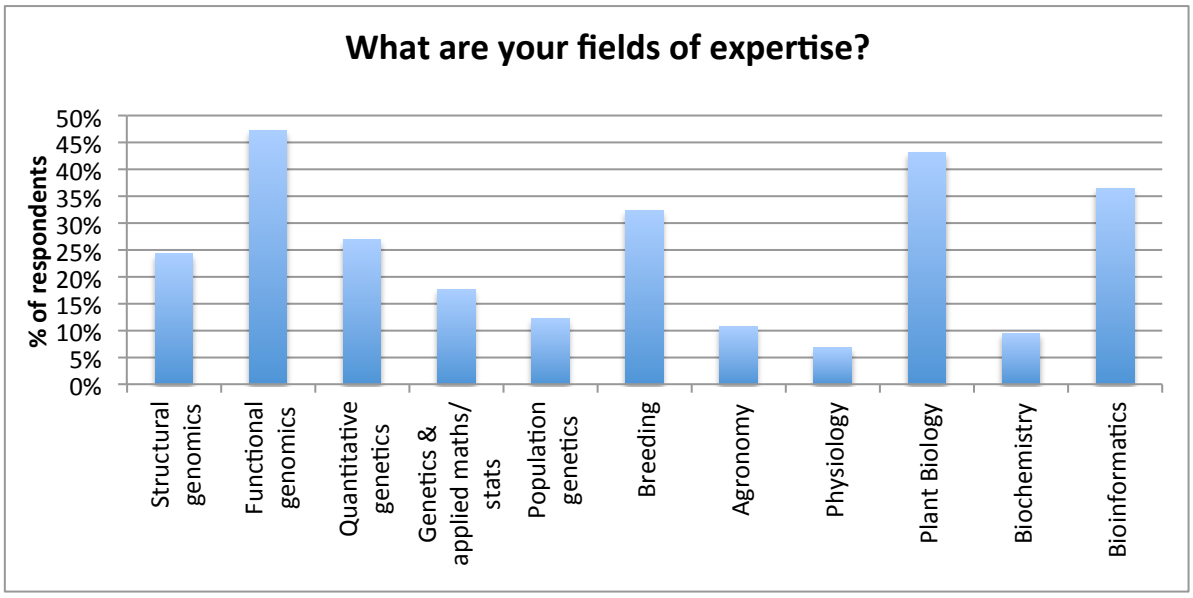
- 74,3% of respondents were research scientists, 16% were professor, 9,5% were graduate students.



- 48% % were older than 40 years, 34,7% between 31 and 40 and 15% under 30 years.
  See graphics below:



- **47% had expertise in functional genomics, 43% in plant biology**
- 36,5% in bioinformatics, 32,4% in breeding,
- 27% in quantitative genetics, 24% in structural genomics.
  See figure below:



- **They are interested in several categories of data that can be grouped in 3 classes:**
  - **genome sequences (NGS included) 49%, sequences variation (SNP, CNV, PAV) 35%, phenotyping (experiments) 33%**
  - genetics maps markers 27% or QTLS 22%, segregating populations or LD panels 22%, curated gene models 22%, resources for natural genetic diversity 22%, phenotypes (ontology) 22%
  - metabolic profiles 13% and protein profiles 11%

    See full table below

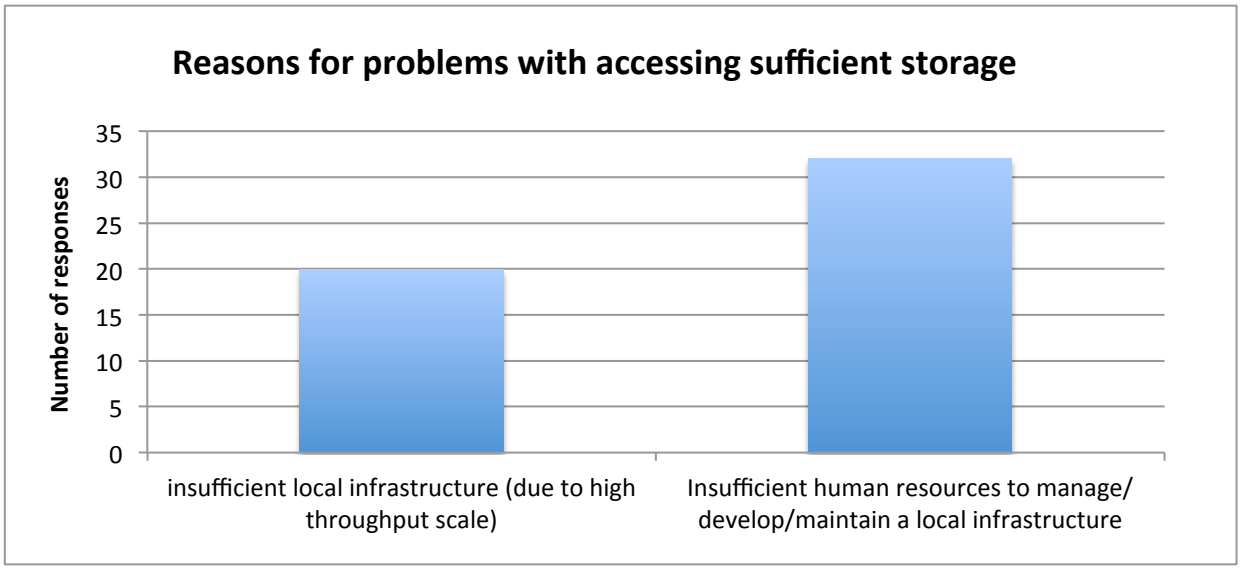
Answer Options	1 : maximum priority
genome sequences (NGS included)	49
sequence variation (SNP, CNV, PAV...)	35
genetics maps (markers)	27
genetics maps (QTLs)	22
curated gene models	22
metabolic profiles	13
protein profiles	11
resources for natural genetic diversity	22
segregating populations or linkage desequilibrium panels	24
phenotypes (classification, ontologies)	22
phenotypes (experimental measurements)	33
data citation	16
long term preservation	14

- **Concerning data analysis**, 4 classes could be identified:
  - **A first class corresponding to users involved in GWAS and QTL analysis at 51,4%, in comparative genomics at 45,9%, on Genome structure and evolution at 45,9%, in genetic diversity or population genetic at 43,2%**

- A second one on gene/protein/metabolic network inferences at 32,4%
- A third on breeding analysis at 24,3%, map based cloning 20,3%, developmental biology (20,3%)
- The last one on modelling of plant development and adaptation 10,8%

- **Concerning data storage**, only 40 persons answered to this question. 32 declare to have insufficient human resources to manage/develop/maintain a local infrastructure and 20 have insufficient local infrastructure (due to high throughput scale)

  See figure below:



- **Concerning data submission**: 57,5% submit reads or pre-computed data from analysis software. They are produced by lab or by external platform.

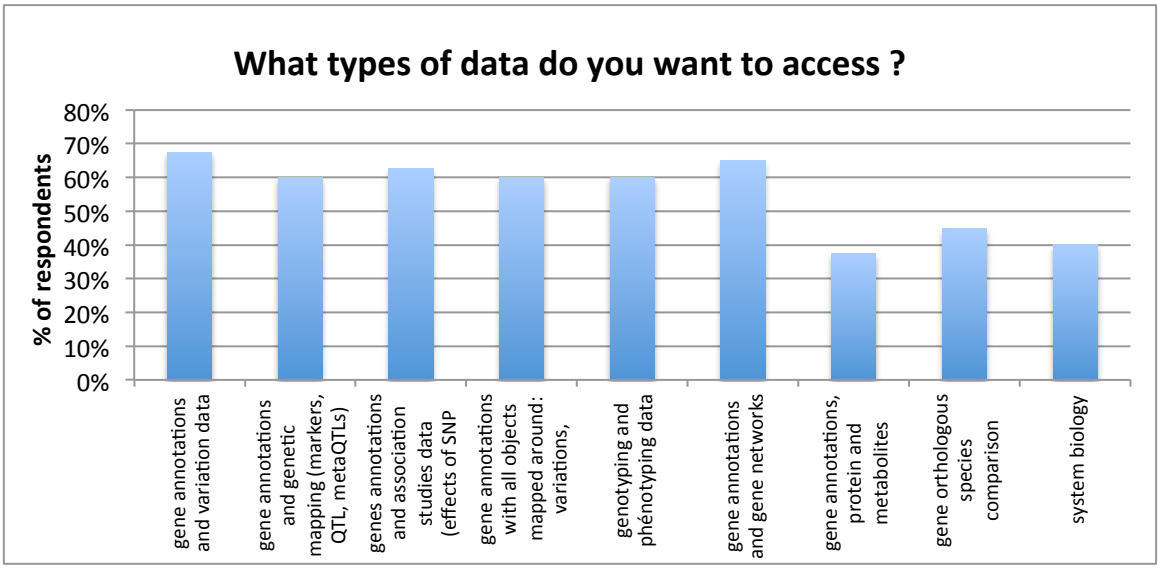
- **Concerning data sharing, they are kean to share many types of data covering genomics and genetics that can be grouped in four classes:**
  - **genome sequences (55%), resequencing data (52,5%), genotyping (50%) and expression data (52,5%)**
  - genetic mapping data (42,5%), phenotyping data (40%) and GWAS data (35%)
  - Exome sequenging (20%), Methylation ChipSeq (20%), Small RNA (20%)
  - Orthologous gene comparisons (15%) and genomic selection data (7,5%)

  See full table below:

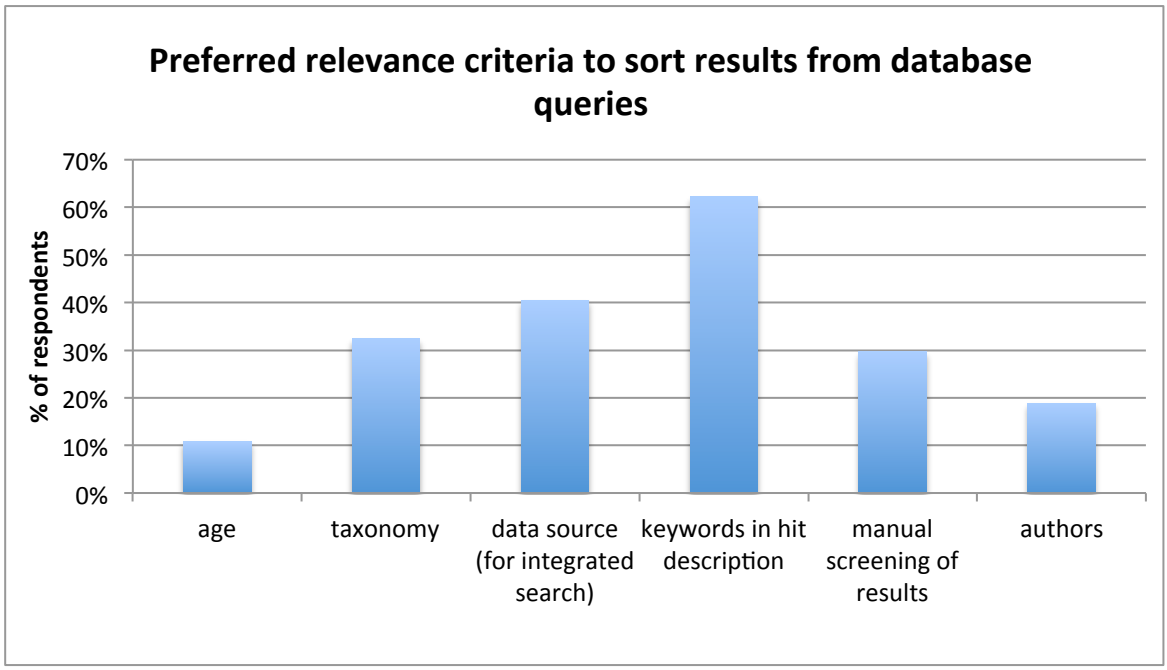
Answer Options	Response Percent
Genome sequences, assembled sequenced	55,0%
Resequencing data: sequence variation (SNPs, CNV, structural variants)	52,5%
Genotyping data (alleles, haplotypes, localisation on genomes, frequency)	50,0%
Annotation data (gene models, gene function prediction, gene ontologies)	37,5%
Genetic mapping data (maps, markers, QTLs, metaQTLs)	42,5%
Exome sequencing data	20,0%
Phenotyping data (traits, phenotypes, experimental conditions)	40,0%
Association studies data (traits, phenotypes, markers effects,	35,0%

statistical values, LD)	
Genomic selection data	7,5%
Expression data (RNASeq)	52,5%
Expression data (Arrays)	35,0%
Methylation data (ChipSeq)	20,0%
Repeats data (transposable elements)	20,0%
Small RNA	20,0%
Orthologous gene comparisons	15,0%
Other (please specify)	

- **Concerning data access:**
  They would like to access to various data types:
  See figure below

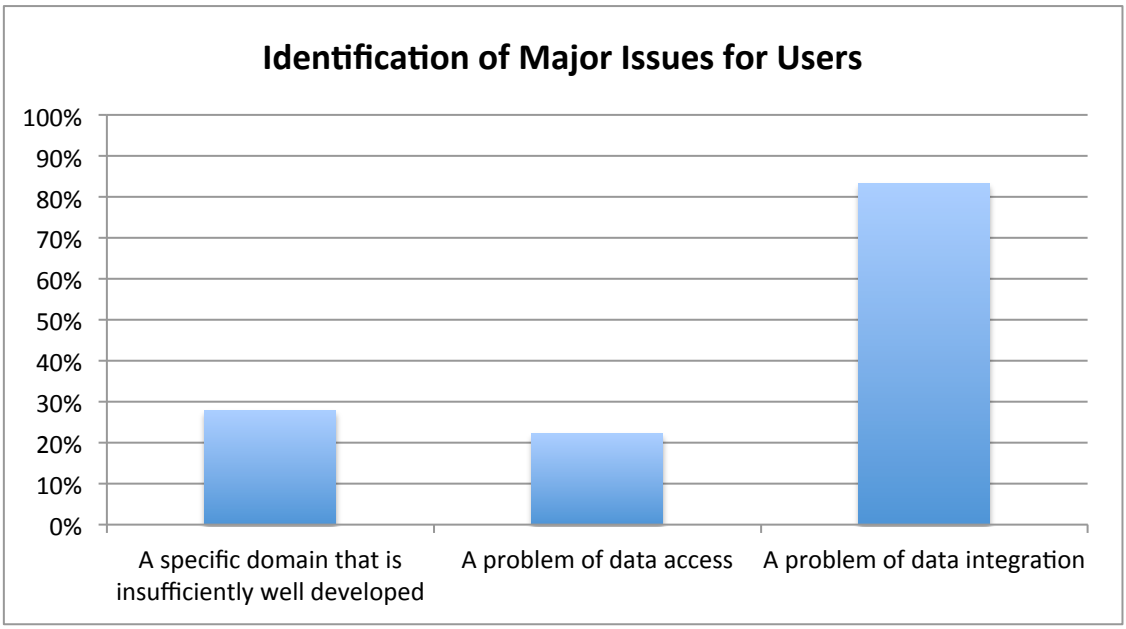


- **Concerning criteria to access data in databases:**
  - **their main criteria at 62,2% concerns keywords in hit description**
  - followed by the data source 40,5%
  - manual screening at 29,7%
  - and author at 18,9%
    See graphics below

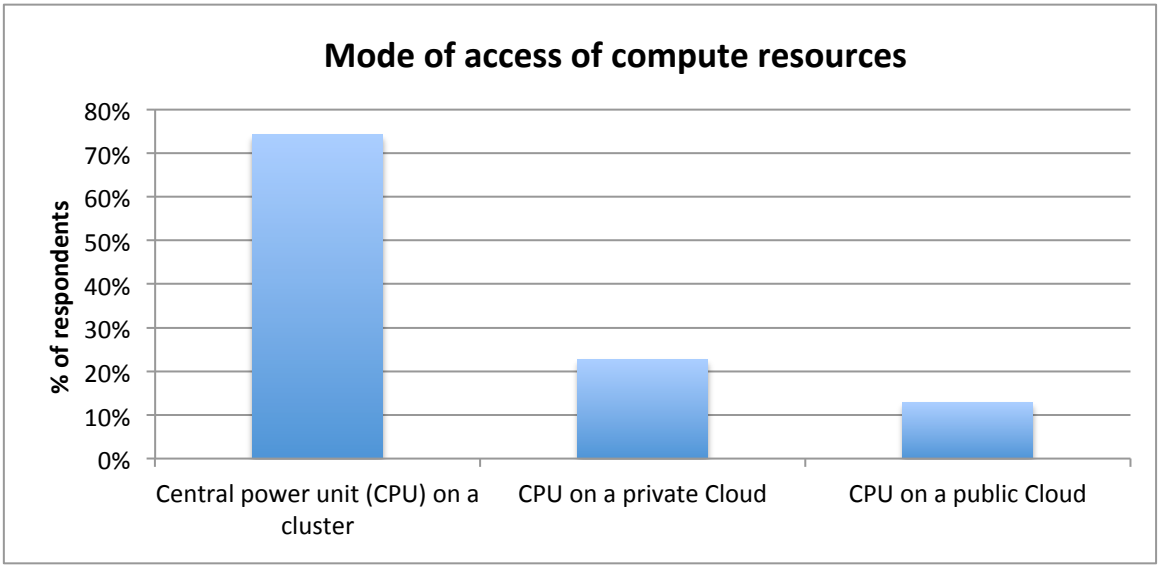


- Concerning tools that could be missing:
  A big problem of data integration (83,3% of survey answers) was discovered. It concerns several data sources and data types. There is also a lack in data access; 22,2% and in some specific domain insufficiently well developed (27,8%).
  Some needs were suggested by users: lots of them concern genome browsers improvements and integration with several data sources
  - To have a connection of polymorphisms / elaborated analysis with genome browser / publication
  - Integration of sequence genome data, SNPs and phenotypes
  - Integrative database (with private and public data)
  - Multidimensional genome browsers for fragmented genomes
  - Integration tools from genome to metabolism and phenotype on plant crop and on lepidoperes
  - Genetic variation analysis in pathway
  - Ultimate omic visualizer
  - Integrated tools for gene ontology and gene network
  - Graphical interface to search RNASeq data and microarray data from specific experiments
  - More integrated genome browser (multiscale, multilevel)

  See figure below:



- **Concerning computing,**
  - **52,5% needs computing resources and 74,2% use CPU on cluster.**
  - 22,6% use a private Cloud and 12,9% a public Cloud.
    See figure below:

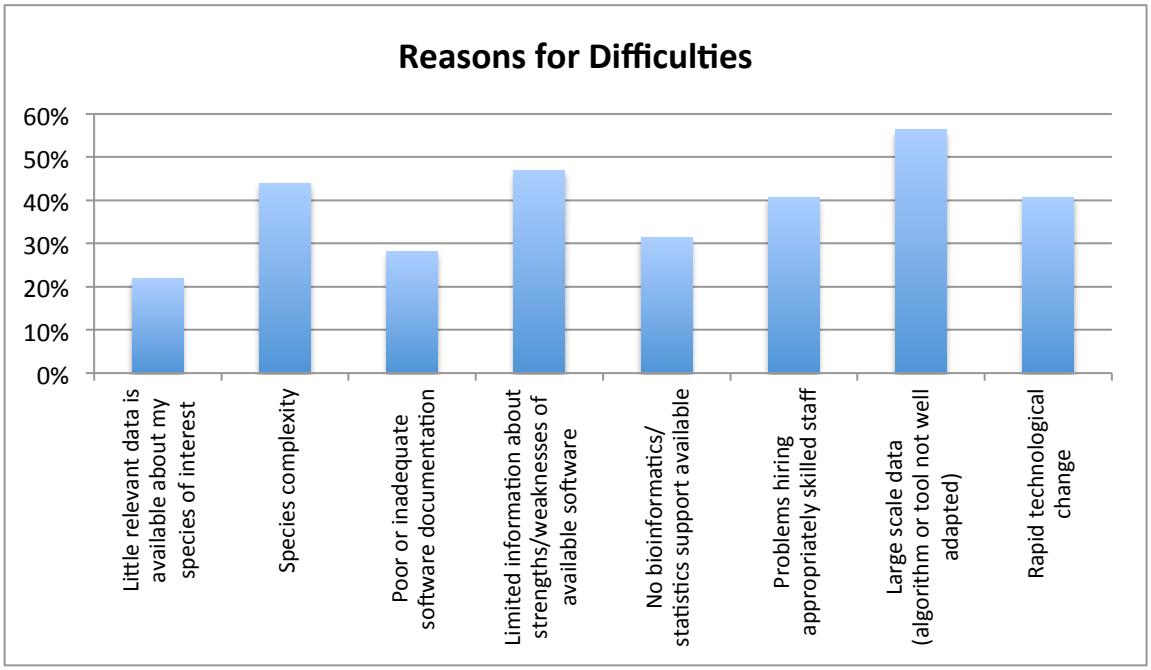


- **Concerning data analysis, 62,5% have problems to analyse their data**

- Several reasons are the causes of these issues:
  - **A lack of algorithm or tools and adapted infrastructure to analyse high throughput data, such as sequences (NGS) was high lightened (56,3%) as well as a problem concerning species complexity (43,8%) that allows difficulties for them to find adapted softwares.** Answers underlined also **the difficulty of having limited information concerning strengths and weaknesses of tools (46,9%)**
  - A second class of answers high lightened problems due to rapid technological change (40,6%), problems in hiring appropriately skilled staff (40,6%), bioinformatics/Statistics lack of support

(31,3%) or poor or inadequate software documentation (28,1%)

See figure below:

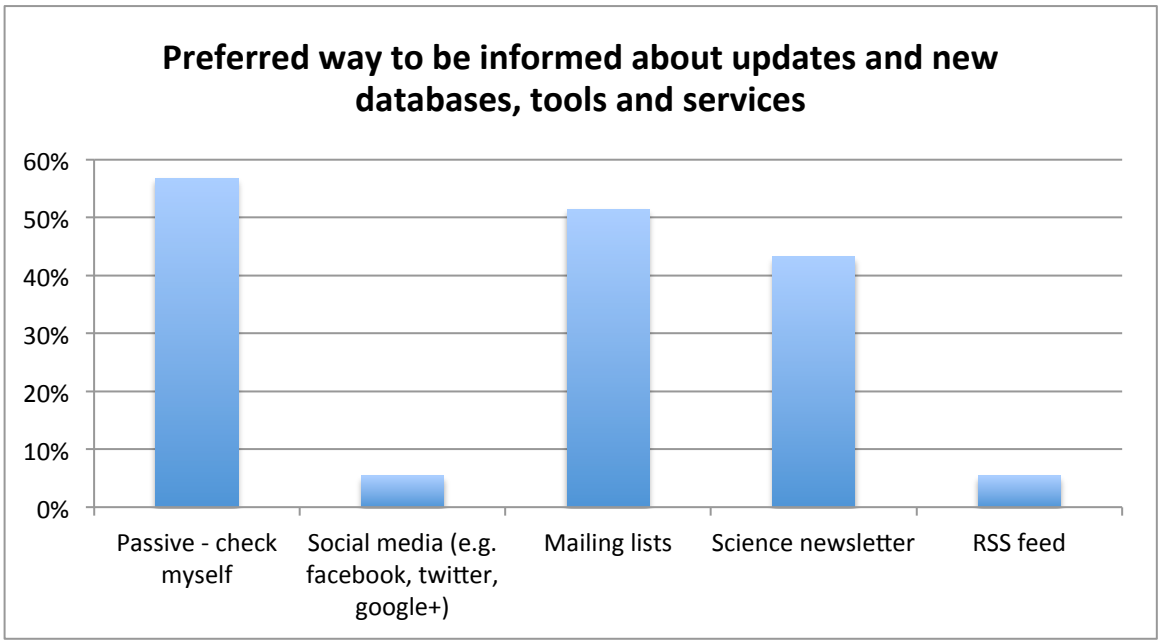


- For 74,4%, users are participating to projects in which bioinformatics are on going.

- **10 users need help in software development**: to update to new technologies, to find the best possible software to apply to a specific problem, for extracting significance from NGS data, for lncRNA identification in plant, for association genetics and quantitative genetics, to understand the concept underlying the programming of the software mainly to debug.

- **41% need help in bioanalysis**: in genomic breeding and QTL detection, in genome to transcriptome associated to NGS, in integration of different data sources for a given gene, gene family, for lnRNAs, for ChIP and expression data, for genomic sequences.

- To help them, from their point of view, transPLANT could undertake for us user training: 85,7 %, a unique web portal: 71,4% and newsletters 28,6%.
- Their preferred way to get informed: by their own (55%), mailing (51%), newsletter (41%). See figure below:



- 15 users describes the name of the project in which they are or were involved with bioinformatics:
  - AmaiZIng (maize), SunRise (sunflower), Rapsodyn (rape), Breedwheat (wheat): French investment for the future running projects, MetaQTL (French funding agency, closed project), Triannot pipeline (INRA fundings), LifeGrid 2013-2015 Région Auvergne / FEDER, PHENOME
  - ARCAD project funded by Agropolis foundation: http://www.arcad-project.org/
  - transPLANT
  - TRITEX
  - PolapGen
  - NCSB; Open PHACTS, various EU FP7
  - SYSFLO (EU Marie Curie ITN)

- One first recommendation was done by one answer: to be careful to integrate biologist and bioinformaticians as well as computer scientist in the same discussion. It will show a lot of difficulties to integrate the working vocabularies of these 3 career expertises but it is indispensable to make the effort from these 3 parts and will teach to coming students and researcher basis in these field that will allow a optimize development of these fields. A second comment was to connect to ESFRI projects like ELIXIR and Dutch DISC.

- 18 answers propose their dreams concerning bioinformatics resources to do things they could or don't know how to, do now:
  - Organize the work to integrate sequence genome data, SNPs and phenotypes
  - easy access to specific public data
  - analyse large size of sequence data
  - enter ngs or microarrays results available on open data source and possibility to visualised them on genomes and with orthologous and paralogous genes information (expression , promoter motifs, phenotypes,..
  - integrate data from quantitative genetics (GWAS, QTL, others) with functional (transcriptomics, metabolomics, mutant phenotypes) and structural data (PAV, CNV)
  - A European iPlant-like for calculation
  - A wheat PLAZA for interoperability
  - Bioinformatic engineers who are avalaible and can understand physiology
  - IT engineers who do not take literally months to format a new computer
  - Access to meta-dataset per plant crop and model system including annotated genome, gene regulation, gene map positions (genetic (incl. integrated maps) and physical), extensive QTL and association peak information, dedicated information to gene function incl. references (map-based cloning, transgenics,

mutants etc.), protein-protein interaction data, evolutionary (phylogentic) data such as orthologues, selection pressure on genes.

- Integrated Pipeline Analysis and Databases
- Have fully annotated genomes for barley and wheat
- a better sequence assembly of large genomes
- Comparative genomics of an high number of genomes (>20)
- Population Genomics

Conclusions

To summarize, a number of the survey's results indicate the importance of issues that were already priorities of the transPLANT project.

Further actions have been prioritised or added to the project's agenda in response to the survey's conclusions.

The results of the survey have been posted on the transPLANT website, at http://www.transplantdb.eu/survey, to stimulate ongoing discussion with the community. It is free to download.

Any comment or suggestions, please contact us at: transplant_help@ebi.ac.uk