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

D8.2
Implementation of a pan-genome viewer

Due date of deliverable: 36
Actual submission date: 29/09/2014

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	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor
BIOGEM

Introduction

Deliverable reference number:D8.2

This deliverable describes the work that has been done in transplant regarding the implementation of a pan-genome viewer, based on existing solutions.

The concept of pan-genome has been first described in prokaryotes. It relies on the fact that genomes at a species level are dynamic, with genes present in every individual (core genome) and genes present in a subset of individuals (dispensable genome). In plants, with the advent of Next Generation Sequencing (NGS) technologies, the pan-genome concept has been observed in model species *Arabidopsis* and rice, but also in several important crop species such as maize (Morgante *et al.* 2007; Springer *et al.* 2009; Lai *et al.* 2010; Hansey *et al.* 2010; Hirsch *et al.* 2014; Chia et al. 2012), barley, sorghum, soybean, and wheat (Saxena *et al.* 2014). These studies demonstrated that a substantial portion of the genomic diversity within one species may lie outside a single reference genome in the form of structural variations (SVs), and that these SVs are often associated with important traits (Figure).

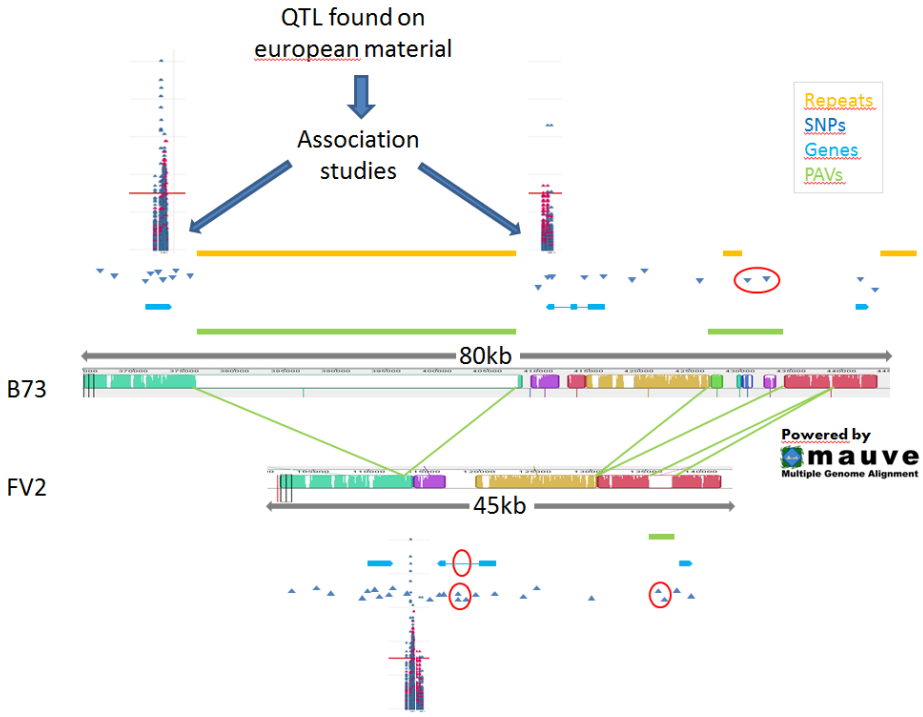


Figure 1. Example of a QTL found on European material and projected on B73 reference. Association studies revealed two regions of interest. But these two regions become a single one when projected on a European line (FV2) because of a 30kb insertion between the two association pics in B73. Furthermore, the second gene misses the internal exon and there is an insertion, a deletion and different SNPs in the promoter region of the third gene in FV2 compared to B73, which shows that it is really important, especially for European breeders, to have another reference than the American B73 cultivar one.

In these studies, the discovered SVs are often available as fasta sequences, and sometimes they are projected on a reference genome. Indeed, most genome browsers (Nielsen *et al.* 2010) follow a model of aligning sequences against a single high quality reference genome. Ensembl goes a little further with its Compara database and pipeline but like other synteny tools it relies on the "single reference genome model". Moreover, comparative genomics tools focus in cross-species comparisons, highlighting homologous regions with sometimes low sequence similarity, and are thus often limited to genes. In contrast, a pan-genome viewer should not be limited

to genes and **focus on differences rather than similarities** (SNPs, INDELs, CNVs, Expression levels), and represent their impact in a useful way for breeding purposes. It should also be able to handle low quality draft genomes as well as high quality reference genomes and to jump smoothly from one to another. Finally it should be NGS compliant and be able to handle large quantities of re-sequenced data, and thus deal with multiple individuals within a single species in a smart way.

This document lists the main components of a pan-genome viewer, reviews existing solutions, and describes the transplant partners' implementations. In addition, it gives recommendation for future developments.

Components of a Pangenome-viewer

1. Pre-computing and data storage

    a. Draft genomes

Using whole genome NGS re-sequencing data, the number of known genetic variants has dramatically increased in the past few years. It has become clear that a single reference genome is not enough to give access to the full genetic diversity of a species. Because of the drop in sequencing costs, new genotypes (i.e. strains, accessions, lines, populations, …) are being sequenced and assembled, giving access to low quality draft assemblies. This is already the case for *Arabidopsis thaliana*, rice, and grapevine (other very important crop species like maize and tomato should follow soon). The particularity in draft assemblies is that they often contain thousands of sequences (contigs/scaffolds) poorly annotated. **So the pan-genome viewer should be able to handle a high number of sequences and include an automatic annotation pipeline.**

    b. Whole genome alignments

Because of the incomplete nature of the draft assemblies, especially in some species having a high proportion of repeats (up to 75% for maize), it could necessary to anchor draft genomes to the reference one in order to fill gaps between unconnected scaffolds. It will be also essential to make all pairwise comparisons of available genotype assemblies within one species in order to make bridges between them. **So the pan-genome viewer should include a whole genome alignment and comparison tool.**

    c. Re-sequencing data and structural variations

In many plant species, low coverage re-sequencing data is available, either genomic (WGS) or transcriptomic (RNA-seq). This data could contribute to better understand variability within one species, and better characterize regions of interest for a given subset of genotypes. This can be achieved by integrating (or linking to) a proper **SV database**, and by providing an **efficient way of storing and visualising NGS data at different zoom levels**. RNA-seq data could also contribute to gene annotation and should also be used to show expression levels under different conditions.

2. Visualization

    a. Switching reference

When trying to explore the pan-genome, one should be able to switch between one reference and another one. Indeed, the reference genome might have large structural variations with other genotypes and this might influence greatly the understanding of a particular region for a QTL. Because of the draft nature of other available genotype assemblies, **this switch might be used several times by the user and should then be very smooth**.

b. Highlighting differences between individuals

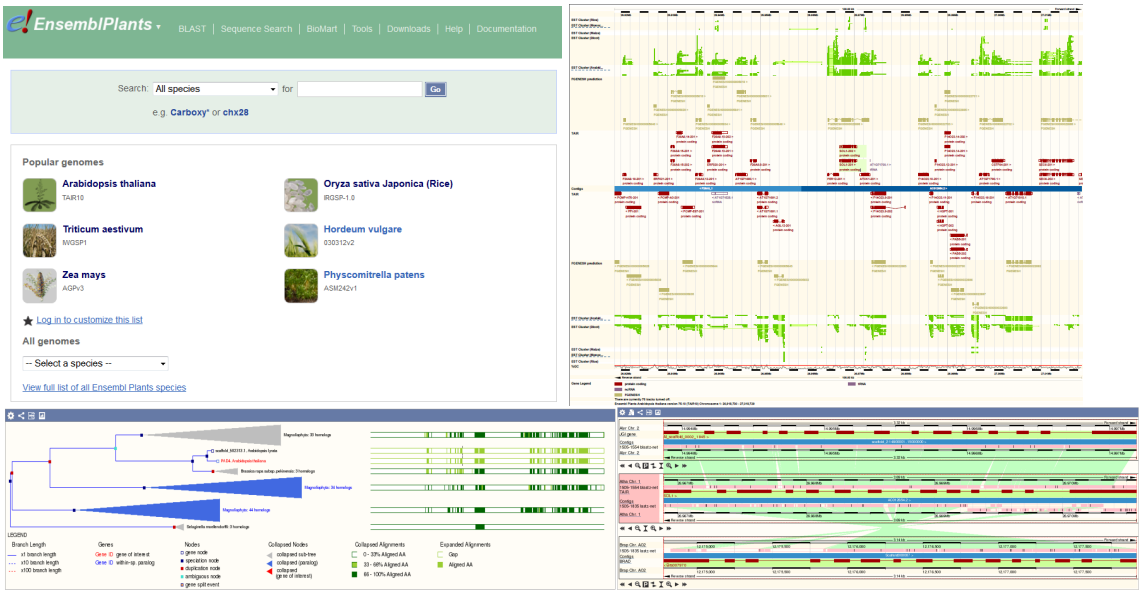
This is the main difference with synteny tools. A **significant development effort** should be made so that the pan-genome viewer not only should be able to identify conserved regions between genotypes, but it should **emphasize divergent regions at different zoom levels** (some structural variations can be as long as a few Mb). Moreover, it should clearly **differentiate unaligned regions which are due, for example, to the presence of repeats which are highly divergent**. It should also **show the impact of the structural variations especially insertions by predicting the presence of new exons or regulatory elements** compared to the reference genome.

Existing solutions and data sets

1. Existing solutions (not exhaustive)

   a. web-based

      i. Ensembl



One of the most well-known genome browser, with strong user support, and many functionalities. It also provides cross-species analyses with the Compara tool, which, could be hijacked to compare draft assemblies from individual genomes from the same species, hence providing **many of the needed features of a pan-genome viewer**. A real test should be made. Link: http://plants.ensembl.org

ii. TASUKE

An interesting way of summarizing NGS re-seq data mapped against a reference genome. It offers a **quick overview of the variability across multiple genotypes at various scales**. Although, it only answers a little part of the needs for a pan-genome viewer.
Link: http://tasuke.dna.affrc.go.jp

iii. JBROWSE

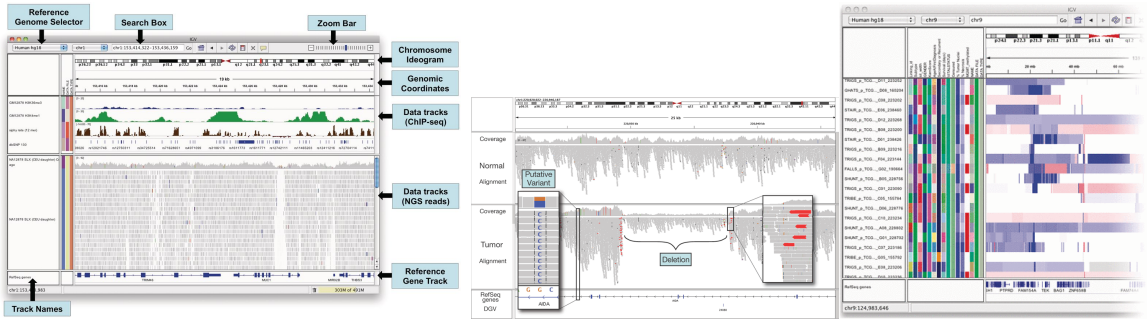
JBrowse is a fast, embeddable genome browser built completely with JavaScript and HTML5, with optional run-once data formatting tools written in Perl. It provides **fast and smooth scrolling and zooming and supports many NGS related formats without any conversion**.
Link: http://jbrowse.org/

iv. The Personal genome browser (PGB)

As expected, the human community is the first one to provide tools to explore individual genomes. The PGB **displays the individual genome variants and associated molecular traits/phenotypes** from the whole genome scale to single nucleotide scale, with reference to genome information simultaneously updated on the background of the same page.
Link: http://www.pgbrowser.org

b. Stand-alone

i. IGV



One of the most used stand-alone **NGS** genome browser. It includes support for many standard formats and as it runs locally, the navigation is **fast and flexible across different zoom levels**. It also offers the possibility to access to reference genomes and data stored in a **remote web repository**. It could be a **good complement to a pan-genome viewer to explore NGS re-sequencing data**.
Link: http://www.broadinstitute.org/igv

Other tools: UCSC Genome browser, ARTEMIS, VISTA, SAVANT, SVA, GBrowse_Syn, SyMap, MisBee.

2. Plant data sets (not exhaustive)

   a. *Arabidopsis thaliana*

   Multiple draft genomes, among which the 19 founders of the MAGIC genetic reference population of recombinant inbred lines (Gan *et al.* 2011, http://mus.well.ox.ac.uk/19genomes/), and many re-sequencing data are available from http://1001genomes.org/.

   b. Rice

   Nine rice genomes are already available at Ensembl Plants and one new draft genome (Kasalath cultivar) was recently assembled and compared to the Nipponbare reference genome (Sakai *et al.* 2014, DDBJ accession DRA000968). Moreover, re-sequencing data for 50 rice accessions are also publicly available (Xu *et al.* 2012, SRA accession SRA023116).

   c. Maize

   Although there are rumours of draft assemblies of American cultivar Mo17 and European cultivar FV2, to our knowledge, none have been made publicly available yet. Nevertheless, there some re-sequencing data available :

   1. WGS re-sequencing data are available for 103 inbred lines (Chia et al. 2012)
   2. RNA-seq data is available for:
      - B73 and Nam parents (available from SRA with the keyword *Zeanome*)
      - 21 inbred lines (Hansey *et al.* 2012)
      - 503 inbred lines (Hirsch *et al.* 2014)

Partners implementation

1. Visualisation of polyploid genomes at EBI

Polyploid genomes (such as bread wheat) throw an interesting light on the idea of "pan-genomes": the presence of a pan genome within a single individual, that contains multiple closely related genomes (homoeologous genomes), with their own differences (inter-homoeologous variants), but also containing classical polymorphisms (i.e. variants specific to one homoeologues segregating in the population), which may be present both homozygously or heterozygously (in which case, there is diversity in the individual within, as well as between, homoeologues) . The problem of clearly visualisation of a polyploid genome, therefore, serves as a prototype of the wider problem of the visualisation of a diverse population. It should be noted that selective breeding in wheat (and also in other crops) has introduced large quantities of "foreign" material into elite lines, meaning that the full pan-genome of such species is likely to be especially complex, compared with a natural population where inheritance is mainly through vertical descent and less due to outcrossing.
Ensembl Plants (http://plants.ensembl.org) is part of the transPLANT infrastructure. To ensure a clear visualization of the bread wheat genome in Ensembl Plants, a number of important criteria were identified: (i) homoeologous relationships should be clearly identified, to make clear how the genic content of one homoeologues relates to another (i.e. whether there has been gene loss, or gene family expansion, in one homoeologue wrt another) (ii) it should be possible to distinguish classical polymorphisms from inter homoeologues variation, but to be able to catalogue both (and infer the functional consequences of each). The task is made more complex by the current, provisional status of the wheat genome assembly, which suffers from fragmentation, missing sequence and (potentially) mis-assembly.
To achieve these aims required various modifications in both the browser software and in the underlying analyses required to support it.

(i) Separation of the assembly into its constituent genomes, and comparative analysis of the three wheat genomes against each other (i.e. A versus B, A versus D, B versus D) to determine homoeology. To support this, the code implementing the chain-net algorithm (Kent *et al*. 2003), one part of the normal Ensembl comparative analysis program, was re-implemented to support bi-directional analysis.

(ii) Improved linking of the output of the DNA and protein-based comparative analyses, to provide a clear visualisation of evidence for assertions of homoeology.

(iii) Co-visualisation of inferred homoeologous regions

(iv) Extension of the Ensembl variation infrastructure to support distinct collections of inter-homoeologous variants and classical polymorphisms (see figure 2).

(v) Development of a pipeline to parse inter-homoeologous variation from the whole genome alignments.

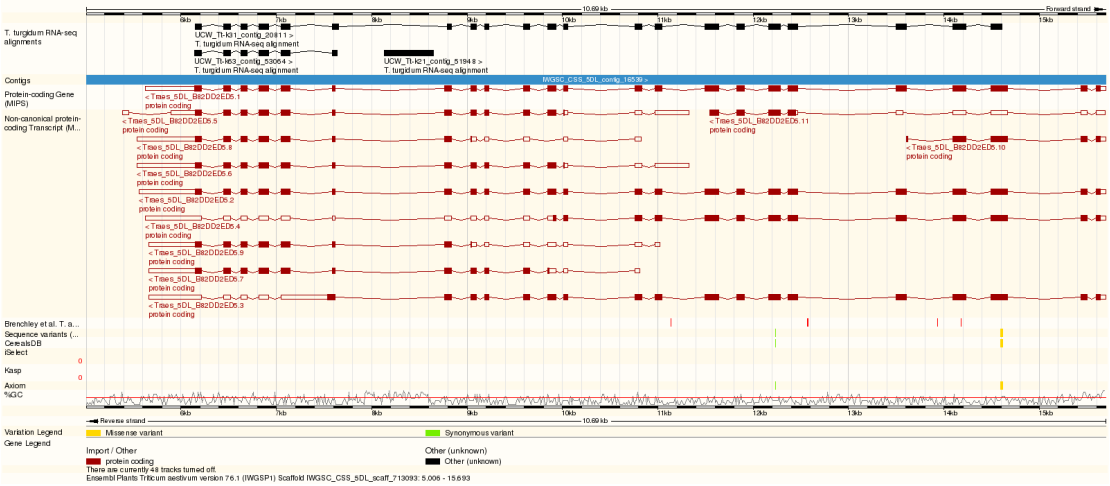


Figure 2. The visualisation of interhomoeologus variants (from the paepr by Brenchley *et al*.), and the Axiom, iSelect and Kaspar – dervied sets of intervarietal SNPs from Cereals DB, as separate tracks in the genome browser.

These improvements have been implemented and used to visualize the International Wheat Genome Consortium's Chromosome Survey Sequence of the genome of *Triticum aestivum* cv. Chinese Spring, which is currently the most complete reference available for this species. An illustration of this genome, showing the tri-partite visualisation of homoeologous regions and the alignment evidence supporting the assertion of homoeology within the Ensembl browser, is shown in figure 3.

Figure 3. The visualisation of the polyploid genome of bread wheat (*Triticum aestivum* cv. Chinese Spring) in Ensembl Plants. The figure shows the tri-partite visualisation of homoeologous regions. Alignment evidence supporting the assertion of homoeology is shown using the green shading. Other aligned regions are identified by the pink bars.

2. The grapevine and *Botrytis cinerea* pan-genome integration at INRA

Today, the quality of resequenced accessions does not generally allow a good assembly. In these cases, only small sequence variation can be found and displayed. This is the case for re-sequenced grapevine accessions. In this context, INRA tested GBrowse over Bioseq::feature to display small-size sequence variations obtained on grapevine accessions and mapped on the genome reference sequence. Links are established with the polymorphism module of GnpIS to display more detailed information on the structural variants, such as synonyms and accession details where the polymorphism has been identified (Ex: http://urgi.versailles.inra.fr/gb2/gbrowse/vitis_12x_pub/, Figure 1).

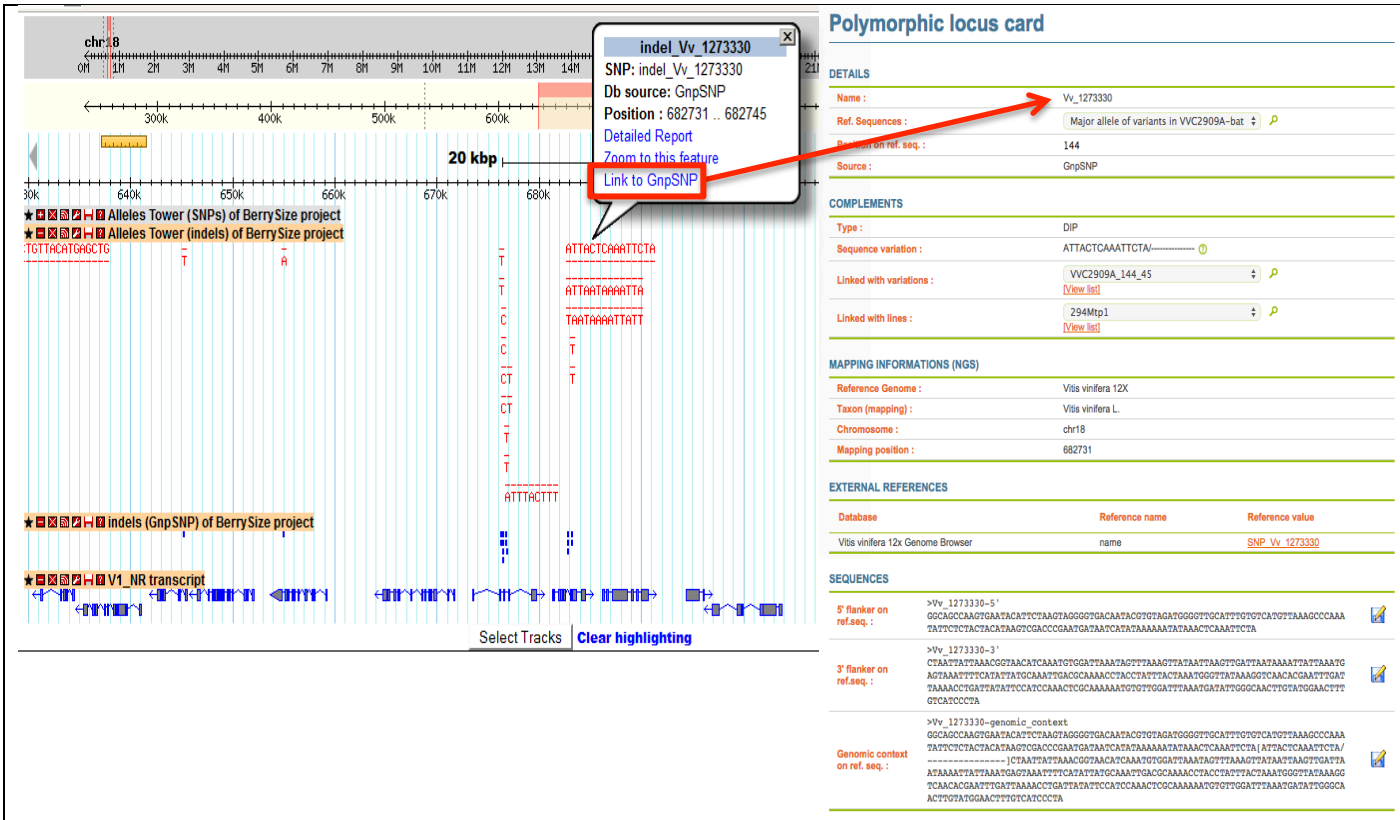


Figure 1: GBrowse grapevine example with link to the GnpIS polymorphism module.

From this GnpIS module, it is also possible to run IGV on BAM and VCF files on the fly to visualize calling evidences of structural variant (Ex: https://urgi.versailles.inra.fr/GnpSNP/snp/genomevariant/form.do, Figure 2)

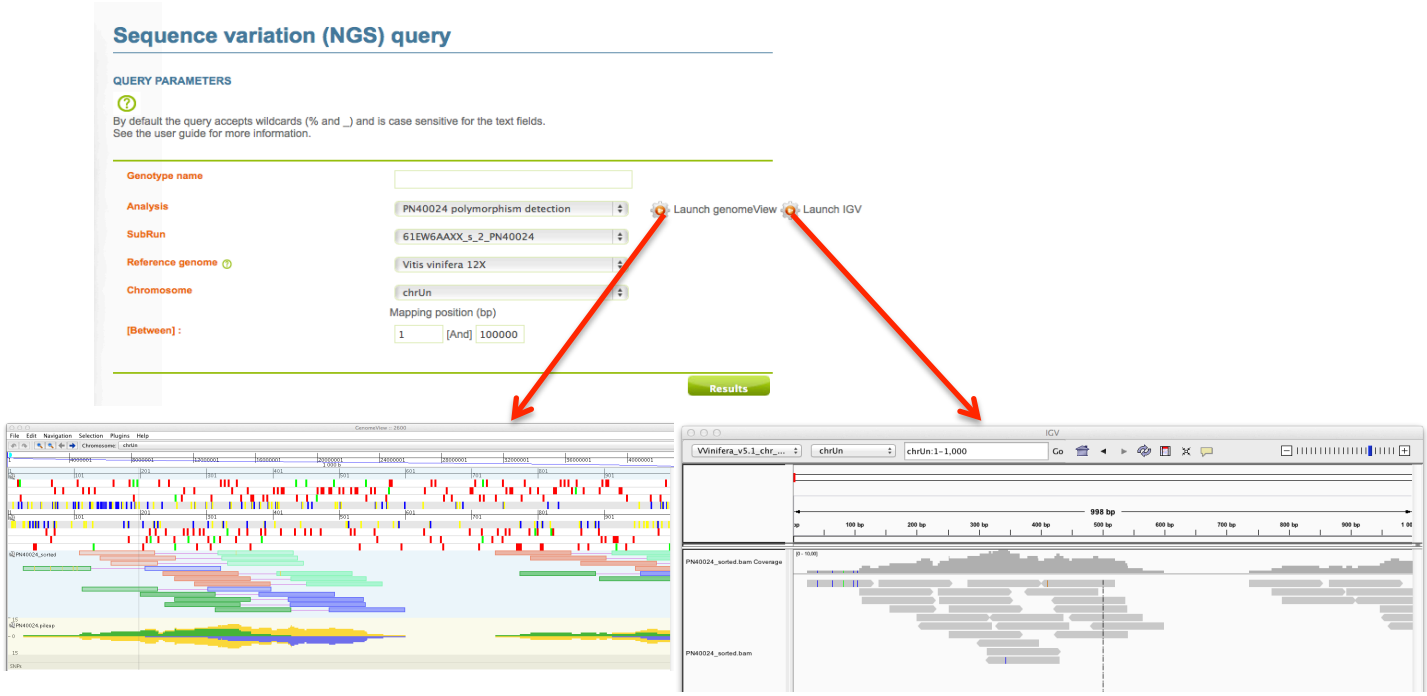


Figure 2: IGV or GenomeView access from GnpIS

The result appears satisfactory for that class of sequence variation. However, large rearrangement cannot be

represented with these tools.

INRA also tested GBrowse_syn to represent the synteny between two *Botrytis cinerea* strains and a close related species *Sclerotinia sclerotiorum* (http://urgi.versailles.inra.fr/gb2/gbrowse_syn/botrytis/ ,Figure 3). Larger rearrangements can be visualized, but to be correctly displayed, it requires good sequence assemblies with long scaffolds for all resequenced strains. Management of more than 5 strains could be a problem to visualize and to store. However, data available for plants and pests are not yet of enougth quality, and this tool appears sufficient today.

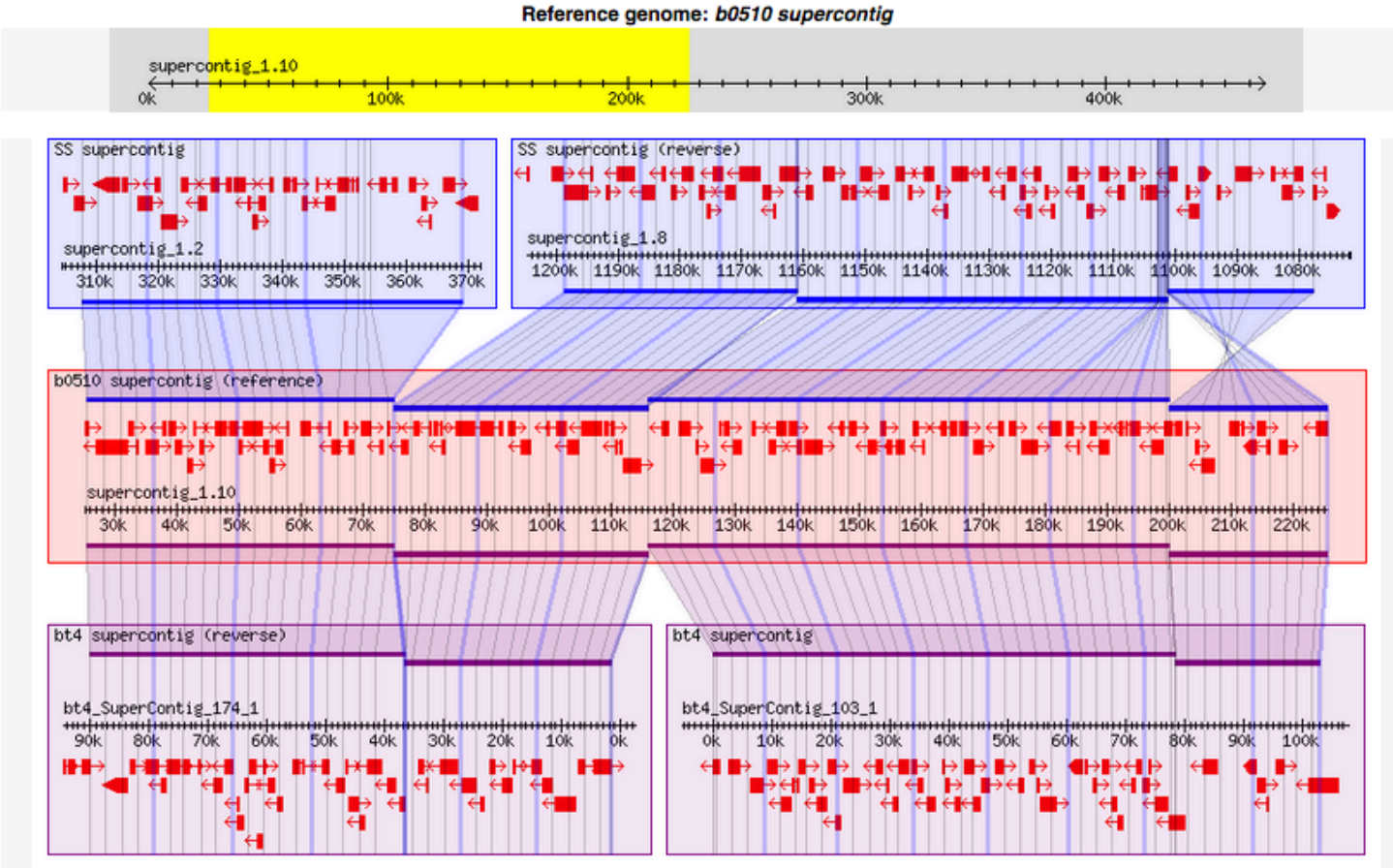


Figure 3: *Botrytis* and *Sclerotinia* synteny view with GBrowse_syn.

Recommendations for future developments

The browsing experience in Ensembl and Gbrowse is not as intuitive and smooth as the one experienced in more recent browsers like JBROWSE, especially with NGS data. A possible alternative (or complement) is IGV which can be launched with java webstart from a web portal and run locally, which makes it very fast (as soon as the computer has enough memory). Moreover, IGV supports a wide variety of data types which it can load from a centralized repository on a remote server.

The Ensembl genome comparison view could be improved as it has very little functionalities (compared to Gbrowse_Syn for example) as well as data associated with. It would be interesting for instance to show known

polymorphisms between the compared genotypes, especially structural variations. One reference genome could also be highlighted and the compared individuals could be ordered according to their pedigree. In fact, this genome comparison view should be the principal entry point and main component in a pan-genome viewer like in most synteny tools, and efforts should be made to make it more flexible, powerful and intuitive. Generally speaking, it would be nice also to simplify the browsing experience in an "Ensembl-like pangenome viewer" by reducing the number of specialized views and integrating more data types in the genome comparison view.

One last point which could be improved is the visualisation of the different types of variations especially in the context of the large number of lines being re-sequenced. TASUKE for instance offers a nice way of representing this type of data and interesting functionalities: the user can choose to view SNPs, INDELs, and/or coverage variations as well as snpEff predictions; he can select accessions to be displayed, and even select one of them to be used as the reference; he can also filter the variants to show according to their quality and read depth (more filters would be nice like the MAF).

Finally, whatever solution is chosen, it should be easily transferable to partners who might want to run locally an instance of this future pan-genome viewer, and thus local support and training should be planned.

Conclusions

This report lists the main functionalities expected by plant breeder users. We looked at available solutions, and among those studied, only the Personal Genome Browser gets close to what could be called a pan-genome viewer. Unfortunately, it can only be used for the human genomes. For plants, it will probably be necessary to develop new approaches to visualize, conveniently for breeding purposes, several individual genomes together under a coordinate-free system to figure out insertions and deletions.

Nevertheless, as shown by transplant partners, we can still use existing software to visualize some of the re-sequencing data deluge that we are already facing in some of the most studied plant species. Ensembl appears the most complete genome browser available for plants and it was already proved that with little development it was able to integrate the 3 genomes of polyploid wheat in a useful way. In the future Ensembl should extend the genome comparison processing and vizualisation that was developed for polyploid genomes to other species which already have multiple good quality assembled genomes (eg. 9 rice genomes are already available in Ensembl Plants). Another good data set could be the 19 genomes of the founders of the Arabidopsis thaliana MAGIC population since all 19 assemblies, bam files and RNA-seq data are available to download from (http://mus.well.ox.ac.uk/19genomes/). This could be a good demo for the transplant partners to evaluate the capabilities of Ensembl and to help better define which improvements should be made to approach the goal of implementing a Pan-genome viewer with all required features.