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

D9.1
Variation repository, first release

Due date of deliverable: 31/9/2013
Actual submission date: 7/10/2013

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: EMBL-EBI

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

EMBL-EBI

Introduction

The large number of genome resequencing projects currently underway are charting the genetic landscape of important crop and model species. The occurrence of genetic variation can be statistically linked to phenotypic traits and used to direct plant breeding programs, an obvious route for the direct application of genomic technologies to benefit society. In addition, such variations provide insights into evolution and function. However, there is currently no central archive capable of efficiently handling the management of such data, which is problematic in part because of the potentially huge size of the data (as sequences are determined for thousands of individual stocks, wild specimens, etc.)

Therefore, transPLANT will develop a 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. This deliverable comprises the completion of the initial development and the availability of the pipeline to accept submissions.

Methods

One key challenge in managing this data comes from the fact that the meaning of a variant locus is provided through its positioning on a reference sequence; but reference sequences for plant genomes are both approximations and abstractions. As a reference sequence are updated, so existing data needs to be migrated forward to be seen in the context of the latest reference and annotation. An often-used approach of doing this involves the alignment of each variant locus (together with its flanking sequence) individually against the new reference. Our 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.

The core components of the data management system we have developed are 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.

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

Components 1-10 have now been completed. The relationship of EBI's services with those of NCBI in the area of variation (point 11 on the above list) is being taken forwards by a new variation team leader recently appointed at EBI, but is not a limiting step for the acceptance of submissions and management of data.

1. The most important meta data for inclusion in any submission is the identity of the genome sequence and version on which variant calls have originally been made (but the system also stores the identity of the reads, to allow for potential re-calling; and of the wider set of sequences against which variants have been called, not just the molecule on which the variant was located). A new system for genome assembly identification has recently been implemented as an extension to the International Nucleotide Sequence Database Collaboration, which we will adopt for use in the transPLANT variation archive. These new identifiers will be the fundamental denominator of sequence identity in the system, and any alternative descriptors/identifiers will be mapped to this.
2. Variant Call Format (VCF) is a text file format (most likely stored in a compressed manner) that has developed as a standard for the representation of variant information in the context of the 1000 genomes project (http://www.1000genomes.org). It contains meta-information lines, a header line, and then a variable number of data lines, each of which describes a position in the genome. The infrastructure will accept submissions in VCF format and allow users to retrieve submitted files in this format.
3. A web-based user interface has been developed, to allow users to submit VCF files and appropriate meta data, and to verify certain elements of the meta data against reference values before deposition in the persistent archive (step 4). A validation infrastructure has been developed to operate behind the web interface to retrieve and verify certain additional information needed for deposition in the archive (e.g. identifiers for reference sequence). At present, the validation and acceptance of submission is performed semi-automatically at EBI, but it could in future be integrated into a more advanced submission interface, allowing users to retrieve and validate their own meta data prior to the automatic acceptance of their submission. Screenshots of the interface are included in Figure 3.

   We will use this interface to accept early submissions to the archive over the course of the next year.

New work (not funded by transPLANT) is underway within the European Nucleotide Archive to provide support for the submissions of VCF format files through the data submission tool Webin, and this will become the primary route for VCF submission once available (and thus integrated with the submission of other data/file types to archival resources hosted at EBI to submission).

4. The archive layer of the infrastructure will be implemented as an extension to the European Nucleotide Archive (http://www.ebi.ac.uk/ena), which already archives the raw sequence data from which both reference sequence and variant calls are generated. ENA already has tools for storage, back-up and fast, indexed retrieval of individual documents, and prior experience of storing huge data volumes (variation call data is large, but a variant call is effectively a reduced form of the sequence reads themselves). The ability to utilise ENA infrastructure has significantly increased the speed of progress on this work package, and allowed a focus on the missing pieces of the infrastructure, i.e. QC tools on data submission, feature mapping pipelines etc. To allow the use of the ENA infrastructure, tools have been written to support the automated submission of quality-checked data into the ENA, and for retrieval using the ENA's data access libraries. Both ENA browser and submission service offer programmatic access via a REST API (http://www.ebi.ac.uk/ena/about/browser), using XML and FASTA.

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

6. The test of the alignment methods was obviously dependent on the existing of a method for feature propagation, although this is in itself straightforward and therefore not effectively under test. All that is required is to process the mappings produced by the genome aligner, and for each mapping to update the position of all features that fall within it. Using the *de novo* calls described in point 5 as a gold standard, precision (or positive predictive value) and recall (or true positive rate/sensitivity) of feature propagation can be measured. When projecting from version 6.1 of the rice genome to version 7, feature propagation based on ATAC mapping reached a precision of 99.41%, and recall of 99.47%; with

flanking sequence alignments reaching 99.14% and 98.60% respectively. When projecting from version 2 to version 7, ATAC reached precision and recall of 98.12% and 96.69%; with flanking sequence alignments reaching 97.60% and 95.81%. For closely as well as distantly related pairs of sequences the flanking sequence alignment approach has a higher rate of false negatives, losing more features between versions; and a higher rate of false positives, placing features incorrectly.

7. The operational data store has been implemented in MongoDB (http://www.mongodb.org), a document-store database system. As such, MongoDB has some advantages applied to this project compared with a classical Relational Database Management System, namely data conceptually belonging together - describing the samples, genotypes and metadata of a position in the genome does not have to split and later re-joined across tables of a relational schema. Furthermore it is better suited to handle very large datasets and allows horizontal partitioning/sharding. The underlying document format consists of indexed JSON files. A schema has been derived to provide access to this data – see figure 3 below.

8. A merging procedure has been developed. Data that has been sharded with MongoDB across multiple physical servers can be processed efficiently in parallel using the MapReduce programming model. (http://research.google.com/archive/mapreduce.html). Each document in the database is processed independently and in parallel during the "map" step and combined in some way to form the output during the "reduce" step. Selecting an appropriate reduce step guarantees that all features sharing the same position are grouped together and are merged. 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: $vc[A\text{-}Z]_1[0\text{-}9A\text{-}Z]_n$ with a n initial n of 4 which will grow as demand for identifiers increases. This gives an initial accession space of 40 million identifiers increasing 36 times with every addition of a extra character. 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, utilising 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 flatfile 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.

In addition to these logical components, 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 overall workflow is summarised in Figure 1. Figure 2 shows the UML class diagram that represents the MongoDB schema.

Figure 1 The transPLANT variation archive, overall data workflow

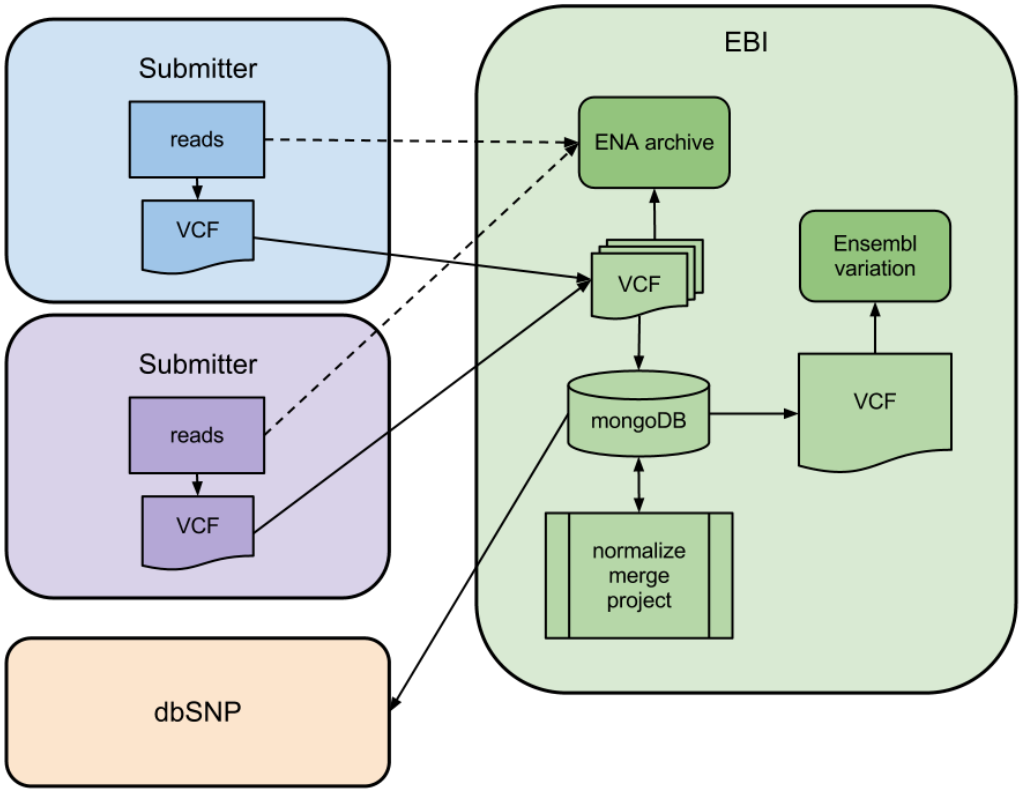


Figure 2. Class diagram for the MongoDB schema.

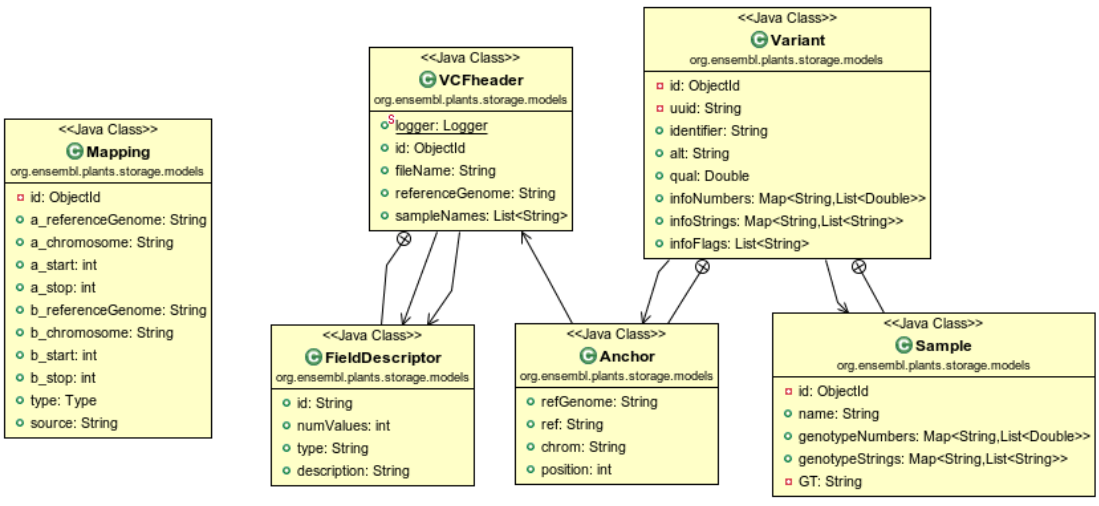
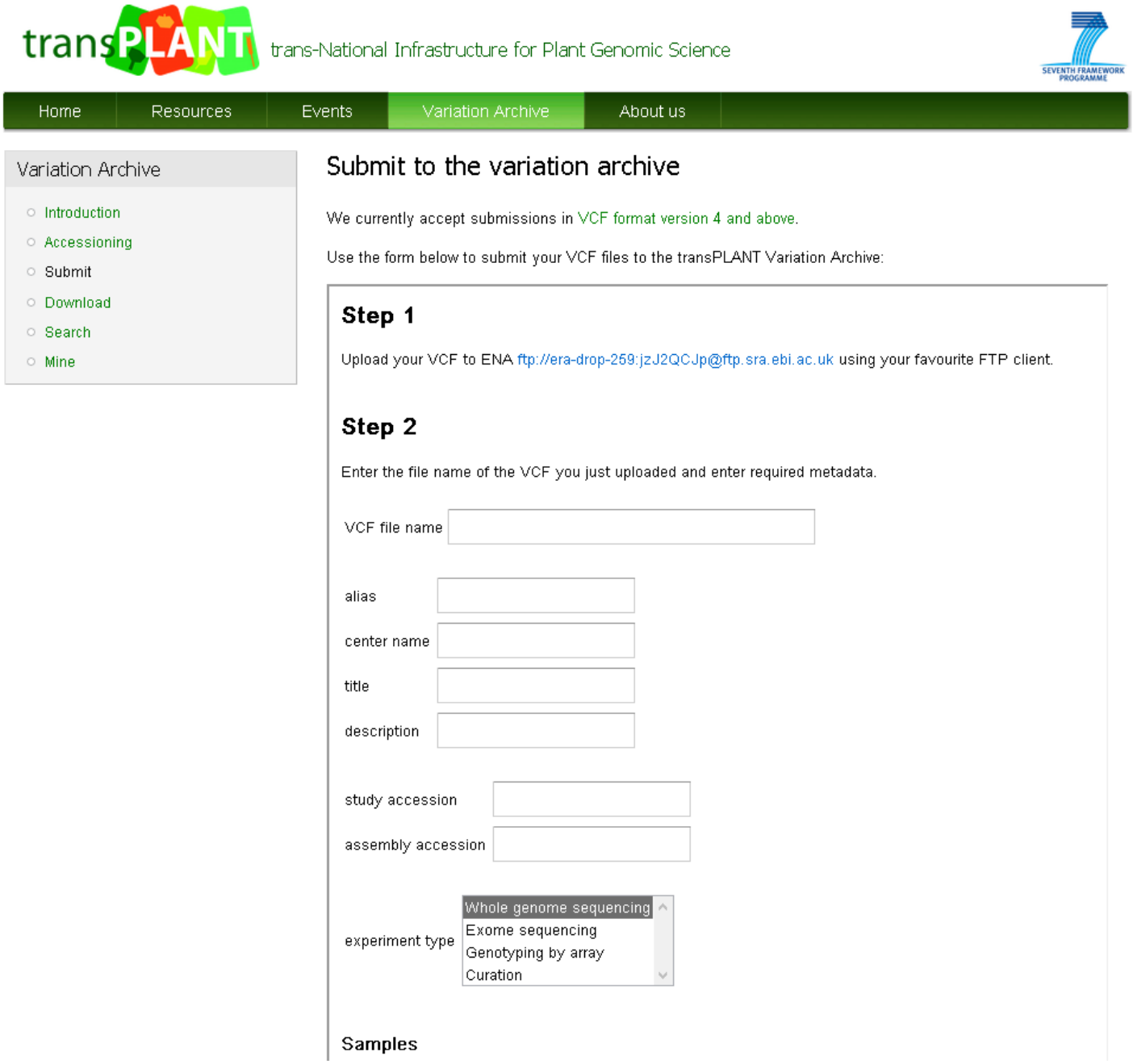


Figure 3. The transPLANT user interface for the submission of variation data:



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 4. 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 to download 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      Lambrusque      Lambrusque      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     .       .       .       .       .       .       .
.       .       .       .       .       .       .       .
```

Results (if applicable, interactions with other workpackages)

Initiation of Public Service

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 http://www.transplantdb.eu/variation/submit, and we are now ready to accept submissions from members of the scientific public. 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.