Theor Appl Genet (2008) 116:701–714
DOI 10.1007/s00122-007-0703-1

ORIGINAL PAPER

Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane

Louis-Marie Raboin · Jérôme Pauquet · Mike Butterfield · Angélique D'Hont · Jean-Christophe Glaszmann

Received: 4 May 2007 / Accepted: 13 December 2007 / Published online: 15 January 2008

Abstract Linkage disequilibrium (LD) in crops, established by domestication and early breeding, can be a valuable basis for mapping the genome. We undertook an assessment of LD in sugarcane (*Saccharum* spp), characterized by one of the most complex crop genomes, with its high ploidy level (≥8) and chromosome number (>100) as well as its interspecific origin. Using AFLP markers, we surveyed 1,537 polymorphisms among 72 modern sugarcane cultivars. We exploited information from available genetic maps to determine a relevant statistical threshold that discriminates marker associations due to linkage from other associations. LD is very common among closely linked markers and steadily decreases within a 0–30 cM window. Many instances of linked markers cannot be recognized due to the confounding effect of polyploidy. However, LD within a sample of cultivars appears as efficient as linkage analysis within a controlled progeny in terms of assigning markers to cosegregation groups. Saturating the genome coverage remains a challenge, but applying LD-based mapping within breeding programs will considerably speed up the localization of genes controlling important traits by making use of phenotypic information produced in the course of selection.

Communicated by J. E. Bradshaw.

Louis-Marie Raboin and Jérôme Pauquet contributed equally to this work.

L.-M. Raboin
CIRAD (Centre de coopération internationale en recherche agronomique pour le développement), UMR PVBMT, Saint-Pierre, 97410 Reunion, France

J. Pauquet · A. D'Hont (✉) · J.-C. Glaszmann
CIRAD, Agropolis, UMR DAP, Avenue Agropolis, 34398 Montpellier Cedex, France
e-mail: dhont@cirad.fr

M. Butterfield
SASRI (South African Sugarcane Research Institute), Mount Edgecombe, South Africa

Introduction

Linkage disequilibrium (LD), i.e., the non-random association of alleles at distinct loci, is a common method to map human disease genes (Cardon and Bell 2001). It has recently become a major focus of interest in plant genetics (Flint-Garcia et al. 2003; Gupta et al. 2005) and its extent has been documented in several species. *Arabidopsis thaliana* as an inbreeding plant species, has LD that extends on average up to 50 kb and thus appears well suited for genome-wide LD mapping (Nordborg et al. 2002, 2005). Crops have undergone domestication, which has generally involved severe bottlenecks in genetic diversity. This has established extensive LD, though at a variable level depending upon the populations under consideration. Recent studies among rice and sorghum landraces, both predominantly inbreeding crop species, point out LD extension up to the 100 kb range (Garris et al. 2003; Hamblin et al. 2005). Maize, as an outbreeding crop species, is characterized by a rapid decline of LD within a few hundred base pairs (Tenaillon et al. 2001), which makes genome wide LD mapping among broad-based landraces nearly impractical because it would necessitate hundreds to thousands of markers. Modern breeding has reinforced LD by the use of a restricted number of parents in the hybridization schemes. Even in maize, modern breeding materials can display LD over distances of several dozen centiMorgans (Stich et al. 2005).

The first instances of whole-genome scans demonstrating association between markers and traits of agricultural value have been reported in sugar beet for the bolting gene (Hansen et al. 2001), in barley for yield, treated as a complex quantitative trait (Kraakman et al. 2004), and in wheat for kernel size and milling quality (Breseghello and Sorrells 2006). This type of study is likely to gain interest among crop geneticists and breeders (Morgante and Salamini 2003; Rafalski and Morgante 2004). The possibility to apply this approach in sugarcane has already been highlighted (Jannoo et al. 1999; Wei et al. 2006).

Sugarcane is an important crop and a remarkable instance of a very efficient physiological set-up, resting on an extremely complex genome and resulting in one of the highest biomass yielding crops (Rahmani et al. 2000). Modern sugarcane cultivars are the product of breeding activities initiated at the end of the nineteenth century. They are derived from interspecific hybridization between the domesticated sugar producing species *Saccharum officinarum* ($x = 10$, $2n = 8x = 80$) and the wild species *Saccharum spontaneum* ($x = 8$, $2n = 5x-16x = 40-128$) followed by repeated backcrossing to *S. officinarum*. Since then sugarcane breeding has relied on the recurrent intercrossing of elite cultivars and clonal selection among the produced progeny. Modern sugarcane cultivars have a complex aneuploid and polyploid genome consisting of 100–130 chromosomes with a total of about 10 Gbp (D'Hont 2005). Genomic in situ hybridisation (GISH) revealed that around 70–80% of chromosomes of modern cultivars are inherited from *S. officinarum*, 10–20% inherited from *S. spontaneum* and 10–20% derived from recombinations between the two ancestral species (D'Hont et al. 1996; Piperidis and D'Hont 2001; Cuadrado et al. 2004). Chromosome assortment at meiosis displays mostly bivalents but pairing seems to be essentially polysomic, with variable ranges of preferential pairing among which a few cases of systematic pairing in some homology groups (Jannoo et al. 2004). The mapping of this genome is facilitated by a highly conserved synteny with sorghum (Dufour et al. 1997; Guimaraes et al. 1997; Ming et al. 1998), but the *S. officinarum* compartment of the genome is highly redundant and remains poorly covered by the best maps (Grivet et al. 1996; Hoarau et al. 2001; Rossi et al. 2003; Ruiz et al. 2004; Aitken et al. 2005; Reffay et al. 2005; Garcia et al. 2006; Raboin et al. 2006). The genetic control of quantitative traits typically involves numerous QTLs with small individual effects (Grivet and Arruda 2001; Ming et al. 2001, 2002a, b) because of the buffering effect of the many alleles segregating simultaneously at the same locus due to polyploidy. Only three major genes have been identified so far, two conferring resistance to brown rust and one controlling stalk colour (Daugrois et al. 1996, Asnaghi et al. 2004, Raboin et al. 2006). The number of ancestral *S. officinarum* clones involved in the genealogy of sugarcane cultivars probably did not exceed 20 among which only a few were extensively used in crossing programs. For *S. spontaneum*, the number of ancestors was even more limited (Arceneaux 1965). Sugarcane breeding is a recurrent process in which each breeding cycle takes between 10 and 15 years. Consequently, only a few generations separate modern cultivars from the first interspecific hybrids and only a few meioses created opportunity to recombine chromosomes inherited from founder sugarcane clones. Our study focussed on a sample of 72 sugarcane cultivars using the potential of the AFLP technique to efficiently cover the large polyploid genome of sugarcane with markers. Our main purpose is to provide a guideline for the practical use of LD to trace alleles of breeding value in sugarcane germplasm.

Materials and methods

Plant material

The sample studied consisted of 72 modern sugarcane "clones" (designation related to the vegetative propagation of sugarcane cultivars) from various breeding stations around the world (Table 1). It encompasses a wide array of relatedness, including clones as related as parent–descendant or full-sib (same parents), clones derived from the same breeding program in Barbados, as well as clones derived from distinct breeding programs around the world, yet resting on the same initial interspecific hybrids. "R570", a modern cultivar from Reunion that has been the focus of genetic mapping with AFLP markers, was used as a repeated control in the DNA analysis.

AFLP analysis

Genomic DNA was extracted and prepared for AFLP analysis according to Hoarau et al. (2001). AFLP analysis (Vos et al. 1995) was performed using the Invitrogen AFLP analysis system I. as recommended by the manufacturer except for slight modifications as in Hoarau et al. (2001). Nine hundred and fifty-six AFLP markers have already been mapped in cultivar "R570" (Hoarau et al. 2001; Raboin et al. 2006, http://www.tropgenedb.cirad.fr). The 72 clones were genotyped using 42 primer pairs of which 40 had been previously used for the genetic mapping of "R570". We used "R570" on the acrylamide gels as a systematic control flanking each of the other 71 samples. Therefore we were able to identify unambiguously many of the AFLP bands previously mapped. However, given the profusion of diverse bands, some R570 bands could not be

Table 1 Parentage and country of origin of the 72 sugarcane clones

Clones	Parent 1	Parent 2	Country	Clones	Parent 1	Parent 2	Country
R570	H 32/8560	R 445	Reunion	FR 84/344	KWT 56/26		FWI
B 47/258	B 39/254	B 34/104	Barbados	FR 84/387	NA 63/90		FWI
B 51/129	B 45/170	B 41/227	Barbados	FR 84/166	B 80/574		FWI
B 63/119	B 49/6	B 49/119	Barbados	H 32/8560	Co 213	POJ 2878	Hawaii
B 66/23	M 147/44	B 49/119	Barbados	H 39/3633	H 32/8560		Hawaii
B 75/524			Barbados	H 39/7028	H 32/8560		Hawaii
B 80/66	B 75/738		Barbados	H 49/5	H 41/3340	H 37/1933	Hawaii
B 80/8	B 73/348	B 74/172	Barbados	H 50/7209	H 44/3098		Hawaii
B 82/288	B 74/142	B 73/428	Barbados	H 61/1721	H 49/3533		Hawaii
B 85/356	WI 73/48	BJ 63/132	Barbados	IAC 64/257	Co 419	IAC 49/131	Brazil
B 86/406	F 146	BR 62/49	Barbados	J 59/3			Jamaica
B 86/409	BJ 74/59		Barbados	Ja 64/19	Ja 55/663	Ja 54/309	
B 86/ 839	B 66/210	CR 68/188	Barbados	LF 53/4789			Fiji
B 87/1172	WI 80/703	CR 63/100	Barbados	LF 53/4825			Fiji
B 77/84	CP 52/43	HJ 57/41	Barbados	M 202/46	Co 281	M 63/39	Mauritius
BJ 78/128			Barbados	Mex 68/P23	Mex 59/89		Mexico
BR 71/48	B 50/135	B 49/119	Barbados	MY 53/53	B 42/231	Co 453	Cuba
BR 75/48	B 63/118	B 56/95	Barbados	MY 55/14	CP 34/79	B 45/181	Cuba
BR 79/4	L 60/14		Barbados	N 12	NCo 376	Co 331	South Africa
BT 72/344			Barbados	N 17	NCo 376	CB 38/22	South Africa
C 227/59	EK 2	POJ 2878	Cuba	NA 56/62	Co 290	CP 43/74	Argentina
CB 56/171	POJ 2961		Brazil	NCo 310	Co 421	Co 312	India
Co 1157	Co 419		India	NCo 376	Co 421	Co 312	India
Co 1177	Co 677	POJ 2961	India	Phil 56/226	POJ 2878	CP 36/105	Philippines
Co 1186	Co 312	Co 617	India	Phil 66/7	Phil 54/60	Co 440	Philippines
Co 1208	Co 312	CoL 9	India	PR 61/632	S 56/287	M 336	Puerto Rico
Co 449	POJ 2878	Co 331	India	Q 84	TROJAN	Co 475	Australia
Co 462	Co 421	Co 313	India	R 526	POJ 2878	R 397	Reunion
Co 842	Co 464	Co 617	India	R 574	H 39/3633	R 567	Reunion
CP 61/37	CP 48/103	CP 55/38	USA	RB 70/96	CB 36/14		Brazil
CP 66/315	CP 52/68	CP 53/17	USA	RB 72/5828	NA 56/79		Brazil
CP 70/1133	CP 56/63	67 P 6	USA	SP 70/1005			Brazil
D 172			Guyana	SP 70/1284	CB 41/76		Brazil
DB 73/419	B 67/128	B 63/118	Barbados	SP 70/3225			Brazil
DB 80/104	B 73/405	WI 73/14	Barbados	SP 70/1423			Brazil
F 160	NCo 310	F 141	Taiwan	SP 71/6113	Co 775		Brazil

scored due to the risk of overlaps and confusion. No repetition was included. Our experience with AFLPs in sugarcane (Hoarau et al. 2001; Asnaghi et al. 2004; Raboin et al. 2006) suggests that genotyping errors, i.e., the frequency of bands that would be scored differently between two repetitions of the same genotype is in the 1% range. With the importance of surveying numerous markers in many cultivars for assessing LD, it is more efficient at this stage to multiply primer pairs or cultivars, and then to double-check critical data for specific purposes (such as fine mapping), than to repeat initial analyses.

Genetic structure of the sample of materials

Associations between unlinked markers or between phenotypes and markers at non-causative genome regions can arise because of population structure (Pritchard et al. 2000). The population homogeneity should therefore be checked before assessing and using LD. No simple method exists for testing structure in cases involving dominant markers in a highly polyploid background. We calculated genetic dissimilarities between all pairwise combinations of clones using the Dice index (Nei and Li 1979) according to the following formula:

$$D_{ij} = b + c/[2a + (b + c)],$$

where D_{ij} is the measure of the genetic dissimilarity between sugarcane cultivars i and j, b is the number of bands present in i and absent in j, c is the number of bands present in j and absent in i and a is the number of bands present in i and j. The matrix of pairwise dissimilarities was then used to build a Neighbour Joining (N-J) tree using the Darwin software (Perrier et al. 2003). This analysis was performed using all the 1,537 polymorphic markers.

The existence of a structure was also assessed using the software Structure (Pritchard et al. 2000) modified to be applied on non-diploid organisms (Pritchard and Wen 2003) and testing the relative likelihoods of having more than one group. This analysis was performed using a subset of 106 independent markers chosen according to their position on distinct chromosomes in the R570 AFLP map (Hoarau et al. 2001). We tested the no admixture model (individuals are discretely from one population or another) recommended for dominant loci. We took the option of independent allele frequencies and a length of Burnin period and MCMC (Monte Carlo Markov Chain) of 10,000 and 100,000, respectively, and ran the analysis ten times for each hypothesis of K varying from 1 to 10.

Testing for significant associations between markers

Sugarcane is highly polyploid. Each basic chromosome is represented by around 10 (a medium number) homologous chromosomes representing a homology group (HG). Each sugarcane cultivar marker pattern displays all markers present on the around 10 homologous chromosomes. All the classical measures of LD (D', r^2, d^2) are related to the standard χ^2 statistics for a 2×2 contingency table (Nordborg and Tavare 2002) but exploit allele frequency or haplotype frequency. Because of the high polyploidy of sugarcane and because of the dominant nature of the markers used, allele frequency or haplotype frequency cannot be calculated. We therefore used the Fisher exact probability to test for associations between markers.

For each pair of markers a 2×2 contingency table (presence versus absence) was established and the Fisher probability was computed (Mehta and Patel 1983; SAS Institute Inc. 1990). The Fisher probability is calculated according to the following formula:

$$P = \Sigma_x p \text{ with } p = n_{1.}!n_{2.}!n_{.1}!n_{.2}!/(n_{..}!n_{11}!n_{12}!n_{21}!n_{22}!),$$

for a given table where x is the set of tables with p less than or equal to the probability of the observed table. n_{ij} is the cell frequency and $n_{i.}$ or $n_{j.}$ are marginal frequencies corresponding to the ith row and the jth column from the 2×2 contingency table.

In order to define a threshold that will help distinguish between "true" linkage-related associations (i.e., associations between markers that are genetically linked on the same chromosome segment and therefore belong to the same haplotype) and other associations between unlinked markers, that we term "spurious" (be they a fortuitous result of the large number of pairwise comparisons performed, due to the existence of undetected structure, or due to complex interactions within the genome such as selection with epistasis), we used two methods. The first consisted of applying the Bonferroni procedure by dividing the significance threshold of 5% (type I error) by the number of comparisons performed. The statistical threshold set using the Bonferroni procedure is generally extremely conservative and the risk of failing to detect "true" associations is rather high (type II error). We therefore also worked out an empirical approach using our knowledge of the meiosis (Jannoo et al. 2004) and the genetic map of "R570" (Hoarau et al. 2001; Rossi et al. 2003; Raboin et al. 2006). The chromosome pairing behaviour at meiosis suggests predominantly polysomic inheritance; thus very few cases of negative (repulsion phase) associations in linkage disequilibrium are expected between markers. Using the subset of markers present in "R570" for which the position on a cosegregation group (CG) as well as assignation to a homology group (HG) is available enabled us to define three classes of associations:

- within-cosegregation group associations (wCG); this class is the one that comprises most "true" LD cases,
- between-cosegregation group but within-homology group associations (wHG); this class may also comprise a limited number of "true" LD instances;
- between-homology group associations (bHG); this class can comprise only "spurious" associations.

The empirical Fisher probability threshold for significance was set so that bHG associations, which do not relate to genetic linkage, would represent only a very low fraction of the total number of retained "significant" associations (e.g., below 5%). The objective was to evaluate the appropriateness of the calculated Bonferroni threshold by comparison with this empirical threshold.

Interpretation framework

Use of dominant markers in a polyploid context

The genetic system we are investigating is very complex. A range of developments for addressing genetics in autopolyploids has been described by Gallais (2003). Here we draw a minimal interpretation framework to identify critical parameters for planning future experiments in sugarcane. It

requires taking several specificities into account, and translating them into simple considerations as follows. Each basic chromosome is represented by around 10 (a medium number) homologous chromosomes in each cultivar. The basic genomes of the two ancestral species are thought to be generally colinear, with only a limited number of major translocations which do not prevent interspecific intra-chromosomal recombination (Grivet et al. 1996; Guimaraes et al. 1997; Ming et al. 1998). For a given HG, a set of 7–900 chromosomes were involved in our results (based on 72 cultivars approximately decaploid), depending on the ploidy level of the HG. This set was sampled from up to 18,000 chromosomes (omitting multiple parents) present in the parents in the previous generation, after usually two recombinations (one per chromosome arm) in a predominantly polysomic pairing system. For each cultivar, the set of homologous chromosomes in a given HG can be considered the result of ten random draws. In these conditions, a marker with an allelic frequency of 0.5, i.e., present on half the homologous chromosomes, will be present on virtually all cultivars ($1-(1-0.5)^{10} > 99.9\%$). For allelic frequencies of 0.01, 0.05, 0.10, 0.20, the expected apparent frequencies among the cultivars are 0.10, 0.40, 0.65 and 0.89, respectively.

Expectations from sugarcane breeding history

The most ancient breeding programmes started a century ago, little after it was realised that sugarcane could be crossed to generate seedlings. There is a pedigree available for most modern cultivars, since the earliest interspecific crosses. Its accuracy is yet uncertain for the male parents, especially in the first few generations, before male fertility in these materials of interspecific origin was properly assessed and before reliable pollen sterilization methods were available. However, the female lineages and the numbers of generations are reliable. Analysis of pedigree records suggests that most cultivars are derived from 19 *S. officinarum* parents and two main *S. spontaneum* parents (Arceneaux 1965). The "oldest" chromosomes may have undergone around seven meioses since their use as founders of the pool of modern breeding materials, whereas those incorporated later for broadening the genetic base may have undergone 2–5 meioses. Moreover, the recurrent use of successful parents results in the fact that some chromosomes today relate to a common ancestor through uneven numbers of meioses. To provide a general framework for our study, we estimated the level of LD expected in the simplest situation: a founder bi-locus haplotype bearing two unique markers (+ ∼ +) on the same chromosome with a recombination frequency r at each generation, assuming a stable marker frequency f across generations (no selection and no drift) and a random "chromosome mating" (predominant polysomy in a high polyploid with low inbreeding). In such conditions, it is possible to predict the typical fate of the frequency of the (+ ∼ +) haplotype and of the frequency of cultivars displaying both markers ([+ ∼ +] phenotype) along generations.

At each generation (g), considering $f_g(+ \sim +)$, $f_g(+ \sim -)$, $f_g(- \sim +)$ and $f_g(- \sim -)$ as the frequencies of the haplotype (+ ∼ +), (+ ∼ −), (− ∼ +) and (− ∼ −) respectively, we have the relationships: frequency of the allele + of the first marker $f_g(+ \sim ?) = f_g(+ \sim +) + f_g(+ \sim -) = f$ and; frequency of allele + of the second marker $f_g(? \sim +) = f_g(+ \sim +) + f_g(- \sim +) = f$. Therefore, the frequencies of the various haplotypes in the whole population of chromosomes follow simple relations (1):

$$
\begin{aligned}
f_g(+\sim+) &= H_g, \qquad (1)\\
f_g(+\sim-) &= f - H_g,\\
f_g(-\sim+) &= f - H_g,\\
f_g(-\sim-) &= 1 - f_g(+\sim+) - f_g(+\sim-) - f_g(-\sim+)\\
&= 1 - 2f + H_g.
\end{aligned}
$$

The haplotypes produced at the next generation may have two origins: they may reproduce the haplotype of the earlier generation without recombination, with probability $(1-r)$, or they may be the product or recombination, with probability r, with the markers coming from different gametes. There results:

$$H_{g+1} = (1-r).H_g + r.f^2.$$

This allows deriving the general formula:

$$H_g = f.(1-r)^g + r.f^2.\left[\Sigma_{i=1\ to\ g}(1-r)^{i-1}\right]. \qquad (2)$$

Relations (1) and (2) enable determination of the frequency of the various possible haplotypes. It is then possible to derive the expected frequencies of the marker combinations exhibited by the cultivars, considering they are the result of a random draw of ten haplotypes, irrespective of the marker doses.

$$
\begin{aligned}
&\text{Frequency of bi - maker phenotype}[-\sim-] = \text{freq}[-\sim-]\\
&\quad = (\text{freq}(-\sim-))^{10}
\end{aligned}
$$

$$
\begin{aligned}
\text{freq}[+\sim-] &= [-\sim+]\\
&= \sum_{i=1\ to\ 10} C_{10}^{i}\text{freq}(+\sim-)^{i}.\text{freq}(-\sim-)^{10-i},
\end{aligned}
$$

$$
\begin{aligned}
\text{freq}[+\sim+] &= 1 - \text{freq}[-\sim-] - \text{freq}[+\sim-]\\
&\quad - \text{freq}[-\sim+].
\end{aligned}
$$

These estimations can be translated into expected distributions among 72 individuals and a Fisher probability of independence can be determined.

Use of a reference map

The availability of the "R570" map is extremely useful to improve the exploration of LD applications. It allows ordering a subset of markers according to genetic linkage, expected to be the main cause of LD. However, this map describes chromosomes that have their own history, whose latest recombination steps are specific to this genotype. Therefore, the projection of the associations on the map helps to understand and visualize LD patterns, but we must expect (1) breaks in the projected association pattern, corresponding to recombination points directly upstream "R570", as well as (2) association between homologous cosegregation groups of the R570 map, corresponding to the haplotypes that have just been broken by these recombinations.

Results

AFLP diversity and population structure

A total of 1537 polymorphic markers were scored from 42 AFLP primer combinations analysed on the 72 sugarcane cultivars. The overall proportion of missing data was 3.5%. The control "R570" displays 807 of the 1,537 markers scored. Among these markers, 463 were previously located on the "R570" AFLP genetic map. The marker frequencies are evenly distributed within a range of 1.4 to 98.6% with a mean of 45.3%. The subset of 463 markers mapped in "R570" display an average frequency of 54% very close to the average frequency of all markers. Among the markers whose species origin had been determined earlier, those derived from *S. spontaneum* have a global frequency distribution that is more favourable to LD detection, with an average frequency of 42%, compared to 60% for *S. officinarum*. Many of the markers of *S. officinarum* origin are highly frequent and therefore could become statistically invisible as far as LD is concerned. The dissimilarity index of Dice ranged from 0.2 to 0.47. The multivariate analysis of the global AFLP data matrix revealed no particular structure in the sample, as illustrated by a star-like NJ tree (Fig. 1). Two clones, LF 53/4789 and LF 53/4825, whose genealogical records are unknown, slightly distinguished themselves from the other clones. The most closely related clones (parent–descendant or full-sib) yield dissimilarities among the lowest and are located in the same area of the tree. However, they do not form outstanding branches. The clones derived from the breeding program in Barbados exhibit a wide distribution, concentrated in the lower left portion of the tree in Fig. 1, but with instances dispersed in the rest of the tree, suggesting that the high ploidy level and the tradition of germplasm exchange among breeders limited the extent of differentiation among breeding programs. The application of the Structure software yielded no indication of any particular structure (data not shown).

LD versus genetic linkage

To discriminate marker associations due to linkage from "spurious" associations, we exploited 396 markers (out of 463) of the "R570" genetic map for which we had full information about their position in their cosegregation group and their homology group (Hoarau et al. 2001; Rossi et al. 2003). The remaining 67 markers belonged to cosegregation groups not yet assigned to a homology group. A total of 78,210 2 × 2 Fisher exact tests were performed corresponding to all possible pairwise combinations between these markers. Using the Bonferroni correction, only association tests that yielded a probability lower than $P = 6.4 \times 10^{-7}$ would be considered as significant (Fig. 2a). Under these conditions, 119 bi-markers associations would be retained, involving 120 distinct markers.

The 78,210 association tests were divided between three sets: wCG (within a cosegregation group), representing 1,488 cases; wHG (within a homology group but on distinct cosegregation groups), representing 11,874 cases and bHG (between different homology groups), representing 64,848 cases. The distribution of marker pairs according to their Fisher probability within each of the sets defined above revealed that potential "true" associations (wCG) are rapidly overwhelmed by "spurious" associations detected within the bHG set (Fig. 2a). To keep the proportion of "spurious" associations (i.e., bHG) below the 5% limit, an empirical threshold of $P = 1.6 \times 10^{-5}$ had to be set. Under these conditions, 163 associations detected within the wCG set (encompassing 156 distinct markers and representing a proportion of 11% in this set) could be considered as genuine and eight associations from the bHG set (15 distinct markers/0.01%) could be considered as spurious. Nine cases of wHG associations encompassing 17 distinct markers are left out the reasoning because of their uncertain nature (genuine versus spurious). Plotting the ratio (number of "significant" bHG associations)/(number of "significant" wCG associations) as a function of the significance threshold (P value) clearly demonstrates the rapid increase of the proportion of "spurious" associations beyond the empirical threshold (Fig. 2b).

It is noteworthy that many cases exist of markers which are closely linked on the R570 map but do not exhibit LD under the empirical threshold of $P = 1.6 \times 10^{-5}$. For example, 116 such cases are recorded among the 167 cases of marker pairs separated by less than 5 cM. In many instances of undetected LD, the frequencies of both markers are either very high, or are unbalanced (Fig. 3).

Fig. 1 Neighbour-joining tree, based on the Dice dissimilarity index calculated from AFLP data (1,537 polymorphic markers), assembling the 72 sugarcane genotypes. Connectors have been drawn between closely related cultivars (parent–descendent or full-sib relationship) of known parentage

Such cases of unbalanced frequencies may correspond to the coexistence of the most frequent marker on several distinct ancestral haplotypes. This may be the case for many of the *S. officinarum* markers which are at high frequency. For cases where both markers are present in comparable frequencies, one explanation can be that they have been placed in close linkage only recently in the parentage of R570.

To estimate LD decay in relation to genetic distance, we exploited all the associations between pairs of markers genetically linked on the map of "R570" (1,484 pair-wise combinations in total). The logarithm (–) of the Fisher probability was used as a measure of LD and plotted as a function of the genetic distance between markers (Fig. 4). A clear decay can be observed up to distances of 30 cM, with a major concentration of strong LD within the first 5 cM.

LD distribution in the genome

The 163 bi-markers associations are distributed in clusters of associated markers scattered over the entire sugarcane genome. This distribution is illustrated with HG VI of the "R570" genetic map in Fig. 5. Some of those clusters span over large genetic distances, occasionally over 50 cM.

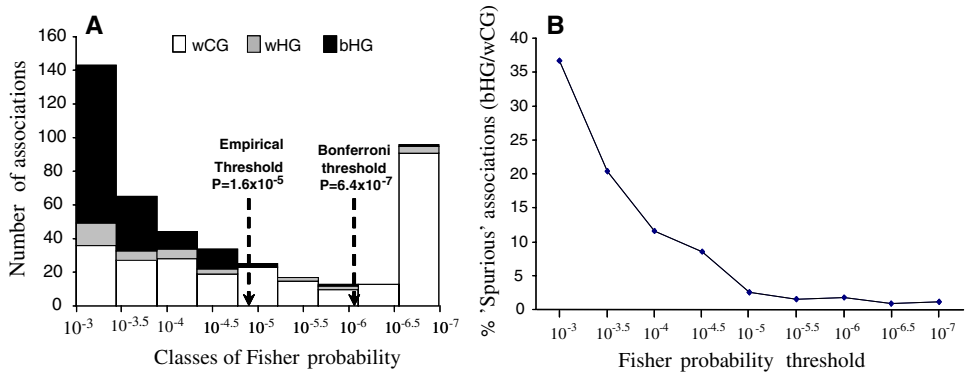

Fig. 2 **a** Distribution of pairwise associations of markers according to the Fisher probability (classes of Fisher probability are not shown for $P > 10^{-3}$). Pair-wise associations of markers are divided in three categories: wCG, wHG and bHG (see "Materials and methods" for details). **b** Percentage of "spurious" associations detected among the 72 sugarcane cultivars surveyed in relation to the Fisher probability threshold retained

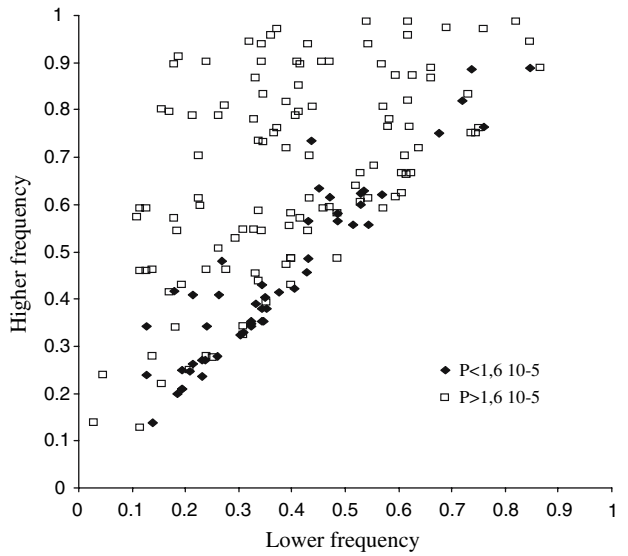

Fig. 3 LD occurence between markers separated by less than 5 cM on the R570 map. Each couple of such markers is represented by a *symbol* according to the respective marker frequencies. *Plain lozenges* identify couples of markers in LD (empirical threshold) whereas *empty squares* identify couples of markers that are not in LD

These clusters of markers in LD correspond to ancestral haplotypes segregating as large blocks in the population of sugarcane cultivars. Although the map coverage obtained in this study remains low (we could score 463 markers out of the 887 already placed on the genetic map), the results clearly indicated that LD occurs over a large range. The pattern of LD along the map of "R570" appeared heterogeneous though, which probably reveals the distinct history of the different haplotypes during modern breeding.

Among the 163 bi-markers associations retained, 51 involved pairs of markers of *S. officinarum* origin, 27 involved pairs of markers of *S. spontaneum* origin, and 22 involved markers of both origin. If we consider the distinct haplotype blocks (i.e., genomic regions encompassing groups of markers significantly associated with one another): 27 had a homogeneous *S. officinarum* constitution, 7 had a homogeneous *S spontaneum* constitution and 10 had a recombined constitution. It is noteworthy that those chromosomal regions recombined between the ancestral species correspond to regions revealing extensive LD.

LD as a mapping basis

A total of 1,180,416 2 × 2 Fisher exact tests, corresponding to all the possible pairwise comparisons between 1,537 polymorphic markers, have been performed. To achieve an overall significance threshold of 5% using the Bonferroni correction, a nominal significance threshold of $P = 4.2 \times 10^{-8}$ must be applied to each test. Using this conservative threshold, 291 associations encompassing 282 distinct markers, were considered as significant. The frequency of these markers is most often (137/282) comprised between 0.30 and 0.50, which is in agreement with the expectations. Indeed, an allelic frequency of 0.05 corresponds to a marker frequency of $1-0.95^{10} = 0.4$ among sugarcane cultivars. Still with the same threshold, a total of 96 haplotypes could be constructed by grouping all significant pairwise combinations of markers by transitivity (i.e., if marker a is associated to marker b and marker b is associated to marker c, all three markers are grouped in the same putative haplotype). These haplotypes were composed of two markers (57/96) to ten markers (Table 2). When the threshold for significance was loosened, the number and the size of the haplotypes increased. But when the threshold was too permissive, many markers were gathered together in a large haplotype containing markers that should not be associated (markers of different HGs for example). A threshold set around $P = 5 \times 10^{-6}$, corresponding to a 1% risk of spurious association (Fig. 2b), yields a distribution of 515 markers into 146 haplotypes, which looks reasonable

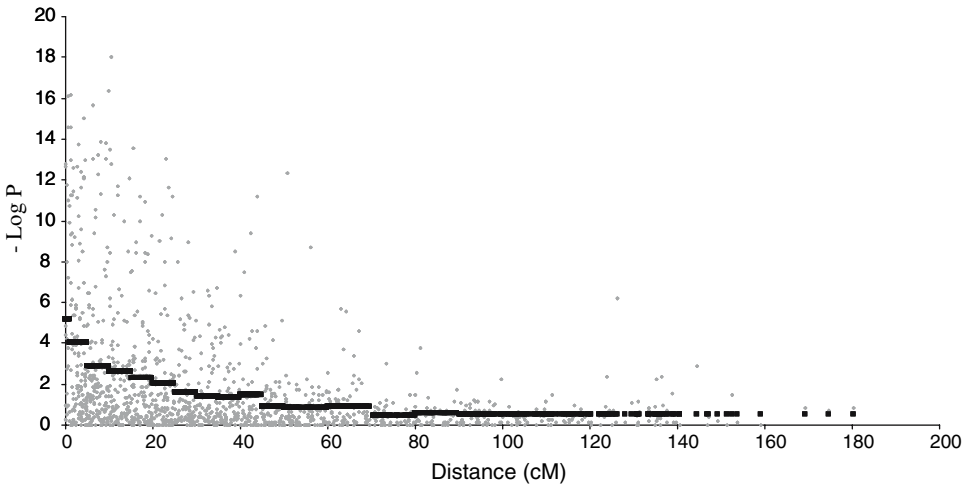

Fig. 4 LD as a function of genetic distance in centimortgan (cM). LD was measured as the Logarithm transformation of the Fisher exact test *P* value (1,484 pair-wise combinations of markers are plotted). The *bold lines* correspond to average LD values for marker pairs in 5 cM bins (5–50 cM), 10 cM bins (50–100 cM) and then over the totality of the remaining 100–200 cM. In the first 5 cM bin, the average was calculated within the first cM, and then the next four cM

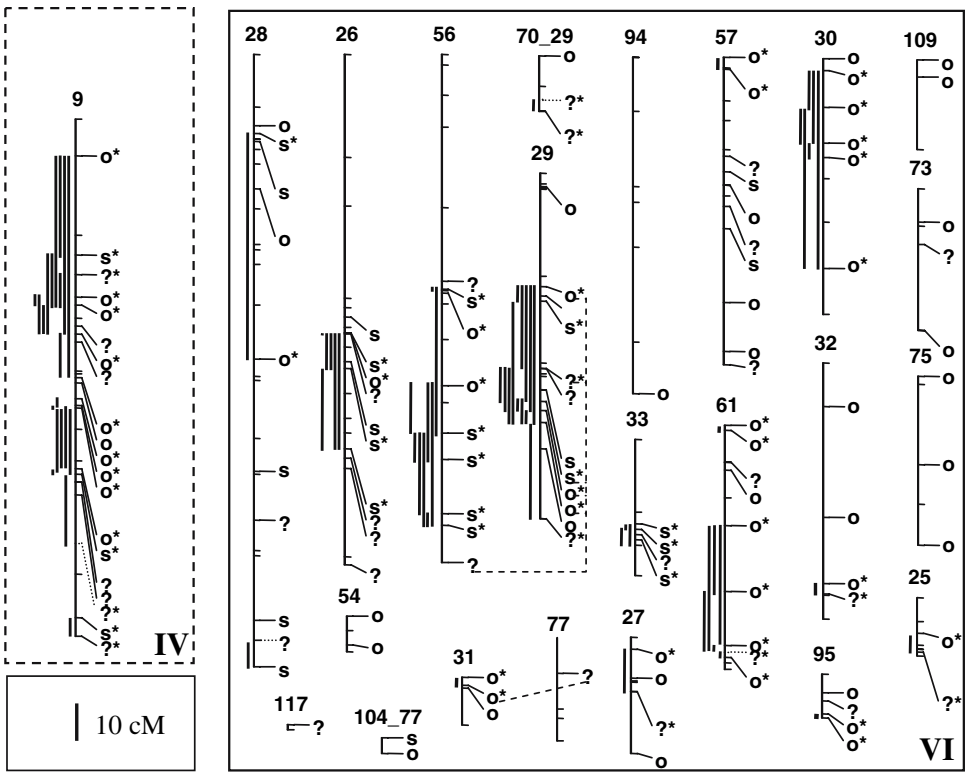

Fig. 5 A sample of the AFLP Map of "R570" (Hoarau et al. 2001, updated with Rossi et al. 2003 and Raboin et al. 2006) with a projection of significant LD as *vertical bars* near the chromosomes. Markers available in our data set have been identified with a letter code corresponding to their putative origin: *o* for *S. officinarum*, *s* for *S. spontaneum* and *?* for unknown origin; the other markers of the map, not scored in our study, are located with *unlabelled dashes*. Markers involved in significant (empirical threshold) associations in our sample of modern cultivars are marked with asterisk and significant LD is indicated by *vertical bars* near the chromosomes

in the sense that no outstandingly large haplotype is identified. At this threshold, the distribution of the markers is highly congruent with the genetic map of "R570". Considering the markers of known map location, 41 haplotypes display only markers from a same CG in "R570" map whereas three haplotypes display markers from distinct CGs from the same HG and only two display markers from distinct HGs. The rest of the cases (100 haplotypes) correspond to haplotypes for which none or only one marker has a known position on "R570" genetic map. By contrast, a threshold less stringent than 10^{-5} does induce fusion of markers from distinct CGs and HGs into a same haplotype. Therefore, our experiment with 72 cultivars and 42 AFLP combinations can be considered as enabling the mapping of over 500 markers.

Comparison with a simple simulation

The application of the simplest model assuming complete initial LD among founder chromosomes, stable marker frequency along the generations, random chromosome assortment at meiosis and absence of population structure (see "Materials and methods") enables estimation of

Table 2 Number and size (number of constituting markers) of the putative haplotypes identified at a given threshold

P	No of associations	No of markers	No of haplotypes	Number of haplotypes per class of haplotype size[a]												
				2	3	4	5	6	7	8	9	10	11	12	13	14
10^{-4}	957	747	150	92	26	12	8	3	3	2	1	–	–	–	–	1
5×10^{-5}	790	650	154	89	27	11	8	3	2	3	2	1	–	–	–	2
10^{-5}	578	533	155	92	28	8	8	4	5	2	2	1	1	1	–	–
5×10^{-6}	480	515	146	84	33	4	9	1	5	3	3	–	1	–	–	–
10^{-6}	401	382	126	76	22	8	9	2	4	1	2	1	1	–	–	–
5×10^{-7}	375	361	122	73	20	11	9	2	4	2	–	1	–	–	–	–
10^{-7}	305	293	99	57	18	11	6	3	1	2	–	1	–	–	–	–
[b]4.2×10^{-8}	291	282	96	57	16	11	5	3	1	2	–	1	–	–	–	–

P	No of associations	No of markers	No of haplotypes	Number of haplotypes per class of haplotype size[a]												
				15	16	17	18	19	21	23	24	27	37	…	88	…	298
10^{-4}	957	747	150	–	–	–	–	–	1	–	–	–	–	…	–	…	1
5×10^{-5}	790	650	154	1	–	–	1	–	–	1	1	1	–	…	1	…	–
10^{-5}	578	533	155	2	–	–	–	–	–	–	–	–	1	…	–	…	–
5×10^{-6}	480	515	146	2	–	–	–	1	–	–	–	–	–	…	–	…	–
10^{-6}	401	382	126	–	–	–	–	–	–	–	–	–	–	…	–	…	–
5×10^{-7}	375	361	122	–	–	–	–	–	–	–	–	–	–	…	–	…	–
10^{-7}	305	293	99	–	–	–	–	–	–	–	–	–	–	…	–	…	–
[b]4.2×10^{-8}	291	282	96	–	–	–	–	–	–	–	–	–	–	…	–	…	–

[a] Haplotype size = number of markers, associated by transitivity, in the same haplotype

[b] Threshold according to the Bonferroni procedure

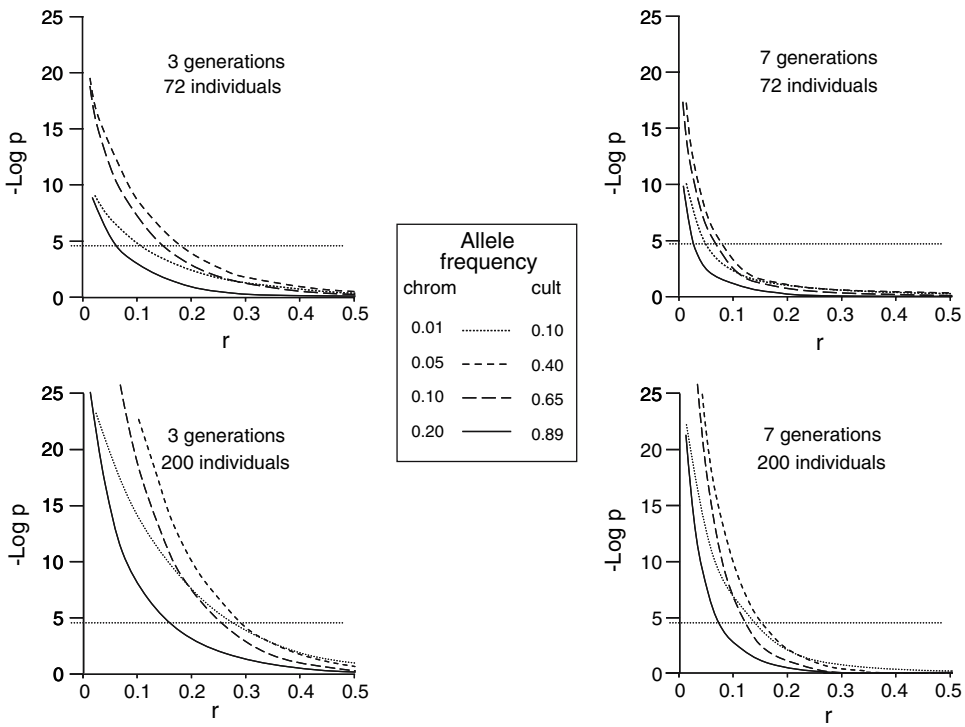

Fig. 6 Evolution of the expected LD significance (transformation minus logarithm of Fisher probability) as a function of recombination rate between the markers according to the simplified model presented in "Materials and methods". Conditions are: 3 or 7 generations of intercrossing after establishment of the founder population, 72 (as in the present study) or 200 sugarcane clones surveyed, markers with allelic frequencies on the chromosomes of 0.01 (*dotted line*), 0.05 (*line with small dashes*), 0.1 (*line with large dashes*) and 0.2 (*plain line*), corresponding to apparent marker frequencies of 0.10, 0.40, 0.65 and 0.89 in the cultivars, respectively

linkage disequilibrium detectability as a function of genetic distance. Examples are given in Fig. 6 within a range of conditions determined by the history of chromosomes (between three and seven meioses), the frequency of markers (between $f = 0.01$ and $f = 0.2$) and the sample size (number of cultivars as 72 and 200). The 1.6×10^{-5} empirical threshold approximately corresponds, using the most favourable allele frequency of 0.05, to a LD detection power close to a recombination frequency of 18% after three generations and 8% after seven generations. Compared to this expectation, our results do exhibit the majority of significant associations within 25 cM, but we frequently observe associations across larger distances, including some beyond 40 cM. Assuming a major part of the genome behaves close to predicted by our model, we can compare the expected resolution power of various combinations of experimental conditions. As an example, the resolution power is likely to be similar between the "three generations × 72 individuals" case and a "seven generations × 200 individuals" case.

Discussion

Our results highlighted a high level of LD among modern sugarcane cultivars. Significant LD has been detected between AFLP markers up to 40 cM apart. The vast majority of LD incidence occurs between 0 and 30 cM with a steady decrease when distance increases. This conclusion was reached using an empirically validated statistical analysis backed-up by a genetic map and applied to a specific set of well-established sugarcane varieties with a wide geographic origin. The conclusion is not surprising given the recent breeding history of this crop. Yet it is particularly useful by providing firm background information in the absence of fully reliable pedigree records in the first generations.

Given the extreme complexity of the sugarcane genome, we needed a simulation to specify the intuitive assumption that the breeding history of sugarcane implies a strong level of LD. The simulation we tried mimics reality with some departures whose consequences must be taken into account for further refinement. Regarding the initial state, the LD among founder chromosomes may not be total (as opposed to our theoretical example), which may make LD less visible; formulating this another way, such favourable cases of linear associations will be rare, and thus sparse along the genome, whereas triangular associations may be more frequent. It will thus be important to be able to develop many markers with the appropriate frequencies in the materials under study. Regarding the breeding process, it is likely that the repeated use of superior parents has increased the level of LD present within the population. This would affect both linked loci and independent loci. The LD between unlinked loci can be addressed by adjusting thresholds as we did using a set of reference

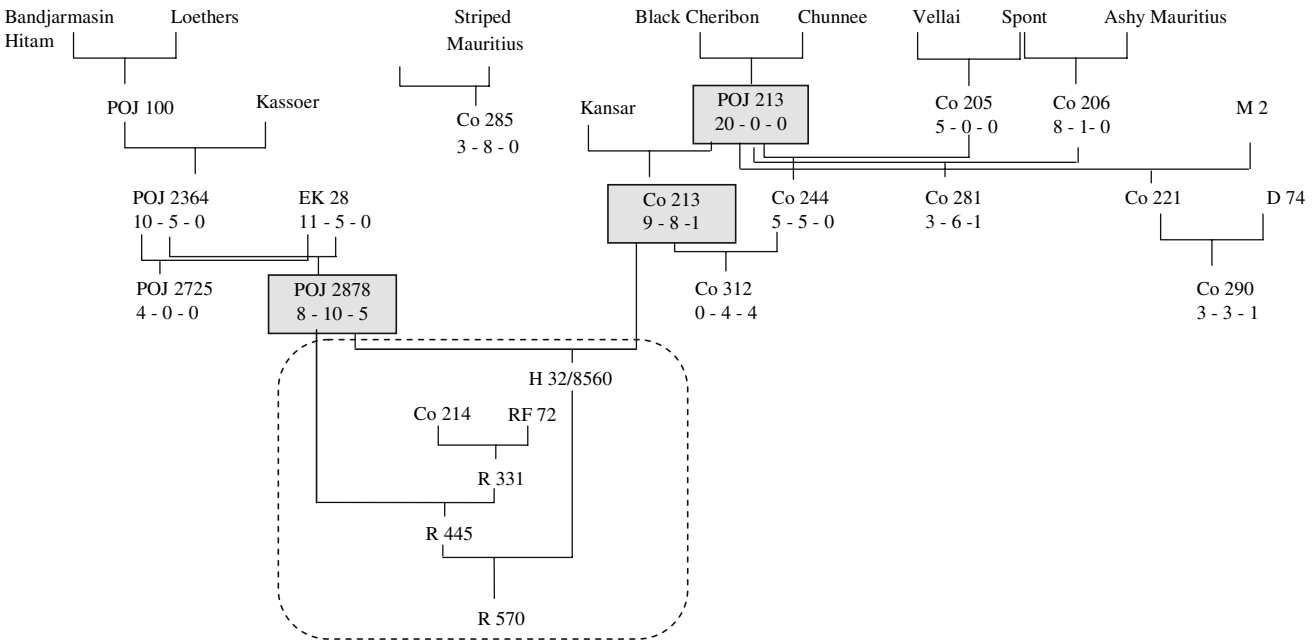

Fig. 7 Genealogy of R570 and number of occurrence of most frequent founder ancestors found in the genealogy of 30 out of the 72 clones studied for which genealogical data were available. The contribution of the parents is indicated below their names as the number of their occurrence as a great grandparent/grandparent/parent within the subset of 30 clones. The major founder ancestors are in boxes

mapped markers. However, within a given homology group, subpopulations of markers borne by haplotypes representative of frequent recurrent parents may show higher LD than others. The analysis of the genealogy (Fig. 7) of the subpopulation of 30 clones for which pedigree records were available, despite their uncertainty, highlights particular cases of frequent parents. "POJ 2878" is found five times as parent/ten times as grandparent/eight times as great grandparent in the genealogy of those cultivars. Another major pathway is revealed through cultivars "POJ 213" and "Co 213". Cultivar "R570" is connected three times to those pathways ("POJ 2878" is found twice as a grandparent and "Co 213" is found once as a grandparent of "R570"). Thus, most of the haplotype blocks visualized on "R570" map are likely to correspond to haplotypes inherited from either "Co 213" or especially "POJ 2878". The differentiation of chromosomes in terms of pairing at meiosis may also impact LD patterns. We highlighted that chromosomes derived from recombination between *S. officinarum* and *S. spontaneum* seem to bear more extensive LD. This can be related to their origin from one of the successful progenitors, as just mentioned above, or to the observation made earlier on R570 that chromosomes of hybrid origin may be less frequently involved in pairing at meiosis and thus recombine less frequently (Jannoo et al. 2004). Similarly, if there is strong affinity among some chromosomes of a particular homology group, thus forming a cluster of more related haplotypes, it may induce LD along the whole chromosome through those markers that are typical to this cluster. All the above can be used to further refine simulations. Despite the current oversimplification, observations conformed to expectations, which suggests than the rationale for the simulation is sound and can tentatively be applied or adapted to other samples of materials, for example larger samples or more advanced materials in terms of generations of recombination.

As a first circle of applications, LD can be used to map markers quite efficiently. Using 42 AFLP primer pairs on 72 cultivars enabled us to reveal more than 1,537 polymorphic markers and map close to 300 markers with high security, and more than 500 with a security that seems to us very acceptable. These figures do not match the number of bands (nearly 900) that could be mapped with a controlled progeny of 300 clones (Hoarau et al. 2001) using nearly the same 42 AFLP. However, when larger populations of cultivars are used, LD-based mapping should become as efficient as mapping in controlled progenies as far as the number of markers is concerned. One must also emphasize that the genotyping investment of a whole population of high-value selected cultivars should be much more cost-effective than genotyping the progeny of one (self-progeny) or two (bi-parental progeny) cultivars.

Further, towards breeding application, LD is likely to enable the localization of genes involved in traits of agricultural interest. One of the main advantages of LD-based mapping in a crop is that it can apply to breeding materials in the normal course of varietal improvement, for which phenotypic information is being systematically produced. It advantageously replaces ad hoc expensive field trials using controlled progenies. The main adaptation it requires from breeders is that records and DNA be taken for all materials, including those that have major faults and will be discarded in the process of selection. Using breeding materials also ensures good coverage of the whole range of adapted diversity and access to the most advanced materials, which have more generations of recombination and offer perspectives for a finer resolution of LD mapping.

Our study can help draw guidelines for future studies. Among specific features, it had the absence of structure among the materials and the back-up by an existing genetic map, which helped to establish relevant statistical thresholds. The number of markers is not expected to strongly affect the thresholds, as long as these markers are not markedly different with regards to their distribution in the genome. This type of difference will be less and less likely since more and more markers are being available and are assembled for best covering the genome. The materials can affect the thresholds essentially through their structure, their number and their advancement in pedigrees. The occurrence of a structure will be limited if the materials belong to the same breeding program. The number of materials is likely to increase if the approach attracts interest from the breeders and if high throughput genotyping becomes generalized. Their advancement in the pedigrees is likely to increase as well, if the approach is applied to materials that passed the first screening steps of the breeding programs and already underwent precise phenotypic characterization. The evolution of the detection power with larger samples or more advanced samples is illustrated in Fig. 6, showing how simulations can help determine the relevant marker density for resolutive mapping studies.

The application of LD for QTL dissection will undoubtedly require a better coverage of the genome with markers. The order of magnitude of LD in sugarcane actually resembles that described in Cattle (Farnir et al. 2000). Although extensive, LD drops sharply when markers are 5 cM or more apart. The minimum number of multi-allelic locus-specific markers required to achieve a density of one or two markers every 5 cM would lie between 300 and 600 (the size of the haploid sugarcane genome is about 1,500 cM). To date, the number of microsatellite markers characterized in sugarcane is far from sufficient to achieve an even and dense enough coverage, although more of them should be provided by

mining EST databases (Pinto et al. 2004). At the level of saturation required, dominant markers are as informative as codominant markers in terms of haplotype tagging. AFLPs have a high throughput potential, but their scoring cannot be easily automated. In this study, we used 42 primer pairs (out of the 64 commercially available) and could produce 1537 markers. This is substantial, yet it is far from sufficient to apprehend the totality of the haplotype diversity in modern sugarcane germplasm. Large numbers of SNPs are available thanks to sugarcane ESTs (http://sucest.lad.ic.unicamp.br/en/, Grivet et al. 2003); however the identification of useful SNPs in sugarcane is difficult because they must be rare in the population of chromosomes surveyed (optimally around $f = 0.05$) to appear at the required frequency in the materials. Diversity arrays Technology (DArT) has a high potential for the application of molecular genotyping to sugarcane breeding. This technology combines the throughput potential of hybridization techniques using DNA arrays, which can detect polymorphism at several hundred loci simultaneously without relying on sequence information (Wenzl et al. 2004), and the flexibility to design genome reduction methods that will favour those marker frequencies required for LD application in sugarcane.

Acknowledgments We gratefully thank J.Y. Hoarau for his helpful revision of the manuscript.

References

Aitken KS, Jackson PA, McIntyre CL (2005) A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. Theor Appl Genet 110:789–801

Arceneaux G (1965) Cultivated sugarcanes of the world and their botanical derivation. Proc Int Soc Sugar Cane Technol 12:844–854

Asnaghi C, Roques D, Ruffel S, Kaye C, Hoarau JY, Télismart H, Girard JC, Raboin LM, Risterucci AM, Grivet L, D'Hont A (2004) Targeted mapping of a sugarcane rust resistance gene (*Bru1*) using bulked segregant analysis and AFLP markers. Theor Appl Genet 108:759–764

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177

Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99

Cuadrado A, Acevedo R, de Moreno Dias la Espina S, Jouve N, de la Torre C (2004) Genome remodelling in three modern S. *officinarum* × S. *spontaneum* sugarcane cultivars. J Exp Bot 55:847–854

Daugrois JH, Grivet L, Roques D, Hoarau JY, Lombard H, Glaszmann JC, D'Hont A (1996) A putative major gene for rust resistance linked with an RFLP marker in sugarcane cultivar R570. Theor Appl Genet 92:1059–1064

D'Hont A (2005) Unravelling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet Genome Res 109:27–33

D'Hont A, Grivet L, Feldmann P, Rao RS, Berding N, Glaszmann JC (1996) Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. Mol Gen Genet 250:405–413

Dufour P, Deu M, Grivet L, D'Hont A, Paulet F, Bouet A, Lanaud C, Glaszmann JC, Hamon P (1997) Construction of a composite sorghum genome map and comparison with sugarcane, a related complex polyploid. Theor Appl Genet 94:409–418

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M (2000) Extensive genome-wide linkage disequilibrium in cattle. Genome Res 10:220–227

Flint-Garcia SA, Thornsberry JM, Buckler IV ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Gallais A (2003) Quantitative genetics and breeding methods in autopolyploid plants. INRA Editions, Paris, 515 pp

Garcia AAF, Kido EA, Meza AN, Souza HMB, Pinto LR, Pastina MM, Leite CS, da Silva JAG, Ulian EC, Figueira A, Souza AP (2006) Development of an integrated genetic map of a sugarcane (*Saccharum spp.*) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. Theor Appl Genet 112:298–314

Garris AJ, McCouch SR, Kresovich S (2003) Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). Genetics 165:759–769

Grivet L, Arruda P (2001) Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr Opin Plant Biol 5:122–127

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum* spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142:987–1000

Grivet L, Glaszmann JC, Vincentz M, da Silva F, Arruda P (2003) ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. Theor Appl Genet 106:190–197

Guimaraes CT, Sills GR, Sobral BWS (1997) Comparative mapping of Andropogoneae: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. Proc Natl Acad Sci 94:14261–14266

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485

Hamblin MT, Fernandez MGS, Casa AM, Mitchell SE, Paterson AH, Kresovich S (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171:12474–1256

Hansen M, Kraft T, Ganestam S, Säll T, Nilsson NO (2001) Linkage disequilibrium mapping of the bolting gene in sea beet using AFLP markers. Genet Res Camb 77:61–66

Hoarau JY, Offmann B, D'Hont A, Risterucci AM, Roques D, Glaszmann JC, Grivet L (2001) Genetic dissection of a modern cultivar (*Saccharum* spp.) I. genome mapping with AFLP markers. Theor Appl Genet 103:84–97

Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC (1999) Linkage desequilibrium among modern sugarcane cultivars. Theor Appl Genet 99:1053–1060

Jannoo N, Grivet L, David J, D'Hont A, Glaszmann JC (2004). Differential chromosome pairing affinities at meiosis in polyploid sugarcane revealed by molecular markers. Heredity 93:460–467

Kraakman ATW, Niks RE, van den Berg PMMM, Stam P, van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446

Mehta CR, Patel NR (1983) A network algorithm for performing Fisher's exact test in r × c contingency tables. J Am Stat Assoc 78:427–434

Ming R, Liu SC, Lin YR, Da Silva J, Wilson W, Braga D, van Deinze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of *Saccharum* and *Sorghum* chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics 150:1663–1682

Ming R, Liu SC, Moore PH, Irvine JE, Paterson AH (2001) QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane cultivars under salinity. Plant Physiol 104:521–526

Ming R, DelMonte T, Moore PH, Irvine JE, Paterson AH (2002a) Comparative analysis of QTLs affecting plant height and flowering time among closely-related diploid and polyploid genomes. Genome 45:794–803

Ming R, Wang YW, Dryer X, Moore PH, Irvine JE, Paterson AH (2002b) Molecular dissection of complex traits in autopolyploid: mapping QTLs influencing sugar yield and related traits in sugarcane. Theor Appl Genet 105:332–345

Morgante M, Salamini F (2003) From plant genomics to breeding practice. Curr Opin Biotech 14:214–219

Nei M, Li W (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:427–434

Nordborg M, Tavaré S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190–193

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J. (2005) The Pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3(7): e196. doi: 10.1371/journal.pbio.0030196

Perrier X, Flori A, Bonnot F (2003) Methods of data analysis. In: Hamon PS, Seguin M, Perrier X, Glaszmann JC (eds) Genetic diversity of cultivated tropical plants, Cirad, Montpellier, pp 31–63

Pinto LR, Oliveira KM, Ulia EC, Garcia AAF, de Souza AP (2004) Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. Genome 47:795–804

Piperidis G, D'Hont A (2001) Chromosome composition analysis of various Saccharum interspecific hybrids by genomic in situ hybridisation (GISH). Int Soc Sugar Cane Technol Congr 11:565

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure from multilocus genotype data. Genetics 155:945–959

Pritchard JK, Wen W (2003) Documentation for structure software: Version 2. http://pritch.bsd.uchicago.edu

Raboin LM, Oliveira KM, Lecunff L, Telismart H, Roques D, Butterfield M, Hoarau JY, D'Hont A (2006) Genetic mapping in the high polyploid sugarcane using a bi-parental progeny; identification of a gene controlling stalk colour and a new rust resistance gene. Theor Appl Genet 112:1382–1391

Rahmani M, Hodges AW, Kiker CF, Shiralipour A (2000) Biomass research and development in Florida: results of 20 years experience. Proceedings of the bioenergy. The nineth biennial bioenergy conference, Buffalo, 15–19 October

Rafalski A, Morgante M (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20:103–111

Reffay N, Jackson PA, Aitken KS, Hoarau JY, D'Hont A, Besse P, McIntyre CL (2005) Characterisation of genome regions incorporated from an important wild relative into Australian sugarcane. Mol Breed 15:367–381

Rossi M, Araujo PG, Paulet F, Garsmeur O, Dias VM, Chen H, van Sluys MA, D'Hont A (2003) Genomic distribution and characterization of EST-derived resistance gene analogs (RGAs) in sugarcane. Mol Gen Genet 269:406–419

Ruiz M, Rouard M, Raboin LM, Lartaud M, Lagoda P, Courtois B (2004) Tropgene-DB, a multitropical crop information system. Nucleic Acids Res 32: D364–D367

SAS Institute (1990) SAS procedures guide, version 6. 3rd edn. SAS Institute Inc, Cary

Stich B, Melchinger AE, Frish M, Maurer HP, Heckenberger M, Reif JC (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111:723–730

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci 98:9161–9166

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Wei X, Jackson PA, Mc Intyre CL, Aitken KS, Croft B (2006) Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. Theor Appl Genet 114:155–164

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci 101:9915–9920