CRISPR/Cas9-based repeat depletion for the high-throughput genotyping of complex plant genomes

Marzia Rossato[1*$], Luca Marcolungo[1*], Luca De Antoni[1], Giulia Lopatriello[1], Elisa Bellucci[2], Gaia Cortinovis[2], Giulia Frascarelli[2], Laura Nanni[2], Elena Bitocchi[2], Valerio Di Vittori[2], Leonardo Vincenzi[1], Filippo Lucchini[1], Kirstin Bett[3], Larissa Ramsay[3], David James Konkin[4], Massimo Delledonne[1$] and Roberto Papa[2$]

*equal contribution

[$]corresponding authors

[1]Department of Biotechnology, University of Verona, Strada Le Grazie 15, 37134, Verona, Italy,

[2]Department of Agricultural, Food and Environmental Sciences, Polytechnic University of Marche, via Brecce Bianche, 60131, Ancona, Italy

[3]Department of Plant Sciences, University of Saskatchewan, 51 Campus Drive, Saskatoon, Saskatchewan S7N 5A8, Canada

[4]National Research Council Canada, 110 Gymnasium Place, Saskatoon, Ontario S7N 0W9

ABSTRACT

High-throughput genotyping facilitates the large-scale analysis of genetic diversity in population genomics and genome-wide association studies that combine the genotypic and phenotypic characterization of large collections of wild and domesticated germplasm. Genotyping by sequencing is progressively replacing traditional genotyping methods due to the lower ascertainment bias. However, genome-wide genotyping by sequencing becomes expensive in species with large genomes and a high proportion of repetitive DNA. Here we describe the use of CRISPR/Cas9 technology to deplete repetitive elements in the 3.76-Gb genome of lentil (*Lens culinaris*), 84% of which consists of repeats, thus concentrating the sequencing data on coding and regulatory regions (unique regions). We designed a custom set of 566,722 gRNAs, each with at least 25 recognition sites, targeting 2.9 Gbp of repeats in 500-bp insert sequencing libraries. We excluded repetitive regions overlapping annotated genes and putative regulatory elements based on ATAC-Seq data. The novel depletion method removed 40% of reads mapping to repeats, increasing those mapping to unique and functional regions by 2.6-fold. This repeat-to-unique shift in the sequencing data increased the number of genotyped bases by up to 17-fold compared to non-depleted libraries. We were also able to identify up to 18-fold more genetic variants in the unique regions and increased the genotyping accuracy by rescuing thousands of heterozygous variants that otherwise would be missed due to low coverage. The method performed similarly regardless of the multiplexing level, type of library or genotypes, including different cultivars and a closely-related species (*L. orientalis*). Our results confirmed that CRISPR/Cas9-driven repeat depletion focuses sequencing data on meaningful genomic regions, helping to improve high-density and genome-wide genotyping in large and repetitive genomes.

KEYWORDS: repetitive elements, CRISPR/Cas9, genotyping by sequencing, next-generation sequencing library

INTRODUCTION

The efficient and accurate determination of genotypes is necessary for large-scale projects investigating the genetic composition of germplasm collections representing wild and domesticated species and inbred lines. One example is the EU H2020 project INCREASE (www.pulsesincrease.eu) (1), which focuses on four legume staples: chickpea, common bean, lentil and lupin. Such projects depend on large cohorts of individuals to enable the comparative analysis of samples with sufficient statistical power. Cost-effective high-throughput genotyping methods are therefore needed to increase the number of samples that can be processed in an economically feasible manner (1). This can only be achieved by reducing the fraction of each individual genome that is sequenced while ensuring that the same homologous regions are examined in each individual (2).

High-throughput low-cost genotyping has largely been achieved by the analysis of single-nucleotide polymorphisms on microarray-based platforms (SNP arrays). These allow up to several thousand SNPs to be tested simultaneously (3). This approach considers a predefined set of markers, resulting in fixed costs per individual regardless of the genome size and fraction of repetitive DNA. However, analysis is restricted to known SNPs that are frequent in the population, while rare and unknown SNPs are ignored. This is a drawback when analyzing diverse landraces and distant wild relatives, as required in the germplasm characterization projects mentioned above (4).

More recently, next generation sequencing (NGS) has provided an opportunity to discover genome-wide variants in an unbiased manner. Genotyping by sequencing (GBS) involves low coverage (5–10x) whole-genome sequencing (lcWGS), allowing the characterization of several million variants (5,6). To reduce costs enough to make WGS affordable even in large germplasm collections, very low coverage (0.5–2x) WGS (ultra-lcWGS) can be combined with imputation to infer positions that are not sequenced or genotyped (7–9). Alternatively, sequencing costs are often minimized by reduced-representation sequencing, which comprises methods such as restriction site-associated DNA sequencing (RAD-Seq) (10) and double-digest RAD-Seq (ddRAD-seq) (2,11). These methods concentrate sequencing data on regions adjacent to restriction sites by exploiting the specificity of restriction endonucleases. Reduced-representation sequencing is suitable for large cohorts, but provides only low-resolution data (0.5–15% in our experience). The resolution can be increased without significantly greater costs by using the Twist 96-Plex Library Prep Kit (formerly iGenomX Riptide Kit) to generate multiplexed libraries, allowing 96–960 samples to be processed simultaneously and resulting in the non-random sampling of millions of genomic positions (12).

Despite the advantages of GBS over SNP arrays, one common disadvantage is that GBS methods do not distinguish between repetitive (low-complexity) and unique (high-complexity) regions, the latter comprising coding and regulatory regions that are the main targets of natural selection and thus the focus of most genotyping projects. In contrast, low-complexity regions of plant genomes mainly comprise transposable elements, simple sequence repeats and tandem repeats. Transposable elements play a key role in genome evolution, but the analysis of such regions is technically challenging and largely uninformative in genotyping studies, unless dedicated analysis workflows are applied (13). Mapping reads to transposable/repetitive elements can result in low-quality alignments that hinder the calling of accurate genotypes, which is a consistent challenge particularly for those plant species with large genomes, where repetitive elements account for up 90% of the total DNA. This includes many domesticated crops such as corn (*Zea mays*), wheat (*Triticum* spp.), lentil (*Lens culinaris*) and soybean (*Glycine max*) (14). One strategy to address this issue is whole exome sequencing (WES), which selects coding regions for preferential sequencing (15) as shown in lentil, wheat and barley (16,17). However, WES is expensive because it requires long and complex protocols, making it unsuitable for the analysis of large populations. It also overlooks regulatory elements, which are equally important as sources of genetic diversity (18–20).

Ideally, lcWGS could be focused on the most complex parts of the genome, avoiding wasted effort on the sequencing of repetitive elements. This could be achieved by using enzymes that enable target enrichment by depleting unwanted sequences from NGS libraries. For example, the duplex-specific nuclease (DSN) selectively digests double-stranded DNA molecules, and can be used to eliminate highly abundant sequences in a controlled denaturation-reassociation reaction (21). This method has been used in RNA-Seq analysis to remove abundant transcripts (22,23) and, just occasionally, also to delete repetitive elements in DNA-Seq libraries generated from plant genomes (24,25). However, the use of this

approach is limited by the need to define an optimal renaturation time for each specific genome, that can vary from some hours to a day (25). DSN also removes informative repetitive elements, such as the coding sequences of abundant gene families, which are particularly relevant in polyploid plants arising from whole genome duplication events (25). More recently, the CRISPR/Cas9 system (clustered regularly interspaced short palindromic repeats and CRISPR-associated nuclease 9) has been used for the selective depletion of unwanted genome fractions from sequencing libraries (26). The Cas9 enzyme can be programmed to cut library fragments by designing specific guide-RNA sequences targeting the unwanted sequences. Subsequently, only intact fragments -retaining adapters at both ends- can be effectively amplified by PCR and generate productive clusters on a sequencing flow-cell. The DASH approach (depletion of abundant sequences by hybridization) involved the use of Cas9 to exclude ribosomal RNA (rRNA) sequences from RNA-Seq libraries and to remove DNA from common pathogens in order to detect rare pathogens in metagenomic samples (26). A similar technology has been recently commercialized under the name "CRISPRclean" by JumpCode Genomics (27).

Here we determined whether CRISPRclean technology could be used to deplete the repetitive elements in libraries prepared from the 3.76-Gbp genome of lentil (*L. culinaris*), 84% of which is repetitive DNA. CRISPRclean technology was combined with Twist multiplexing libraries and we evaluated its performance, focusing on the technical features required for genotyping. Our results will facilitate the large-scale genomic analysis of lentil as well as other plant species with large and highly-repetitive genomes.

RESULTS

Depletion of *L. culinaris* repetitive DNA using CRISPR/Cas9

We designed a custom set of gRNAs to deplete the repetitive DNA content of the *L. culinaris* CDC Redberry genome (28), targeting transposable elements (3.1 Gbp, 83%), simple sequence repeats (303 Mbp, 8%) and tandem repeats (13 Mbp, 0.4%) in the nuclear genome, as well as the entire mitochondrial genome (mtDNA, 489 kbp) and chloroplast genome (chlDNA, 118 kbp) (Table S1). We excluded repetitive DNA that overlapped with functional regions such as annotated genes (185 Mbp, 5%) and putative regulatory regions identified using ATAC-Seq data (78 Mbp, 2%), (Table S1). All nuclear DNA outside the gRNA target regions is hereafter defined as unique. The final design comprised 566,722 gRNAs with at least 25 recognition sites, potentially of targeting 2.9 Gbp (77%) of the *L. culinaris* nuclear genome and 92% of its repetitive regions when using a sequencing library with 500-bp inserts (Table S2). An additional 2,366 gRNAs targeted the mitochondrial and chloroplast genomes (Table S1). The gRNAs were assigned to 11 pools based on their cutting frequency in the *L. culinaris* genome (Table S2).

Pool	Genomic target	Number of individual gRNAs	gRNA relative multiplicity	Number of total gRNAs	Relative frequency	ng per reaction (1x)	Depletion group (based on cutting frequency)
1	Nuclear	60,000	1x	60,000	1	40	Group 1
2	Nuclear	60,000	1x	60,000	1	40	Group 1
3	Nuclear	60,000	1x	60,000	1	40	Group 1
4	Nuclear	39,463	1x	39,463	1	60	Group 2
	Mitochondrial	1,870	4x	7,480	N/A		
	Chloroplast	261	50x	13,050	N/A		
5	Nuclear	60,000	1x	60,000	2	40	Group 1
6	Nuclear	60,000	1x	60,000	2	40	Group 2
7	Nuclear	49,574	1x	49,574	2	40	Group 2
	Nuclear	7,394	1x	7,394	3		
8	Nuclear	60,000	1x	60,000	3	40	Group 2
9	Nuclear	34,198	1x	34,198	4	80	Group 3
	Nuclear	20,486	1x	20,486	5		
	Chloroplast	100	50x	5,000	N/A		
10	Nuclear	33,613	1x	33,613	6-9	120	Group 3
	Nuclear	5,139	5x	25,695	20-99		
11	Nuclear	16,799	3x	50,397	10-19	80	Group 3
	Nuclear	54	50x	2,700	≥100		
	Chloroplast	137	50x	6,850	N/A		

Table S2. Composition of gRNA pools. For each pool, the table shows the number of individual and total gRNAs, the genomic target, the gRNA abundance, relative frequency, quantity (ng) per reaction, and the subdivision of pools into three groups based on cutting frequency. The multiplicity indicates the number of individual gRNA molecules present in the pool relative to the least abundant ones. The relative frequency is the ratio between the number of cut sites for each gRNA and the minimum number of cut sites (25). N/A, not applicable, as the exact number of organelle genomes is not known.

The custom gRNAs were tested on three Twist 8-plex libraries, allowing the reproducible sampling of the same genomic regions by random priming during first-strand DNA synthesis. Each multiplex library comprised eight replicates of three distinct *L. culinaris* samples (cv. Castelluccio) (**Table S3**). Cas9/gRNAs ribonucleoprotein (RNP) complexes (1:10 protein/gRNA ratio) were generated, and gRNAs with more target sites in the genome were used at higher relative concentrations in the final reaction (**Table S2**). Depletion reactions in the presence of RNP complexes were conducted either using all gRNAs simultaneously or by splitting the gRNA pools into three groups based on cutting frequency (**Table S2**) and using the groups sequentially, starting with the lowest cutting frequency. Depleted and non-depleted libraries were sequenced, generating 91 million fragments on average (**Table S4**). The sequencing data were normalized at ~50 million fragments per library in order to compare the proportion of reads mapping on repetitive and unique regions of the nuclear genome and on the organelle genomes (**Figure 1**). The number of reads mapping to repetitive regions (total repeats, nuclear repeats, mtDNA and chlDNA) was significantly lower in the depleted libraries compared to the non-depleted libraries, with the sequential depletion strategy using three gRNA groups performing best and depleting 37.7% of the repetitive DNA (**Figure 1A**). The results were similar when considering only the nuclear repetitive regions (37.2% depletion) (**Figure 1B**). A small fraction of total reads mapped to the organelle genomes (~1%). Both depletion strategies were similarly effective in the chloroplast genome, resulting in ~78.5% depletion (**Figure 1C**). In contrast, no significant depletion was observed in the mitochondrial genome (**Figure 1D**). In parallel, the sequential depletion strategy achieved a 130% increase in the number of reads mapping to unique regions, from 17.5 to 40.4 million (**Figure 1E**). Given that the concentration of Cas9 RNPs influences the cutting efficiency (26), we repeated the sequential depletion strategy using double amount of Cas9 and gRNA. This modified the read distribution further, achieving 41.2% depletion of nuclear repeats and a 160% increase in reads mapped to unique regions. This double sequential depletion strategy was therefore the most efficient, and was applied in all subsequent experiments. Overall, our results demonstrated that the custom

gRNA set and Cas9 effectively targeted fragments containing repetitive DNA sequences and depleted them in the resulting sequencing libraries. Manual inspection of mapped reads confirmed that less sequencing data was assigned to regions of repetitive DNA and more reads were mapped to unique parts of the genome (**Figure 2A-B**).

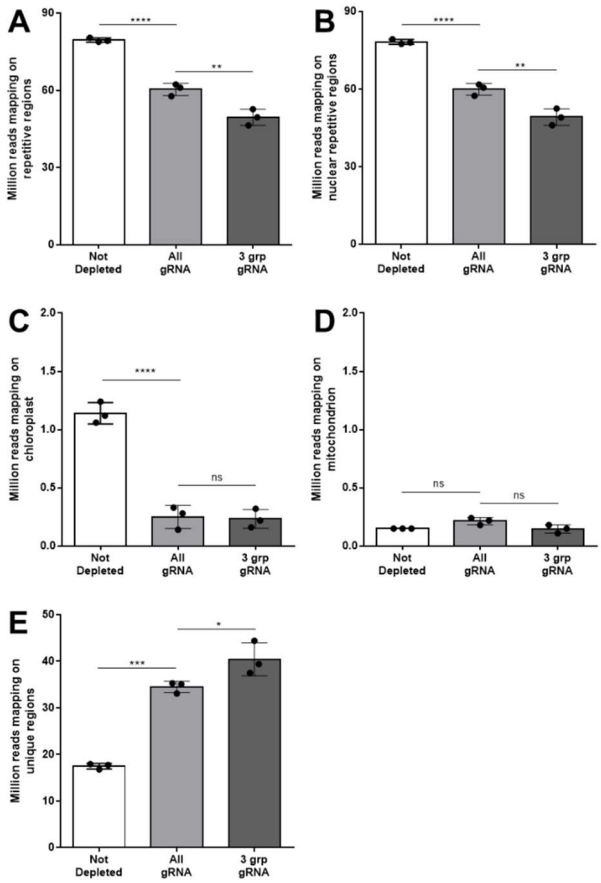

Figure 1. Distribution of mapped reads after CRISPR/Cas9-mediated repeat depletion. Libraries of *L. culinaris* cv. Castelluccio DNA (8-plex) were depleted using the custom gRNA set and Cas9. The gRNAs were used simultaneously (All) or were split into three groups that were used sequentially in order of increasing cutting frequency (3 grp). Bar graphs show the number (in millions) of reads mapping to repetitive DNA (**A**), to nuclear repetitive DNA (**B**), to the chloroplast genome (**C**), to the mitochondrial genome (**D**), and to the unique regions of the nuclear genome (**E**). Data are means ± SD (n = 3 for each condition; *p-adj < 0.05, **p-adj < 0.01, ****p-adj < 0.0001; one-way ANOVA plus Tukey's multiple comparisons test; ns, not significant).

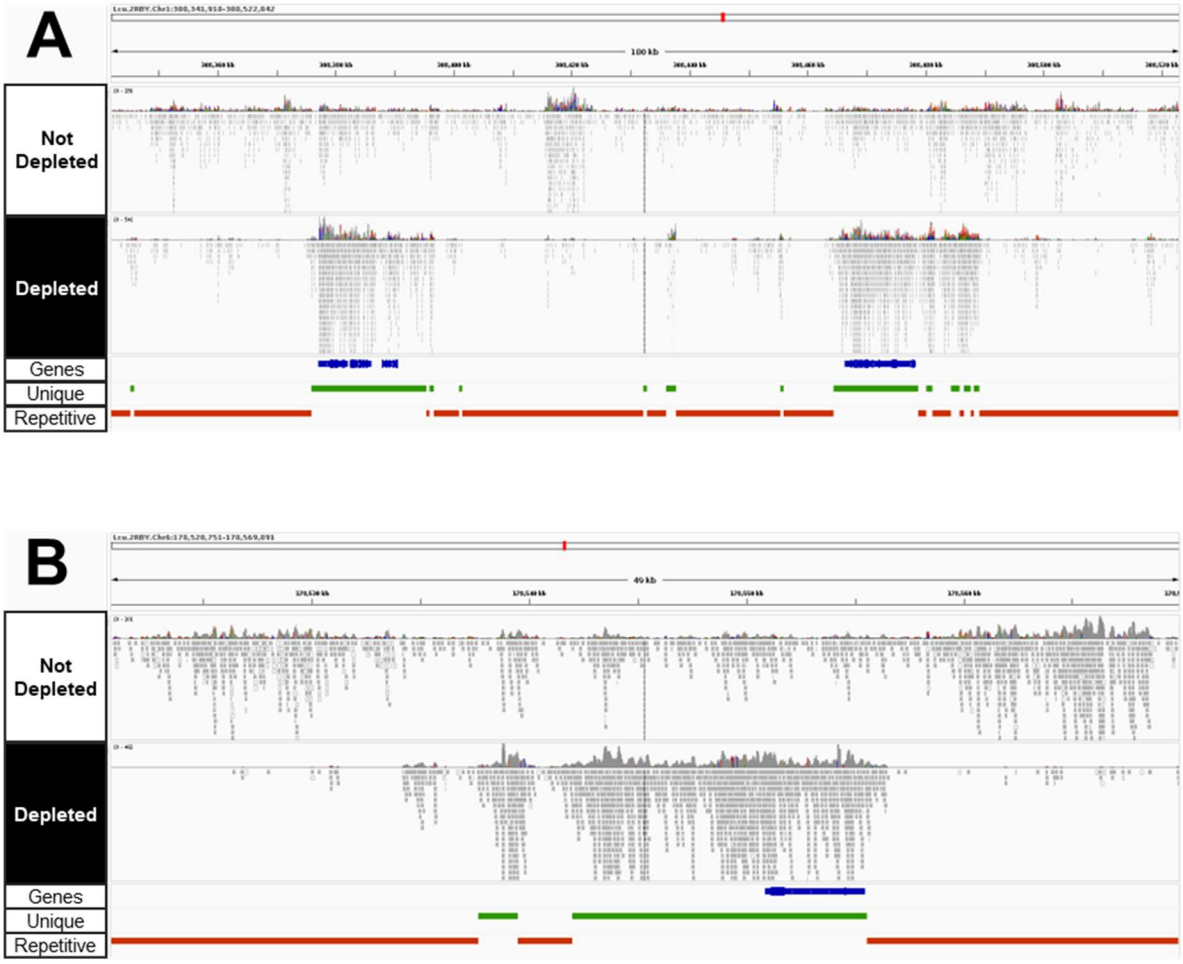

Figure 2. Reads mapped to repetitive or unique regions of the lentil genome before and after CRISPR/Cas9-mediated repeat depletion. Integrative Genome Browser Visualization (IGV) of Illumina sequencing data mapped to two genomic sites of ~180 and ~50 kbp, before and after CRISPR/Cas9-mediated repeat depletion. Tracks in blue, green and red represent annotated genes, unique regions and repetitive regions, respectively.

Efficiency of CRISPR/Cas9-mediated depletion for different classes of nuclear repeats

We next examined the depletion of different classes of repetitive sequences in the *L. culinaris* genome. Reads mapping to the most abundant retroelements, namely the Ty3-Gypsy family (64% of the genome (28), were reduced by 47% in the depleted libraries, whereas those mapping to the Ty3-Copia family (15% of the genome) and other long terminal repeat (LTR) elements (3% of the genome) were depleted by 3% and 38%, respectively (**Figure 3A** and **Table S5**). In contrast, there was no decrease in the abundance of other transposable elements (Line, CACTA, mu, hAT, Helitron, Harbinger, mariner and Sine), each representing < 1% of the genome (**Table S5**), and there was no reduction in the number of reads mapping to tandem repeats (0.4% of the genome) or simple sequence repeats (8% of the genome) (**Figure 3A**). We observed a significant correlation between the variation in mapped reads after depletion and the abundance of these repeat classes in terms of overall repeat length and occurrence in the genome (**Figure 3B-C** and **Table S5**). Given that the number of gRNAs targeting each repeat class increased proportionally with the repeat size and occurrence, the most efficiently depleted repetitive elements also featured a higher density of gRNA targets (**Figure 3D**). There was a significant correlation between the variation of mapping reads following depletion and the gRNA density over the whole target region when considering each single cut site in the genome (**Figure S1**). In particular, target regions with a density > 8 gRNAs/kbp showed a negative read variation in 85% of cases (**Figure S1**) whereas regions targeted by < 8 gRNAs/kbp usually showed limited or no depletion (**Figure S1**). We therefore concluded that the depletion efficiency across different repeat classes was dependent on the density of gRNA targets.

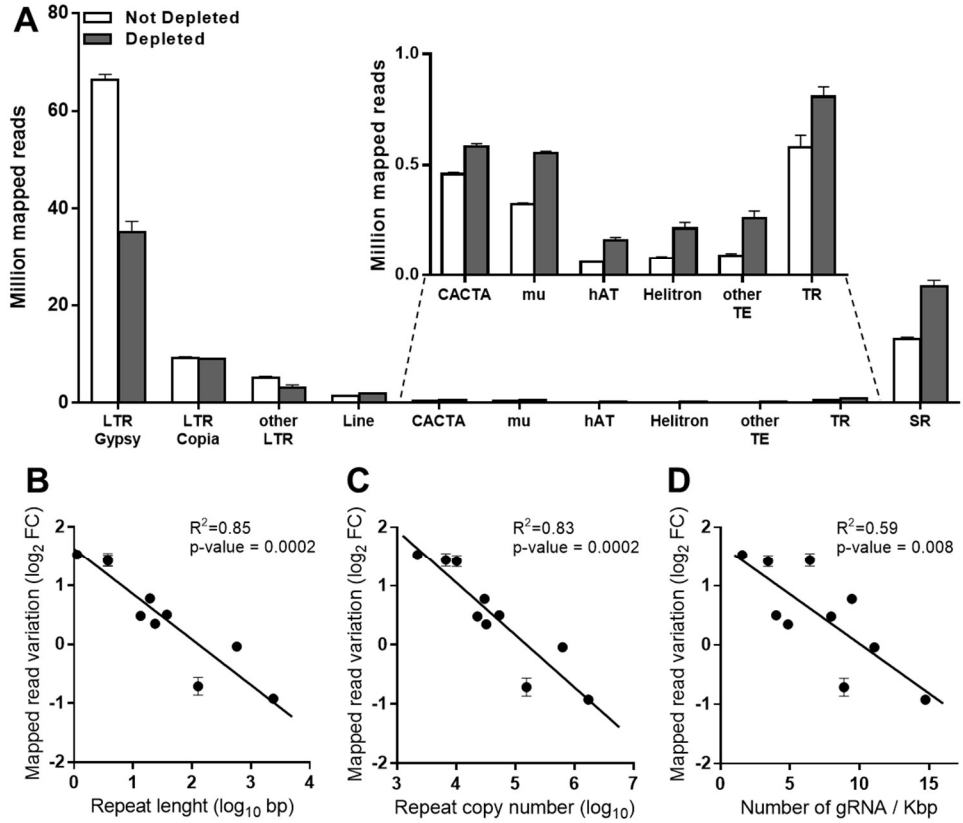

Figure 3. Sequencing data distribution after CRISPR/Cas9-mediated repeat depletion according to the class of nuclear repeat. (**A**) Number of reads (in millions) mapping to different nuclear repeat classes before and after CRISPR/Cas9-mediated repeat depletion: Ty3-Gypsy (LTR Gypsy), Ty1-Copia (LTR Copia), other LTR, Line, CACTA, mu, hAT, Helitron transposons, other transposable elements with abundance < 0.1% (Harbinger, mariner, Sine), tandem repeats (TR) and simple repeats (SR). Correlation between the variation of mapped reads following CRISPR/Cas9-mediated repeat depletion on the same repeat classes versus the repeat length (**B**), repeat copy number (**C**) and density of gRNAs targeting each repeat class (**D**). Data are means ± SE (n = 3). FC – fold change.

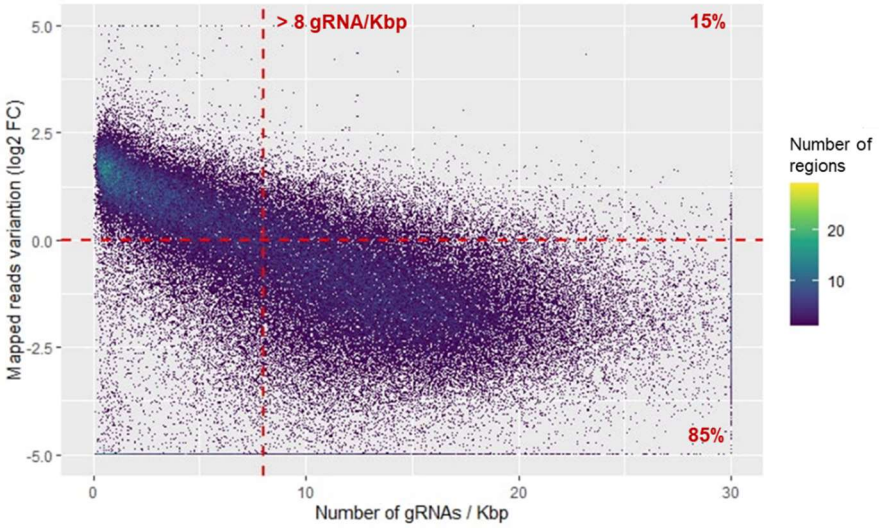

Figure S1. Variation of mapped reads in relation to gRNA density. Density plot shows the variation of mapped reads over 1-kbp windows of nuclear repeats in relation to the number of gRNAs targeting the same region in one representative experiment. Only regions with at least one gRNA cut are shown. Percentages indicate the fraction of regions with more than eight gRNAs showing either positive or negative read variations. For plotting simplicity, coverage variation values < –5 were set to –5 and cut frequency values > 30 were set to 30. FC – fold change.

Impact of CRISPR/Cas9-mediated repeat depletion on genotyping accuracy

Next, we investigated the impact of CRISPR/Cas9-mediated repeat depletion on the number of genomic positions in the unique regions where a base can be reliably genotyped (% PASS at a depth of ≥ 5 reads). For this analysis, sequencing data generated from depleted and non-depleted *L. culinaris* cv. Castelluccio samples were downsampled from 62 to 6 million fragments to mimic lcWGS and ultra-lcWGS. Consistently more bases were genotyped within the unique regions of the depleted sample over the whole range considered (**Figure 4A**). Because a genotyped position does not necessarily allow the variant to be identified (this also depends on allele coverage), we also determined the impact of repeat depletion on variant calling. Following depletion, the total number of variants identified in the unique regions increased significantly by ~10-fold on average, with a delta of ~20,000 to ~300,000 more variants identified in the depleted sample (**Figure 4B**). This allowed us to identify up to ~50,000 variants that would not be detected without the depletion strategy, despite occupying genotypable positions in the non-depleted sample (**Figure 4C**). These false negative variants in the non-depleted sample were not called due to allelic imbalance caused by the low coverage (**Figure S2**). Consistently, most of them (~97%) were heterozygous variants (**Figure 4C**).

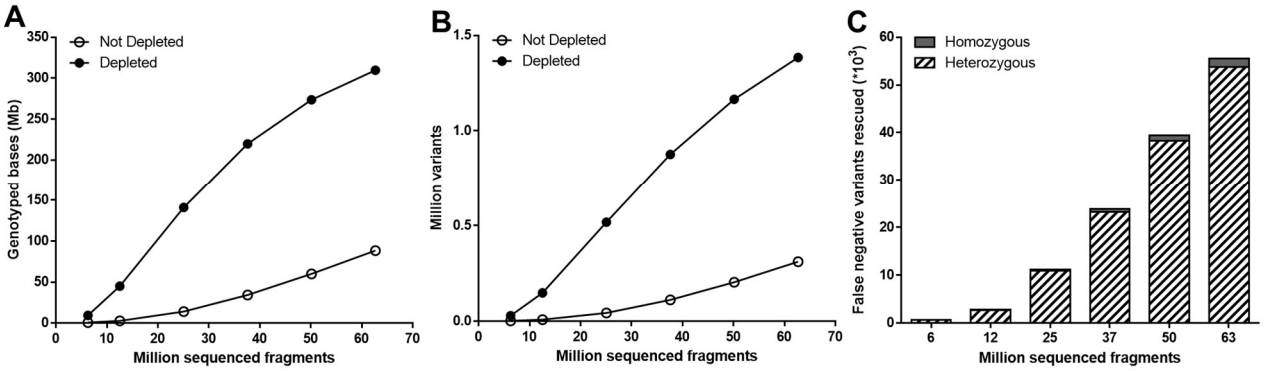

Figure 4. Genotyping performance on unique regions after CRISPR/Cas9-mediated repeat depletion. (A) Number of PASS positions within unique regions achieved with or without depletion, starting from different amounts of sequencing data. **(B)** Number of variants identified within unique regions with or without depletion, starting from different amounts of sequencing data. **(C)** Number of variants within unique regions that were located in a PASS position and identified in the depleted sample (1/1 or 1/0) but not in the non-depleted sample (0/0).

Figure S2. Examples of false negative variants rescued by CRISPR/Cas9-mediated repeat depletion. Integrative Genome Browser Visualization (IGV) of Illumina sequencing reads mapped at two genomic sites of ~40 bp (**A** and **B**) before (upper tracks) and after (lower tracks) CRISPR/Cas9-mediated repeat depletion. Two heterozygous variants have been called in the depleted sample but were overlooked in the depleted sample, despite their location in a genotypable position (PASS, DP ≥ 5).

Performance of CRISPR/Cas9-mediated repeat depletion on different samples and library types

Finally, we assessed the performance of CRISPR/Cas9-mediated repeat depletion on different lentil genotypes, multiplexing levels and library types (**Table S4**). Similar variations in the coverage of repetitive/unique regions and the number of mapped reads were observed when depleting multiplex libraries generated from a different cultivar (RB, Redberry) or from the closely-related species *L. orientalis* (**Figure 5A-B**) when compared to the original Castelluccio

cultivar (**Figure 1**). Furthermore, there was no significant difference in performance when library multiplexing was increased from 8-plex to 96-plex while maintaining the 1:10 Cas9:gRNA ratio and 1 ng of treated library per sample (**Figure 5C-D**), or when treating standard singleplex WGS libraries (**Figure 5E-F**). Overall, these results demonstrated that CRISPR/Cas9-mediated repeat depletion using the same gRNA set is at least equally effective when applied to a group of three lentil cultivars and one close wild relative, with the possibility to process individual samples or multiple samples simultaneously.

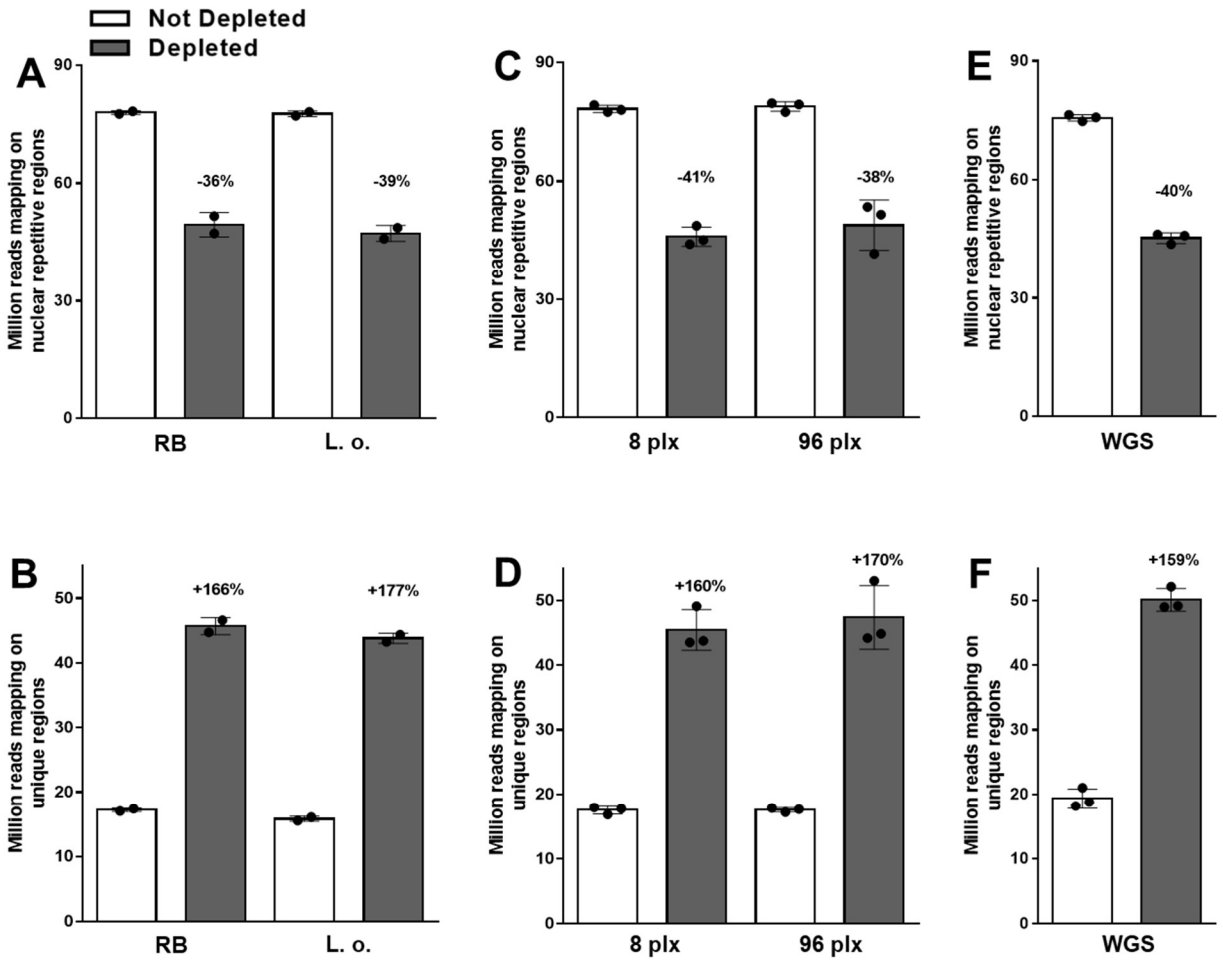

Figure 5. Variations of sequencing data distribution after CRISPR/Cas9-mediated repeat depletion in different lentil samples, multiplexing and library types. Sequencing reads mapping to the repetitive or unique regions before and after CRISPR/Cas9-mediated repeat depletion in (**A-B**) multiplex libraries generated from different lentil samples, namely *L. culinaris* cv. Redberry (RB) or *L. orientalis* (L. o.), (**C-D**) multiplex libraries containing 8 or 96 samples (plx) or (**E-F**) standard singleplex WGS libraries generated from the same samples. Data are means ± SE (n = 2–3) normalized for the same sequencing input (50 million fragments). Variation percentages observed following CRISPR/Cas9-mediated repeat depletion are reported above each condition.

DISCUSSION

In a typical GBS experiment, the data derived from repetitive DNA is directly proportional to the repeat content of the genome, representing ~40% and ~80% for ddRAD-seq and lcWGS in *L. culinaris*, respectively. However, these data are largely uninformative. The bigger the genome, the more sequencing costs are therefore wasted on repeats. The traditional solution is the capture and sequencing of coding regions (WES). However, we approached the problem from the opposite perspective by depleting repetitive elements from the large genome of *L. culinaris* using CRISPR/Cas9 technology. Similar methods have been used for RNA-Seq library normalization (29), metatrascriptomics (26), pathogen detection (26) and single-cell analysis (30). Here, the main challenge is that 85% of the ~3.2-Gb *L. culinaris* genome is repetitive DNA. The design of the gRNA array therefore required stringent multi-step filtering resulting in a set of ~600,000 gRNAs. To our knowledge, this is the first time CRISPR/Cas9 has been used with such a large number of gRNAs either *in vitro* or *in vivo*, representing a fundamental advance in the technology platform beyond the specific goals of our project. Despite these technical challenges, CRISPR/Cas9-mediated repeat depletion produced sequencing libraries with consistently lower proportions of repetitive DNA (~40% depletion) and enriched the unique regions, thus allowing the generation of more meaningful sequencing data. The depletion efficiency was comparable or even superior to the DSN method (25), but is also easier to standardize and allows a more specific targeting of repeats. CRISPR/Cas9-mediated repeat depletion allowed us to genotype up to 17-fold more bases on unique regions as compared to non-depleted libraries, and consequently we identified an average of 10-fold more genetic variants (up to 18-fold more). The shift in the distribution of sequencing data increased the coverage of unique regions by > 2.5-fold, and this was beneficial especially for the identification of heterozygous variants that would otherwise be missed due to unbalanced allelic sampling. Allele drop-out is a well-known cause of errors in GBS experiments (31). CRISPR/Cas9-mediated repeat depletion therefore improves the accuracy of genotyping experiments, especially in plants with highly heterozygous genomes.

Repetitive DNA in plant genomes can be divided into two broad categories: dispersed mobile elements and tandem/simple repeats. Dispersed mobile elements are made up of DNA transposons and retrotransposons, the most abundant of which are the LTR retrotransposons (32). Although mobile elements are not under the same selection pressure as genes, the degree of conservation across multiple copies of the same element is sufficiently high to allow the targeting of multiple copies with single gRNAs. The most efficient depletion (47%) was achieved for the most abundant LTR retrotransposon family (Gypsy, ~64% of the genome), followed by all the other LTR families (–15%). The cutting of LTR elements by Cas9 was responsible for almost the entire depletion observed at the genome-wide level, whereas the depletion of other mobile elements was negligible. This was probably due to the lower repetition of such elements, which translated into a poor cutting frequency in the final gRNA design, comprising only gRNAs with 25 targets. These targets were present at a typical density of < 8 gRNA/kbp, reducing the likelihood of depletion at less abundant repeat classes. A similar observation was reported for RNA-Seq libraries, where a gRNA every 50–100 nucleotides (10–20 gRNAs/kbp) achieves excellent depletion results (26). Simple and tandem repeats were also not depleted efficiently, although the gRNA density was close to 8. In these cases, the highly repetitive motifs of such sequences may have reduced the cutting efficiency of Cas9 (33). Given that LTR retrotransposons make up the majority of repeats in plant genomes (94% in *L. culinaris*) and are the principal cause of plant genome size variation (32,34), the design of gRNAs to target only LTR sequences may be the most efficient strategy to reduce the genome size in sequencing experiments. Future gRNA designs in other species and further optimization of the *L. culinaris* design should maximize the gRNA number on these most abundant elements, instead of dispersing the effort across the remaining repetitive fraction (< 10%). Another factor influencing the efficiency of CRISPR/Cas9-mediated repeat depletion was the dose of Cas9 and gRNAs. Doubling the dose improved repeat depletion by ~10%, so the gRNA target density and RNP concentration are two factors that can improve depletion efficiency, albeit with a slight increase in overall costs.

Organelle genomes often constitute a large fraction of DNA derived from plants (35). Although the mitochondrial and chloroplast genomes are smaller than the nuclear genome, they are present in multiple copies per cell, and they can represent > 20% of the total sequence data (36,37). CRISPR/Cas9-mediated repeat depletion has been shown to reduce the fraction of sequencing libraries derived from organelle genomes in ATAC-Seq experiments (38). Our method was efficient for the depletion of chloroplast DNA (by 67%) while the depletion of mitochondrial DNA was only marginal,

probably reflecting the 50-fold gRNA multiplicity used for the chloroplast genome compared to 4-fold for the mitochondrial genome. Still, in lentil, the fraction of sequencing data attributable to organelles was largely due to chloroplasts (88%), whose depletion was therefore sufficient to decrease the data mapping on organelles after Cas9 treatment (-65% overall). Although in the case of lentil the total sequencing data attributable to organelles was rather low (~1.3% in the not depleted libraries), the depletion of organelle DNA from sequencing libraries will be highly beneficial for organisms with a strongly unbalanced ratio of organelle vs nuclear DNA, such as *Cypripedium calceolus* (36) and *Haematococcus pluvialis* (37).

Although the efficiency of CRISPR/Cas9-mediated repeat depletion could be improved, our results demonstrated that depleted libraries were more informative than standard ones when normalized for the amount of sequencing data. The coverage of unique regions increased, allowing us to genotype more bases and to discover more variants. Alternatively, CRISPR/Cas9-mediated repeat depletion can be used to reduce sequencing costs (by ~75% in lentil), because the same number of genotyped bases (or detected variants) in unique regions can be detected with much less sequencing data. The investment required to produce the gRNA set targeting *L. culinaris* repeats was similar to that necessary for WES custom probe synthesis, but the depletion approach is not restricted to coding regions and can also seek variants in regulatory regions, which are excluded by WES.

The cost of library preparation for a GBS project can easily exceed the cost of sequencing in the case of small genomes and/or ultra-lcWGS, especially given the steadily falling price of sequencing. More recent library-preparation kits circumvent several lengthy steps that require expensive reagents, and allow large sample sets to be processed in multiplex reactions. We used the Twist 96-Plex Library Prep Kit (formerly iGenomX Riptide) that constructs Illumina NGS libraries by polymerase-mediated extension of barcoded random primers. This type of library is beneficial for GBS in general because random priming reduces uniform genome coverage but allows more reproducible sampling of the same sites across multiple samples (12). Most importantly, the kit is designed to process large numbers of samples (up to 96 simultaneously) at low costs and without advanced equipment (just a multichannel pipette). We estimated that the net cost to achieve 5-fold average coverage of the unique regions of the lentil genome by combining CRISPR/Cas9-mediated repeat depletion with a 96-plex Twist library is approximately US$90, made up of US$15 for library preparation, US$10 for depletion and US$65 for sequencing on a NovaSeq6000 S4 flowcell. This is similar to standard ddRAD-Seq costs (US$ 40–60) but can genotype at least an order of magnitude more bases by focusing on unique regions. CRISPR/Cas9-mediated repeat depletion combined with Twist multiplex libraries is therefore an effective strategy for genotyping projects involving hundreds or thousands of samples.

The repeat-to-unique shift in the profile of sequencing data allows to concentrate data on the same -relevant- portions across different libraries, thereby detecting a larger number of differences between samples. Overall this improves genotyping accuracy and reduces the fraction of missing data at the population level. This approach was successful in different cultivars of *L. culinaris*, and also in the closely-related species *L. orientalis.* This is important because genotyping experiments typically include distant/wild relatives and related species, from which it is possible to develop evolutionary studies and plan breeding experiments, including the introgression of characters of interest. For example, the INCREASE project features a collection of 2000 lentil single-seed descent (SSD) derived accessions that includes both cultivated varieties and local landraces (39). Further experiments could determine whether the gRNAs designed in this study are also suitable for the depletion of repeats in other leguminous species (Fabaceae) with very large and repetitive genomes, such as the pea (*P. sativum*, 3.92 Gb, 83% repetitive)(40) and faba bean (*Vicia faba* L., 12 Gb 79% repetitive)(41). Population studies, eQTL analysis, GWAS and pre-breeding programs are just some of the approaches that can benefit from CRISPR/Cas9-mediated repeat depletion. This method therefore has the potential to increase our genetic knowledge of plant species that are currently difficult to analyze without a significant economic investment due to the large genome size and high proportion of repetitive DNA.

MATERIAL AND METHODS

Multiplex-library preparation. We prepared 8-plex and 96-plex multiplex libraries according to the Twist 96-Plex Library Preparation Kit protocol (Twist Bioscience, South San Francisco, CA, USA) with the following modifications. For each sample, we denatured 100 ng of genomic DNA (25 ng/µl) at 98 °C for 1 min. Ultra-low (30%) GC random primer set A was used for the extension and termination reaction (Reaction A) followed by 8 and 9 cycles of PCR amplification for the 96-plex and 8-plex libraries, respectively. Final libraries were purified using Twist DNA Purification Beads (0.65x volume) and a second round of purification was applied to the supernatant using 10 µl of beads to achieve a median insert size of 500 bp. Libraries were quantified using the Qubit BR DNA kit and a Qubit device (Thermo Fisher Scientific, Waltham, MA, USA) and size distributions were assessed using a Tape Station System (Agilent Technologies, Santa Clara, CA, USA). Non-depleted libraries were pooled at equimolar concentrations and sequenced on a NovaSeq6000 instrument (Illumina, San Diego, CA, USA) to generate 150-bp paired-end reads.

Design of gRNAs. The gRNA set was designed by JumpCode Genomics (San Diego, CA, USA) against the repetitive regions of the *L. culinaris* CDC Redberry v2.0 reference genome (28) (https://knowpulse.usask.ca/Lens/culinaris). The available repeat annotation (transposable elements and tandem repeats) was integrated with the annotation of simple and tandem repeats identified by RepeatMasker v4.0.6 and Tandem Repeat Finder v4.9 using 2 7 7 80 10 50 2000 -d -h parameters to identify intervals for gRNA design (**Table S1**). Adjacent or overlapping intervals were collapsed into single intervals before design. As a first step, all 20 nt sequences with adjacent PAM sites for Cas9 (NGG) were identified in the target intervals. Second, the guides were filtered to exclude secondary structure, high and low GC content, homopolymers, dinucleotide repeats and low *in vitro* cleavage efficiency prediction scores (Azimuth algorithm; (42)). Third, the resulting guides were filtered to minimize off-target cleavage in unique regions of the genome by excluding guides that have complementary sites in genomic regions corresponding to genes and open-chromatin regions identified by ATAC-Seq (allowing for up to 3 mismatches). As a final step, and to reduce the number of guides in the set, guides were selected to have no less than 25 cleavage sites each and to maintain an inter-guide spacing of at least 500 bp. The final guide set, comprising 569,088 unique guides, was split into 11 pools for the purpose of synthesis. The number of copies of each guide varied and reflected the number of on-target cleavage sites for each guide. DNA oligonucleotides, consisting of the 5' bacteriophage T7 RNA polymerase promoter sequence followed by variable numbers of "G" nucleotides, the target-specific 20 nt guide sequence and, at the 3'end, the invariant single gRNA sequence, were synthesized in pools of up to 60,000 unique sequences using a microarray-based method (Agilent Technologies, Santa Clara, CA). The oligonucleotide pools were amplified by PCR, then converted to RNA by *in vitro* transcription using T7 RNA polymerase. The products of transcription were treated with DNase I and column purified to generate the final gRNA material. Pools 1–3, 5–8 and 10 contain only gRNAs targeting the nuclear genome. Pools 9 and 11 contain both nuclear and chloroplast genome gRNAs, and pool 4 is the only pool containing gRNAs that target the nuclear, chloroplast and mitochondrial genomes (**Table S2**).

Repeat depletion with JumpCode CRISPRclean. Repetitive regions were depleted using the Cas9 protein and the custom gRNA set described above, according to the Jumpcode CRISPRclean Ribosomal RNA Depletion from Human RNA-Seq Libraries for Illumina Sequencing protocol (Jumpcode Genomics, San Diego, CA, USA) with the following modifications. The input was 10 and 100 ng for the 8-plex and 96-plex libraries, respectively. Depletion was carried out either using all gRNA simultaneously or by splitting the gRNA pools into three groups based on cutting frequency, which were used sequentially in order of increasing cutting frequency (**Table S2**). The sequential depletion strategy was also conducted using the double amounts of gRNAs and Cas9. The reaction volume was 20 µl when using all gRNAs simultaneously or 26 µl for the sequential and double sequential protocols. The quantity of each gRNA pool per reaction is shown in **Table S1** and amounted to 620 ng in the simultaneous and sequential depletion reactions or 1240 ng in the double sequential depletion reaction. The Cas9 enzyme was diluted 1:5 in 1x Cas9 Buffer and 0.0029 µl was used per ng gRNA. The reactions were incubated at 37 °C and libraries were treated in the presence of gRNAs for a total of 1 h (simultaneous depletion protocol) or 3 h (sequential and double sequential depletion protocols, with the sequential gRNA pools added at 1-h intervals). The depleted samples were then size selected using 0.6x volume of AMPure XP Beads (Beckman Coulter, Brea, CA, USA). Libraries were amplified with 10 and 6 PCR cycles for the 8-plex and 96-plex libraries, respectively before final purification with 60µl (0.6x volume) AMPure XP Beads. The concentrations of depleted libraries were measured using the

Qubit system and size distributions were assessed on a Tape Station System as described above. Depleted libraries were pooled at equimolar concentrations and sequenced on a NovaSeq6000 instrument to generate 150-bp paired-end reads.

Data analysis and variant calling. Raw read quality was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the multiplex libraries were demultiplexed using fgbio v1.3.0 DemuxFastqs, assigning fragments by exploiting the unique sample identifier included during first-strand synthesis. The raw reads were then quality filtered and the Illumina sequencing adapters removed using scythe v0.991 and syckle v1.33, respectively. Filtered reads were aligned to the *L. culinaris* v2.0 reference genome using bwa-mem v2.2.1 and the resulting alignments were converted to bam files and sorted using samtools v1.13. PCR-derived duplicates were removed using the GATK MarkDuplicates tool v4.1.7.0 and overlapping portions of the paired-end reads were clipped using the fgbio v1.3.0 ClipBam tool. The cleaned bam files were used to calculate coverage depth, breadth and fraction of PASS bases (at ≥ 5x) using bedtools v2.30.0 genomecov and GATK v3.8 CallableLoci, respectively. The number of reads aligning to the reference genome and to different regions of interest was calculated using samtools v1.13 with option -c to discard reads with a 2308 sam flag in order to consider only the primary alignment, thus omitting repetitive counts of the same multimapping reads. When necessary, sequencing data were normalized to a pre-defined number of input fragments using seqtk sample v1.3.

The coverage variation between the depleted and non-depleted libraries was calculated for target regions with at least one target site using the following formula:

$$Coverage_variation = \log_2 \frac{Mean_cov(depleted)}{Mean_cov(not\ depleted)}$$

In those regions with a mean coverage of 0 in the depleted libraries, the value was set to the lowest calculated value of –12.1325. The coverage variation plot was generated using the ggplot package in R.

Genomic variants were identified using GATK HaplotypeCaller v4.1.7.0 with the parameters "--min-base-quality-score 20 -ERC GVCF". Individual gVCF files were merged using GATK GenomicsDBImport v4.1.7.0 and the final VCF file was generated using GATK GenotypeGVCFs v4.1.7.0. Variant filtration was achieved using GATK hard filters (https://gatk.broadinstitute.org/hc/en-us/articles/360037499012?id=3225).

Acknowledgments. This research was funded by the EU H2020 project INCREASE (www.pulsesincrease.eu). We acknowledge the Twist and Jumpcode Genomics teams for excellent technical support.

Author Contributions. Conceptualization MR, MD and RP; Methodology MR, LM and MD; Software LM and GL; Investigation LDA, EBE, GC and FL; Formal analysis MR, LM and LDA; Validation GF, LN and LV; Resources KB, LR and DJK; Data Curation LM and GL; Writing - Original Draft MR; Writing - Review & Editing LM, LDA and MD; Visualization MR and LM; Supervision MR and MD; Project administration MR; Funding acquisition EBI, MD and RP.

Conflict of Interest. Authors MR and MD are partners of Genartis srl. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1. Bellucci E, Mario Aguilar O, Alseekh S, Bett K, Brezeanu C, Cook D, et al. The INCREASE project: Intelligent Collections of food-legume genetic resources for European agrofood systems. Plant Journal. 2021 Nov 1;108(3):646–60.

2. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012 May 31;7(5).

3. Pavan S, Delvento C, Ricciardi L, Lotti C, Ciani E, D'Agostino N. Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies. Vol. 11, Frontiers in Genetics. Frontiers Media S.A.; 2020.

4. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays. 2013 Sep;35(9):780–6.

5. Tanaka N, Shenton M, Kawahara Y, Kumagai M, Sakai H, Kanamori H, et al. Investigation of the Genetic Diversity of a Rice Core Collection of Japanese Landraces using Whole-Genome Sequencing. Plant Cell Physiol. 2021 Feb 4;61(12):2087–96.

6. Friel J, Bombarely A, Fornell CD, Luque F, Fernández-Ocaña AM. Comparative analysis of genotyping by sequencing and whole-genome sequencing methods in diversity studies of olea europaea l. Plants. 2021 Nov 1;10(11).

7. Deng T, Zhang P, Garrick D, Gao H, Wang L, Zhao F. Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. Front Genet. 2022 Jan 3;12.

8. Zan Y, Payen T, Lillie M, Honaker CF, Siegel PB, Carlborg Ö. Genotyping by low-coverage whole-genome sequencing in intercross pedigrees from outbred founders: A cost-efficient approach. Genetics Selection Evolution. 2019 Aug 14;51(1).

9. Wang J, Sun G, Ren X, Li C, Liu L, Wang Q, et al. QTL underlying some agronomic traits in barley detected by SNP markers. BMC Genet. 2016 Jul 7;17(1).

10. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Vol. 12, Nature Reviews Genetics. 2011. p. 499–510.

11. Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJA, Huvenaars KHJ, et al. Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One. 2012 May 25;7(5).

12. Siddique A, Suckow G, Ordoukhanian P, et al. RipTide High Throughput NGS Library Prep for Genotyping in Populations. J Biomol Tech. 2019;30(Suppl):S35-S36.

13. Yan H, Haak DC, Li S, Huang L, Bombarely A. Exploring transposable element-based markers to identify allelic variations underlying agronomic traits in rice. Plant Commun. 2022 May 9;3(3).

14. Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. Crop genome sequencing: Lessons and rationales. Vol. 16, Trends in Plant Science. 2011. p. 77–88.

15. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007 Dec;39(12):1522–7.

16. Ogutcen E, Ramsay L, von Wettberg EB, Bett KE. Capturing variation in Lens (Fabaceae): Development and utility of an exome capture array for lentil. Appl Plant Sci. 2018 Jun 1;6(7).

17. He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, et al. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet. 2019 May 1;51(5):896–904.

18. Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, et al. Widespread long-range cis-regulatory elements in the maize genome. Nat Plants. 2019 Dec 1;5(12):1237–49.

19. Wang P, Xiong Y, Gong R, Yang Y, Fan K, Yu S. A key variant in the cis-regulatory element of flowering gene Ghd8 associated with cold tolerance in rice. Sci Rep. 2019 Dec 1;9(1).

20. Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: Charting functional regulatory maps in plants. Nucleic Acids Res. 2020 Jan 1;48(D1):D1104–13.

21. Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, et al. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 2004;32(3).

22. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling [Internet]. 2014. Available from: http://www.biomedcentral.com/1471-2164/15/419

23. Miller DFB, Yan PS, Buechlein A, Rodriguez BA, Yilmaz AS, Goel S, et al. A new method for stranded whole transcriptome RNA-seq. Methods. 2013 Sep 15;63(2):126–34.

24. Ichida H, Abe T. An improved and robust method to efficiently deplete repetitive elements from complex plant genomes. Plant Science. 2019 Mar 1;280:455–60.

25. Matvienko M, Kozik A, Froenicke L, Lavelle D, Martineau B, Perroud B, et al. Consequences of Normalizing Transcriptomic and Genomic Libraries of Plant Genomes Using a Duplex-Specific Nuclease and Tetramethylammonium Chloride. PLoS One. 2013 Feb 8;8(2).

26. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. 2016 Mar 4;17(1).

27. Technology Overview Version 1.2 Harnessing CRISPR to boost NGS sensitivity with CRISPRclean ™. 2021.

28. Ramsay L, Koh CS, Kagale S, Gao D, Kaur S, Haile T, et al. Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. Available from: https://doi.org/10.1101/2021.07.23.453237

29. Prezza G, Heckel T, Dietrich S, Homberger C, Westermann AJ, Vogel J. Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. 2020; Available from: http://www.rnajournal.org/cgi/doi/10.1261/rna.

30. Parchman TL, Jahner JP, Uckele KA, Galland LM, Eckert AJ. RADseq approaches and applications for forest tree genetics. Vol. 14, Tree Genetics and Genomes. Springer Verlag; 2018.

31. Cooke TF, Yee MC, Muzzio M, Sockell A, Bell R, Cornejo OE, et al. GBStools: A Statistical Method for Estimating Allelic Dropout in Reduced Representation Sequencing Data. PLoS Genet. 2016 Feb 1;12(2).

32. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Vol. 65, Annual Review of Plant Biology. Annual Reviews Inc.; 2014. p. 505–30.

33. Müller Paul H, Istanto DD, Heldenbrand J, Hudson ME. CROPSR: an automated platform for complex genome-wide CRISPR gRNA design and validation. BMC Bioinformatics. 2022 Dec 1;23(1).

34. Lee SI, Kim NS. Transposable Elements and Genome Size Variations in Plants. Genomics Inform. 2014;12(3):87.

35. Sakamoto W, Takami T. Chloroplast DNA dynamics: Copy number, quality control and degradation. Vol. 59, Plant and Cell Physiology. Oxford University Press; 2018. p. 1120–7.

36. Gargiulo R, Kull T, Fay MF. Effective double-digest RAD sequencing and genotyping despite large genome size. Mol Ecol Resour. 2021 May 1;21(4):1037–55.

37. Ren Q, Wang Y chu, Lin Y, Zhen Z, Cui Y, Qin S. The extremely large chloroplast genome of the green alga Haematococcus pluvialis: Genome structure, and comparative analysis. Algal Res. 2021 Jun 1;56.

38. Montefiori L, Hernandez L, Zhang Z, Gilad Y, Ober C, Crawford G, et al. Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. Sci Rep. 2017 Dec 1;7(1).

39. Guerra-García A, Gioia T, von Wettberg E, Logozzo G, Papa R, Bitocchi E, et al. Intelligent Characterization of Lentil Genetic Resources: Evolutionary History, Genetic Diversity of Germplasm, and the Need for Well-Represented Collections. Curr Protoc. 2021 May 1;1(5).

40. Kreplak J, Madoui MA, Cápal P, Novák P, Labadie K, Aubert G, et al. A reference genome for pea provides insight into legume genome evolution. Nat Genet. 2019 Sep 1;51(9):1411–22.

41. Jayakodi M, Golicz AA, Kreplak J, Fechete LI, Angra D, Bednář P, et al. The giant diploid faba genome unlocks variation in a global protein crop. Available from: https://doi.org/10.1101/2022.09.23.509015

42. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016 Feb 1;34(2):184–91.