Theor Appl Genet (2006) 113:344–356
DOI 10.1007/s00122-006-0301-7

ORIGINAL PAPER

Anchoring of a large set of markers onto a BAC library for the development of a draft physical map of the grapevine genome

Didier Lamoureux · Anne Bernole · Isabelle Le Clainche · Sarah Tual · Vincent Thareau · Sophie Paillard · Fabrice Legeai · Carole Dossat · Patrick Wincker · Marilyn Oswald · Didier Merdinoglu · Céline Vignault · Serge Delrot · Michel Caboche · Boulos Chalhoub · Anne-Françoise Adam-Blondon

Received: 23 January 2006 / Accepted: 21 April 2006 / Published online: 18 May 2006

Abstract Five hundred and six EST-derived markers, 313 SSR markers and 26 BAC end-derived or SCAR markers were anchored by PCR on a subset of a Cabernet Sauvignon BAC library representing six genome equivalents pooled in three dimensions. In parallel, the 12,351 EST clusters of the grapevine UniGene set (build #11) from NCBI were used to design 12,125 primers pairs and perform electronic PCR on 67,543 nonredundant BAC-end sequences. This *in silico* experiment yielded 1,140 positive results concerning 638 different markers, among which 602 had not been already anchored by PCR. The data obtained will provide an easier access to the regulatory sequences surrounding important genes (represented by ESTs). In total, 1,731 islands of BAC clones (set of overlapping BAC clones containing at least one common marker) were obtained and 226 of them contained at least one genetically mapped anchor. These assigned islands are very useful because they will link the genetic map and the future fingerprint-based physical map and because they allowed us to indirectly place 93 ESTs on the genetic map. The islands containing two or more mapped SSR markers were also used to assess the quality of the integrated genetic map of the grapevine genome.

Electronic Supplementary Material Supplementary material is available to authorised users in the online version of this article at http://dx.doi.org/10.1007/s00122-006-0301-7.

Communicated by P. Langridge

Didier Lamoureux and Anne Bernole contributed equally to this work.

D. Lamoureux · A. Bernole · I. Le Clainche · S. Tual · V. Thareau · S. Paillard · M. Caboche · B. Chalhoub · A. F. Adam-Blondon (✉)
UMR INRA-CNRS-UEVE de Recherches en Génomique Végétale, 2 rue Gaston Crémieux, BP5708, 91057 Evry Cedex, France
e-mail: adam@evry.inra.fr

Present address: A. Bernole
Gip GEVES Magneraud, Le Magneraud St-Pierre-d'Amilly, BP 52, 17700 Surgères, France

Present address: V. Thareau
Institut de Biotechnologie des Plantes, Université Paris-Sud, Bat. 630, 91 405 Orsay cedex, France

F. Legeai
Unité de Recherches en Génomique-Informatique, Infobiogen, 523 place des Terrasses, 91 034 Evry cedex, France

C. Dossat · P. Wincker
Génoscope, 2 rue Gaston Crémieux, BP5708, 91057 Evry Cedex, France

M. Oswald · D. Merdinoglu
UMR Vigne et Vins d'alsace, 28 rue Herrlisheim, 68021 Colmar cedex, France

C. Vignault · S. Delrot
UMR Transport des Assimilats, Bât. Botanique, Université de Poitiers, 40 Avenue du Recteur Pineau, 86022 Poitiers, France

Introduction

Grapevine (*Vitis vinifera* L.) breeding often focuses on two main goals: better resistance levels to various pathogens and improved berry quality. These two goals are intertwined since resistant genotypes are usually found in wild species also exhibiting poor fruit quality, while commercial cultivars are often susceptible to several pathogens. Moreover, two main obstacles are slowing down genetical progress. First, grape is a perennial woody plant, which is synonymous with a relatively long cycle from seed to seed. Second, its genome exhibits a high level of heterozygosity due to preferential allogamous mating (Siret 2001; Aradhya et al. 2003; Salmaso et al. 2004).

High levels of heterozygocity are making difficult the construction of fingerprint-based physical maps and also sometimes PCR walking on genes and their regulatory sequences. The latter problem can be solved by using BACs carrying the gene of interest as a template for PCR walking whereas the former requires both to establish links between the genetic map and the contigs of BACs and between the BAC clones contained in a contig. Several large-insert comprehensive genomic libraries have also been developed for the grape genome (Tomkins et al. 2001; Adam-Blondon et al. 2005; http://www.vitaceae.org) using the bacterial artificial chromosome system (BAC, Shizuya et al. 1992). One of them used the Cabernet Sauvignon cultivar and contains 44,500 clones with a mean insert size of 142 kb, thus representing about 12.3 genome equivalents (Adam-Blondon et al. 2005). This library is one of the two reference libraries chosen by the International Grape Genome Program (IGGP) network (http://www.vitaceae.org) for grape genomics projects. A subset of this library representing six genome equivalents has been pooled in three dimensions to allow easy PCR screening (Adam-Blondon et al. 2005). On the other side, thanks to this high level of heterozygosity, genetic maps covering the whole genome are now available. Several of them were built with simple sequence repeats (SSR) markers (e.g. Riaz et al. 2004; Adam-Blondon et al. 2004). These highly polymorphic markers, transferable from one species to another close one, allow the comparison between maps from different populations, or even their combination into an integrated map (Doligez et al. 2006).

The objectives of our work were (1) to start to establish a connection between existing genetic maps of the grape genome and a physical map under construction using a fingerprinting-based method (Luo et al. 2003), (2) to establish links between overlapping BACs through the anchorage of a large number of markers and (3) to provide an easier access to the regulatory sequences surrounding important genes (represented by EST Vitis sequences). For these purposes, we decided to anchor on the Cabernet Sauvignon BAC library, either by PCR or *in silico*, genetically mapped markers (mainly SSRs) and a large set of ESTs. All theses resources are made available to the grapevine scientific community.

Materials and methods

BAC library 3D pooling

The construction of the pools has been described in Adam-Blondon et al. (2005). Briefly, for each of the six superpools, the clones from eight 384-well plates were grown on solid medium and pooled by plate, line and column dimensions. The suspensions in water were centrifuged, resuspended in $T_{10}E_{10}$ then boiled at 96°C for 30 min and centrifuged again. The supernatant was used as PCR template after dilution to 1/400 in $T_{10}E_{0.1}$. The BAC library is public and available for ordering at the National Resource Center for Plant Genomics (http://www.cnrgv.toulouse.inra.fr).

Choice of SSRs and ESTs for PCR anchoring

SSRs were mainly chosen from the VVI set (Merdinoglu et al. 2005; data deposited at NCBI dbSTS under accession numbers BV140581 to BV140613 and BV140613 to BV140771) and the Vitis Microsatellite Consortium set coordinated by Agrogene. The others are described in Riaz et al. (2004) and Adam-Blondon et al. (2004). Most of them have been mapped on the integrated *Vitis vinifera* genetic map (Doligez et al. 2006).

Around 1,000 ESTs were chosen from the French UniGene set (Terrier et al. 2005). Primers were designed using the GENOPLANTE®SPADS v 1.1.4 software (Thareau et al. 2003) to improve the specificity of the primers regarding the known EST Vitis sequences (that were used as the reference sequence to test the specificity of the primers) with the following parameters: amplicon size min = 100, opt = 200 and max = 300, primer size min = 18, opt = 20 and max = 25, T_m min = 55, opt = 60 and max = 65, GC% = 40–80 and T_m difference between the two primers = 3°C. All ordered primers were tested on Cabernet Sauvignon genomic DNA. The subset of primers that allowed an amplification was further analysed on 3D pools.

The whole set of markers, later on referred to as 'STS', is described in supplementary data S1 that can be downloaded at the following address: http://urgi.infobiogen.fr/projects/CT_Cible_Importante/CI2001002//index.php.

PCR anchoring on 3D pools

The pooled DNA was used to complement a 8-µl reaction mix containing 0.5 U AmpliTaq® DNA polymerase (PE Applied Biosystems), 1× PCR buffer, 1.6 mM $MgCl_2$, 400 µM dNTP, 12 pmol of each primer and 20% v/v of loading buffer (60% w/v sucrose, 5 mM cresol red in water). The final volume of the reaction was 10 µl. For each analysed marker, the experimental design included three controls: water, pIndigoBAC vector carrying *E. coli* total DNA (called "vector control" below) and Cabernet Sauvignon genomic DNA.

PCR was conducted with the following parameters: 94°C—5 min; then 15 cycles of 94°C—20 s, 65°C—20 s with a decrease of 1°C per cycle, 72°C—20 s; then 35 cycles of 94°C—20 s, 50°C—20 s, 72°C—20 s; and finally 72°C—7 min. This program was run on a GeneAmp® PCR System 9700 (PE Applied Biosystems). PCR amplifications were loaded in 1× TAE-buffered Seakem® LE (Cambrex Bio Science Rockland Inc.) 3% agarose gels, and electrophoresis was done at 300 mA for 40 min at 10°C. Gels were stained with ethidium bromide and UV pictures of the gels were taken with a Gel Doc system (Bio Rad).

PCR checking of 3D coordinates on individual clones

STS, which were also genetically mapped by Doligez et al. (2006), were assayed by PCR on individual clones to assess the accuracy of '3D coordinates' determination. PCR reactions were conducted the same way as above except that a 15-µl reaction mix was used. Template DNA was added by simply dipping a toothpick in the cultured medium of a specific clone then in the reaction mix. The whole volume of the reaction was loaded on the gel for electrophoresis.

In silico anchoring

Primers were designed using Primer3 v0.9 software (Rozen and Skaletsky 2000) on the best representative sequence of each cluster from the grape UniGene set (build #11 with 12,351 clusters) downloaded from NCBI (http://www.ncbi.nlm.nih.gov/UniGene) with the following parameters: product size from 100 to 500 bp (optimal: 200 bp), primer size from 18 to 25 bp (optimal: 20 bp), primer GC content from 40 to 80%, primer T_m from 50 to 65°C, maximum difference in T_m for paired primers of 5°C.

The ends of the 44,544 BAC clones from the Cabernet Sauvignon library were one pass sequenced. As a result we obtained 77,237 exploitable sequences with an average size of 671 bp. Those sequences have been deposited in EMBL database under accession number CT486010 to CT563247. In order to avoid the most redundant sequences, we only considered the sequences that behave as singletons when clustered with Biofacet v2.4 (Glémet and Codani 1997) with the following parameters : 95% nucleic identity on the total length of the smallest sequence. We ended up with a set of 67,543 sequences.

The anchoring was done with electronic PCR (e-PCR; Schuler 1997) in its latest version known as me-PCR v1.0.5c (Murphy et al. 2004). Running parameters were as follows: word size of 4 bp, two mismatches allowed outside of the word, product default size of 400 bp, margin of 250 bp.

Draft assembly

The software SAM v2.5 (Soderlund and Dunham 1995) was used to process the data set. We considered as an 'island' any set of clones with at least one marker, and as a 'contig' any combination of at least one clone harbouring two different markers plus a second clone harbouring at least one of these markers. Contigs were assembled linkage group by linkage group by first loading known mapped markers, then adding step by step new clones sharing at least one marker with loaded clones ('Follow map' command). Finally, the clones from all identified assigned islands were removed and the remaining clones were used to build contigs unassigned to linkage groups.

Results

The total data set of coordinates of the positive BACs obtained both in PCR and *in silico* anchoring experiments is given in electronic supplementary materials S2 that can be downloaded at the following address: http://urgi.infobiogen.fr/projects/CT_Cible_Importante/CI2001002//index.php.

PCR anchoring

We tested 828 different primer pairs for anchoring on the Cabernet Sauvignon BAC library subset. While 257 of them were SSRs, 545 were ESTs and the 26 other were SCAR or markers derived from BAC end sequences. A small proportion (15 primer pairs, 1.81%) failed to amplify anything in our experimental conditions, including the genomic control despite the results of previous tests. Seven pairs (0.85%) amplified vector sequences or water controls or amplified weakly and were thus impossible to score on 3D pools. Twenty

others (2.42%) amplified every pool DNA but not the vector control and were therefore suspected of being highly repetitive sequences or chloroplastic markers since chloroplastic contamination in the BAC library was estimated to be about 2.6% (Adam-Blondon et al. 2005). Interestingly, 14 primer pairs (11 ESTs and 3 SSRs, 1.69%) did not amplify any pool DNA although they amplified the genomic control DNA. These primer pairs may thus target sequences that were not represented in the 6× part of the source library (Table 1). In summary there were 772 useful primer pairs divided into 502 ESTs and 244 SSRs and 26 of other origins.

Because 82 primer pairs, mainly SSRs (77), amplified several distinct bands (which may represent either different alleles for the same locus or different loci; Adam-Blondon et al. 2005), we scored each band independently and got a total of 506 EST markers, 313 SSR markers and 26 BAC end-derived and SCAR markers.

PCR checks were performed on individual clones for more than 300 marqueurs to increase the number of unique BAC clone coordinates (see S1 and S2). We used this dataset to estimate the reliability of the results obtained with 3D pools. Seven hundred and fifty-four 3D 'unique' coordinates (only one positive clone in the super-pool) for 286 markers were compared with data obtained after individual checks: 86 anchors were not confirmed, which would suggest a 11% rate of 3D false positives for the 3D pools screening. However, it should be noted that this error rate also include false negatives from the PCR on individual clones and is thus an overestimation. The number of observed positive clones for a single marker ranged from 1 to 45. The mean number of positive clones for each marker was 6.60 overall, but was 7.34 for ESTs as opposed to only 5.35 for SSRs (5.9–6.5 and 4.8, respectively, taking into account the rate of false positive discussed above). The distribution of the number of markers according to the number of positive clones (Fig. 1) was significantly different for the two types of markers, as shown by a P value of $2.2e^{-16}$ for the Wilcoxon test. Moreover, we compared the two types of markers for the number of markers with five or less hits as opposed to the number of markers with more than five hits in a chi-square test. The actual classes were significantly different from theoretical classes (P value of $1.4e^{-14}$). The same result was achieved when defining the classes with 'seven or less hits' as opposed to 'more than seven hits' (P value of $2.2e^{-11}$).

The average number of clones presented above takes into account 'unique' coordinates obtained on each superpool or checked on individual clones as well as 'ambiguous' coordinates obtained when several clones were positive for a given superpool (in this case the maximum of the number of hits obtained in the plate pools, line pools or column pools was considered). With the same hypothesis as above regarding the Poisson distribution applied to each superpool with a $\lambda = 0.94$ parameter (the approximate genome coverage of each superpool), the probability of getting no hit on a given superpool is 39%, the probability of getting exactly one hit is 37% and the probability of getting more than one hit is 24%. With a subset of 4,338 more deeply analysed tests (723 markers by six superpools), we obtained 1,764 (40.6%) results with one hit (1,570 hits taking into account the rate of false positives, 36.2%) and 1,472 (34.9%) results with more than one hit (1,310 hits taking into account the rate of false positives, 36.2%).

Table 1 List of the STS for which no hit was found in the 6× subset of the BAC library that was screened

Marker name	GeneBank accession number	NCBI build #11 cluster number	Linkage group
CM004A05	BQ792640	Vvi.6800	
GB007D01	BQ798707		16
GT173E07	BQ799100	Vvi.6744	
GT182H06	BQ799299	Vvi.412	
PT011A02	BQ800588		
RB005C10	BQ795366	Vvi.466	
RT043D04	BQ796542	Vvi.7826	
RT052H03	BQ796771	Vvi.7573	
RT061C10	BQ797050	Vvi.6501	
RT082C03	BQ797702	Vvi.779	
RT092G12	BQ797982	Vvi.6997	
VMC3B7-2			19
VMC8G9			12
VVIO61			1

In silico anchoring

In order to gain additional markers, we used the grapevine UniGene set (build #11) from NCBI. From this source of 12,351 EST clusters, we successfully designed 12,125 primer pairs. With these primer pairs and the SSR primer pairs previously used in PCR anchoring experiments, we scanned in an electronic PCR experiment the 67,543 nonredundant BAC end sequences we obtained from our library. This step yielded 1,140 positive results concerning 638 different markers (among which 605 ESTs, see electronic supplementary data S1 and S2), each one obtaining from 1 to 44 hits with an average of 1.83 positive clones per marker. However, some of the markers were already present in our 'wet' anchoring experiments (list given in Table 2). The exact number of new markers added by the *in silico*

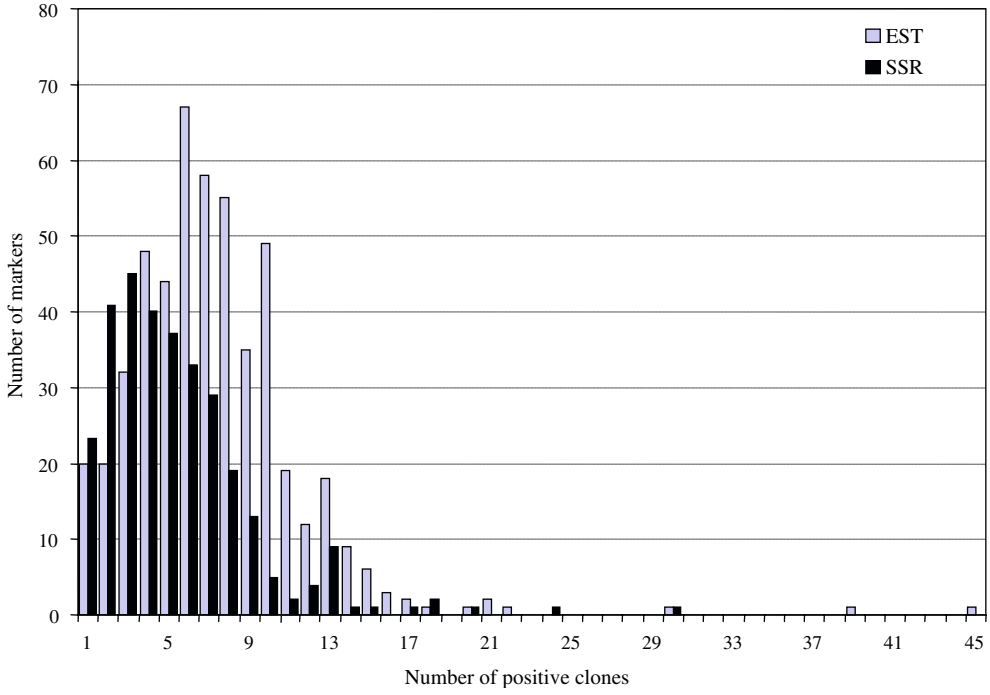


Fig. 1 Distribution of the number of markers according to the number of positive clones for each category of marker

approach was therefore 602. Table 2 shows the difference in the average number of BACs anchored for 36 markers for which we obtained results by both PCR (7.55 hits in average) and ePCR (1.14 hits in average) approaches.

Physical map draft

A total number of 1,398 markers anchored with unique coordinates of a total of 2,681 BAC clones, i.e. 6% of the complete Cabernet Sauvignon BAC library and 14.5% of the clones included into the screened pools. With this data set, we assembled 1,731 anchored islands (same notation as Ewens et al. 1991). The 1,731 islands were divided into 226 assigned islands (i.e. with at least one genetically mapped anchor, Fig. 2) and 1,521 unassigned islands. Sixteen of these had to be discarded because their genetically mapped anchor appeared to be multilocus on the grapevine integrated genetic map (Fig. 2). The number per linkage group of the remaining 210 islands ranged from 3 (linkage group 3) to 27 (linkage group 1), with an average of 11 islands per linkage group (Fig. 2).

Among the islands, we distinguished what we called contigs (see definition in Materials and methods section). There were 170 contigs divided into 77 assigned contigs and 93 unassigned contigs. Each one of the 19 linkage groups contained at least one contig, the maximum being eight on linkage groups 1 and 8. An average of four contigs per linkage group was obtained. These contigs allowed us to indirectly place 93 ESTs on the genetic map thanks to their physical link with a genetically mapped SSR on a contig (Table 3).

The contigs containing two or more SSR markers (indicated by black boxes on Fig. 2) were also useful to assess the quality of the integrated genetic map of the grapevine genome (Doligez et al. 2006). They contain adjacent markers on the genetic map in most of the cases with three exceptions (Fig. 2): in linkage group 10 one contig was made of BACs anchored by both VMC4F9.1 and VVIR21, in linkage group 12 several contigs could be built with BACs anchored by VVIB10, VVIV05, VMC7F1 and VMCNG2H7, but no BAC of these contigs was anchored by VMC4F3.1 and VMC8G9 and finally in linkage group 18 several BACs were anchored by ADH3, ADH1 and VVIU04 but none of the BACs of the contig contained VVIP08. In group 10, none of the maps used to construct the integrated map contained both VMC4F9.1 and VVIR21 : this problem may be due to the difficulty in obtaining a clear order in some areas of the maps in the integrated map approach for reasons that are extensively discussed by Doligez et al. (2006). In group 12, all the markers in the contigs except VMC7F1 were duplicated thus hampering the assignment of a group of anchored BACs to a specific map position. In group 18, the order of the markers on the genetic map was chosen as the most probable one, but other orders could

Table 2 Comparison of the number of BAC anchored using the same primer pairs either by PCR screening of the 3D pools or by ePCR screening of the BAC-end sequences

Marker name	Number of BAC anchored by ePCR	Number of BAC anchored by PCR
GT171G08	1	12[a]
GT171H11	1	7[a]
GT172F12	1	10[a]
GT184F02	1	7[a]
PT007A04	2	All[a]
RB000A24	2	45[a]
RB006D06	1	13[a]
RT012B04	1	5
RT024C05	1	12
RT081A11	1	1[a]
RT084F10	1	8[a]
ST004C08	2	4[a]
TB000A08	2	12[a]
TB005E07	1	10[a]
TT283C09	1	1[a]
VMC3C7	1	7
VMC3E12	1	10
VMC3F8	1	5
VMC5B3	1	7
VMC5E9	1	3
VMC5H11	1	12
VMC5H2	2	6
VMC7F1	1	14
VVGLUSTR	1	2
VVIH54	1	2
VVII52	1	3
VVIM01	1	3
VVIM04	1	6
VVIM25	1	1
VVIN78	1	5
VVIP08	1	7
VVIP09	1	7
VVIQ22.2	1	5
VVIQ57	1	7
VVIT30	1	1
VVIT65	2	10
VVIV35	1	2
	Average[b] 1, 14	Average[b] 7, 55

[a] The number of positive clones was estimated directly from the results obtained after the 3D pools screening as the sum on the six super pools of the maximum of the numbers of coordinates obtained for the plate, line and column

[b] The marker PT007A04 has been excluded from the calculation

not be discarded (LOD difference with the most probable order < 2).

Discussion

We present here the largest set of gene-related and SSR markers anchored on a grapevine BAC library to date (1,447 markers) by two different approaches (PCR and ePCR) and their use to build a draft physical map of the grapevine genome. In order to set up a resource useful for both mapping-based projects and functional analysis projects of the grapevine community, we took care to establish links between the BAC islands and the available grapevine genetic maps and to anchor as many genes as possible using the available unigene sets.

Anchoring markers on BACs with two strategies

Out of 828 primer pairs 772 were successfully used to anchor by PCR 506 ESTs, 313 SSRs and 26 other markers on a set of BAC clones representing six grapevine genome equivalents organised in 3D pools. Only a small proportion of the tested primers (14) amplified none of the BACs of the subset. Assuming that with a library coverage of 6×, the probability of getting exactly n positive clones for any marker follows a Poisson law with a parameter $\lambda = 6$, according to the assumptions of Arratia et al. (1991), then $P(n = 0)$ was close to 0.0025 and when considering the number of tested markers, the expectation for 'no hits' results was about two. The difference between the predicted and the observed values probably reflected the fact that the 6× subset used for the 3D pool construction contains 75% of *Hin*dIII clones. This may have favoured a slightly biased sampling of the genome.

The average number of positive clones per marker was 5.9 to 6.6 in the entire dataset, which is consistent with the screening of a set of BAC representing six genome equivalents using unique sequences. However, we also obtained a significantly higher average number of hits with EST markers than with SSR markers. The proportion of markers anchoring more than one BAC clone in the superpools was also higher than expected and might reflect the fact that part of the primers we used were amplifying duplicated sequences. SSR primers were chosen to be mapped at a unique locus on the available genetic maps when we started the work (Riaz et al. 2004; Adam-Blondon et al. 2004). The EST-based primers were designed from a UniGene set deduced from a set of 3′ end sequences (Terrier et al. 2005) which are more likely to contain untranslated gene-specific sequences. The risk of amplifying multigene families was thus minimized but our results showed that it was probably not completely avoided. The higher mean number of hits per EST-derived marker could also be explained by a non-homogeneous coverage of the grape genome by our library, which would show a slight bias favouring genic regions as compared to SSR-containing regions. However, at the present time, no data allow us to state that these regions might be distinct. Moreover, SSR motifs tend to be more

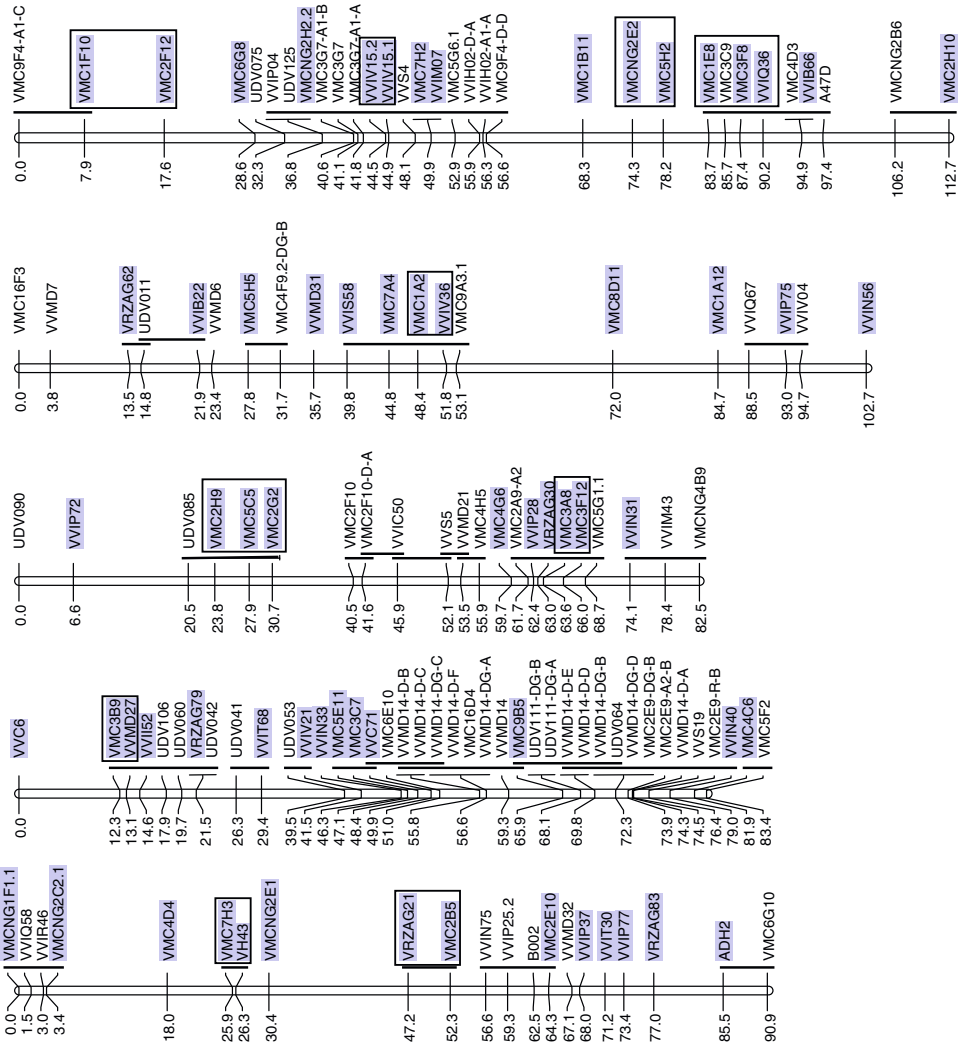
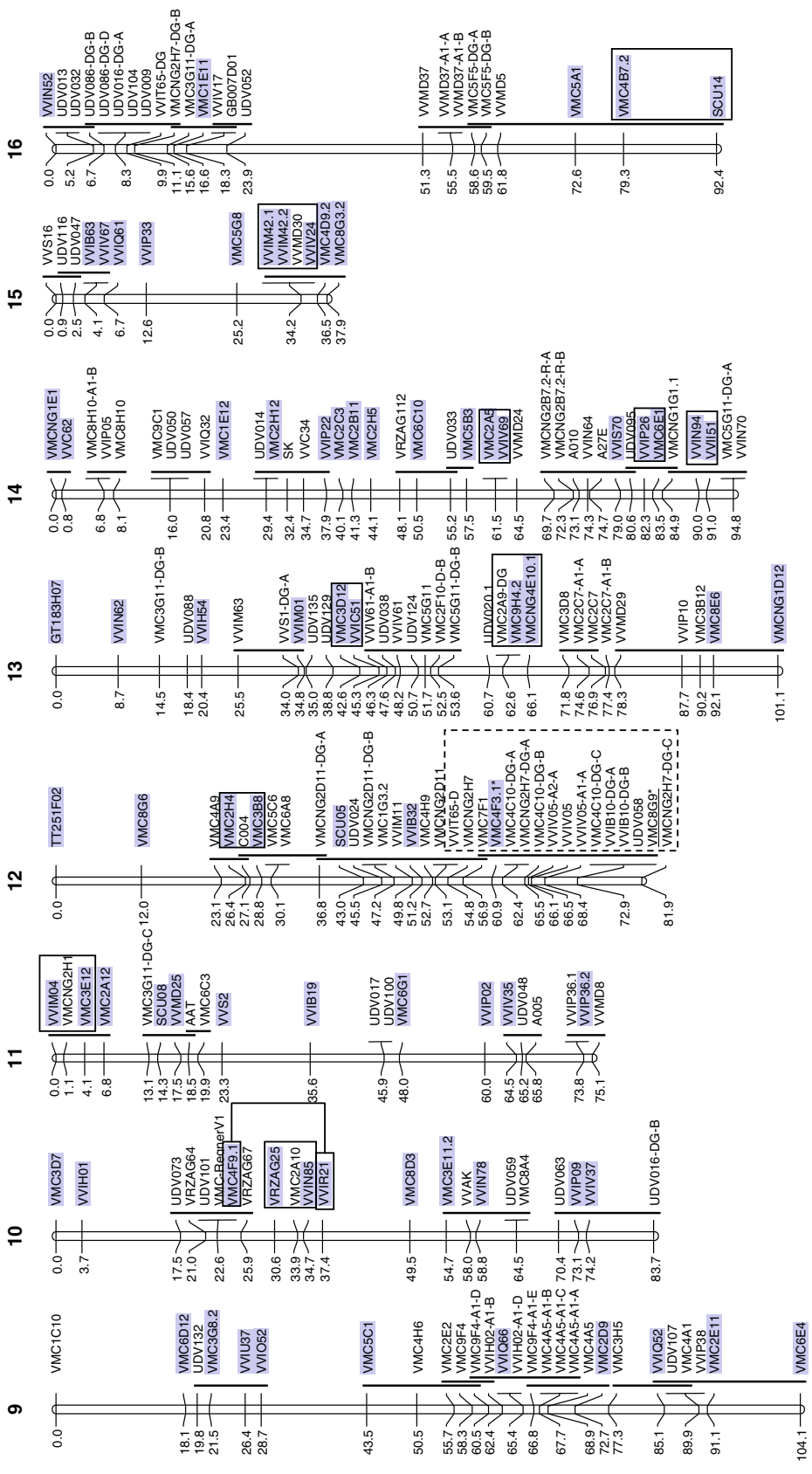


Fig. 2 Complete integrated map built from five different grapevine populations by Doligez et al (2006). Distances are in cM Kosambi. *Vertical bold lines* indicate groups of loci with orders unsure at LOD 2. Markers anchored on BAC islands are in *grey boxes* when they correspond to a unique locus and in *hatched grey* when they correspond to duplicated loci. Markers that anchored the same contig of BACs are in a *black square*. *Hatched black square* indicates that the contig was built only with duplicated markers. Markers in a *black square* that are not part of the contig are indicated by a *star*. Markers for which no hit were found on BACs are *underlined*

Fig. 2 continued



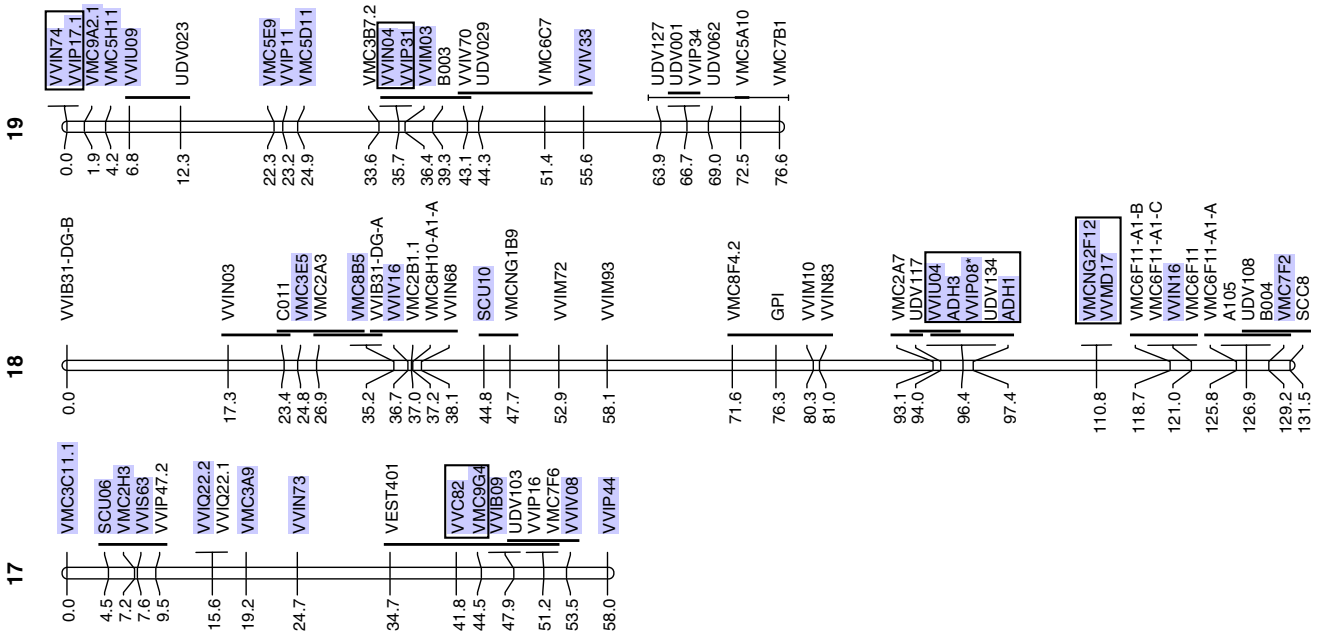


Fig. 2 continued

frequent in the low-copy fraction of the genomes (Morgante et al. 2002) and although it has been shown that some plant genomes exhibit gene-rich and gene-poor regions (Barakat et al. 1999), no similar trend has been reported to date on grape. Finally, SSRs seem more or less homogeneously scattered along the grape genetic linkage groups (Riaz et al. 2004; Adam-Blondon et al. 2004; Doligez et al. 2006). Data specific to the grape genome are obviously needed to properly address this issue.

The *in silico* PCR anchoring experiment was conducted with a larger number of 12,125 primer pairs on a set of BAC-end sequences and we tried to estimate what could reasonably be expected from such an experiment. The 67,543 BAC-end sequences added up to a total of about 45.2 Mb of sequence, which roughly corresponds to 0.1 genome equivalents. Assuming an even distribution of genes along the genome and an even sampling of our sequences, 12,464 tested sequences with a 0.1× coverage would yield 1,246 anchored genes. The number of anchored genes was actually 638, which is around 50% of the predicted value. The difference could easily be explained by (1) the non-random sampling of the BAC-end sequences, since they are paired because corresponding to both ends of a same BAC and (2) the short size of the sequences (671 bp in average) that may leave out a whole set of genes. The ePCR approach, although the number of BAC screened was twice the number of BACs screened in the 3D pool screening approach, was seven times less efficient. This can be explained by the fact that only the two ends of these BACs are screened by ePCR, versus the whole sequence by PCR. However, this strategy of *in silico* anchoring was almost immediate as compared to PCR anchoring and it allowed us to anchor more ESTs in a single experiment than in the previous months of lab work. Finally, with this strategy, the position of the gene on the BAC (at one extremity) is known, which can be helpful for physical mapping purposes.

Building a draft physical map, linked to the integrated grapevine genetic map

We finally obtained a dataset of 2,681 BAC clones anchored with unique coordinates with 1,398 markers. This allowed us to assemble 1,731 anchored islands, divided into 210 assigned islands and 1,521 unassigned islands. In order to estimate the expected number of islands for our system, we used calculations from Ewens et al. (1991) with our parameters: formula [7] with $G = 475$ Mb (Lodhi and Reisch 1995), $M = 1{,}398$, $N = 18{,}432$ and $L = 145$ kb. The expected number of islands was then 982, which is less than the observed value of 1,731. This gap was probably because (1) we disregarded multiple coordinates obtained on 3D pools to avoid false positives; (2) in the *in silico* anchoring step, the BAC-end sequences we used cannot replace the whole BAC sequences; and (3) we scored both alleles and duplicate loci the same way when we detected

Table 3 Co-localization of anchored and genetically mapped markers

LG	Contig number	Genetic marker	EST-based marker name	gb accession number	NCBI build #11 cluster number	Predicted function[a]
1	glg35	VMC9D3	TT264D11	BQ794361	Vvi.2154	Xyloglucan endotransglycosylase XET2
	glg36	VMC4F8	GB007G08	BQ798743		Unknown
			RT051E08	BQ796784	Vvi.7284	Weakly similar to a ribulose biphosphate carboxylase/oxygenase, chloroplast precursor
	glg37	VVIF52	RT074B06	BQ797506	Vvi.7912	Weakly similar to a nodulin MtN3 family protein
	glg38	VVIB94	GB004H07	BQ798508	Vvi.309	Unknown
	glg39	VVIQ57	CF210948	CF210948	Vvi.9603	Unknown
			CF373321	CF373321	Vvi.4650	Unknown
			RB005B09	BQ795359	Vvi.523	Weakly similar to xyloglucan: xyloglucosyl transferase
	glg40	VMC8E8	CB971776	CB971776	Vvi.12913	Weakly similar to polyprotein
	glg41	VMC4D2	RT064C10	BQ797290	Vvi.1073	Moderately similar to cell elongation protein/DWARF1/DIMINUTO (DIM)
2	glg42	VVIB23	CB969297	CB969297	Vvi.6988	Moderately similar to beta-fructosidase (BFRUCT4)/beta-fructofuranosidase/ invertase, vacuolar
			RT021C10	BQ795806		Multi-copper oxidase-related protein
	glg43	VVIB01	GT192H04	BQ799634	Vvi.242	Moderately similar to ribosomal protein L11 family protein
3	glg44	VMC8F10	RT043C07	BQ796586	Vvi.7002	Moderately similar to isoflavone reductase, putative
			ST007F11	BQ793223	Vvi.6697	Moderately similar to CBL-interacting protein kinase 6
	glg89	VVIB59	CM001E06	BQ792463	Vvi.1772	Weakly similar to a isoflavone reductase
			GB000A64	BQ798137	Vvi.7810	Calmodulin
			RT074F08	BQ797535	Vvi.761	Moderately similar to a putative ubiquitin-conjugating enzyme
4	glg46	VrZAG83	RT083C04	BQ797635	Vvi.7564	Weakly similar to thaumatin, putative
	glg47	VNG1F1.1	CB915696	CB915696	Vvi.28	Weakly similar to serine hydroxymethyltransferase 2 (mitochondrial)
	glg48	VVADH2	VVADHR			Alcohol dehydrogenase (pseudogene)
5	glg49	VrZAG79	CF514168	CF514168	Vvi.11876	Weakly similar to DNAJ heat shock N-terminal domain-containing protein
6	glg4	VMC2G2, VMC2H9, VMC5C5	PT013F02	BQ800568	Vvi.2021	Moderately similar to triosephosphate isomerase, cytosolic, putative
	glg50	VrZAG30	CF514542	CF514542	Vvi.11944	Unknown
7	glg30	VMC1A2, VVIB18, VVIV36	CT005D12	BQ792325	Vvi.559	Unknown
			PT007B07	BQ800315	Vvi.1994	Moderately similar to protein phosphatase 2C, putative/PP2C, putative
			RT021F11	BQ795834	Vvi.896	Weakly similar to a putative glutathione S-transferase
	glg51	VVIN56	PT003B03	BQ800139	Vvi.7795	Chlorophyll A-B binding protein CP29 like
			RT021D07	BQ795813	Vvi.7795	Chlorophyll A-B binding protein CP29 like
	glg52	VVIB22	RT022G05	BQ795933	Vvi.7134	Moderately similar to NAD-dependent epimerase/dehydratase family protein
			RT081G10	BQ797676	Vvi.1109	Weakly similar to putative beta-glucosidase
	glg55	VMC1A12	CM001H09	BQ792497	Vvi.1811	Weakly similar to BSD domain-containing protein
			CD798708	CD798708	Vvi.8419	Unknown

Table 3 (Contd.)

LG	Contig number	Genetic marker	EST-based marker name	gb accession number	NCBI build #11 cluster number	Predicted function[a]
			TT251E11	BQ794005		Alpha-6-galactosyltransferase
	glg54	VVIS58	RT052G06	BQ796768	Vvi.1000	Unknown
	glg56	VMC5H5	RT072D03	BQ797408	Vvi.6760	Weakly similar to protein phosphatase 2C, putative
8	glg1	VMC1E8	GT172A02	BQ798958	Vvi.6796	Strongly similar to 1 2-cys peroxiredoxin, chloroplast, putative
		VMC1E8, VMC3F8, VVIQ36	PT011A11	BQ800592	Vvi.13169	Ribosomal protein L10A like
		VMC1E8	RT033B12	BQ796287	Vvi.958	Moderately similar to a putative CCR4-NOT like transcription complex protein
	glg14	VVIB66	BQ792976	BQ792976	Vvi.1926	Weakly similar to a exostosin family protein
			GB007G07	BQ798742	Vvi.337	Weakly similar to a putative pyruvate kinase
			ST004B12	BQ792976	Vvi.1926	Weakly similar to a exostosin family protein
	glg57	VMC2H10	RT074G09	BQ797544	Vvi.7374	Unknown
	glg58	VMC7H2	CA808926	CA808926	Vvi.1384	Unknown
	glg59	VVIM07	CT006G06	BQ792414	Vvi.7508	Lipid transfer protein isoform 4
			VVPNLTP1	AF467946	Vvi.7508	Lipid transfer protein isoform 4
9	glg60	VMC6D12	CF415510	CF415510	Vvi.11517	Weakly similar to a transducin family protein
			GB001B11	BQ798211	Vvi.60	Unknown
	glg61	VVIQ52	CF518571	CF518571	Vvi.12350	Unknown
10	glg20	VrZAG25, VVIN85	GT192C05	BQ799558	Vvi.441	Moderately similar to a ribosomal protein L34 family protein
	glg20		RT092D12	BQ797966	Vvi.1448	Moderately similar to a magnesium-chelatase subunit chlI
	glg23	VMC4F9.1	RT023C07	BQ796085	Vvi.581	Weakly similar to ethylene-responsive element-binding factor 4
			RT043E07	BQ796599	Vvi.6624	Putative phosphate/triose-phosphate translocator
			RT094E08	BQ798086	Vvi.7878	Weakly similar to a dormancy-associated protein
	glg62	VMC3D7	CA818803	CA818803	Vvi.2563	Moderately similar to a MATE efflux family protein
11	glg11	VMC3E12, VVIM04	GT184A05	BQ799341	Vvi.30	Weakly similar to a invertase/ pectin methylesterase inhibitor family protein
			PT003H05	BQ800182	Vvi.7525	Unknown
			RT083H06	BQ797666	Vvi.459	RBX1-like protein
	glg63	VVMD25	CF371951	CF371951	Vvi.10485	Moderately similar to a ethylene-responsive family protein
			PT007B04	BQ800312	Vvi.8000	Weakly similar to a putative thioredoxin peroxidase
	glg64	VVIP02, VVIC05	RB001G09	BQ795075	Vvi.6703	Unknown
12	glg25	VVIB10, VVIV05	GB006G08	BQ798660	Vvi.160	Weakly similar to a 6-phosphogluconate dehydrogenase NAD-binding domain-containing protein
	glg69	VMC2H4	TB000A97	BQ793242	Vvi.7619	Putative ripening-related protein (grip31 gene)
	glg70	VMC8G6	GT193E05	BQ799672	Vvi.449	60S ribosomal protein L11
13	glg13	VMC3D12, VVIC51	GB009D01	BQ798799	Vvi.2399	Moderately similar to a tropinone reductase
	glg71	GT183H07	CB971776	CB971776	Vvi.12913	Polyprotein, transposable element
			GT183H07	BQ799396	Vvi.226	Unknown
			RT054H08	BQ797027	Vvi.700	60S acidic ribosomal protein P0

Table 3 (Contd.)

LG	Contig number	Genetic marker	EST-based marker name	gb accession number	NCBI build #11 cluster number	Predicted function[a]
	glg72	VVIH54	PT013A02	BQ800538	Vvi.3095	Weakly similar to a putative protein phosphatase 2C
			RT051E08	BQ796784	Vvi.7284	Unknown
			RT082H10	BQ797749	Vvi.787	Unknown
	glg73	VNG1D12	CD798080	CD798080	Vvi.8376	Unknown
14	glg7	VMCNG1E1, VVC62	CM004D01	BQ792661	Vvi.1148	Weakly similar to a dormancy/ auxin associated family protein
	glg74	VMC6C10	GB004H07	BQ798508	Vvi.309	Unknown
	glg75	VMC5B3	CB980047	CB980047	Vvi.6385	Unknown
	glg9	VMC6E1, VVIP26	GB003C04	BQ798368		Unknown
			TB007C05	BQ793740	Vvi.1979	ADP-ribosylation factor
15	glg76	VVIV67	PT003D05	BQ800154	Vvi.2044	Weakly similar to a senescence-associated protein-related
16	glg77	VMC1E11	RT071A03	BQ797375	Vvi.8137	Putative eukaryotic translation initiation factor 1A
			TT261G10	BQ794189	Vvi.1360	Moderately similar to a putative beta-amylase
17	glg15	VMC9G4, VVC15 VVC82	CA816775	CA816775	Vvi.7276	Unknown
			CF212850	CF212850	Vvi.10315	Unknown
	glg78	VVIB09	RB005G05	BQ795399	Vvi.858	Unknown
	glg81	SCU10	TT282F09	BQ794683	Vvi.3294	Weakly similar to a cryptochrome 2 apoprotein
	glg81		VVHT2a	AY663846	Vvi.550	Hexose transporter HT2
19	glg16	VVIN74, VVIP17.1	PT001G12	BQ800073		Unknown
	glg29	VVIN04, VVIP31, SCU11	CB971111	CB971111	Vvi.5423	Weakly similar to a F-box family protein
	glg82	VVIM03	GB009H06	BQ798848	Vvi.7958	Catalase
			RT052G11	BQ796859	Vvi.1672	Unknown
			RT073H08	BQ797452		Unknown
	glg84	VMC5H11	GB009D05	BQ798803	Vvi.191	Unknown
	glg86	VMC9A2.1	RT084H07	BQ797853	Vvi.7613	Moderately similar to a putative tubulin
	glg87	VMC5D11	CF511891	CF511891	Vvi.1180	Moderately similar to a auxin-responsive GH3 protein

SSR markers are named while GenBank accession numbers, NCBI build #11 cluster numbers and a predicted function are shown for ESTs

[a] Retrieved from UniGene (NCBI). Basically, there are three distinctions of similarity: "highly similar to" means > 90% in the aligned region, "moderately similar to" means 70–90% similar in the aligned region and "weakly similar to" means < 70% similar in the aligned region

the bands of different size with a single primer pair. The combination of these phenomena may have prevented the joining of a fraction of the islands. However, this lack of joining was probably a consequence of the early stages of this effort, where each new analyzed marker increases the number of islands because it creates a new island (Barillot et al. 1991).

The 210 assigned islands are especially important since they represent a link between the genetic map and the future fingerprint-based physical map (11 islands per linkage group in average) and because they allowed us to indirectly place 93 ESTs on the genetic map. This provides direct links between data of physical, functional and genetical nature, respectively, BAC clones containing regulatory sequences, expressed genes and their possible role in genetic variation.

Finally, the contigs containing two or more SSR markers (indicated by black boxes in Fig. 2) were also useful to assess the quality of the integrated genetic map of the grapevine genome, suggesting better orders for markers in some areas.

Acknowledgements This research was funded by the Génoplante grant CI2001002-SEP20 and the Institut National de la Recherche Agronomique. The authors warmly thank D. Forest and M. Tabet for technical help, Dr M-L. Martin-Magniette for help with statistics and P. Abbal, N. Terrier and A. Ageorges for providing us with their very first unigene set of EST sequences.

References

Adam-Blondon A-F, Roux C, Claux D, Butterlin G, Merdinoglu D, This P (2004) Mapping 245 SSR markers on the *Vitis vinifera* genome: a tool for grape genetics. Theor Appl Genet 109:1017–1027

Adam-Blondon A-F, Bernole A, Faes G, Lamoureux D, Pateyron S, Grando MS, Caboche M, Velasco R, Chalhoub B (2005) Construction and characterization of BAC libraries from major grapevine cultivars. Theor Appl Genet 110:1363–1371

Aradhya MK, Dangl GS, Prins BH, Boursiquot JM, Walker MA, Meredith CP, Simon CJ (2003) Genetic structure and differentiation in cultivated grape, Vitis vinifera L. Genet Res 81:179–192

Arratia R, Lander ES, Tavaré S, Waterman MS (1991) Genomic mapping by anchoring random clones : a mathematical analysis. Genomics 11:806–827

Barakat A, Tran Han D, Benslimane A-A, Rode A, Bernardi G (1999) The gene distribution in the genomes of pea, tomato and date palm. FEBS Lett 463:139–142

Barillot E, Dausset J, Cohen D (1991) Theoretical analysis of a physical mapping strategy using random single-copy landmarks. Proc Natl Acad Sci USA 88:3917–3921

Doligez A, Adam-Blondon A-F, Cipriani G, Di Gaspero G, Laucou V, Merdinoglu D, Meredith CP, Riaz S, Roux C, This P (2006) An integrated SSR map of grapevine based on five different populations. Theor Appl Genet (in press)

Ewens WJ, Bell CJ, Donnelly PJ, Dunn P, Matallana E, Ecker JR (1991) Genome mapping with anchored clones: theoretical aspects. Genomics 11:799–805

Glémet E, Codani JJ (1997) LASSAP, a large scale sequence comparison package (note: LASSAP is now Biofacet™). Comput Appl Biosc 13:137–143

Lodhi MA, Reisch BI (1995) Nuclear content of *Vitis* species, cultivars, and other genera of the Vitaceae. Theor Appl Genet 90:11–16

Luo M-C, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SnaPshot labelling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82:378–389

Merdinoglu D, Butterlin G, Bevilaqua L, Chiquet V, Adam-Blondon A-F, Decroocq S (2005) Development and characterization of a large set of microsatellite markers in grapevine (*Vitis vinifera L.*) suitable for multiplex PCR. Mol Breed 15:349–366

Morgante M, Hanafey M, Powell W (2002) Microsatellite are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Murphy K, Raj T, Winters RS, White PS (2004) me-PCR: a refined ultrafast algorithm for identifying sequence-defined genomic elements. Bioinformatics 20:588–590

Riaz S, Dangl GS, Edwards KJ, Meredith CP (2004) A microsatellite based framework linkage map of *Vitis vinifera* L. Theor Appl Genet 108:723–726

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, pp 365–386

Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, Zyprian E, Toepfer R, Grando M. S, Velasco R (2004) Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera* L.), as revealed by single nucleotide polymorphisms. Mol Breed 14:385–395

Schuler GD (1997) Sequence mapping by electronic PCR. Genome Res 7:541–550

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci USA 89:8794–8797

Siret R (2001) Etude du polymorphisme génétique de la vigne cultivée (*Vitis vinifera* L.) à l'aide de marqueurs microsatellites : application à la caractérisation des cépages dans les vins. Thèse de l'Université de Montpellier I, 11 janvier 2001, 148p

Soderlund C, Dunham I (1995) SAM: a system for iteratively building marker maps. Comput Appl Biosci 11:645–655

Terrier N, Glissant D, Grimplet J, Barrieu F, Abbal P, Couture C, Ageorges A, Atanassova R, Léon C, Renaudin J-P, Dedaldechamp F, Romieu C, Delrot S, Hamdi S (2005) Isogene specific oligo arrays reveal multifaceted changes in gene expression during grape berry (*Vitis vinifera* L.) development. Planta 6:1–16

Thareau V, Déhais P, Serizet C, Hilson P, Rouzé P, Aubourg S (2003) Automatic design of gene specific sequences tags for genome-wide functional studies. Bioinformatics 19:2191–2198

Tomkins MR, Peterson DG, Yang TJ, Ablett ER, Henry RJ, Lee LS Holton TA, Waters D, Wing RA (2001) Grape (*Vitis vinifera* L.) BAC library construction, preliminary STC analysis, and indentification of clones associated with flavonoid and stilbene biosynthesis. Am J Enol Vitic 52:287–291