

Report of VIGNA-VIGNE Joint Meeting Paris, 24 November 2006

Participants :

ITALY

University of Milano: Enrico Pè and David Horner

University of Bari: Graziano Pesole

University of Padova: Giorgio Valle

University of Udine: Alberto Policriti, Nicola Vitacolonna, Cristian Del Fabbro

FRANCE

INRA

Anne-Francoise Adam-Blondon

Sébastien Aubourg

Michel Caboche

Christian Clépet

Fabrice Legeai

GENOSCOPE

Francois Artiguenave

Jean-Marc Aury

Nathalie Choisne

Olivier Jaillon

Claire Jubin

Francis Quetier

Jean Weissenbach

Patrick Wincker

Mailing list of participants :

(in order of the round table)

Francis Quétier : quetier@genoscope.cns.fr

Patrick Wincker : pwincker@genoscope.cns.fr

Enrico Pè : enrico.pe@unimi.it

Olivier Jaillon : ojailon@genoscope.cns.fr

François Artiguenave : artigue@genoscope.cns.fr

Claire Jubin : cjubin@genoscope.cns.fr

Michel Caboche : caboche@versailles.inra.fr

Giorgio Valle : valle@cribi.unipd.it

Graziano Pesole : graziano.pesole@biologia.uniba.it

Nicola Vitulo : nicolav@cribi.unipd.it

David Horner : david.horner@unimi.it

Jean Weissenbach : jsbach@genoscope.cns.fr

Cristian Del Fabbro : delfabbro@appliedgenomics.org

Nicola Vitacolonna : vitacolonna@appliedgenomics.org

Alberto Policriti : policriti@appliedgenomics.org
Fabrice Legeai : fabrice.legeai@versailles.inra.fr
Christian Clépet : clepet@evry.inra.fr
Sébastien Aubourg : aubourg@evry.inra.fr
Nathalie Choisne : choisne@genoscope.cns.fr
Anne-Françoise Adam-Blondon : adam@evry.inra.fr

1/ Current status of the French-Italian Consortium for the sequencing of the grape genome

Genome sequencing project in France

Genome sequencing and genome assemblies (Patrick Wincker)

PN40024 BAC library : sequenced
Fosmid library : to be constructed; to be sequenced
3kb plasmid libraries : constructed, active sequencing
10kb plasmid libraries : constructed, active sequencing
31st October 2006 : 4.6 million validated reads => about 6 genome equivalents.

Three assemblies have been done using Arachne when the coverage reached 2.9, 3.7 and 4.99X. These assemblies provide only raw results as no elimination of internal redundancy has been yet performed and no extension of the contig sequences. In the final assembly, in addition to these two points, some methods developed to deal with residual heterozygosity will be applied.

In the raw assemblies, no inflation of the genome size has been observed. The genome size in the last assembly is already close to expectations. This led us to confirm a good level of homozygosity of the sequenced genotype, PN40024. This aspect was further confirmed by :

- comparing the finished sequence of 9 random PN40024 BACs with the corresponding contigs of shotgun sequence : no polymorphism was observed in those regions.
- Whole genome => less than 20% of assembled regions display a residual heterozygosity (probably far less).
- Whole genome count of regions where bi-allelism can still be found : less than 20% of the regions.

Chloroplast « contaminations » (Francis Quetier)

In the rice sequencing project : development of a strategy (MUMmer algorithm and Blast) to distinguish between integrated pieces of chloroplast (Cp) genome in the nuclear genome and the true Cp genome. It showed that insertions can be found in all the chromosomes, with some chromosomes poorer in insertion events than others. The biggest inserted stretch, whose size is close to the total size of the Cp genome, is 99.98% identical to the true Cp genome.

Genoscope has carried out a purification of mitochondrial and Cp DNA from both organelles purified from PN40024 leaves in order to have a true reference for the two genomes for grape. The libraries are under validation by sequencing.

Full length cDNA (Christian Clépet)

Goal : 5 libraries, 50000 clones : 300 reference transcripts for the training of annotation software.

- Library A : Cabernet-Sauvignon suspension cells submitted to various abiotic stresses (normal, salt, water, hot, cold, anaerobic, Hygromycine) => Done
- Library B : Pinot Noir (PN162) leaves => Done
- Library C : Pinot Noir (PN177) flower buds => Done
- Library D : PN40024 leaves and petioles => Done
- Library E : PN40024 *in vitro* plant roots => To be done

The size of the clones obtained for the A, B, C, D libraries has been analysed by PCR. A subset of clones (1920) of the A, B, C libraries has been sequenced in order to assess their quality. Only 1.7% of bad sequences or empty clones were observed. Both reads were available for 1756 clones, among which 1409 clones were fully covered. 10% of these 1409 clones are more than 1.4 kb in size, 50% are between 0.9 and 1.2 kb in size and 30% are under 1.2 kb in size.

The complexity of the library was analysed after clustering of the 5' reads with CAP3. 895 different genes were obtained with an average of 1.9 clones/gene.

Regarding the completeness of the inserts, for 1672 sequences that match with Arabidopsis proteins, in 82% of the cases, the clones contained the sequence for the full protein. However no check for the ATG was done at this stage : it will be performed for the 50% of sequences fully included in a contig of the assembly. These gene will then be used for the constitution of the training set for the annotation of the grape genome.

Genome sequencing project in Italy

Padova group (Giorgio Valle)

This group has three 3730XL sequencing machines and 2.5 out of the three available are dedicated to the grape genome sequencing.

Number of runs per day : 24 (1 hour per run per sequencer) => 5600 reads

Average run length : 710 bases

Rate of success : 94%

Mainly 3kb clone libraries have been sequenced : there is still some improvement needed for the 10kb clones libraries.

Check for the "chloroplast contamination"

Problem of double sequences => well-to-well contamination?

November 2006 : 0.2X coverage has been reached.

April 2007 : 1.41 coverage

August 2007 : 2.34 coverage : 112 millions bases for the end of the project.

Udine Group (Alberto Policriti)

This group has 2 sequencers full time for sequencing and 1 for the fingerprinting of the PN40024 genotype until April 2007. After April, 3 sequencers will be available for sequencing the grape genome.

1 hour per run.

The read average length is 750 bases and the efficiency 94%.

They started in June 2006 : 362 plates have been processed between June and August 2006 and 878 plates since September. They have sequenced 1240 plates in total.

On 22nd November, 2006, the status of the sequencing project is:

- 335 151 360 bp sequenced
- 0.69X genome coverage
- 475 392 reads
- 1238 sequencing plates
- 619 library plates

The completion of the work is expected around July 2007 at the latest.

Throughput :

Sequencing plates/day : 12

Sequencing plates/week : 84

Daily output : 3 248 640 bp/day

Weekly output : 22 740 480 bp/week

Monthly output : 90 961 920 bp/month

Genome coverage/month : 0.19X

475 000 sequences were submitted ten days ago to the Trace repository but are still not accessible.

Tests of assembly using both shotgun sequences and physical map data and PCAP assembler are in progress using *Mus musculus* data.

Full length cDNA (Enrico Pé)

A CAP trapper method has been used to produce a first library from mixed tissues of Corvina berries at several stages of development (100 000 clones). The quality of the library is being evaluated (transcripts are 500-2000 nt in length). Sequencing is in progress.

Future plans consist in constructing other libraries :

- leaves => to be skipped?
- *In vitro* cells treated with different pathogens (powdery and downy mildew) and nitric oxide.

2/ Genetic and physical mapping

Genetic and physical mapping (Anne-Françoise Adam-Blondon)

The aim is to order and orient the contigs of shotgun sequences along the *Vitis* chromosomes.

The biggest public genetic map available (Doligez *et al.* 2006 Theor Appl Genet, 113 : 369-382) is used as a reference in the project : it contains 515 loci among which 257 are framework. Two improvements are necessary :

- Increase the number of loci. For this purpose, 286 SSR markers found in Cabernet-Sauvignon BES have been tested : 172 have amplified on genomic DNA, among which, 149 have been tested for polymorphism. Only 63% of the tested markers have proved to be useful. Many of them were multi-loci and thus useless to make links between the genetic map and the contigs of sequence.

- Improve the robustness of the marker ordering by using a biggest population (many markers of the Doligez *et al.* map were scored on less than 100 individuals). 250 non-framework markers of the Doligez *et al.* map are currently being tested for polymorphism.

Another tool is available : a Cabernet-Sauvignon physical map, which consists of 1763 fingerprint contigs (Sophie Paillard, unpublished) anchored with 1641 markers (Lamoureux *et al.* 2006, Theor Appl Genet, 113 : 344-356, MR Thomas, unpublished). Only 322 of these markers were genetically mapped : 71 additional mapped markers have been anchored.

In parallel a database of genetic and physical data has been developed. The C-map viewer has been tested for the visualisation of the links between the genetic map, the physical map and the contigs of sequence.

The future is to use the next available assembly already arranged into super-contigs using the BAC end sequences and to realize :

- Blasts of BES and ePCR of markers from the physical map in order to organize the super-contigs into meta-contigs

- ePCR the genetic markers on the sequence contigs in order to link the genetic map and the meta-contigs of sequences.

- Design markers on unmapped or un-orientated meta-contigs and genetic mapping of these markers.

Happy mapping (Enrico Pé)

In this approach, pools of genomic DNA representing less than 1X genome are used like radiation hybrids in mammals for mapping.

Several steps are to be completed :

- extraction of high molecular weight DNA => OK

- gamma ray irradiation of the DNA => range chosen

- panel preparation => OK

- Pre-amplification of panel DNA by degenerate 50-mers (?).=> Done

- Second pre-amplification with multiplexed set of primers corresponding to the markers to be mapped => underway with a first set of test markers

- Individual amplification of the markers => underway

After this first attempt a decision will be made as to whether it is worth going on.

Two panels of markers are possible if the approach works :

- Large range panel (100-800 kb) => STS and BES (1200 markers) linking physical and genetic maps.

- Short range panel (50-400 kb) => assist the local ordering of sequence contigs.

Physical map of the PN40020 (Alberto Policriti)

BAC fingerprinting :

70656 BAC clones

8 plates/week

3072 fingerprints/week

At the 22nd November, 2006, the status is:

Processed BAC library plates (384) : 49

Fingerprinting of : 18816 clones

Coverage : 5.5X

The completion of the work is expected in April 2007 (FPC assembly comprised).

3/ Annotation strategy in France (Olivier Jaillon)

Use of GAZE to integrate several layers of data to produce gene models (based on dynamic programming, Howe KL *et al.* 2002 Gen Res 12 : 1418-1427) :

- *Ab initio* predictions : Geneid, SNAP, GlimmerHMM, Eugene
- Comparative genomics : ExoFISH (*A. thaliana*, Poplar, Rice...)
- Alignment with expressed sequences : genewise, EST2Genome

Computing at Genoscope : one month.

Functional annotation : Interproscan, Gene Ontology, Kegg, orthologous genes, paralogous genes.

Visualisation : Generic Genome Browser (GGB).

4/ Plan for bioinformatics in Italy (Graziano Pesole)

VIGNA : VitisGeNomeAnalysis

New assembly by the Udine group using PCAP with physical map constraints.

Development of specific tools for genome annotation :

- identification of repeats and mobile elements : A. Policriti => 2 or 3 months
- gene finding (comparative approaches , SGP2) : G. Pesole => 2 or 3 months
- Detection and analysis of conserved sequence tags and ultra-conserved elements (CSTminer, GenoMiner) : G. Pesole => 2 months.
- Large scale genome alignments : G. Valle or A. Policriti => 6 months.
- Prediction of microRNAs and other non-coding RNAs (Universities of Bari and Milano)
- Alternative splicing prediction (ASPIC)
- Detection and analysis of transcription factor binding sites (WEEDER) and RNA motifs (RNAprofile) involved in post-transcriptional regulation (prediction of 5' or 3' or non-coding RNAs).

At Christmas, we will know how to construct the pipeline.

Complementarity and priorities :

Repeated sequences : Verona has a tool for counter filtering repeated sequences (ready in a couple of weeks) and Udine the tools for annotation of TE elements.

5/ Data exchange within the Consortium

Policy on reads

Each participant will deposit sequences in batches identified with the name of the laboratory and the date of the batch deposition on a ftp website accessible by the three sequencing laboratories with a password. All the sequences should be deposited, without any filtering.

Policy on assembly

The raw assemblies will be very helpful. The 5X assembly will thus be deposited on the ftp website.

The access to this ftp website by other groups of the VIGNA Consortium is possible but with the signature of a MTA (Materiel Transfer Agreement).
Two weeks are necessary to set up this system.

Policy on cDNA

The training set of genes will be distributed to the Consortium as soon as it is completely finished and evaluated.

Policy on physical map data

A meeting will be organized at Udine.

Data exchange format

C-map for genetic and physical map. For sequences it has been discussed in the former meeting and already agreed.

6/ Public data release

A demand for a massive use of the contig sequences (5X assembly) has been submitted to F. Quétier from Black Meyers : the Consortium agrees to sign a MTA but it must be clear that it is a preliminary assembly that may contain errors.

A demand of the group of USDA Geneva has been made to A-F Adam-Blondon on the way to acknowledge the Consortium in a publication using the data available on the web sites of the sequencing project : exchanges by mail will be done to set up a sentence that will be posted on each web site.

Next meeting : mi-February at Padova.