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

D9.2
Data from Arabidopsis 1001 genomes project integrated in central hub

Due date of deliverable:
Actual submission date:

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable:

Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)		
Dissemination Level		
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)	

Contributor

EBML-EBI

Introduction

Deliverable reference number: D9.2

Three factors are essential for continued improvement of crop species by plant breeding: tools to identify adequate genetic variation, large-scale phenotyping (to allow variants to be associated with desirable traits), and technology to efficiently (re)combine useful alleles in new breeding lines. Material from wild relatives, ancestors and landraces held in germplasm collections of crop species contains an underexploited wealth of genetic variation, and will therefore offer a useful gene pool to cope with existing and new breeding challenges. Exploiting wild and early domesticated resources 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. Once adequate genetic variation has been identified, the efficiency and success rate of breeding programs that make use of it can be greatly increased by DNA based selection of lines and markers associated with traits of interest.

In this work package, we have already developed a new mechanism for the analysis and archiving of genomic variation data from plants. In this deliverable, we report on the population of this infrastructure, with data sets from large public resequencing projects.

Methods

Data is submitted to the archive in Variant Call Format (VCF), a standard format for the representation of variation data (see http://www.1000genomes.org/wiki/analysis/variant%20call%20format/vcf-variant-call-format-version-41 for specification). When submitting variants into the variation archive one or many VCF files must be supplied with mapping metadata, which is required to keep track of reference sequence assembly versions and samples. Both sample and sequence fields in VCF are free from, so to ensure correct identification they are mapped to accessions provided by other databases. For example, in the case of barley the sample name "MAPPING_barkeMOREX_BWA_v1.sorted.rmdup.bam" is mapped to the ENA accession ERS140558, and the sequence name "contig_1" is mapped to the ENA accession CAJW010000001.1. This mapping is entered by the submitter on the VCF submission website (see WP9 report), and is stored in JSON format. After quality control, the VCF and meta data are deposited in the European Nucleotide Archive, which is used as the ultimate persistent store of the submitted data. New submissions are merged with previous submissions, and unique locus identifiers assigned, through a process implemented using the document database MongoDB as an operational data store.

The process for loading MongoDB from VCF, and details of the, merging, accessioning, and MongoDB data model, was described in the deliverable D9.1 with subsequent improvements described in the latest annual report for work package 9. It can be started either programmatically or from the command line.

Results (if applicable, interactions with other workpackages)

Several large scale datasets (up to 12 billion variants) have been processed by our system.

Dataset	VCF size	Samples	Genotypes	Unique loci/

	in GB			Accessions
Arabidopsis 1001	153.4	1,211	12,148,751,881	13,629,424
Tomato	38.4	84	308,965,530	71,156,450
Barley	2.5	14	25,001,485	15,282,555

Arabidopsis 1001

The 1001 Genomes Project (http://1001genomes.org) was launched at the beginning of 2008 to discover the whole-genome sequence variation in 1001 strains of the reference plant Arabidopsis thaliana. The resulting information is paving the way for a new era of genetics that identifies alleles underpinning phenotypic diversity across the entire genome. Each of the strains in the 1001 Genomes project is an inbred line with seeds that are freely available from stock centres the scientific community. One of the useful features of the project is that unlimited numbers of plants with identical genotype can be grown and phenotyped for each strain, in as many environments as desired, and so the sequence information can be used directly in association studies at biochemical, metabolic, physiological, morphological, and whole plant-fitness levels. The analyses enabled by this project will have broad implications for areas as diverse as evolutionary sciences, plant breeding and human genetics.

A pre-release of the release dataset has been provided to us in VCF format, containing 1,211 samples and 13.6 million loci. A single locus describes the variations of a subset of varieties, on average 891 samples or 73.6% of all samples - this results in a total of 12.1 billion genotypes described.

A single mongoDB node backed by NFS (Network File System) was used to obtain the following results, giving room for future improvement by scaling mongDB nodes horizontally via sharing, and storing data locally on each node instead of relying on comparatively slow access times over the network.

Table 1 Load performance, Arabidopsis

	Time in hours	Genotypes / second	Unique loci / second
Load	73	46,000	52
Export	45	75,000	84

In total 13.6 million accessions (unique loci) have been assigned sequentially, from ‘vcZOH30Y’ to ‘vcZWL7K1’.

If it becomes necessary to remove very low quality samples, or add additional samples it is important that this can be done efficiently. To assess the performance of sample manipulation in the variation repository sample “6030”, which is mentioned in 8 million loci was removed from the repository, and later added back in. Removing all mentions of the sample took 42 hours, while the re-addition took 10 hours.

The data has been processed in the archive, but is not yet cleared for final public release by the1001 genomes consortium. Update, if necessary, and release will occur on publication.

Tomato

The aim of the 150 Tomato Genome ReSequencing project (http://www.tomatogenome.net, http://onlinelibrary.wiley.com/doi/10.1111/tpj.12616/abstract) is to reveal and explore the genetic variation available in tomato. The project selected tomato as target crop because it is economically one of the most important crop species for the Dutch breeding industry, and is one the most important vegetables globally. However, since the tomato shows only limited genetic diversity in commercial breeding lines, valuable alleles will be available in wild tomato relatives. Since breeding and selection was targeted at only a narrow range of desirable agricultural traits, also old breeding material could be source of interesting alleles that have been lost

during domestication.

The published dataset has been downloaded from ENA (Study accession PRJEB5235) as 84 VCF files, containing 84 samples, 309.0 million genotypes in total and 71.2 million unique loci. After merging, every unique locus describes on average 4 samples or 5.2% of all samples.

Table 2 Load performance, Tomato

	Time in hours	**Genotypes / second**	**Unique loci / second**
Load	24	3,500	820
Export	5.6	15,000	3,500

Contrasted with the throughput of Arabidopsis above we can see that there is a trade-off between the number of genotypes and loci that can be processed per time unit. Arabidopsis requires writing very large loci to the database, which can be done efficiently as all genotypes for one locus are stored together in the VCF file. For tomato on the other hand each locus has to be updated many times throughout the load process, which limits the throughput of genotypes, but allows a higher throughput of loci.

In total 71.2 million accessions (unique loci) have been assigned sequentially, from 'vcZWL7K2' to 'vcZ22YC83' in the accession space described in the first work package report.

Barley

The International Barley Sequencing Consortium (IBSC) (http://mips.helmholtz-muenchen.de/plant/barley/, http://www.nature.com/nature/journal/v491/n7426/full/nature11543.html), a multi-national team of scientists from many countries aims to develop new and better barley varieties able to cope with the demands of climate change.

Cultivated barley, derived from its wild progenitor *Hordeum vulgare* ssp. spontaneum, is among the world's earliest domesticated crop species1 and today represents the fourth most abundant cereal in both area and tonnage harvested (http://faostat.fao.org). Approximately three-quarters of global production is used for animal feed, 20% is malted for use in alcoholic and non-alcoholic beverages, and 5% as an ingredient in a range of food products. Barley is widely adapted to diverse environmental conditions and is more stress tolerant than its close relative wheat. As a result, barley remains a major food source in poorer countries, maintaining harvestable yields in harsh and marginal environments. In more developed societies it has recently been classified as a true functional food. Barley grain is particularly high in soluble dietary fibre, which significantly reduces the risk of serious human diseases including type II diabetes, cardiovascular disease and colorectal cancers that afflict hundreds of millions of people worldwide.

The dataset has been provided to us by the IBSC as six VCF files, containing 14 samples, 24.5 million loci in total and 15.3 million unique loci. Every unique locus describes on average 1.6 samples or 11.4% of all samples.

Table 3 Load performance, barley

	Time in hours	**Genotypes / second**	**Loci / second**
Load	**4.2**	**1,600**	**1,000**
Export	1.2	6,000	3,600

In total 15.3 million accessions (unique loci) have been assigned sequentially, from 'vcZ00001' to 'vcZOH30X'.

All publicly released data is available via FTP at ftp://ftp.ebi.ac.uk/pub/databases/transplant/variation/, or via the transPLANT website at http://www.transplantdb.eu/variation/download. On public release, data is additionally exposed through the Ensembl Plants database.

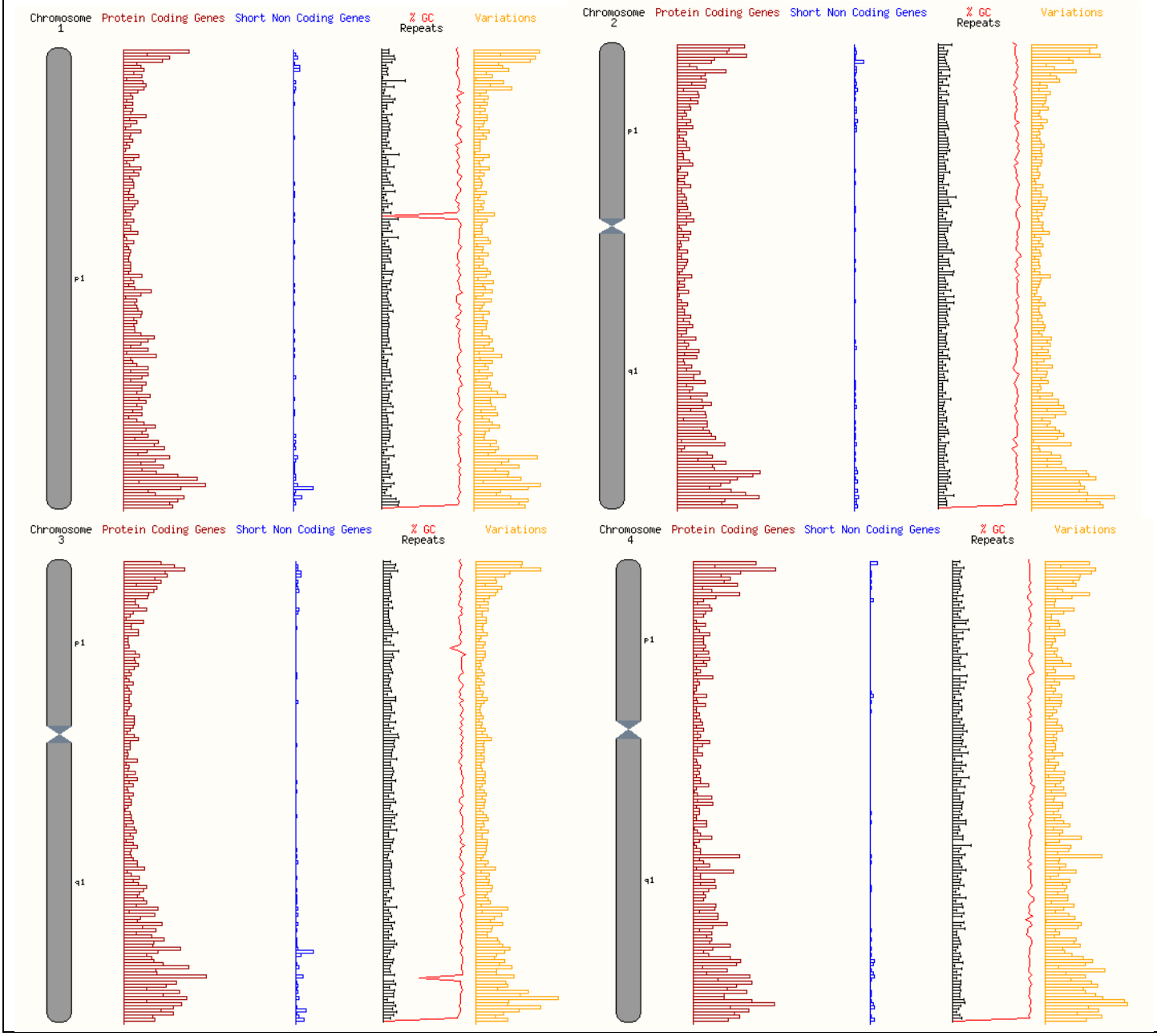
Future Actions

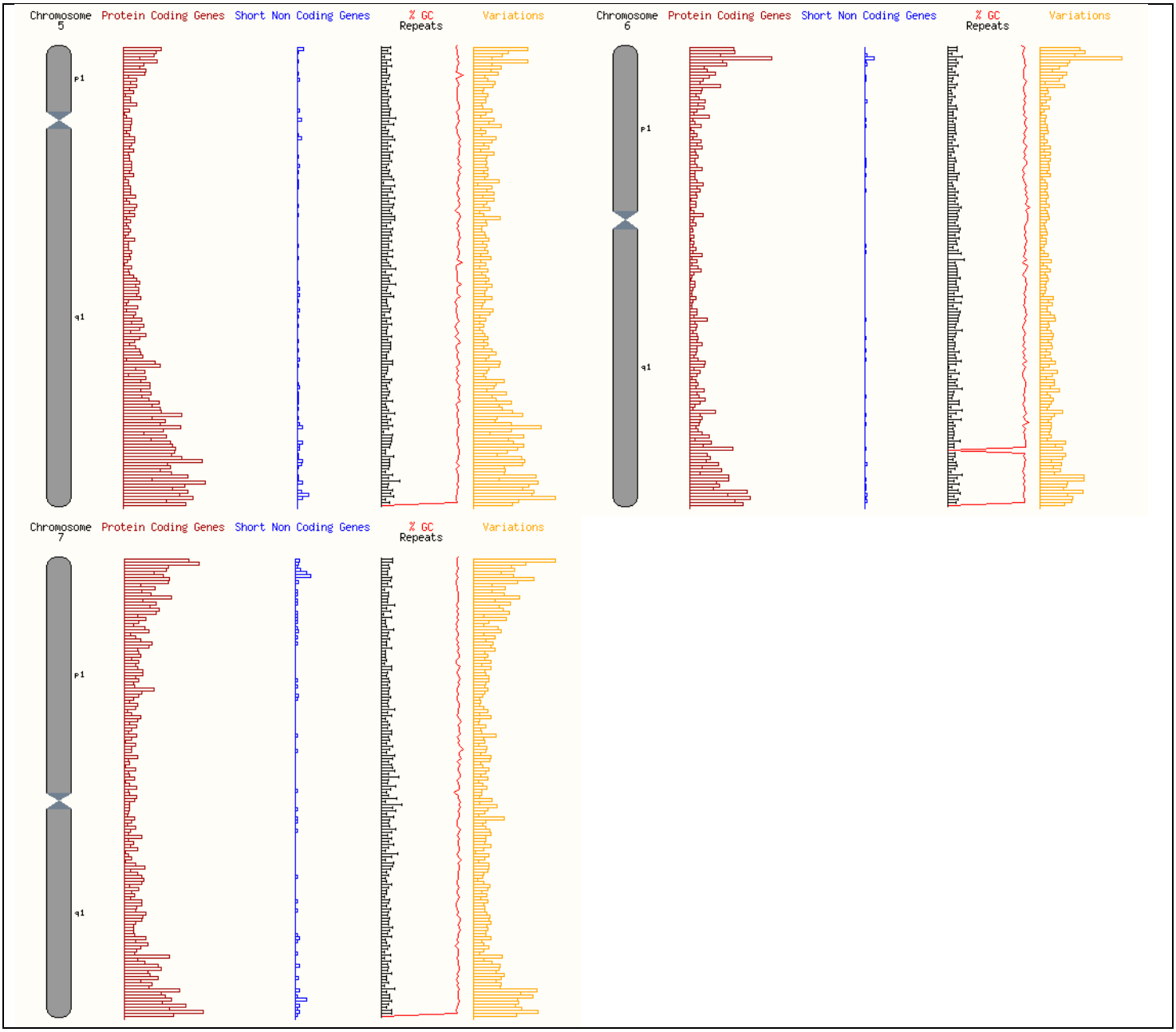
We are currently engaging in a process of community outreach, together with our US collaborators at the Gramene database, to secure the submission of additional data sets. Data sets/communities we are currently targeting are summarised in **table 4**.

Table 4 Data sets currently targeted for inclusion in the archive.

Species/data set name	**Data types**	**Notes**
Grape	Array, GBS (Genotyping by sequencing)	5000 individuals
Maize	Resequencing	Panzea project
Peach		
Poplar		
Rice (*Oryza sativa*)	Resequencing	3000 rice project from IRRI
	Array	20,000 X 770K array
Rice (other species)	Resequencing	OGE project: 6 species, 14 individuals per species
Rye	Resequencing	
Soybean	GBS	10K individuals
	Resequencing	1 K individuals
Tomato	Resequencing	Data from US to supplement existing European data
Wheat	GBS	From CIMMYT

Figure 1 Distributions of variant loci for the barely genome, as visualised in the Ensembl Plants interface.




Chromosome 5
Protein Coding Genes
Short Non Coding Genes
% GC
Repeats
Variations
p1
q1
Chromosome 6
Protein Coding Genes
Short Non Coding Genes
% GC
Repeats
Variations
p1
q1
Chromosome 7
Protein Coding Genes
Short Non Coding Genes
% GC
Repeats
Variations
p1
q1