Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it?

N. Scarcelli, C. Mariac, T. L. P. Couvreur, A. Faye, D. Richard, Francois Sabot, C. Berthouly-Salazar, Y. Vigouroux

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 434–445 doi: 10.1111/1755-0998.12462

Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it?

N. SCARCELLI,* C. MARIAC,* T. L. P. COUVREUR,*,† A. FAYE,* D. RICHARD,* F. SABOT,* C. BERTHOULY-SALAZAR*,‡ and Y. VIGOUROUX*

**UMR DIADE, IRD Montpellier, 911 avenue Agropolis, 34394 Montpellier Cedex 5, France, †Département des Sciences Biologiques, Laboratoire de Botanique Systématique et d'Ecologie, Ecole Normale Supérieure, Université de Yaoundé I, BP 047 Yaoundé, Cameroon, ‡Route des Hydrocarbures, Centre de Recherche de Bel-Air IRD/ISRA, BP 1386 – 18524 Dakar, Senegal*

Abstract

Next-generation sequencing allows access to a large quantity of genomic data. In plants, several studies used whole chloroplast genome sequences for inferring phylogeography or phylogeny. Even though the chloroplast is a haploid organelle, NGS plastome data identified a nonnegligible number of intra-individual polymorphic SNPs. Such observations could have several causes such as sequencing errors, the presence of heteroplasmy or transfer of chloroplast sequences in the nuclear and mitochondrial genomes. The occurrence of allelic diversity has practical important impacts on the identification of diversity, the analysis of the chloroplast data and beyond that, significant evolutionary questions. In this study, we show that the observed intra-individual polymorphism of chloroplast sequence data is probably the result of plastid DNA transferred into the mitochondrial and/or the nuclear genomes. We further assess nine different bioinformatics pipelines' error rates for SNP and genotypes calling using SNPs identified in Sanger sequencing. Specific pipelines are adequate to deal with this issue, optimizing both specificity and sensitivity. Our results will allow a proper use of whole chloroplast NGS sequence and will allow a better handling of NGS chloroplast sequence diversity.

Keywords: Chloroplast, GATK, intra-individual polymorphism, NGS, SAMTOOLS, SNP

Received 29 May 2015; revision received 7 August 2015; accepted 21 August 2015

Introduction

Chloroplast sequences are important molecular markers for phylogeography and phylogeny studies or to understand seed gene flow in plants (Petit & Vendramin 2007). In fact plastid data represent an important part of phylogeography data and continue to do so (Garrick *et al.* 2015). Thanks to next-generation sequencing (NGS) approaches, sequencing full plastomes for dozens or hundreds of individuals is now easily achievable either through whole-genome sequencing methods (Straub *et al.* 2012; Bock *et al.* 2014) or through targeted enrichment strategies (Mariac *et al.* 2014).

With such an increase in whole chloroplast sequences, various studies showed numerous polymorphic positions when calling SNPs (e.g. Sabir *et al.* 2014) on a single individual. Chloroplasts have a haploid genome, and consequently, SNPs are expected to be homozygous. Those results raise two fundamental questions: (i) Why do we detect polymorphic sites on single individual? and (ii) How do we account for them when undertaking bioinformatic analyses?

Correspondence: Nora Scarcelli, Fax: +33(0)4 67 41 62 22; E-mail: nora.scarcelli@ird.fr

Because isolating and extracting chloroplast DNA is long and tedious, many studies do not specifically extract chloroplast DNA. Instead, generally the total DNA extraction and consequently the next-generation sequencing library constructed contain chloroplast, mitochondrial and nuclear DNA. Consequently, different hypotheses can explain the detection of intra-individual polymorphic positions when sequencing the chloroplast genome with next-generation sequencing approaches and libraries construction based on total DNA (Fig. 1):

1 Intra-individual polymorphism results strictly from errors associated with NGS. Errors can happen at different stages: at each PCR during library construction, during the sequencing process *per se* (synthesis, signal processing) and during the SNP calling. Sequencing errors range from 0.0009% for INDELS to 0.094% for substitutions for an Illumina MiSeq (Jünemann *et al.* 2013) and 0.0014% and 0.16%, respectively, for an Illumina HiSeq 2000 (Minoche *et al.* 2011). Moreover, even

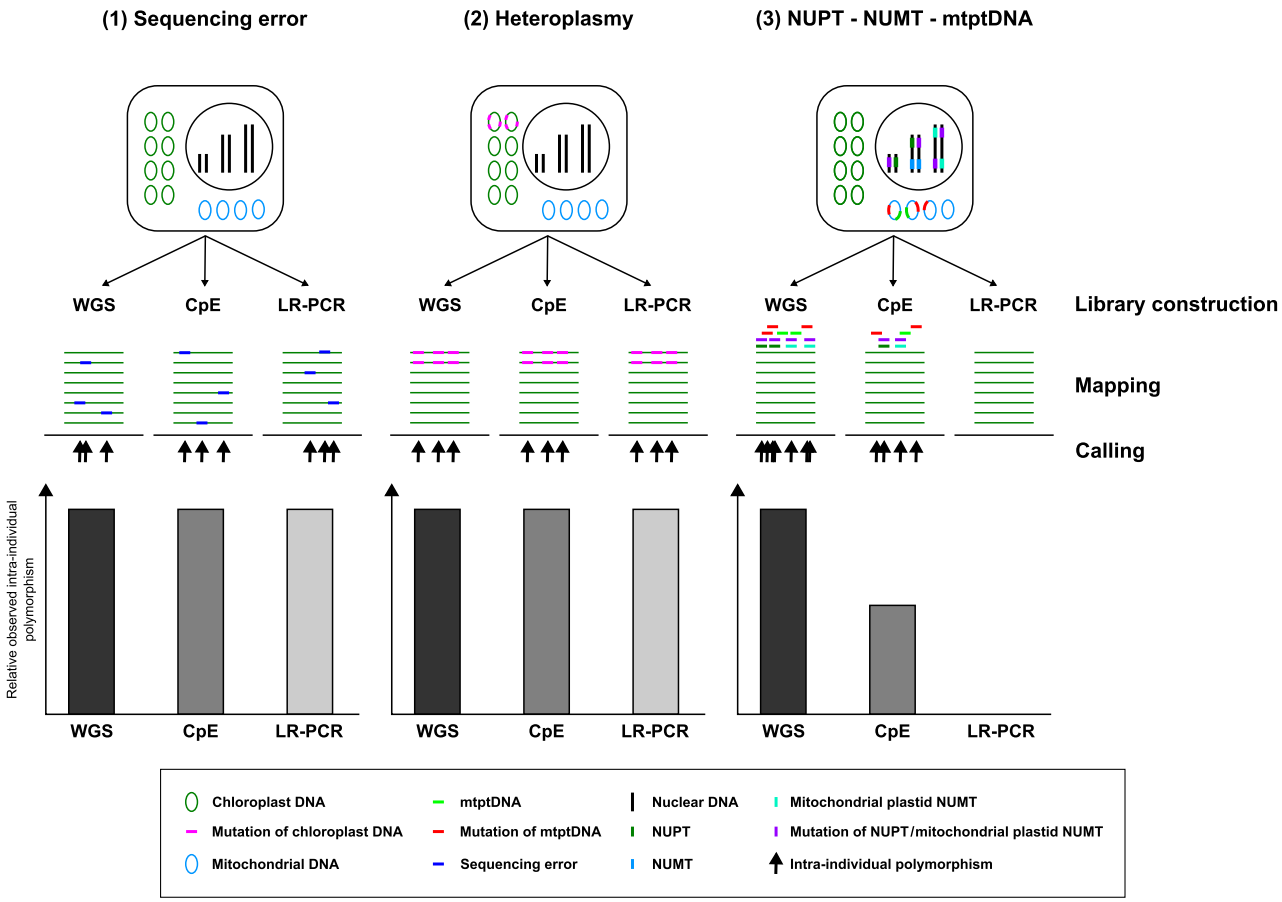

Fig. 1 Hypotheses on the origin of intra-individual polymorphic positions in plastomes. Expected number of intra-individual polymorphic positions based on three different library construction methods. Observed polymorphisms are hypothesis related and therefore cannot be compared between them. WGS = whole-genome sequencing; CpE = chloroplast enrichment; LR-PCR = long-range PCR; NUPT = nuclear plastid DNA; NUMT = nuclear mitochondrial DNA; mtptDNA = mitochondrial plastid DNA. Errors and heteroplasmy are expected to affect intra-individual polymorphism in the same way for the 3 methods. NUPT, NUMT and mtptDNA should produce many intra-individual polymorphic positions in WGS because mitochondrial and nuclear genomes are sequenced. When using CpE, the ratio chloroplast/(nuclear + mitochondrial) DNA increases, thus less NUPT, NUMT and mtptDNA are expected to be sequenced. Finally, LR-PCR amplifies very long fragments while mainly short fragments are integrated in the nuclear genome. Therefore, intra-individual polymorphism is expected to be low in LR-PCR.

if the calling errors can be handled, for example by filtering on base quality, the error rate is low but will still be significant compared to the amount of data generated (at least 0.1% according to Nielsen *et al.* 2012).

2 Intra-individual polymorphism results strictly from heteroplasmy and NGS allow it to be uncovered. Heteroplasmy is defined as the existence of nonidentical chloroplast molecules in a cell or an organism (Wolfe & Randle 2004). Considering that multiple copies of the chloroplast genome are present in each single cell (200–18 000 copies, Bendich 1987; Kumar *et al.* 2014), it is reasonable to expect a certain level of diversity. Heteroplasmy can occur because of mutations among copies of the plastome and if the persistence of that mutation is such that, it is detectable by sequencing methods. It could also occur because of biparental inheritance of chloroplasts in certain organisms. Even if chloroplasts are mostly uniparentally inherited (maternally in most angiosperms), biparental transmission of chloroplast is observed among flowering plants and can sometimes be quite frequent (Corriveau & Coleman 1988). However, unless the different chloroplast copies provide an evolutionary advantage to the plant, heteroplasmy is expected to disappear quickly by genetic drift to homoplasmy (Birky 2001; Greiner *et al.* 2015).

3 Intra-individual polymorphism results from plastid DNA transferred into the mitochondrial or the nuclear genome or both. Next-generation sequencing based on untargeted sequencing allows chloroplast DNA molecules as well as nuclear and mitochondrial DNA

derived from chloroplast sequences to be retrieved. Nuclear plastid DNA, or NUPT, is chloroplast DNA sequences transferred to the nuclear genome (Wolfe & Randle 2004; Leister 2005). They can represent a significant part of the nuclear genome from 0.01% of the nuclear genome for *Arabidopsis thaliana* to 0.27% for *Oryza sativa* (Michalovova *et al.* 2013; Yoshida *et al.* 2014). NUPT are generally small fragments, ranging from 200 bp to 600 bp, but some fragments can be very long (5000–20 000 bp; Richly & Leister 2004; Michalovova *et al.* 2013; Yoshida *et al.* 2014). Once integrated into the nuclear genome, NUPT evolve according to their length, with small NUPT being highly divergent (Richly & Leister 2004; Yoshida *et al.* 2014), potentially creating intra-individual polymorphic positions when aligned with the chloroplast genome. Mitochondrial plastid DNA, or mtptDNA, represents plastid DNA fragments integrated into the mitochondrial DNA (Wolfe & Randle 2004; Leister 2005; Bock & Timmis 2008). They represent up to 10% of the mitochondrial DNA (Wang *et al.* 2012; Zhang *et al.* 2012). Once integrated, mtptDNA evolves rapidly and neutrally, especially by deletions and by mutations biased to A and T conversion (Sloan & Wu 2014). Moreover, multiple copies of the mitochondrial genome are present in a single cell (100–3300 copies, Bendich 1987). Therefore, divergent positions can occur when aligning mitochondrial and chloroplast genomes. Finally, mitochondrial nuclear DNA, or NUMT, is mitochondrial sequences transferred to the nuclear genome. They are also very common, highly duplicated and can include mtptDNA (Leister 2005; Bock & Timmis 2008; Hazkani-Covo *et al.* 2010; Michalovova *et al.* 2013). Indeed, up to 40% of organelle DNA integrated in the nuclear genome maps indistinctly on chloroplast and mitochondrial DNAs (Yoshida *et al.* 2014). Considering 200 and 1800 plastome copies in a single cell (Bendich 1987), a polymorphism in 5% frequency corresponds to 11 and 95 copies of the alternative allele (respectively). It is therefore highly likely to observe polymorphism due to chloroplast transfers to the mitochondrial and nuclear genome.

Finally, it is also possible that intra-individual polymorphism might result from a mixture of all three previous hypotheses.

To assess these different hypotheses, several next-generation sequencing approaches from nonchloroplast specific to highly chloroplast targeted are available as follows: (i) whole-genome sequencing (WGS or genome skimming): no pretreatment occurred, plastomes and other regions (nuclear and mitochondrial) are simultaneously sequenced; (ii) hybridization capture enrichment (CpE): short plastome-specific probes are designed and used to capture through hybridization the chloroplast genome; and (iii) long-range PCR (LR-PCR): the plastome sequence is specifically amplified prior sequencing using long-range PCRs (7- to 22-kbp-long fragments).

These different approaches are expected to give different intra-individual polymorphism levels depending of the predominant hypotheses explaining the observed intra-individual polymorphism (Fig. 1):

1. If we assume that errors are a random process, they will affect the three sequencing methods in a similar way. The intra-individual polymorphism level will depend on the error rate.
2. If heteroplasmy occurs, it will affect the three methods in the same way. The intra-individual polymorphism level will depend on the frequency of the alternate alleles.
3. If intra-individual polymorphism results from plastid DNA integrated to the mitochondrial and the nuclear genome (NUPT, NUMT and mtptDNA), the level of intra-individual polymorphism is expected to be dependent on the quantity of nuclear and mitochondrial genomes sequenced. Using the WGS method, both the nuclear and mitochondrial genomes are sequenced, leading to a high level of intra-individual polymorphic positions. Using CpE, chloroplast sequences are enriched, but this approach will also lead to hybridization of chloroplast sequences integrated in different genomes. We therefore expect a lower level of intra-individual polymorphism when compared to WGS. Finally, using LR-PCR, we expect only chloroplast sequences as LR-PCR amplifies very long plastid-specific fragments, leading to very low or null intra-individual polymorphism levels.
4. If intra-individual polymorphism results from a mix of errors, heteroplasmy and mitochondrial/nuclear genome integration, the pattern observed will depend on which hypothesis is predominant.

In this study, we first compared the observed intra-individual polymorphism levels based on three different NGS approaches (whole-genome sequencing, hybridization capture enrichment and long-range PCR) to gain insight into the origin of the intra-individual polymorphism observed when sequencing chloroplast genomes. Second, we provide guidelines about how to deal with these intra-individual polymorphic sites during bioinformatic treatment. For that, we compared the results of nine different SNP calling methods between NGS and traditional Sanger sequencing as a reference. Finally, we recommend the most appropriate bioinformatics treatment to deal with chloroplast NGS data.

Materials and methods

Plant materials and sequencing

In this study, we analysed 32 different samples (Table 1): 1 rice sample (*Oryza glaberrima*), 10 pearl millet samples (*Pennisetum glaucum*), 11 yam samples (three species: *Dioscorea rotundata*, *D. abyssinica* and *D. praehensilis*) and 10 *Podococcus* samples (two species: *P. barteri* and *P. acaulis*). Total DNA (nuclear, chloroplast and mitochondrial DNA) extraction was performed using leaves as previously described (Mariac *et al.* 2006; Scarcelli *et al.* 2006).

Some of the sequences were already published (Mariac *et al.* 2014), while the remaining were generated for this study (Table 1). Full protocols are described elsewhere for NGS (Mariac *et al.* 2014) and Sanger sequencing (Scarcelli *et al.* 2011) and, for clarity, a brief explanation is provided in each case.

Whole-genome sequencing (WGS). DNA samples were sheared using a Covaris E220 (Covaris, Woburn, USA) to yield ~400-bp fragments. DNA was then repaired and tagged using 6-bp barcodes for multiplexing (Mariac *et al.* 2014). Real-time PCR was then performed to generate the libraries. Paired-end sequencing (2 × 150 bp) was performed on an Illumina MiSeq with reagent kit V2 at CIRAD, Montpellier, France.

Chloroplast enrichment (CpE). The protocol was similar to the WGS protocol, but libraries were hybridized to chloroplast-specific probes designed prior to sequencing (Mariac *et al.* 2014).

Long-range PCR (LR-PCR). Long-range PCR was performed using the LongAmp® Taq PCR Kit (New England Biolabs). A total of 11 primer pairs were used (Scarcelli *et al.* 2011; Mariac *et al.* 2014). Amplified DNA was used as template and sequencing followed the WGS protocol.

Sanger sequencing. Sanger sequencing was performed on a total of 89 yam fragments, approximately representing 50% of the total yam chloroplast genome (Scarcelli *et al.* 2011, Supporting information). Amplification was performed with Failsafe enzyme mix (Epicentre), and sequencing PCRs were performed using BigDye terminator kit (Applied Biosystems). Sequencing was performed on an ABI prism 3130 (Applied Biosystems, at INRA, Montpellier, France) using both forward and reverse primers.

All sequences are available on either GenBank (Sanger fasta sequences), NCBI-SRA (.fastq files) or DRYAD (.fastq and aligned .bam files) as listed in Supporting information.

Bioinformatics analysis

Demultiplexing, data cleaning and mapping. Demultiplexing based on the 6-bp barcodes was performed using the freely available PYTHON script DEMULADAPT (https://github.com/Maillol/demultadapt), using a 0-mismatch threshold. Adapters and low-quality bases were removed using CUTADAPT 1.2.1 (Martin 2011) with the following options: quality cut-off = 20, minimum overlap = 7 and minimum length = 35. Reads with a mean quality lower than 30 were discarded afterwards using a freely available PERL script (https://github.com/SouthGreenPlatform/arcad-hts/blob/master/scripts/arcad_hts_2_Filter_Fastq_On_Mean_Quality.pl). Mapping was performed using BWA MEM 0.7.5a-r405 (Li & Durbin 2009) with -M and -B 4 options, using the appropriate chloroplast reference genomes (rice: NC_001320.1, pearl millet: NC_024171.1, yam: NC_024170.1 and *Podococcus*: KR_347117).

Impact of sequencing methods on observed intra-individual polymorphism. We compared the level of intra-individual polymorphism according to the different sequencing methods used (WGS, chloroplast enrichment and long-range PCR) and the allele frequency of the alternate allele.

First, we only analysed the rice sample, because the three sequencing methods (WGS, CpE and LR-PCR) were performed on the exact same DNA sample (Table 1, sample TOG6208). To avoid biases in sequencing depth, we normalized the number of reads to get a similar average coverage (100×) for the CpE and the LR-PCR approaches, resulting in 104 962 mapped reads. For the WGS approach, we used all the reads available (93 223), resulting in a 88× average coverage. SAMTOOLS 1.1 with option -B (Li *et al.* 2009) was used to generate an *mpileup* file. VARSCAN v2.3.7 (Koboldt *et al.* 2012) was used to call SNPs and genotypes using this *mpileup* file. Using VARSCAN, SNP and genotype calling do not rely on any assumptions compared to the Bayesian statistics implemented in GATK (McKenna *et al.* 2010) and SAMTOOLS (Li & Durbin 2009). After filtering for a minimum quality read, VARSCAN considers a variant allele frequency threshold over all sample reads to call a variant position. Similarly, for each variant position, the genotype of a sample is called homozygote if the alternate allele does not reach a threshold fixed by the user. In our case, we fixed this option to 50% of sample reads (–min-freq-for-hom option). Using the VARSCAN option minimum variant allele frequency threshold (–min-var-freq), one can tune the alternate allele frequency and thus test its impact on the observed intra-individual polymorphism. We tested different values for this minimum frequency for 0%, 5%, 10% and 15% (methods BVar0, BVar05, BVar10 and

Table 1 Description of samples and the sequencing methods used to test (A) the impact of sequencing methods on observed intra-individual polymorphism and (B) the impact of SNP calling methods on chloroplast diversity assessment

Species	Sample name	Sequencing method	No. libraries	No. sequencing	No. fragment sequenced	Origin
(A) Impact of sequencing methods on observed intra-individual polymorphism						
Oryza glaberrima	TOG6208	WGS	1	1	–	Mariac *et al.* (2014)
		LR-PCR	1	1	–	Mariac *et al.* (2014)
		CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	PE08106-E1	WGS	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	PE11356	LR-PCR	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	18311	CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	18945	CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	19529	CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	9024	CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	PE01514	CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	PE02747	CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	PE05720	CpE	1	1	–	Mariac *et al.* (2014)
Pennisetum glaucum	PE05727	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea rotundata	CR629	WGS	1	1	–	Mariac *et al.* (2014)
Dioscorea praehensilis	P603	LR-PCR	1	1	–	Mariac *et al.* (2014)
Dioscorea rotundata	CR634	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea rotundata	CR654	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea rotundata	CR3546	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea abyssinica	A241	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea abyssinica	A466	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea abyssinica	A564	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea abyssinica	A571	CpE	1	1	–	Mariac *et al.* (2014)
Dioscorea praehensilis	P458	CpE	1	1	–	Mariac *et al.* (2014)
Podococcus barteri	Podo 5-1	WGS	1	1	–	This study
Podococcus barteri	Podo 5	LR-PCR	1	1	–	This study
Podococcus acaulis	Pa-Ndjo9	CpE	1	1	–	This study
Podococcus acaulis	Pa-Ndjo3	CpE	1	1	–	This study
Podococcus barteri	Pb-Ndjo11	CpE	1	1	–	This study
Podococcus barteri	Pb-Oyem	CpE	1	1	–	This study
Podococcus barteri	Pb-Kola5	CpE	1	1	–	This study
Podococcus acaulis	Pa-Mayoko9	CpE	1	1	–	This study
Podococcus barteri	Pb-Aloum8	CpE	1	1	–	This study
Podococcus barteri	Pb-Campo1	CpE	1	1	–	This study
(B) Impact of SNP calling methods on chloroplast diversity assessment						
Dioscorea rotundata	CR659	WGS	1	1	–	This study
		Sanger	–	–	88	Scarcelli *et al.* (2011)
Dioscorea abyssinica	A571	WGS	3	4	–	This study
		Sanger	–	–	89	Scarcelli *et al.* (2011)
Dioscorea praehensilis	P458	WGS	1	1	–	This study
		Sanger	–	–	88	Scarcelli *et al.* (2011)

WGS, whole-genome sequencing; CpE, chloroplast enrichment; LR-PCR, long-range PCR.

BVar15; details in Table 2). The percentage of observed intra-individual polymorphism was calculated over the total number of polymorphic positions. The full script used to perform the bioinformatics analysis is available as Supporting information.

Then, we kept only the reads mapped to the chloroplast genome and we mapped them to the closest available mitochondrial genome (*O. sativa*, BA000029.3) and to the nuclear genome (ADWL00000000.1). We report the number of reads mapping on both the chloroplast and the mitochondrial and on both the chloroplast and the nuclear genome.

To assess whether our results can be applied to other species, we performed the same analysis on pearl millet, yam and the palm genus *Podococcus*. WGS and LR-PCR were undertaken for one sample per species. For the CpE approach, we had access to data of eight different samples per species (Table 1). This allowed us to assess the variance in intra-individual polymorphism between samples. For these species, we only calculated the

Table 2 Summary of the different softwares used with associated options to call SNPs. The full scripts with all detailed options are available as Supporting information

Name	Software	Options
Hap	GATK HaplotypeCaller	-stand_emit_conf 10
Uni	GATK UnifiedGenotyper	-stand_emit_conf 10
Uni1	GATK UnifiedGenotyper	-ploidy 1 -stand_emit_conf 10
Bcf	SAMTOOLS mpileup BCFTOOLS view	-g -vcg
BcfB	SAMTOOLS mpileup BCFTOOLS view	-gB -vcg
Var15	SAMTOOLS mpileup VARSCAN mpileup2snp	Default –min-var-freq 0.15
Var50	SAMTOOLS mpileup VARSCAN mpileup2snp	Default –min-var-freq 0.5
BVar0	SAMTOOLS mpileup VARSCAN mpileup2snp	-B –min-var-freq 0
BVar05	SAMTOOLS mpileup VARSCAN mpileup2snp	-B –min-var-freq 0.05
BVar10	SAMTOOLS mpileup VARSCAN mpileup2snp	-B –min-var-freq 0.1
BVar15	SAMTOOLS mpileup VARSCAN mpileup2snp	-B –min-var-freq 0.15
BVar50	SAMTOOLS mpileup VARSCAN mpileup2snp	-B –min-var-freq 0.5

intra-individual polymorphism fraction for a minimum variant allele frequency of 15% (BVar15).

To assess whether the observed differences between the three sequencing methods are significant, intra-individual polymorphism levels were compared using a binomial test.

Impact of SNP calling methods on chloroplast diversity assessment. Evaluating NGS data analysis methods and pipelines requires 'true' reference sequences. In a previous study (Scarcelli *et al.* 2011), 50% of three yam chloroplast genomes was sequenced using the traditional Sanger approach. Here, we sequenced the same three individuals using NGS and analysed the data using bioinformatics pipelines (Table 1). By comparing the results with the Sanger data, we assessed the impact of SNP calling methods on diversity results. For this, we made the assumption that the Sanger sequencing gives the correct calling and that any discrepancy between Sanger and NGS data is due to NGS sequencing or calling errors.

Sanger data were aligned to the available yam chloroplast reference (NC_024170.1) using GENEIOUS PRO 4.7.6 (Kearse *et al.* 2012) with default values of the GENEIOUS ALIGNMENT tool. For NGS data from WGS, SNP calling was performed using a combination of three different software programs with different options, leading to nine different SNP calling methods:

1. GATK v3.3-0-g37228af (McKenna *et al.* 2010) was used to call SNPs with HaplotypeCaller and UnifiedGenotyper. Because the chloroplast genome is haploid, it was unclear whether we needed to adjust the ploidy option. Therefore, we tested the default parameters (ploidy = 2) with HaplotypeCaller and UnifiedGenotyper, and we used -ploidy 1 with UnifiedGenotyper (methods Hap, Uni and Uni1; details Table 2). By default, UnifiedGenotyper performs a join SNP calling, while HaplotypeCaller performs an individual SNP calling.
2. SAMTOOLS 1.1 (Li & Durbin 2009) was used to generate an *mpileup* file. By default, SAMTOOLS uses a probabilistic realignment for the computation of base alignment quality. It is possible to disable this realignment using option -B. We tried both methods. We then used BCFTOOLS 1.1 (Li 2011) to call SNPs (methods Bcf and BcfB; details in Table 2).
3. SAMTOOLS 1.1 (Li & Durbin 2009) was used to generate an *mpileup* file with and without option -B. We then used VARSCAN v2.3.7 (Koboldt *et al.* 2012) to call SNPs. The option –min-var-freq was set to 15% or 50% (methods Var15, Var50, BVar15 and BVar50; details in Table 2).

The full scripts used to perform the bioinformatics analyses are available in Supporting information.

We only kept the positions where a Sanger sequence was available, and we recorded all the discrepancies between Sanger and NGS data for each individual. We refer to a false negative when NGS data do not see an alternate homozygote found with Sanger sequence and inversely a false positive when NGS data identify an alternate homozygote not observed by Sanger sequence. False polymorphism refers to a intra-individual polymorphic genotype from NGS data instead of a homozygote alternate or reference allele from Sanger sequence. We calculated the percentage of error as the number of disagreements between Sanger and NGS divided by the total number of genotypes (identical + different) over all three samples and SNPs positions. We also calculated a special case of the previous measure as the percentage of intra-individual polymorphism error, that is the number of disagreements with intra-individual polymorphic calls divided by the total number of genotypes.

Finally, NGS sequencing is known to be error prone (Nakamura *et al.* 2011), and we wanted to assess how reproducible SNPs found in several sequencing of a single DNA are. We used a single yam sample (A571) and generated three independent whole-genome libraries. The three libraries were sequenced on a MiSeq and one also on a HiSeq2000 (Illumina, Genotoul, Toulouse, France). As a result, we analysed four independent runs of the same sample (Table 1). We used the same 9

bioinformatics methods previously described (Hap, Uni, Uni1, Bcf, BcfB, Var15, Var50, BVar15 and BVar50; Table 2), and we calculated the mean error number between the same sequenced samples.

Results

Impact of sequencing methods on observed intra-individual polymorphism

In the rice sample, the number of intra-individual polymorphic positions was high for the whole-genome sequencing (WGS), intermediate for the chloroplast enrichment (CpE) and low for the long-range PCR-based sequencing (LR-PCR) (Fig. 2, detailed results are available in Supporting information). The number of intra-individual polymorphic positions decreases when increasing the minimum variant allele frequency from 5% to 15%. Thus, intra-individual polymorphic positions are characterized by a variant allele with modest frequency. All differences are highly significant ($P < 0.01$), except when comparing CpE/LR-PCR with a minimum allele frequency of 15% ($P = 0.14$). To assess whether part of this result could be explained by differential mapping depending of the methodology, we mapped the reads that mapped to the chloroplast to the mitochondrial and nuclear genomes as well. The percentage of reads mapping simultaneously on the chloroplast and the nuclear genome was high for WGS (97%), intermediate for CpE (67%) and null for LR-PCR sequencing approaches (Supporting information). The percentage of reads mapping simultaneously on the chloroplast and the mitochondrial genome was similar for the three methods (~15%) (Supporting information).

The results observed in rice are also observed for yam and *Podococcus* species. (Fig. 3, detailed results are available in Supporting information); WGS produces more intra-individual polymorphic positions than CpE method and LR-PCR. All differences are highly significant ($P < 0.01$), except for *Podococcus* CpE/WGS ($P = 0.22$). Finally, for pearl millet, no marked differences between the three methods were observed (all comparisons $P > 0.05$), but pearl millet has a very low number of SNPs and consequently a low number of intra-individual polymorphic SNPs. For these three plants, we kept the number of mapped reads necessary

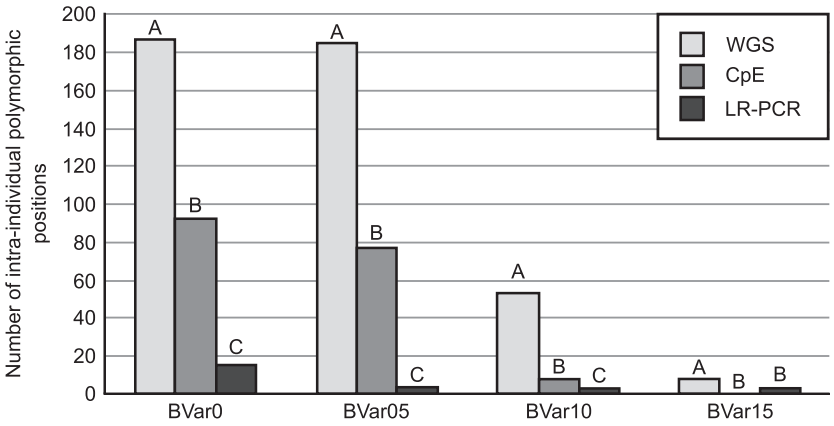

Fig. 2 Effects of the NGS approaches and different calling methods on the intra-individual polymorphic positions found for the rice sample (see Table 2 for details). For each method, letters indicate the significantly different groups ($P < 0.05$). WGS = whole-genome sequencing; CpE = chloroplast enrichment; LR-PCR = long-range PCR.

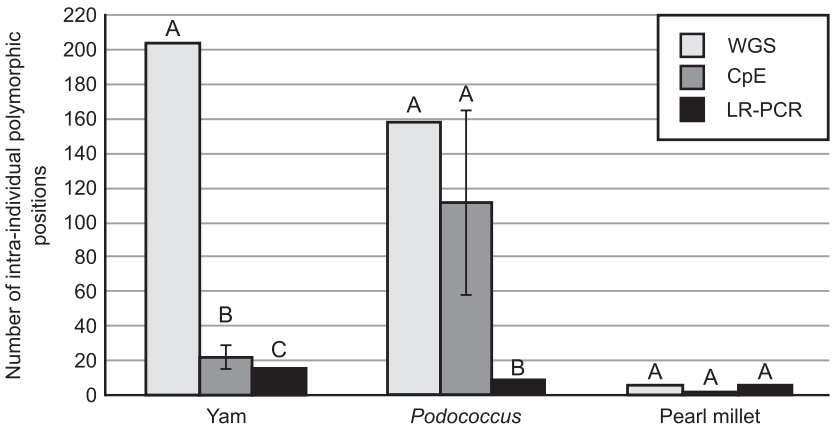

Fig. 3 Effects of the different NGS approaches on the intra-individual polymorphic positions found for yam, pearl millet and *Podococcus* samples with the calling method BVar15. For CpE, values are the mean over the eight samples and bars represent the standard deviation. For each species, letters indicate the significantly different groups ($P < 0.05$), based on mean comparison. WGS = whole-genome sequencing; CpE = chloroplast enrichment; LR-PCR = long-range PCR.

to fit the rice average coverage (100×). As we do not have a mitochondrial and genome references for these species, we could not control differential mapping to mitochondrial and nuclear genomes.

VARSCAN is the most effective SNP calling method for chloroplast diversity

We compared bioinformatics pipelines to assess their power to retrieve SNPs identified by Sanger sequencing for 50% of three yam chloroplast genomes (A571: 81 146 bp, P458: 79 388 bp and CR659: 80 383 bp). The average sequencing coverage was high for the three samples with a mean of 317×. VARSCAN methods considering a minimum variant allele frequency of 15% or 50% manage to find all the correct SNPs and to generate little to no false intra-individual polymorphic positions (Fig. 4a, b,d). Moreover, these methods create very few false positives (Fig. 4c). However, the effectiveness of VARSCAN is conditioned on the use of SAMTOOLS option -B to disable the probabilistic realignment for the computation of base alignment quality. When the -B option was not selected, a low number of true SNPs are retrieved, while many false negatives are found (Fig. 4a,b), thus percentage of errors are high (Fig. 4e,f). GATK (HaplotypeCaller and UnifiedGenotyper) and SAMTOOLS/BCFTOOLS combinations found most of the true SNPs, but they also generated many false intra-individual polymorphic positions (Fig. 4d) coupled with a significant proportion of error (Fig. 4e,f). The haploid option for GATK UnifiedGenotyper removes all intra-individual polymorphic positions (Fig. 4d,f) but leads to an increase of false positives compared to the Sanger reference (Fig. 4c). Detailed results are available in Supporting information.

For yam, the whole chloroplast genome was sequenced four independent times using the same sample (individual A571), generating 690 879, 529 666 and 191 051 reads (mean coverage: 712×) for the illumina miseq sequencing (2 × 150 bp) and 9 944 879 reads (mean coverage: 1565×) for the Illumina HiSeq sequencing (2 × 100 bp). When comparing the number of errors to the number of total nucleotide positions of the chloroplast (Fig. 4h), it appeared that 99.999% of the positions are correctly recovered using the VARSCAN approach (with 50% minimal allele frequency). The highest error rate was observed with GATK UnifiedGenotyper, where only 99.972% of the positions are correctly identified. Consequently, the number of errors produced depends on the calling method used (Fig. 4g). Our analyses suggest that VARSCAN with option variant allele frequency equal to 50% is the most appropriate method to deal with these intra-individual polymorphic positions. Detailed results are available in Supporting information.

Discussion

Where do plastid intra-individual polymorphisms come from?

Although NGS methods have significantly increased our ability to sequence more data, these approaches also come with problems of their own. Despite the fact that the chloroplast genome is haploid, we found a significant number of intra-individual polymorphic positions when NGS methods were used. Our analyses provide important data towards understanding the presence of intra-individual polymorphism in haploid organelles. The different NGS approaches investigated here lead to different predictions of observed intra-individual polymorphism (Fig. 1). It is important to note that none of the three methodologies is expected to be more biased than the others. Indeed, WGS may represent accurately the proportions of chloroplast vs. mitochondrial and nuclear genomes because each genome coverage is expected to be proportional to their frequencies. No strong bias is expected in CpE method because it requires only few (6) PCR steps and the primers used are compatible to sequences inserted on each side of the targeted fragment. Consequently, the primers themselves might not be associated with a particular bias. Finally, to be amplified with LR-PCR, a nonchloroplast fragment must be larger than the PCR target (7–22 kb). In the case a NUPT, NUMT or mtptDNA is amplified, its initial proportion must be very low compared to the original chloroplast fragment and the amplification biases have to be very important so that the nuclear or mitochondrial version appears preferentially.

Our results show that in all species except pearl millet (see below), WGS approaches resulted in high levels of observed intra-individual polymorphism, enrichment capture (CpE) to medium levels, while long-range PCR to few observed intra-individual polymorphic sites. Moreover, on rice, the number of reads mapping simultaneously on the chloroplast and the mitochondrial or nuclear genomes was high with WGS, medium for CpE and low for LR-PCR. Thus, we suggest that intra-individual polymorphism levels observed when using WGS or CpE approaches are explained mostly by plastid sequences transferred to the nuclear or mitochondrial genomes. Indeed, because WGS and CpE approaches are not plastome-specific approaches, they will have a tendency of sequencing plastid DNA in the nuclear and/or mitochondrial genomes resulting in the observed intra-individual polymorphism. In contrast, the LR-PCR approach is plastome specific, eliminating any contamination from the nucleus or the mitochondrion. Thus, this observed intra-individual polymorphism is mainly a genome reorganization artefact that will create several

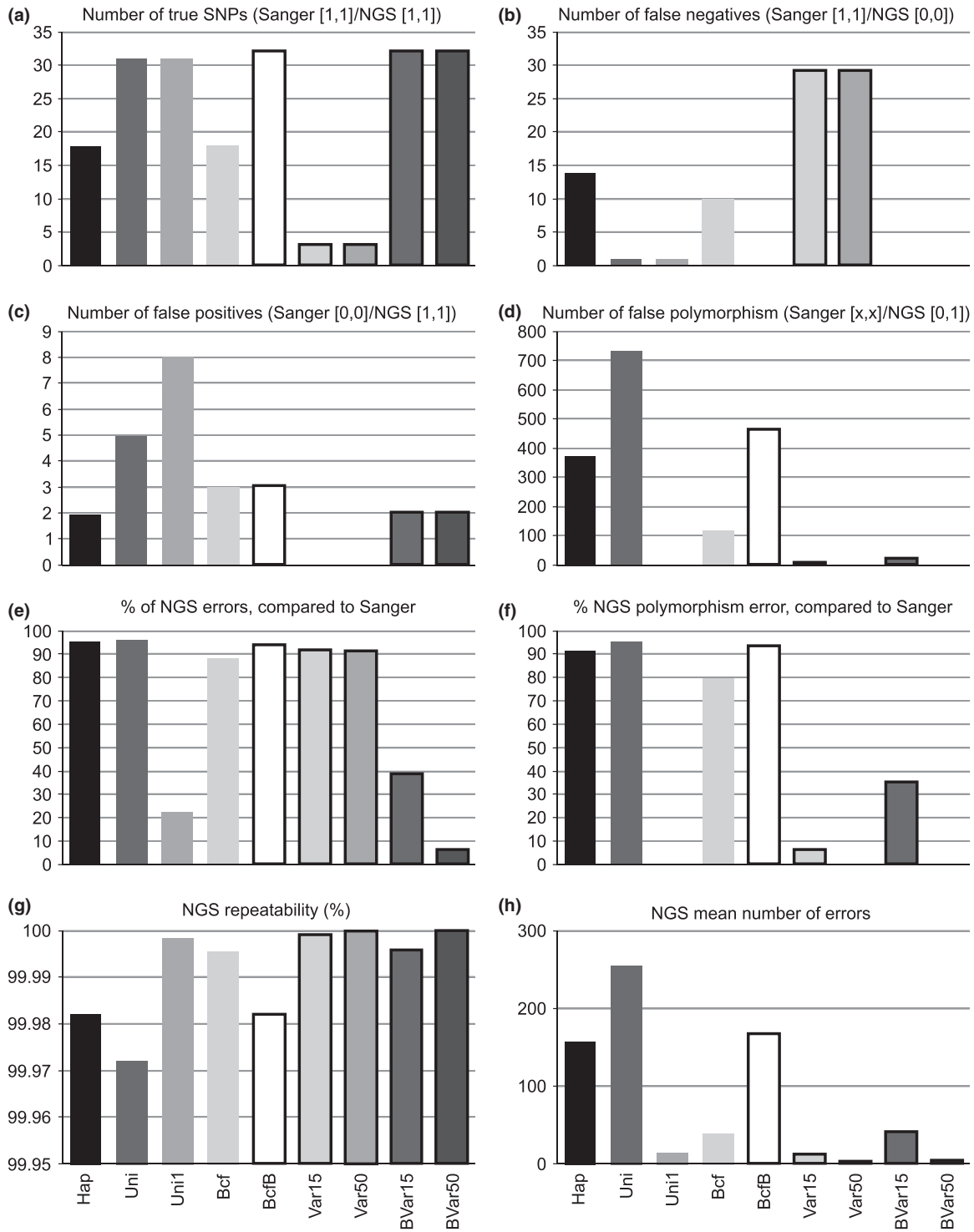

Fig. 4 Comparison of the different calling methods used (see Table 2 for details). We made the assumption that the Sanger sequencing gives the correct calling and that any discrepancy between Sanger and NGS data is due to NGS sequencing or calling errors. (a) The number of true SNPs, that is SNPs found by both Sanger and NGS; (b) the number of false negatives, that is SNPs found by Sanger only; (c) the number of false positives, that is SNPs found by NGS only; (d) the number of false intra-individual polymorphic positions, that is the number of polymorphic positions found by NGS only; (e) the percentage of error observed on NGS data, compared to Sanger data; (f) the percentage of intra-individual polymorphic error observed on NGS data, compared to Sanger data; (g) the NGS repeatability and (h) the NGS mean number of errors, compared to Sanger data. 0 = identical to the reference; 1 = different from the reference; [x, x] = [1,1] or [0,0]. Based on Sanger method, 33 [1,1] SNPs are expected.

analytical problems that have to be properly dealt with when using plastome diversity for phylogeography or population genetic studies.

Different studies showed that once included in the nuclear or mitochondrial genome, NUPT and mtptDNA evolve quickly (Michalovova *et al.* 2013; Sloan & Wu 2014). Therefore, just after a transfer, no polymorphism will be generated. The polymorphism will increase with the evolution of the transferred DNA and finally decrease when the fragment evolved so much that it will not be possible to map it on the chloroplast. However, because the transfer from the chloroplast genome is a continuous process, different levels of intra-individual polymorphism may be found for different individuals, according to the time and rate of mutation of the insert. This could explain why no significant differences were found between the three approaches for pearl millet. Moreover, pearl millet showed an overall low level of SNP diversity compared to the other species (Fig. 3) and consequently a low number of intra-individual polymorphic SNPs. Thus, conclusions about the presence of intra-individual polymorphism in pearl millet remain inconclusive.

Our findings do not exclude sequencing and calling errors as they are inherent to NGS whatever the sequencing approaches used. Errors and DNA transfers share one important characteristic: the diversity they 'create' is not useful when analysing the chloroplast genome as they do not reflect its evolution. It is therefore necessary to eliminate the intra-individual polymorphism produced by genome transfers and errors when using plastomes for phylogeny, phylogeography or population genetic studies. On the other hand, heteroplasmy is part of the chloroplast evolution and reflects mutations and biparental inheritance. However, even if heteroplasmy is certainly present in some species, there remains little evidence of its existence (Mason *et al.* 1994; Sabir *et al.* 2014) and it has been shown that chloroplast heteroplasmy can evolve quickly to homoplasmy (Birky 2001; Greiner *et al.* 2015). Moreover, even if heteroplasmy is present, we could consider only one chloroplast sequence for the study. Therefore, we think that it is preferable to remove all intra-individual polymorphic sites when using chloroplast data generated from NGS.

Our findings about intra-individual polymorphism in chloroplast NGS sequences might also apply to mitochondrial NGS sequences. As the chloroplast genome, the mitochondrial genome shows evidences of important transfers to the nuclear genome (Hazkani-Covo *et al.* 2010). However, mitochondrial heteroplasmy is more commonly documented in plants and animals and is even linked to genetic diseases (Kmiec *et al.* 2006; He *et al.* 2010; Woloszynska 2010; Chinnery & Hudson 2013). Mitochondrial heteroplasmy results from paternal leakage, mutations, but also from recombination (Kmiec *et al.* 2006; Woloszynska 2010). There are also evidence that heteroplasmy in mitochondria is under a certain nuclear control (Kmiec *et al.* 2006; Woloszynska 2010). Therefore, without more in-depth studies, it is unclear how the ratio errors/heteroplasmy/genome transfer will affect the levels of observed intra-individual polymorphism in mitochondria.

What is the best bioinformatic approach to use when dealing with chloroplast intra-individual polymorphism?

Overall, when comparing SNP calling methods to Sanger data, we found that VARSCAN gave the best results in terms of finding the correct SNPs and optimizing specificity and sensitivity (reducing the false positives and the false negatives, respectively). SNP calling is a critical step when analysing NGS data. Different methods have been proposed and tested to minimize the number of false positives and false negatives (Nielsen *et al.* 2011; Liu *et al.* 2013; Warden *et al.* 2014). Several studies have already compared the use of different software for SNP calling, but results are contentious. For human data, for example, GATK usually outperformed SAMTOOLS but the more accurate algorithm is either UnifiedGenotyper (Cornish & Guda in press) or HaplotypeCaller (Pirooznia *et al.* 2014).

We could not generalize human DNA analysis to our specific case, because we are dealing with a haploid genome. Still, we need to obtain as few intra-individual polymorphic positions as possible because, as we showed earlier, these intra-individual polymorphic positions are mainly an artefact of plastid DNA transfers.

SAMTOOLS mpileup has been designed to manage diploid data (Li & Durbin 2009). Despite this, our data showed that SAMTOOLS provided accurate results, but only when using the -B option.

GATK HaplotypeCaller and UnifiedGenotyper's genotype calling methods are based on a Bayesian model where the prior used is the diploid genotype probabilities (McKenna *et al.* 2010). Since the 3.3 version, both algorithms are able to deal with haploid genomes. Here, GATK UnifiedGenotyper was more efficient than GATK HaplotypeCaller, but both algorithms produced more false intra-individual polymorphic positions. However, using the diploid option with GATK UnifiedGenotyper and deleting intra-individual polymorphic positions seems like a more efficient method than using the haploid option, because of the observed number of false positives.

In contrast, the VARSCAN calling method uses base quality, read depth and variant allele frequency only and can therefore be used whatever the ploidy level (Koboldt

et al. 2012). In the VARSCAN manual, it is recommend to use the -B option of SAMTOOLS when generating the *mpileup* file which recalibrates the quality of the base alignment. In this study, we confirm this recommendation as we noticed that using VARSCAN without the -B option generated very few variant positions and therefore a large number of false negatives. The combination of SAMTOOLS mpileup -B with VARSCAN allows to finely tune the variant allele frequency and thus to allow or disallow intra-individual polymorphic positions according to our specific goals. When choosing –min-var-freq = 0.15 and –min-freq-for-hom = 0.5, only variant alleles with a high frequency will be taken into account. When choosing –min-var-freq = 0.5 and –min-freq-for-hom = 0.5, no intra-individual polymorphic genotypes will be called and the genotype will be a homozygote for the most frequent allele. Interestingly, when forcing genotypes to be homozygous, we did not increase the calling of false negatives and false positives. Therefore, we recommend using the combination of SAMTOOLS -B/VARSCAN –min-var-freq 0.5 –min-freq-for-hom 0.5 for SNP calling when analysing chloroplast genome data generated by NGS approaches.

Acknowledgements

This project was supported by Agropolis Foundation through the "Investissements d'avenir" program (ANR-10-LABX-0001-01) under the reference ID 1202-040 (CHLORODIV) to T. Couvreur and the Agence Nationale de Recherche (ANR, project AFRICROP ANR-13-BSV7-0017) to Y. Vigouroux.

References

Bendich AJ (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays*, **6**, 279–282.

Birky CW (2001) The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics*, **35**, 125–148.

Bock R, Timmis JN (2008) Reconstructing evolution: gene transfer from plastids to the nucleus. *BioEssays*, **30**, 556–566.

Bock DG, Kan NC, Ebert DP, Rieseberg LH (2014) Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. *New Phytologist*, **201**, 1021–1030.

Chinnery PF, Hudson G (2013) Mitochondrial genetics. *British Medical Bulletin*, **106**, 135–159.

Cornish A, Guda C (In press) A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Research International* ID456479, (in press).

Corriveau JL, Coleman AW (1988) Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 Angiosperm species. *American Journal of Botany*, **75**, 1443–1458.

Garrick RC, Bonatelli IA, Hyseni C *et al*. (2015) The evolution of phylogeographic data sets. *Molecular Ecology*, **24**, 1164–1171.

Greiner S, Sobanski J, Bock R (2015) Why are most organelle genomes transmitted maternally? *BioEssays*, **37**, 80–94.

Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics*, **6**, e1000834.

He Y, Wu J, Dressman DC *et al.* (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, **464**, 610–614.

Jünemann S, Sedlazeck FJ, Prior K *et al.* (2013) Updating benchtop sequencing performance comparison. *Nature Biotechnology*, **31**, 294–296.

Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kmiec B, Woloszynska M, Janska H (2006) Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. *Current Genetics*, **50**, 149–159.

Koboldt D, Zhang Q, Larson D *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, **22**, 568–576.

Kumar RA, Oldenburg DJ, Bendich AJ (2014) Changes in DNA damage, molecular integrity, and copy number for plastid DNA and mitochondrial DNA during maize development. *Journal of Experimental Botany*, **65**, 6425–6439.

Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends in Genetics*, **21**, 655–663.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2989.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.*, 1000 Genome project data processing subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu X, Han S, Wang Z, Gelernter J, Yang BZ (2013) Variant callers for next-Generation Sequencing data: a comparison study. *PLoS ONE*, **8**, e75619.

Mariac C, Robert T, Allinne C *et al.* (2006) Genetic diversity and gene flow among pearl millet crop/weed complex: a case study. *Theoretical and Applied Genetics*, **113**, 1003–1014.

Mariac C, Scarcelli N, Pouzadou J *et al.* (2014) Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources*, **14**, 1103–1113.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, **17**, 10–12.

Mason RJ, Holsinger KE, Jansen RK (1994) Biparental inheritance of the chloroplast genome in Coreopsis (Asteraceae). *Journal of Heredity*, **85**, 171–173.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Michalovova M, Vyskot B, Kejnovsky E (2013) Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity*, **111**, 314–320.

Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, **12**, R112.

Nakamura K, Oshima T, Morimoto T *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, **39**, e90.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, **7**, e37558.

Petit RJ, Vendramin GG (2007) Plant phylogeography based on organelle genes: an introduction. In: *Phylogeography of southern European refugia* (eds Weiss S & Ferrand N), pp. 23–97. Springer, Dordrecht.

Pirooznia M, Kramer M, Parla J *et al.* (2014) Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, **8**, 14.

Richly E, Leister D (2004) NUPTs in sequenced Eukaryotes and their genomic organization in relation to NUMTs. *Molecular Biology and Evolution*, **21**, 1972–1980.

Sabir JSM, Arasappan D, Bahieldin A *et al.* (2014) Whole mitochondrial and plastid genome SNP analysis of nine date palm cultivars reveals plastid heteroplasmy and close phylogenetic relationships among cultivars. *PLoS ONE*, **9**, e94158.

Scarcelli N, Tostain S, Vigouroux Y, Agbangla C, Daïnou O, Pham JL (2006) Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. *Molecular Ecology*, **15**, 2421–2431.

Scarcelli N, Barnaud A, Eiserhardt W *et al.* (2011) A set of 100 chloroplast DNA primer pairs to study population genetics and phylogeny in Monocotyledons. *PLoS ONE*, **6**, e19954.

Sloan DB, Wu Z (2014) History of plastid DNA insertions reveals weak deletion and AT mutation biases in Angiosperm mitochondrial genomes. *Genomes Biology and Evolution*, **6**, 3210–3221.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip pf the genomic iceberg: next-generation sequencing for plant systematic. *American Journal of Botany*, **99**, 349–364.

Wang D, Rousseau-Gueutin M, Timmis JN (2012) Plastid sequences contribute to some plant mitochondrial genes. *Molecular Biology and Evolution*, **29**, 1707–1711.

Warden CD, Adamson AW, Neuhausen SL, Wu X (2014) Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ*, **2**, e600.

Wolfe AD, Randle CP (2004) Recombination, heteroplasmy, haplotype polymorphism, and paralogy in plastid genes: implications for plant molecular systematics. *Systematic Botany*, **29**, 1011–1020.

Woloszynska M (2010) Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes—though this be madness, yet there's method in't. *Journal of Experimental Botany*, **61**, 657–671.

Yoshida T, Furihata HY, Kawabe A (2014) Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species. *DNA Research*, **21**, 127–140.

Zhang T, Fang Y, Wang X *et al.* (2012) The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. *PLoS ONE*, **7**, e30531.

N.S., C.M., T.C., Y.V. designed research; N.S., C.M., A.F., D.R. performed research and analysed data, N.S., C.M., F.S., C.B.-S., Y.V. participate to bioinformatics analyses; N.S., T.C., Y.V. wrote the study with input from all authors.

Data accessibility

All sequences are available on either GenBank (Sanger fasta sequences), NCBI-SRA (.fastq files) or DRYAD DOI 10.5061/dryad.31733 (aligned.bam files) as listed in Supporting information. Moreover, the.vcf files with SNP calling are also available on DRYAD DOI: 10.5061/dryad.31733 (as listed in Supporting information).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 List of GenBank (.fasta), NCBI-SRA (.fastq) and DRYAD (.fastq, .bam and .vcf) ID resources used in this paper.

Table S2 Scripts used to prepare raw data, map to the reference and call SNPs (12 different methods). List of adaptors and tags used during the libraries' construction.

Table S3 Number of reads mapping on the chloroplast, mitochondrial and nuclear genomes for the rice sample.

Table S4 Total number of SNP [1,1] and polymorphic positions [0,1] observed for each species and for each calling method.