SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids

Antonio A. F. Garcia*1, Marcelo Mallinari*1, Thiago G. Marcon1,2, Oliver R. Serang1, Renato R. Silva1, Maria L. C. Vieira1, Renato Vicentini2, Estela A. Costa2, Melina C. Mancini2, Melissa O. S. Garcia2, Maria M. Pastina1, Rodrigo Gazaffi1, Eliana R. F. Martins3, Nair Dahmer4, Danilo A. Sforça2, Claudio B. C. Silva2, Peter Bundock5, Robert J. Henry6, Glaucia M. Souza7, Marie-Anne van Sluys8, Marcos G. A. Landell9, Monalisa S. Carneiro10, Michel A. G. Vincentz2,4, Luciana R. Pinto9, Roland Vencovsky1 & Anete P. Souza2,4

1Departamento de Genética, Escola Superior de Agricultura “Luiz de Queiroz” Universidade de São Paulo, Brazil, 2Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, SP, Brazil, 3Department of Neurobiology, Harvard Medical School and Proteomics Center, Children’s Hospital Boston, USA, 4Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil, 5Department of Neurobiology, Harvard Medical School and Proteomics Center, Children’s Hospital Boston, USA, 6Department of Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil, 7Department of Neurobiology, Harvard Medical School and Proteomics Center, Children’s Hospital Boston, USA, 8Department of Botânica, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil, 9Centro de Cana, Instituto Agronômico de Campinas, Ribeirão Preto, SP, Brazil, 10Departamento de Biotecnologia de Plantas, Centro de Ciências Agrárias, Universidade Federal de São Carlos, São Carlos, SP, Brazil.

Many plant species of great economic value (e.g., potato, wheat, cotton, and sugarcane) are polyploids. Despite the essential roles of autopolyploid plants in human activities, our genetic understanding of these species is still poor. Recent progress in instrumentation and biochemical manipulation has led to the accumulation of an incredible amount of genomic data. In this study, we demonstrate for the first time a successful genetic analysis in a highly polyploid genome (sugarcane) by the quantitative analysis of single-nucleotide polymorphism (SNP) allelic dosage and the application of a new data analysis framework. This study provides a better understanding of autopolyploid genomic structure and is a sound basis for genetic studies. The proposed methods can be employed to analyse the genome of any autopolyploid and will permit the future development of high-quality genetic maps to assist in the assembly of reference genome sequences for polyploid species.

Common marker systems, such as Amplified Fragment Length Polymorphism (AFLP) and Simple Sequence Repeat (SSR), have been successfully used in the last few decades for several types of genetic studies, including diversity analysis, genetic mapping, quantitative trait locus (QTL) mapping, synteny (co-linearity) definition, co-ancestry estimation, and more. However, most of these applications have been developed in diploid plant species in which the theoretical foundation for analysis and interpretation of the results has already been established. These tools are less developed for autopolyploids, i.e., organisms that have more than two sets of chromosomes of the same type and origin. Despite the fact that great progress has been made using marker systems in autotetraploids (e.g., potato), other, more complex polyploid species, such as sugarcane, strawberry, and some forage crops, have not yet fully benefited from molecular marker information.

This is because several unrealistic and simplified assumptions need to be made. AFLP and SSR (and even RFLP) do not allow a straightforward estimation of the number of copies of each allele at a given polymorphic locus in complex polyploids (species with more than four chromosomes per homology group). For example, in sugarcane, there are approximately 22 linkage maps, and only a few of these maps include loci with high allelic doses. The scenario is similar for QTL studies. Some models have attempted to consider the effects of QTL dosage, but these models still rely on marker data that are not fully informative. In Saccharomyces cerevisiae,
microarray studies have demonstrated that gene expression and gene regulation may depend upon the ploidy level, emphasising that allele dosage should be included in marker-assisted selection or genome-wide association studies. Similar conclusions were reached by 3.

The development of modern genotyping technologies, such as the Sequenom iPLEX MassARRAY®9, Illumina GoldenGate¹⁰, and protocols such as Genotyping by Sequencing (GBS)¹¹ or RAD seq¹², allows the evaluation of single-nucleotide polymorphisms (SNP) throughout the genome. One interesting feature of these novel approaches is the possibility of evaluating the relative abundance of each allele, i.e., the allelic dosage. This significantly increases the information embodied in each locus and provides several advantages for genetic analysis, such as mapping mutants via quantitative bulked segregant analysis¹² and the possibility of estimating ploidy level for polyploids¹³,¹⁴. For complex polyploids, such as sugarcane, this is essential because each marker locus needs to be positioned in a homology group. What is remarkable is that the sugarcane homology groups have different numbers of chromosomes⁵. This makes the estimation of ploidy level for each SNP an essential step for further analysis. Furthermore, less studied polyploid species with unknown ploidy levels could directly benefit from this modern marker approach.

To illustrate one of the advantages of using SNPs for these purposes, let us assume a hypothetical population of an autoploid species having the following genotypes for a given locus: \texttt{aaaaaa}, \texttt{Aaaaaa}, \texttt{AAaaaaa}, \texttt{AAAaaa}, and so on up to \texttt{AAAAAAA}. Using the A allele as reference, these individuals are said to have between zero (nulliplex) and six copies (hexaplex) of the allele. The number of copies of the reference allele is the allele dosage. If the individuals are evaluated with a marker system, such as AFLPs or SSRs, they are scored as 0 (gel band absent) for \texttt{aaaaaa} or 1 (gel band present) for all the other individuals due to one intrinsic limitation of the method that is associated with overlapping ploidy level. Thus, a result of “1” in a binary marker system indicates the presence of at least one copy of allele A. However, if SNPs are evaluated, the scores will be \texttt{0A} : 6a, \texttt{1A} : 5a, \texttt{2A} : 4a, and so on up to \texttt{6A} : 0a (this allelic dosage notation will be used throughout this manuscript). A marker system that allows for the direct observation of all genotypes is therefore much more informative and should be preferred. Nevertheless, this raises new challenges because new statistical methods must be developed to allow for the comprehensive analysis and interpretation of data in this new scenario.

In this work, we have evaluated the use of SNPs and novel statistical methods for SNP calling and ploidy level estimation in sugarcane using high mass spectrometry-based procedures and the SuperMASSA software¹³,¹⁴. We demonstrate that it is possible to estimate the ploidy level and the dosage of SNPs, providing useful insights into the sugarcane genome interpretation. Sugarcane is an excellent test case because it is a complex polyploid with an unknown ploidy level and frequent aneuploidy¹⁵. This work will make studies on linkage and QTL mapping, association mapping, and genomic selection possible by bringing the advantages of molecular markers to complex polyploids that, with the exception of a few well-studied autotetraploids (such as potato), have poorly understood genomes. We explored two different scenarios. First, 271 SNPs generated using the Sequenom iPLEX MassARRAY technology were used to analyse a population of 180 individuals from a biparental cross between the varieties IACSP95-3018 and IACSP93-3046. Second, 1034 SNPs were analysed in a panel of 142 relevant sugarcane genotypes. The panel consisted of important commercial varieties in addition to ancestral and parental genotypes that have been frequently used in a wide spectrum of breeding programs.

**Results**

Figure 1, panels A.1, B.1, and C.1 show examples of scatter plots of genotypes in the segregating population for a selected SNP (SugSNP382). It is clear that there are three clusters of points, each corresponding to one genotype. The data are shown together with dotted lines indicating the expected angles where the individuals would be placed if the ploidy level were 8, 10, and 12. The results suggest that the ploidy level was 10 because the clouds of points deviated slightly from the lines to other ploidy levels. When observing the data from the parents (Figure 1, A.2, B.2, C.2) and considering the closest distance to the expected genotype, the deduced configurations should be \texttt{8T} : \texttt{0G × 6T} : \texttt{2G}, \texttt{10T} : \texttt{0G × 8T} : \texttt{2G}, and \texttt{12T} : \texttt{0G × 9T} : \texttt{3G}. We must note that to be consistent with the number of observed clusters (three), the expected genotype distributions in the population were set to assume that the locus had a double dosage in one parent and was nulliplex in the other. The deduced value for ploidy 12 was not consistent with the putative number of observed clusters (three) in the progeny because a triple-dosage locus would allow for four clusters in the progeny. The expected population ratios (Figure 1, A.3, B.3, C.3) were slightly different for each ploidy level, with \texttt{3 : 8 : 3}, \texttt{2 : 5 : 2}, and \texttt{5 : 12 : 5} values for octa-, deca-, and dodecaploidy, respectively. It must be emphasised that it would be extremely difficult to distinguish these levels only by inspection or even by a simple statistical test with reasonable sample sizes.

The results described above help to explain the complex scenarios involved in determining ploidy and dosage. These issues have recently been analysed using the statistical procedures included in the SuperMASSA software¹⁴. The model simultaneously considers all available information and the genetic constraints that the derived results must fulfil, i.e., the possible genotypes to be observed given the ploidy level and the parental genotypes, the ratio between allele intensities, and the expected complete polysomic segregations. This allowed the exclusion of a triple dosage for ploidy 12. Because the expected segregations are similar, the classification relies on the ratio of the alleles (indicated by dotted lines on Figure 1), and this is one of the reasons why the choice of a technology with less bias for ratios is essential. These issues have been thoroughly discussed in¹⁴. Those authors analysed how to address situations where some bias is present. In our previous experience with Sequenom and Illumina data¹³, we observed that the former experimental approach is much less likely to produce an allele ratio bias.

We present a deeper analysis of SNPs using SuperMASSA¹⁴ in Figure 2, where the statistical results for three selected SNPs are depicted. For SugSNP382 (described previously in Figure 1), the results indicate that the posterior probability of ploidy 10 is close to 1; all individuals were allocated to clusters with individual posterior probabilities no smaller than 0.6 (almost all these probability values were close to 0.9). There was also a good agreement between the observed and expected distribution of the genotypes in the biparental population. In addition, we can deduce that the parental genotypes must have been \texttt{8T} : \texttt{2G × 10T} : \texttt{0G}. The preliminary visual inspection of the scatter plot described in Figure 1 is consistent with our statistical results.

For SugSNP151 and SugSNP715, the other two SNPs shown in Figure 2, the analysis is more complicated. Although it was possible to find models with high posterior probability for ploidy levels 18 and 16, the individual posterior probabilities in both cases were all smaller than 0.6. This means that if a small naive posterior threshold of 0.65 were used, none of the individuals would be classified as having a specific genotype. This clearly shows that, as reported previously¹⁴, the posterior probability cannot be used as a single criterion to interpret the results. There were also differences between the observed and expected distributions. Although this result may not be considered reliable enough to interpret the available laboratory data, the most likely configuration for these SNPs is ploidy 18 and 16, with parental genotypes of \texttt{15G} : \texttt{3A × 12G} : \texttt{6A} and \texttt{10T} : \texttt{6C × 7T} : \texttt{9C}, respectively.

The estimates of ploidy level for the 249 SNPs evaluated in the biparental population fell between 2 and 20 (Figure 3a). An examination of three loci classified as having a ploidy of 2 (SugSNP_0004, SugSNP_0033 and SugSNP_0036) showed that these results are
clearly associated with data of poor quality. The ratios between the masses of these alleles did not follow any expected pattern and were quite different from what was observed for all other SNPs. Therefore, these SNPs were not included in the final presentation of the results (Figure 3); for the same reason, five loci with ploidy 4 were also discarded (SugSNP_0011, SugSNP_0017, SugSNP_0018, SugSNP_0048, and SugSNP_0083); note that another two SNPs with ploidy 4 (SugSNP_0008 and SugSNP_0061) are presented in Figure 3; both had a single allelic dosage in one of the parents.

The procedure to develop the SNPs must not, in principle, exclude or favor any homology group. In our analysis, only 2 out of 249 loci were classified as having a ploidy of 4 and a single dosage, but there are no reports of such ploidy levels in the sugarcane literature. We must conclude that it is unlikely that sugarcane has homology groups with four chromosomes (autotetraploid). One possible explanation is that the observed results were caused by some bias in the angles of the scatter plots. If the PCR amplification has a different efficiency for each chromosome, the ratio between the allele intensities may be slightly different from the real ratio and therefore the angles of lines in the scatter plots could be biased by these differences (please see the additional simulations examining this bias in the Supplementary Material). As explained for Figure 1, for small dosages, the differences in the expected segregations are virtually indistinguishable and rely heavily on the scattered plot angle estimation; therefore, if this bias was present, some loci may have been misclassified as autotetraploids. We applied the same reasoning when analyzing the association mapping panel; consequently, loci with an estimated ploidy of 2 or 4 were not included in the final results (Figure 3b), and of the 987...
SNPs that were initially available (after quality control), 855 were taken into account. For all other ploidy levels, the number of loci within each ploidy class suggests that our results are reliable. The ploidy levels fell between 6 and 20, showing that the number of chromosomes within the homologous groups is not constant in sugarcane, which is in agreement with previous results.  

The distribution of loci within each ploidy level and category (A, B, and C) was similar for both the biparental population (Figure 3a) and the panel of sugarcane genotypes (Figure 3b), with the exception of those loci with ploidy 20, which were more frequent in the panel. All of the category A ploidy levels seemed to be present in about the same proportions (except ploidy 4, which was likely to be a misclassification) in both scenarios (Figures 3a and 3b). For category B, there was a trend of having more loci with higher ploidy levels; this was even clearer for loci of category C, particularly for the biparental population (Figure 3a), where none of the loci had a ploidy level smaller than 12.

It is important to mention that the analysis of the 142 sugarcane genotypes within the panel (Figure 3a) was much more complicated

---

**Figure 2 | Results of statistical analysis for three selected SNPs in a biparental sugarcane population.** Each panel of three graphs correlates to one SNP. The scatter plots show the classification of each individual in a cluster (genotype), indicated with a different colour; the centre of each circle has a small grey dot, whose colour intensity indicates its posterior probability of being allocated in the cluster. Expected (in yellow) and observed distributions for the estimated ploidy level and dosage on the parents are indicated on the histograms; the same colours used on the scatter plots were considered for the observed distribution. The posterior probabilities for each ploidy evaluated in the range 2 to 20 (only even numbers) are also indicated.
because there were no parental genotypes available to guide the analysis, as there were in the biparental population. For this group, we assumed Hardy-Weinberg equilibrium and that all individuals had the same ploidy level for a given locus. Given the complexity of the sugarcane genome, this assumption may seem rather strong, but the final genetic results are consistent with a ploidy level distribution similar to that of the biparental population. Again, we observed that the number of chromosomes within homologous groups was not constant.

The number of single-dose loci for SNPs in categories A and B (Figure 4) indicates that these are more frequent for ploidy levels up to 12. At the ploidy level of 20, only 5% of the SNPs were single dose. It is remarkable that so few SNPs overall had single-dose alleles. Interestingly, if these SNPs were used to build a linkage map using the conventional approach (single-dose markers), only loci classified as single-dose loci in IACSP93-3046 and as nulliplex in IACSP95-3018 (or vice-versa) or those classified as single-dose loci in both parents would be considered. Only 76 (30.5%) SNPs would meet these criteria if all ploidy levels were considered altogether.

The results presented in Figures 3 and 4 are interesting and informative, but because they are based only on the posterior probability of a ploidy level, they need to be interpreted together with individual probabilities. Figure 5 shows that the analysis of ploidy levels 6, 8, and 10 was more reliable, as most of the loci had medians for the individual posterior probabilities in the range 0.80 to 1.00. The opposite was observed for ploidy levels 18 and 20, as almost all of the loci (both in the biparental population and the panel) had medians in the range 0.40 to 0.60. Most of the individual medians at ploidy levels 12 and 14 were between 0.60 and 0.80, whereas the individual means at ploidy level 16 was evenly distributed in the ranges of 0.60 to 0.80 and 0.40 to 0.60.

Discussion

Developing a consistent and self-contained analysis depends on being able to estimate ploidy in species that have complicated genome structures, such as sugarcane. Due to its particular domestication process that involves the unequal participation of the parental species’ genomes (Saccharum officinarum and S. spontaneum, known to have high chromosome numbers), cytogenetic studies may not be reliable under some circumstances. The approach used in this study combined mass spectrometry and the computer program SuperMASSA.

Figure 3 | Representation of the estimates of ploidy level (in bold font) for the configurations with highest posterior probabilities for the biparental population (a) and association mapping panel (b). The areas of the rectangles are proportional to the number of SNPs that have the same ploidy level, indicated within each rectangle in parenthesis. According to the posterior probabilities calculated for each even-numbered ploidy level in the range 2 to 20, each SNP was classified into one category, using the following ad hoc criteria: Category A (green), when the highest posterior probability is greater than or equal to 0.80; Category B (yellow), when no single value of the posterior probability is higher than 0.80 but the sum of the two highest ones is greater than or equal to 0.80; and Category C (red): all other cases. In parentheses: the number of SNPs within the given ploidy level and category. The total SNP number for (a) was 241, and the total SNP number for (b) was 855.

Figure 4 | Proportion of loci with a single dose in the biparental population. Loci were classified as single dosage when they had a SNP with only a single copy of one of the alleles in one parent, being a nulliplex in the other (thus segregating in a 1 : 1 fashion in the progeny), or when both parents had a single copy of the same allele (segregating in a 3 : 1 ratio). The areas of the rectangles are proportional to the number of SNPs of each ploidy level, indicated in bold font. SNPs with single doses are represented in dark blue, with the proportion of the respective ploidy in parenthesis. Only SNPs within categories A and B (see Figure 3) were considered.
markers3, and our results show that the actual portion of the genome constraints and increases the accuracy of ploidy estimation. Furthermore, the availability of parental data adds further con-
because the exact allele dosage of a locus is frequently unknown. The primary advantage of the approach used by SuperMASSA is that it makes use of the distribution of alleles in the population in addition to the relative intensities of each allele. Using both types of information is important for resolving cases in which similar relative allele intensities could be produced. For instance, tetraploid and octoploid individuals can both produce relative allele intensities of 0:4, 1:3, 2:4, 3:1, and 4:0; however, if no distinct clusters of indi-
genotypic classes in a segregating population because all clusters that have at least one copy of the allele will collapse into a single cluster (i.e., a dominant action). This also suggests that the identification of single-dose loci using AFLP is strongly biased.

What can be said about the sugarcane ploidy level? Our results suggest that the most likely ploidy levels are between 6 and 14 (Figure 5), and several lines of evidence support our findings. The genetic maps that have already been published using different sugarcane population types (e.g., biparental crosses, selfings, and others) all have recognised homo(eo)logy groups; interestingly, most homo(eo)logy groups were established with particular numbers of co-segregation groups, which also supports the mixed-ploidy nature of the sugarcane genome, consistent with the results presented here. Our estimates for ploidies 6–14 showed high (or intermediate) individual posterior probabilities. Furthermore, the proportion of loci with single dosages for these ploidy levels in the biparental popul-
ution (Figure 4) is in agreement with previous reported results (e.g.16), with the exception of ploidy 6. The proportion of loci with ploidy levels between 6 and 14 was approximately the same for loci within category A, both in the biparental population and in the genotype panel (Figure 3). This was expected because sugarcane chromosomes are approximately the same size and the markers were in principle chosen to evenly cover the genome. There is also bi-
Figure 5 | Distribution of SuperMASSA individual posterior probabilities. For each locus, the median of all individual posterior probabilities was calculated. For instance, a median of 0.80 indicated that 50% of the individual posterior probabilities were greater than 0.80. The graphs show the distribution of the medians of each SNP locus that were classified with a specific ploidy. Only loci of category A (see Figure 3) were considered in this analysis.
We speculate that this class of SNPs belongs to the subgenome (or haplotype) of *S. robustum* that persists in the sugarcane genotypes after breeding. The vast majority of *S. officinarum* clones display 2n = 80 chromosomes. The species is stated to have eight sets (or copies) of 10 chromosomes (x = 10), i.e., octoploid.

Currently, it is supposed that modern sugarcane cultivars could exhibit 2n (S. officinarum) + n (S. spontaneum) constitution; when hybrids with *S. spontaneum* are produced, the chromosomes of *S. officinarum* double their number and form pairs of homologues, and those of *S. spontaneum* pair among themselves. This point was considered in classical publications22,23. Subsequent in situ hybridisation-based studies have confirmed the basic chromosome numbers (x) in the genus *Saccharum*24 and suggested that the genomes of modern hybrids are composed of 10–20% *S. spontaneum* chromosomes, 5–17% recombinant chromosomes and the remainder composed of *S. officinarum* chromosomes25,26. Therefore, one would expect to find 8 as the most frequently estimated ploidy level, all derived from *S. officinarum*. This particular value was found in 26.7% of SNPs classified in Category A (considering only ploidy levels 6–14) in the biparental population (Figure 3a) and 10.1% SNPs used in the panel of genotypes and belonging to category A (Figure 3b). A possible explanation is that almost all genotypes analysed here were commercial varieties (mainly interspecific hybrids) with a modified chromosomal composition from the ancestors as a result of domestication.

For *S. spontaneum*, which displays a wide range of chromosome numbers (from 2n = 40 to 2n = 128), a basic chromosome number of x = 8 was suggested. The five major cytotypes with 2n = 64, 80, 96, 112, and 12827 have 8, 10, 12, 14 and 16 sets (or copies) of eight chromosomes, respectively. These are consistent with the values observed in this study. We may suppose that all these SNP-containing loci were inherited from *S. spontaneum* (maybe as haplotypes) or that they are located on the chromosomes that were identified as recombinants between the two species *S. officinarum* and *S. spontaneum*. Alternatively, when looking at ploidy level 8, all chromosomes could be inherited only from *S. officinarum*. It is also important to mention that the repeated cycles of backcrosses to *S. officinarum* applied by early breeders, combined with the double transmission phenomenon28,29, could result in high ploidy levels because the contribution of the recurrent parent will be prevalent.

Chromosomal rearrangements are reported to be a rapid response to the formation of allopolyploid genomes30; intergeneric translocations occur predominantly between homo(eo)logous chromosome2s, and homo(eo)logous shuffling and chromosome compensation maintain genome balance in re-synthesised allopolyploids30. All the rearrangements may have occurred in the early evolutionary process of modern sugarcane. Supposedly, there is a most regular ploidy level, and all variations represent chromosome rearrangements that were herein observed.

The observation of 18 or 20 copies of a SNP-containing locus does not mean that this extreme figure represents the actual ploidy level. One could suggest reasonable cytological explanations for these high numbers; for example, for at least some of these loci, we may be detecting polysomic loci as a consequence of chromosomal segment copy number due to chromosomal rearrangements. On the other hand, the presence of univalents as a result of intergeneric pairing is well documented in sugarcane varieties. One should assume that bivalent pairing is not random but rather involves the same homo (eo)logous chromosomes30; therefore, two (or more) copies of the same univalent can be inherited from ascendants and pair during meiosis. The detection of certain high-copy SNP-containing loci may be a consequence of additional non-homologous pairing. However, it is important to mention that high values of ploidy were associated with some loci that did not have a reliable classification in our study. They were also more frequent in the panel, which is more difficult to analyse. Loci with ploidy 16 fall between these two scenarios (ploidy 6–14 and 18–20).

A recent review of the quantitative genotyping of polyploids31 reported that, even when data are difficult to analyse (i.e., presenting high variance or strong allele-specific bias), the SuperMASSA software can still provide useful information to help to interpret the results and allow the evaluation of the reliability of those results. In that review, the authors evaluated the posterior probabilities of extremely high ploidies (in the range 2 to 100). It is obvious that most of these ploidies do not have biological support, but the study revealed that when the locus displays a high variance, the generated model tends to attribute a cluster to each point in the diffuse cloud, resulting in a very high estimate of ploidy level. We have not tried to adjust our models with ploidy levels above 20 due to computing-time limitations, but we have deliberately included ploidy values without biological support (2 and 4) or with weak evidence (18 and 20). The results show that this was a good strategy because the resulting individual posterior probabilities were rather small, indicating that our observations of high ploidy values (18 and 20) are likely to be explained as discussed above31.

We have also performed some simulations to better understand the SuperMASSA output under normal and extreme situations (Figures S1 to S10, Supplementary Material). We observed that the software performed well when no extreme violations of its underlying assumptions were considered (for example, skew on the expected angles of clusters and segregation distortion). However, in the presence of high segregation distortion (for example, due to preferential pairing at meiosis) or some bias in the allele ratios, the estimated ploidy could be rather high (18 or 20).

The *in situ* hybridisations also helped us interpret the SuperMASSA estimations (Figure 6). The number of observed blocks (or signals) in these hybridisations could be taken as a rough estimate of ploidy level for IACSP 93-3046 (P2 of the biparental population). For SugSNP382, which yielded good and reliable results in the ploidy analysis (Figure 3), the number of observed blocks has been 8, which is close enough to the estimate of 10 provided by SuperMASSA. It is important to mention that SuperMASSA uses segregation ratios as an important feature to estimate ploidy; this is not necessarily the same as estimating ploidy by chromosome counting. For example, a homo(eo)logy group could have 10 chromosomes: 6 from *S. officinarum* and 4 from *S. spontaneum*. If there is preferential pairing at meiosis and the polymorphism is present in the genome of *S. officinarum*, the locus will behave like a hexaploid in the segregating population; in contrast, the results from cytological studies revealed a ploidy of 10. For SugSNP715 and SugSNP151, the number of blocks was 10 for both. This is clear evidence that, as previously explained, high ploidy estimates (16 and 18) combined with small individual posterior probabilities are likely to be statistical artefacts. Moreover, SugSNP382 yielded the same estimate for the ploidy level in the biparental population and the panel of genotypes, which was not the case for the other two SNPs analysed in Figure 3.

In conclusion, the results derived from the two different scenarios presented here (a biparental population and a panel of genotypes) provide extremely useful insights. First, as expected, it is clear that the sugarcane genome is complex and that the number of chromosomes in each hom(e)o)logy group varies depending on the SNP-containing locus. Second, our results agree with previous sugarcane cytogenetic data32 and demonstrate the robustness of analysing SNP markers in autopolyploid species. Third, the ploidy level of each SNP locus was also estimated; it must be emphasised that this estimation cannot be performed with common marker systems.

In the light of our results, the ploidy of sugarcane commercial varieties (interspecific hybrids) was estimated to be in the range from 6 to 14 for each homo(eo)logy group; this has biological and statistical support. Several factors may explain the observation of estimates in the 16–20 ploidy range, a) they are actual results; b) they were caused by a combination of preferential pairing at meiosis and a lack of bivalent pairing or segregation distortion; c) there are intrinsic
difficulties in analysing loci with high ploidy and allelic dosage; or d) MassARRAY technology did not perform well for some loci, causing bias in the allele ratio and/or high variance for clusters. The results reported in the literature and our own in situ hybridisations for the three selected SNPs suggest that reason (a) is very unlikely. However, if these high estimates were actual results, further linkage studies will show that these loci with high ploidy will show evidence of linkage with other loci of the same ploidy level and, also, will not be linked with the ones in the ploidy range 6–14. It is important to mention that linkage studies based on genetic maps will require the development of new statistical approaches, such as the ones presented by 32 and 33 for autotetraploids, that would not be straightforward to use for our results. Current ideas that put strong emphasis on single-dose loci are not appropriate. Concerning point (b), this argument may be verified by further cytological information, which will help us understand the meiotic behaviour of this complex species and subsequently make modifications to the underlying assumptions in the statistical model. For explanations (c) and (d), specific procedures should be developed to optimise the methodology for dealing with complex polyploids. It is reasonable to assume that if most of these high ploidy values are true, these loci will co-segregate and will not be linked with loci with small ploidy; this will result in homo(e)logous groups for the corresponding ploidy level. It is important to perform linkage studies in the biparental population to determine if loci with high/unknown ploidy are co-segregating with others showing high posterior probability for ploidy level; then the ploidy of these loci could be indirectly inferred.

None of the other currently available approaches are suitable to investigate polyploid genome structures as comprehensively as this approach. Therefore, we anticipate that the shaping of polyploid genomes by evolutionary processes will be better understood by applying this SNP genotyping method. Considering that most of the angiosperms are polyploid 34 and recent sequenced genomes also applying this SNP genotyping method will not only in terms of new genetic understanding, statistical genetic modelling, and prediction capabilities but also in terms of understanding the biological aspects of evolutionary and domestication processes. Finally, it is interesting to note that our study unveiled the genomic structure of a complex polyploid species by exploiting the simplest manifestation of genetic variation, the SNP. This approach should provide an important tool for developing high-quality genetic maps that will assist in QTL mapping and the assembly of reference genome sequences for the large proportion of plant species that are polyploid or have duplicated chromosomal regions.

**Methods**

**Molecular and cytological analysis.** Two representative scenarios were considered: a) a progeny of 180 individuals from a sugarcane F1 biparental population derived from the cross between two commercial varieties, IACSP 95-3018 (female, named P1 along the text) x IAC93-3046 (male, named P2); and b) an association mapping panel with 142 relevant sugarcane genotypes (Table 1), representing commercial varieties and important ancestors of modern cultivars. Sugarcane genomic DNA was obtained from young leaves using standard techniques. A total of 1034 sugarcane SNPs were developed; 91 were derived from previously reported sequence data 36 (Table S1), and the remaining 943 were developed from 2908 cluster sequences with differential expression 37 that were selected from the SUCEST database 38 (Table S2). SNPs were discovered using QualitySNP software 39 with minor modifications, and primers were designed using the MassARRAY Assay Design package. All 1034 sugarcane SNPs (Tables S1 and S2) were genotyped in the association mapping panel (iPLEX GOLD chemistry, Sequenom Inc., San Diego/CA, USA) (Table 1), and 271 SNPs from these (SUCEST database, Table S1) were evaluated in the progeny of the biparental population. Due to data quality control (especially due to very low signal), the data from 22 and 47 SNPs were discarded from the biparental population and from the panel of genotypes, respectively. Therefore, for the statistical analysis, 249 and 987 SNPs were used in the biparental population and in the panel, respectively. The SNP genotyping method was based on MALDI-TOF analysis performed on a mass spectrometer platform from Sequenom Inc. Both parents from the biparental population were scored 12 times for each SNP.

The SNP assay is based on the single-base extension of locus-specific primers followed by mass spectrometry to detect polymorphisms, yielding allele-specific information 36. Assuming equal ionisation efficiency for all alleles, equal PCR amplification of alternate alleles, and equal nucleotide incorporation accuracy/equilibria, the mass intensities should be proportional to the abundances of each allele.

Three selected SNPs (SugSNP382, SugSNP151, and SugSNP715) were analysed with FISH to check their hybridisation with IACSP93-3046. Leaf genomic DNA was isolated using the DNeasy Plant Mini Kit (Qiagen) and amplified using a pFru DNA Polymerase kit (Thermo Scientific) and specific primers (Table S3). The fragments of DNA were cloned using Escherichia coli DH10b as host and pGEM-T Vector Systems (Promega) as vector. Colonies containing recombinant plasmids were identified for selection on LB agar medium supplemented with X-gal and IPTG. Recombinant plasmids were isolated using the alkaline minipreparation procedure, and the insert nucleotide sequences were determined with an ABI3500 automated DNA sequencer (Applied Biosystems). DNA Sequences were analysed with Lasergene 7 (DNASTar, Madison, WI, USA) and aligned by using the ClustalW option of the MegAlign program. The clones were used to amplify the probes for FISH using Taq DNA polymerase (Inviagen) and purified using Wizard® SV Gel and PCR Clean-Up System (Promega). Chromosome preparations were made from root tips collected from culms grown in a plastic box containing filter paper with the regular application of water. Cytological preparations were carried out as previously described 40. All probes were labelled by nick translation (Invitrogen). SugSNP715 and SugSNP151 were labelled with digoxigenin-11-dUTP (Life Technologies) and detected with Anti-DIG-rodamine; SugSNP382 was labelled with Biotin-14dATP (Roche) and detected with avidin-FITC. The procedure and conditions for FISH were previously described 40.

**Statistical analysis.** The output of Sequenom iPLEX MassARRAY technology is a scatter plot D with quantitative alleles intensities for individuals 1, 2, up to n1, n2. Because each SNP was bi-allelic, two intensities are presented, x, and y, usually

---

**Figure 6** | In situ hybridisation of IACSP93-3046 chromosomes. (A) SugSNP715 (10 blocks, arrows); (B) SugSNP151 (10 blocks; grey arrow, nucleus; white arrow, metaphase nucleus) and SugSNP382 (8 blocks; grey arrowhead, interphase nucleus; white arrowhead, metaphase nucleus).
represented in bi-dimensional scatter plots (see Figure 1 for an example of a loci with alleles $T$ and $G$). For data quality, all data points with small intensities for both alleles were removed; they were then clustered within a circle. The radius of each cluster, a graphical Bayesian method was used. The model can be described in two parts. First, a Gaussian model based on the relative dosage is used to model the probability that a given set of genotypes will occur given the population structure. Second, a multinomial distribution is used to model the probabilities that an individual with a known genotype will produce certain intensities for each allele; ideally, the relative intensities would be proportional to the relative dosages of the respective alleles. Second, a multinomial distribution is used to model the probability that an individual with a known genotype will produce certain intensities for each allele; ideally, the relative intensities would be proportional to the relative dosages of the respective alleles. Second, a multinomial distribution is used to model the probability that an individual with a known genotype will produce certain intensities for each allele; ideally, the relative intensities would be proportional to the relative dosages of the respective alleles.

Following the recommendation reported in 10, to find the maximum a posteriori (MAP) solution for the estimates of the parameters in the model, all even-numbered ploidy levels in the range of 2 to 20 were tested. The SuperMASSA min posterior report threshold was set to 0, and the values of individual posterior probability (which indicates the maximum threshold that will allow the individual to be assigned to a given genotype) were also calculated. For example, if two individuals have posterior probabilities 0.55 and 0.65 and the naive posterior report threshold is set to 0, both of them will be assigned to genotypes; changing the threshold to 0.60, only the latter will be included; with a threshold of 0.90, both will be excluded. This was shown to be important when interpreting the results of the SNP calling.
34. Masterson, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**, 421–424 (1994).
35. Bundock, P. C. *et al.* Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotech. J.* **7**, 347–354 (2009).
36. Papini-Terzi, F. S. *et al.* Sugarcane genes associated with sucrose content. *BMC Genomics* **10**, 120 (2009).
37. Vettore, A. L. *et al.* Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* **13**, 2725–2735 (2003).
38. Tang, J., Vosman, B., Voorrips, R. E., Gerard van der Linden, C. & Leunissen, J. A. M. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* **7**, 438 (2006).
39. Moraes, A. P. & Guerra, M. Cytological differentiation between the two subgenomes of the tetraploid *Emilia fosbergii* Nicolson and its relationship with *E. sonchifolia* (L.) DC. Asteraceae). *Plant Syst. Evol.* **287**, 113–118 (2010).
40. Schwartzacher, T. & Heldt-Harrison, J. S. *Practical in situ Hybridization*. (BIOS Scientific Publishers, 2000).

**Acknowledgments**
The authors wish to thank Dr Daniel Ugarte for his invaluable suggestions for elaboration of the manuscript. This work was supported by grants from Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) - Bioen Program, grant no. 2008/52197-4, and Conselho Nacional do Desenvolvimento Científico e Tecnológico (CNPq) – INCT- Bioetanol and CeProBio Project. CBCS, DAS, EAC, MCM, MM, and MMP received graduate fellowships from FAPESP. TGM and BRG received a fellowship from CNPq. ND, MOGS, and RG received post-doctoral fellowships from FAPESP. AAFG, APS, ERFM, GMS, MLCV, MAVL, MV, and RVe received research fellowships from CNPq.

**Author contributions**
R.Vi. and T.G.M. identified and developed the sugarcane SNPs. A.A.F.G., M.M., M.M.P., ORS, R.G. and R.R.S. performed the statistical analysis. MOGS and T.G.M. carried out the SNP genotyping. L.R.P. and M.G.A.L. carried out the biparental cross. L.R.P., M.G.A.L. and M.S. were responsible for the field trials and the collection of the plant material. E.A.C., C.B.C.S., G.M.S., L.R.P., M.A.V.S., M.C.M., M.S.C., M.V., P.B., R.H. and R.Ve. participated in the molecular genetic studies. E.R.F.M., N.D. and D.A.S. performed in situ hybridisation experiments. M.L.C.V. provided the cytogenetic interpretation of the results. A.A.F.G. and A.P.S. conceived the study and participated in its design and coordination. A.A.F.G., M.M. and A.P.S. drafted the manuscript. All authors read and approved the final manuscript.

**Additional information**
Supplementary information accompanies this paper at http://www.nature.com/scientificreports

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Garcia, A.A.F. *et al*. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep.* **3**, 3399; DOI:10.1038/srep03399 (2013).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0