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

Marzia Rossato[1,2*$], Luca Marcolungo[1*], Luca De Antoni[1], Giulia Lopatriello[1], Elisa Bellucci[3], Gaia Cortinovis[3], Giulia Frascarelli[3], Laura Nanni[3], Elena Bitocchi[3], Valerio Di Vittori[3], Leonardo Vincenzi[1], Filippo Lucchini[1], Kirstin E. Bett[4], Larissa Ramsay[4], David James Konkin[5], Massimo Delledonne[1,2$] and Roberto Papa[3$]

*equal contribution

[$]corresponding authors

Marzia.rossato@univr.it; Luca.marcolungo@univr.it; Luca.deantoni@univr.it; giulia.lopatriello@univr.it; e.bellucci@staff.univpm.it; gaia.cortinovis93@gmail.com; g.frascarelli@pm.univpm.it; l.nanni@staff.univpm.it; e.bitocchi@staff.univpm.it; v.divittori@staff.univpm.it; leonardo.vincenzi@univr.it; filippo.lucchini@univr.it; k.bett@usask.ca; l.ramsay@usask.ca; David.Konkin@nrc-cnrc.gc.ca; massimo.delledonne@univr.it; r.papa@staff.univpm.it

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

[2]Genartis s.r.l., Via IV Novembre 24, 37126, Verona, Italy

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

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

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

Running Title: Reduced genome representation by repeats exclusion

ABSTRACT

High-throughput genotyping enables 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 accessions. Sequencing-based approaches for genotyping are progressively replacing traditional genotyping methods due to the lower ascertainment bias. However, genome-wide genotyping based on 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% consisting of repeats, thus concentrating the sequencing data on coding and regulatory regions (single-copy regions). We designed a custom set of 566,766 gRNAs targeting 2.9 Gbp of repeats and excluding 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 single-copy regions by ~2.6-fold. When analyzing 25 million fragments, this repeat-to-single-copy shift in the sequencing data increased the number of genotyped bases of ~10-fold compared to non-depleted libraries. In the same condition, we were also able to identify ~12-fold more genetic variants in the single-copy 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 demonstrated that CRISPR/Cas9-driven repeat depletion focuses sequencing data on meaningful genomic regions, thus improving high-density and genome-wide genotyping in large and repetitive genomes.

KEYWORDS: repetitive elements, CRISPR/Cas9, sequencing-based genotyping, 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) (Bellucci et al. 2021), 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 (Bellucci et al. 2021). 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 (Peterson et al. 2012).

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 (Pavan et al. 2020). 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 (Lachance and Tishkoff 2013).

More recently, next generation sequencing (NGS) has provided an opportunity to discover genome-wide variants in a less biased manner. Sequencing-based approaches for genotyping involves low coverage (5–10x) whole-genome sequencing (lcWGS), allowing the characterization of several million variants (Tanaka et al. 2021; Friel et al. 2021). 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 (Deng et al. 2022; Zan et al. 2019; Wang et al. 2016). Alternatively, sequencing costs are often minimized by reduced-representation sequencing, which comprises methods such as genotyping by sequencing (GBS) (Elshire et al. 2011), restriction site-associated DNA sequencing (RAD-Seq) (Davey et al. 2011; Baird et al. 2008) and double-digest RAD-Seq (ddRAD-seq) (Truong et al. 2012; Peterson et al. 2012). 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, with a small fraction of analysed and genotyped bases (Pavan et al. 2020) that may not provide sufficient marker density and depth to confidently identify variants under selection in large genomes (Guerra-García et al. 2021). The resolution can be increased without significantly greater costs in sample prep by using

78 the Twist 96-Plex Library Prep Kit (formerly iGenomX Riptide Kit) to generate multiplexed libraries, allowing 96–960
79 samples to be processed simultaneously and resulting in the non-random sampling of millions of genomic positions
80 (Siddique et al. 2019).

81 Despite the advantages of sequencing-based genotyping over SNP arrays, one common disadvantage is that sequencing
82 methods generally do not distinguish between repetitive (low-complexity) and single-copy (high-complexity) regions, the
83 latter comprising coding and regulatory regions that are the main targets of natural selection and thus the focus of most
84 genotyping projects. In contrast, low-complexity regions of plant genomes mainly comprise transposable elements,
85 simple sequence repeats and tandem repeats. Transposable elements play a key role in genome evolution, but the
86 analysis of such regions is technically challenging and largely uninformative in genotyping studies, unless dedicated
87 analysis workflows are applied (Yan et al. 2022). Mapping reads to transposable/repetitive elements can result in low-
88 quality alignments that hinder the calling of accurate genotypes, which is a consistent challenge particularly for those
89 plant species with large genomes, where repetitive elements account for up 90% of the total DNA. This includes many
90 domesticated crops such as corn (*Zea mays*), wheat (*Triticum* spp.), lentil (*Lens culinaris*) and onion (*Allium cepa*) (Feuillet
91 et al. 2011). One strategy to address this issue is whole exome sequencing (WES), which selects coding regions for
92 preferential sequencing (Hodges et al. 2007) as shown in lentil, wheat and barley (Ogutcen et al. 2018; He et al. 2019).
93 However, WES only focuses on coding sequences and thus overlooks regulatory elements, which are equally important
94 as sources of genetic diversity (Ricci et al. 2019; Wang et al. 2019; Tian et al. 2020).

95 Ideally, lcWGS could be focused on the most complex parts of the genome, avoiding wasted effort on the sequencing of
96 repetitive elements. This could be achieved by using enzymes that enable target enrichment by depleting unwanted
97 sequences from NGS libraries. For example, the duplex-specific nuclease (DSN) selectively digests double-stranded DNA
98 molecules, and can be used to eliminate highly abundant sequences in a controlled denaturation-reassociation reaction
99 (Zhulidov et al. 2004). This method has been used in RNA-Seq analysis to remove abundant transcripts (Zhao et al. 2014;
100 Miller et al. 2013) and, just occasionally, also to delete repetitive elements in DNA-Seq libraries generated from plant
101 genomes (Ichida and Abe 2019; Matvienko et al. 2013). However, (Matvienko et al. 2013)DSN can also remove
102 informative repetitive elements, such as the coding sequences of abundant gene families, which are particularly relevant
103 in polyploid plants arising from whole genome duplication events (Matvienko et al. 2013). More recently, the CRISPR/Cas9
104 (clustered regularly interspaced short palindromic repeats and CRISPR-associated nuclease 9) system has been used for
105 the selective depletion of unwanted genome fractions from sequencing libraries (Gu et al. 2016). The Cas9 enzyme can

106 be programmed to cut library fragments by designing specific guide-RNA sequences targeting the unwanted sequences.
107 Subsequently, only intact fragments -retaining adapters at both ends- can be effectively amplified by PCR and generate
108 productive clusters on a sequencing flow-cell. The DASH approach (depletion of abundant sequences by hybridization)
109 involved the use of Cas9 to exclude ribosomal RNA (rRNA) sequences from RNA-Seq libraries and to remove DNA from
110 common pathogens in order to detect rare pathogens in metagenomic samples (Gu et al. 2016). A similar technology has
111 been recently commercialized under the name "CRISPRclean" by JumpCode Genomics (JumpCode Genomics 2021).

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

117

118 RESULTS

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

120 We designed a custom set of gRNAs to deplete the repetitive DNA content of the *L. culinaris* CDC Redberry genome
121 (Ramsay et al. 2021), targeting transposable elements (totaling 3.1 Gbp of sequence across the whole genome,
122 corresponding to 82.5% of its size), simple sequence repeats (58 Mbp, 1.5%) and tandem repeats (13 Mbp, 0.4%) in the
123 nuclear genome, as well as the entire mitochondrial genome (mtDNA, 489 kbp) and chloroplast genome (cpDNA, 118
124 kbp) (Supplemental Table S1). We excluded repetitive DNA that overlapped with functional regions such as annotated
125 genes (185 Mbp, 5%) and putative regulatory regions identified using ATAC-Seq data (78 Mbp, 2%) (Supplemental Table
126 S1). All nuclear DNA outside the gRNA target regions is hereafter defined as single-copy . The final design comprised
127 566,766 gRNAs with at least 25 recognition sites, potentially targeting 2.9 Gbp (77%) of the *L. culinaris* nuclear genome
128 and 93.5% of its repetitive regions when using a sequencing library with 500-bp inserts (Supplemental Table S2 and
129 Supplemental File S1). An additional 2,366 gRNAs targeted the mitochondrial and chloroplast genomes (Supplemental
130 Table S2). The gRNAs were assigned to 11 pools based on their cutting frequency in the *L. culinaris* genome (Supplemental
131 Table S2).

132 The custom gRNAs were tested on three Twist 8-plex libraries, allowing the reproducible sampling of the same genomic
133 regions by random priming during first-strand DNA synthesis. Each 8-plex library comprised replicates of three distinct *L.*
134 *culinaris* samples (cv. Castelluccio) (**Supplemental Table S3**). Cas9/gRNAs ribonucleoprotein (RNP) complexes (1:2.5
135 protein/gRNA ratio) were generated, and gRNAs with more target sites in the genome were used at higher relative
136 concentrations in the final reaction (**Supplemental Table S2**). Depletion reactions in the presence of RNP complexes were
137 conducted either using all gRNAs simultaneously or by splitting the gRNA pools into three groups based on cutting
138 frequency (**Supplemental Table S2**) and using the groups sequentially, starting with the lowest cutting frequency.
139 Depleted and non-depleted libraries were sequenced, generating 91 million fragments on average (**Supplemental Table**
140 **S4**). The sequencing data were normalized at ~50 million fragments per library in order to compare the proportion of
141 reads mapping on repetitive and single-copy regions of the nuclear genome and on the organelle genomes (**Figure 1**).
142 The number of reads mapping to repetitive regions (total repeats, nuclear repeats, mtDNA and cpDNA) was significantly
143 lower in the depleted libraries compared to the non-depleted libraries, with the sequential depletion strategy using three
144 gRNA groups performing best and depleting 37.7% of the repetitive DNA (**Figure 1A**). The results were similar when
145 considering only the nuclear repetitive regions (37.2% depletion) (**Figure 1B**). A small fraction of total reads mapped to
146 the organelle genomes (~1%). Both depletion strategies were similarly effective in the chloroplast genome, resulting in
147 ~78.5% depletion (**Figure 1C**). In contrast, no significant depletion was observed in the mitochondrial genome (**Figure**
148 **1D**). In parallel, the sequential depletion strategy achieved a 130% increase in the number of reads mapping to single-
149 copy regions, from 17.5 to 40.4 million (**Figure 1E**). Given that the concentration of Cas9 RNPs influences the cutting
150 efficiency (Gu et al. 2016), we repeated the sequential depletion strategy using double amount of Cas9 and gRNAs. This
151 modified the read distribution further, achieving 41.2% depletion of nuclear repeats and a 160% increase in reads mapped
152 to single-copy regions. The sequential depletion strategy with double RNPs was therefore the most efficient, and was
153 applied in all subsequent experiments. Overall, our results demonstrated that the custom gRNA set and Cas9 effectively
154 targeted fragments containing repetitive DNA sequences and depleted them in the resulting sequencing libraries. **Figure**
155 **2A-B** shows the alignments of reads at two representative genomic regions, confirming that less sequencing data was
156 assigned to regions of repetitive DNA and more reads were mapped to single-copy parts of the genome ().

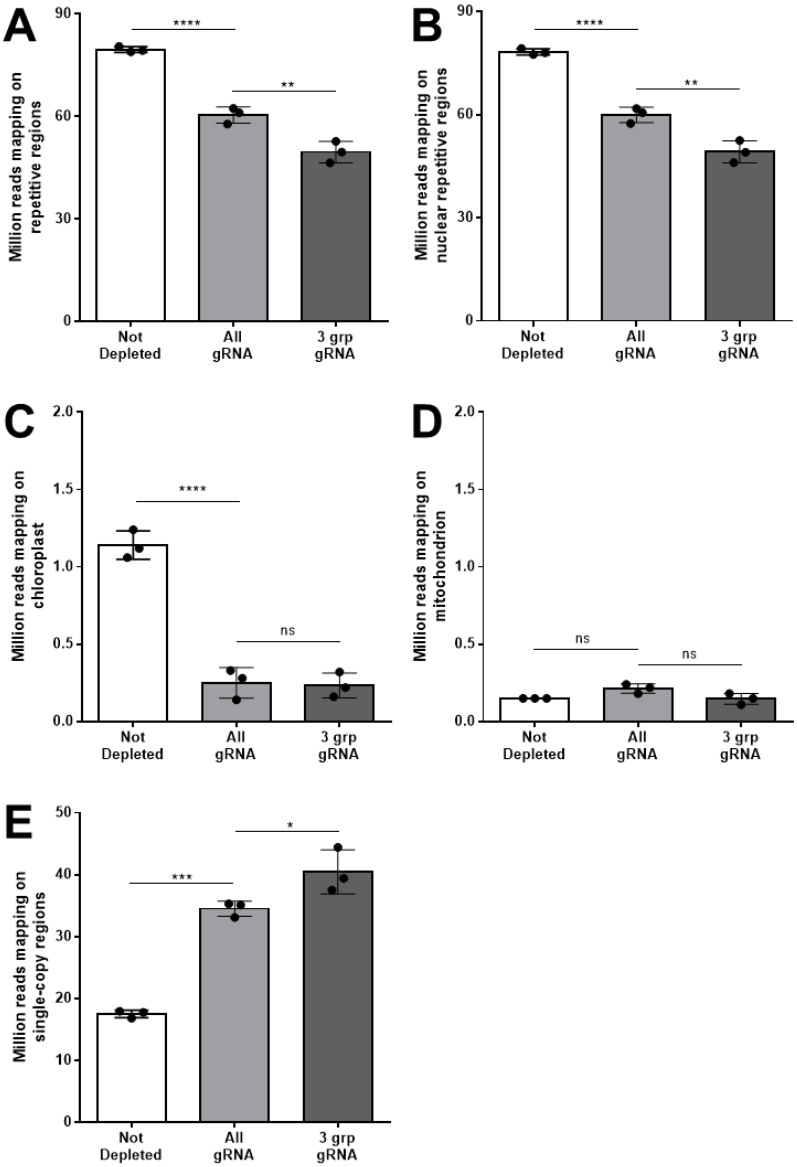


Figure 1. Distribution of mapped reads after CRISPR/Cas9-mediated repeat depletion. Libraries of *L. culinaris* cv. Castelluccio DNA (8-plex) were depleted using the custom gRNA set and Cas9. The gRNAs were used simultaneously (All) or were split into three groups that were used sequentially in order 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 single-copy 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).

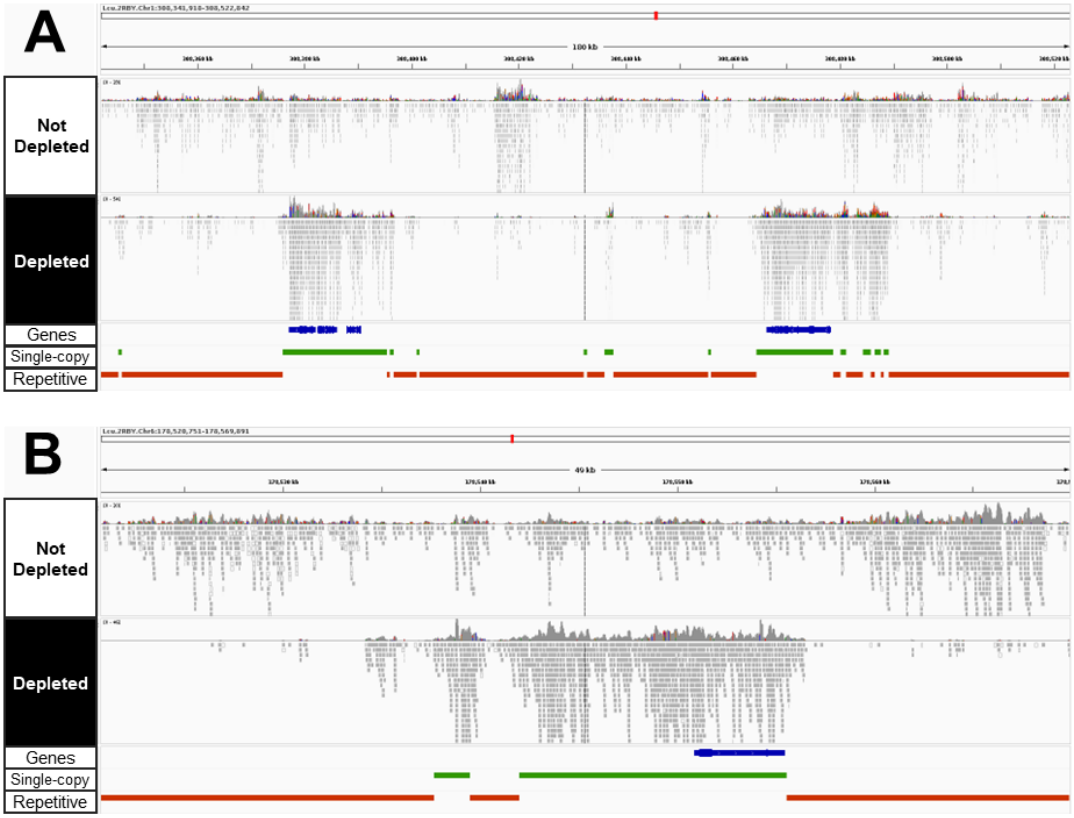
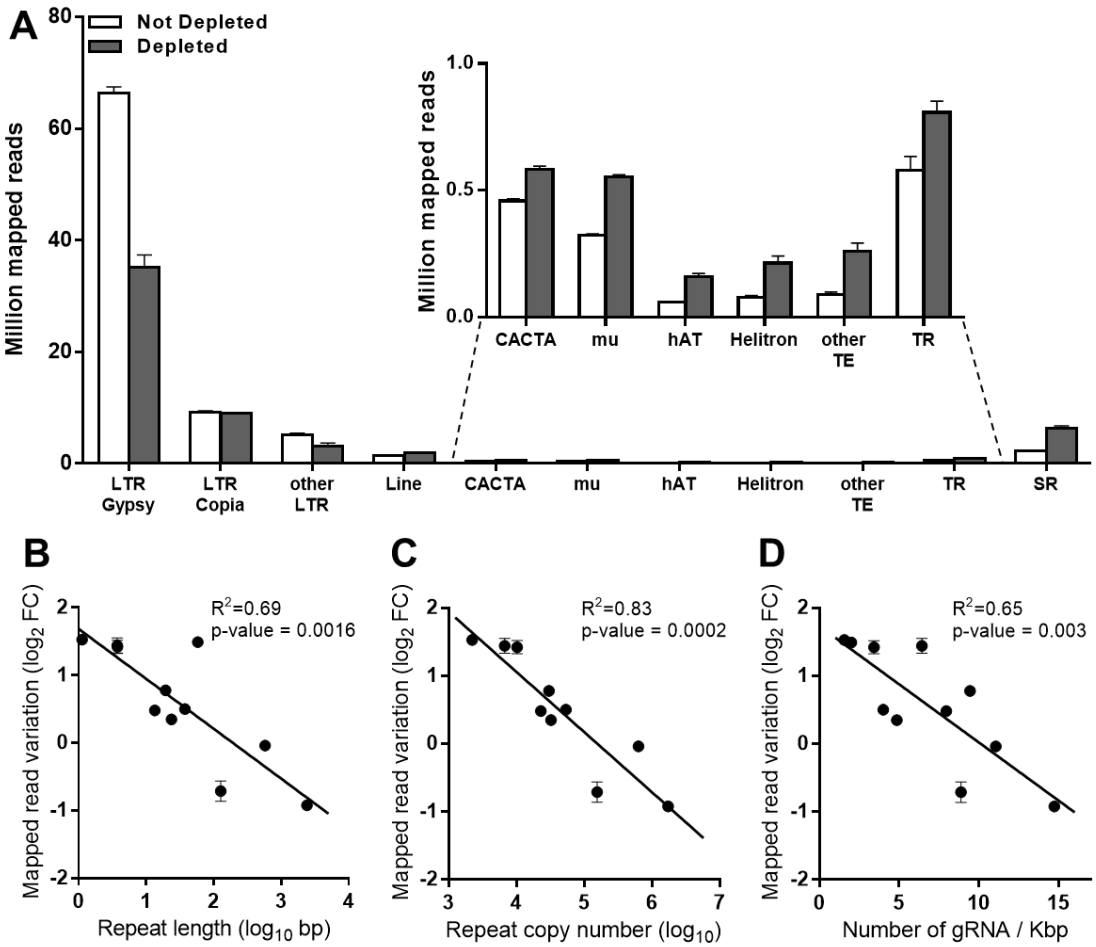


Figure 2. Reads mapped to repetitive or single-copy regions of the lentil genome with or without CRISPR/Cas9-mediated repeat depletion. Integrative Genome Browser Visualization (IGV) of Illumina sequencing data mapped to two representative genomic sites of ~180 and ~50 kbp. Tracks in blue, green and red represent annotated genes, single-copy 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 (Ramsay et al. 2021), 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 **Supplemental 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 (**Supplemental 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

179 (**Figure 3B-C** and **Supplemental Table S5**). Given that the number of gRNAs targeting each repeat class increased
180 proportionally with the repeat size and occurrence, the most efficiently depleted repetitive elements also featured a
181 higher density of gRNA targets (**Figure 3D**). There was a significant correlation between the variation of mapping reads
182 following depletion and the gRNA density over the whole target region when considering each single cut site in the
183 genome (**Supplemental Figure S1**). In particular, target regions with a density > 8 gRNAs/kbp showed a read reduction in
184 85% of cases (**Supplemental Figure S1**) whereas regions targeted by < 8 gRNAs/kbp usually showed limited or no depletion
185 (**Supplemental Figure S1**). We therefore concluded that the depletion efficiency across different repeat classes was
186 dependent on the density of gRNA targets.



187

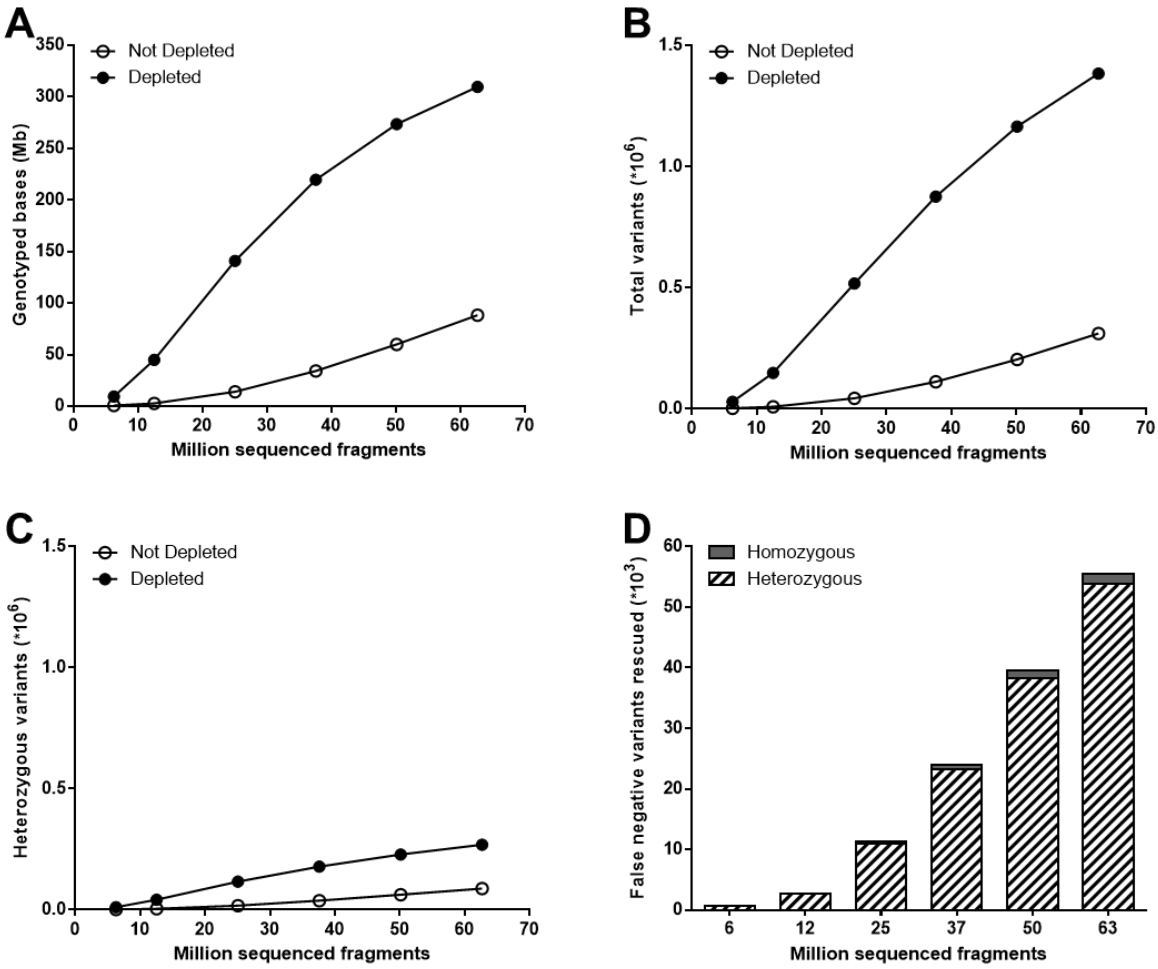
188 **Figure 3. Sequencing data distribution after CRISPR/Cas9-mediated repeat depletion according to the class of nuclear repeat.** (**A**) Number of reads (in
189 millions) mapping to different nuclear repeat classes with or without CRISPR/Cas9-mediated repeat depletion: Ty3-Gypsy (LTR Gypsy), Ty1-Copia (LTR
190 Copia), other LTR, Line, CACTA, mu, hAT, Helitron transposons, other transposable elements with abundance < 0.1% (Harbinger, mariner, Sine), tandem
191 repeats (TR) and simple repeats (SR). Correlation between the variation of mapped reads following CRISPR/Cas9-mediated repeat depletion on the

192 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
193 = 3). FC – fold change.

194

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

196 Next, we investigated the impact of CRISPR/Cas9-mediated repeat depletion on the number of genomic positions in the
197 single-copy regions where a base can be reliably genotyped (PASS at a depth of ≥ 5 reads). For this analysis, sequencing
198 data generated from depleted and non-depleted *L. culinaris* cv. Castelluccio samples were downsampled from 62 to 6
199 million fragments to mimic lcWGS and ultra-lcWGS. Consistently more bases were genotyped within the single-copy
200 regions of the depleted samples over the whole range considered (from ~3.5 to ~13-fold), with the highest gains at the
201 lowest amounts of sequencing data (6 to 25 million fragments) (**Figure 4A**). Because a genotyped position does not
202 necessarily allow the variant to be identified (this also depends on allele coverage), we also determined the impact of
203 repeat depletion on variant calling. Following depletion, the total number of variants identified in the single-copy regions
204 increased significantly from ~4.5 to ~18-fold, with a delta of ~25,000 to ~1 million more variants identified in the depleted
205 sample (**Figure 4B**). Also the number of heterozygous variants identified was increased, although these constituted a
206 minor fraction of total variants, as lentil is an autogamous species (**Figure 4C**). This allowed us to identify from ~650 up
207 to ~50,000 variant positions that would be erroneously classified as reference without the depletion (**Figure 4D**). These
208 false negative variants in the non-depleted sample were not called due to allelic imbalance caused by the low coverage
209 (**Supplemental Figure S2**). Consistently, most of these false negative variants rescued by depletion were heterozygous
210 (~97%, **Figure 4D**).



211

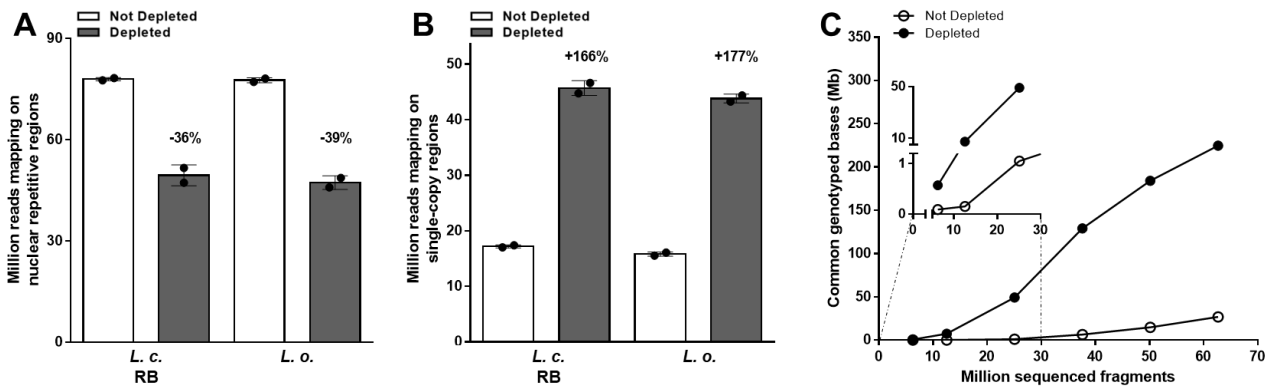
212 **Figure 4. Genotyping performance on single-copy regions with or without CRISPR/Cas9-mediated repeat depletion starting from different amounts of**
213 **sequencing data. (A)** Number of genotyped positions . **(B)** Number of total variants identified . **(C)** Number of heterozygous variants identified. **(D)**
214 Number of variants that were genotyped as reference (0/0) in the non-depleted sample and identified as homozygous (1/1) or heterozygous (1/0)
215 alternative in the depleted sample . N=3 at 6, 12 and 25 million fragments, N=2 at 37, 50 and 63 million fragments.

216

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

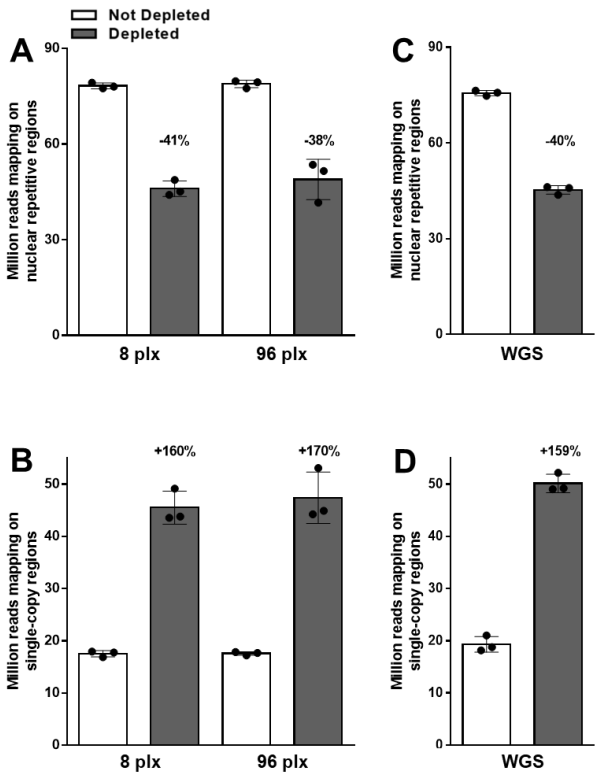
218 Finally, we assessed the performance of CRISPR/Cas9-mediated repeat depletion on different lentil genotypes,
219 multiplexing levels and library types (Supplemental Table S3 and Supplemental Table S4). Similar variations in the
220 coverage of repetitive/single-copy regions and the number of mapped reads were observed when depleting multiplex
221 libraries generated from a different cultivar (RB, Redberry) or from the closely-related species *L. orientalis* (Figure 5A-B)
222 when compared to the original Castelluccio cultivar (Figure 1). To assess the impact of depletion when comparing
223 different samples, these data were downsampled as described above, genotyped positions were identified and

224 intersected with those of Castelluccio samples. Depletion improved the genotyping reproducibility, as the number of
225 genotyped positions in common between all analyzed samples, within the single-copy regions, was consistently higher
226 than in the condition without depletion (**Figure 5C**). Finally, there was no significant difference in the performance of
227 CRISPR/Cas9-mediated repeat depletion when library multiplexing was increased from 8-plex to 96-plex while
228 maintaining the 1:2.5 Cas9:gRNA ratio and 1 ng of treated library per sample (**Figure 6A-B**), or when treating standard
229 singleplex WGS libraries (**Figure 6C-D**). Overall, these results demonstrated that CRISPR/Cas9-mediated repeat depletion
230 using the same gRNA set is at least equally effective when applied to a group of two lentil cultivars and one close wild
231 relative, providing an improved genotyping reproducibility and the possibility to process individual samples or multiple
232 samples simultaneously.



233

234 **Figure 5. Performances of CRISPR/Cas9-mediated repeat depletion in different lentil samples.** (**A-B**) Sequencing reads mapping to the repetitive or
235 single-copy regions with or without CRISPR/Cas9-mediated repeat depletion in multiplex libraries generated from different lentil samples, namely *L.*
236 *culinaris* cv. Redberry (*L. c.* RB) or *L. orientalis* (*L. o.*). Data are means ± SE (n = 2) normalized for the same sequencing input (50 million fragments).
237 Variation percentages observed following CRISPR/Cas9-mediated repeat depletion are reported above each condition. (**C**) Number of genotyped
238 positions in common between all samples analyzed in the study (3 distinct samples of *L. culinaris* cv. Castelluccio, 1 sample of *L. culinaris* cv. Redberry,
239 and 1 samples of *L. orientalis*), within the single-copy regions. N=5 at 6, 12 and 25 million fragments, N=4 at 37, 50 and 63 million fragments.



240

241 **Figure 6. Performances of CRISPR/Cas9-mediated repeat depletion at different multiplexing level and library types.** Sequencing reads mapping to the
242 repetitive or single-copy regions with or without CRISPR/Cas9-mediated repeat depletion in (**A-B**) multiplex libraries containing 8 or 96 samples (plx) or
243 (**E-F**) standard singleplex WGS libraries generated from one *L. culinaris* cv. Castelluccio, one *L. culinaris* cv. Redberry and one L. orientalis sample. Data
244 are means ± SE (n = 3) normalized for the same sequencing input (50 million fragments). Variation percentages observed following CRISPR/Cas9-
245 mediated repeat depletion are reported above each condition.

246

247 DISCUSSION

248 In a typical genotyping experiment based on sequencing, the data derived from repetitive DNA is directly proportional to
249 the repeat content of the genome, however these data are largely uninformative. The bigger the genome, the more
250 sequencing costs are therefore wasted on repeats. The traditional solution is the capture and sequencing of coding
251 regions (WES). However, we approached the problem from the opposite perspective by depleting repetitive elements
252 from the large genome of *L. culinaris* using CRISPR/Cas9 technology. Similar methods have been used for RNA-Seq library
253 normalization (Prezza et al. 2020), metatrascriptomics (Gu et al. 2016), pathogen detection (Gu et al. 2016; Le et al. 2021)
254 and single-cell analysis (Homberger et al. 2023; Le et al. 2021). Here, the main challenge is that 84% of the 3.7-Gb *L.*
255 *culinaris* genome is repetitive DNA. The design of the gRNA array, therefore, required stringent multi-step filtering

256 resulting in a set of ~566,766 gRNAs. To our knowledge, this is the first time CRISPR/Cas9 has been used with such a large
257 number of gRNAs either *in vitro* or *in vivo*, representing a fundamental advance in the technology platform beyond the
258 specific goals of our project. Despite these technical challenges, CRISPR/Cas9-mediated repeat depletion produced
259 sequencing libraries with consistently lower proportions of repetitive DNA (41.2% depletion) and enriched the single-
260 copy regions (160% increase), thus allowing the generation of more meaningful sequencing data.

261 Repetitive DNA in plant genomes can be divided into two broad categories: dispersed mobile elements and
262 tandem/simple repeats. Dispersed mobile elements are made up of DNA transposons and retrotransposons, the most
263 abundant of which are the LTR retrotransposons (Bennetzen and Wang 2014). Although mobile elements are not under
264 the same selection pressure as genes, the degree of conservation across multiple copies of the same element is
265 sufficiently high to allow the targeting of multiple copies with single gRNAs. The most efficient depletion (47%) was
266 achieved for the most abundant LTR retrotransposon family (Gypsy, ~64% of the genome), followed by all the other LTR
267 families (–15%). The cutting of LTR elements by Cas9 was responsible for almost the entire depletion observed at the
268 genome-wide level, whereas the depletion of other mobile elements was negligible. This was probably due to the lower
269 repetition of such elements, which translated into a poor cutting frequency in the final gRNA design, comprising only
270 gRNAs with 25 targets. These targets usually featured < 8 gRNA/kbp, namely a density associated with a low depletion
271 rate. A similar observation was reported for RNA-Seq libraries, where a gRNA every 50–100 nucleotides (10–20
272 gRNAs/kbp) achieves excellent depletion results (Gu et al. 2016). Simple and tandem repeats were also not depleted
273 efficiently, although the gRNA density was close to 8. In these cases, the highly repetitive motifs of such sequences may
274 have reduced the cutting efficiency of Cas9 (Müller Paul et al. 2022). Given that LTR retrotransposons make up the
275 majority of repeats in plant genomes (94% in *L. culinaris*) and are the principal cause of plant genome size variation
276 (Bennetzen and Wang 2014; Lee and Kim 2014), the design of gRNAs to target only LTR sequences may be the most
277 efficient strategy to reduce the genome size in sequencing experiments. Future gRNA designs in other species and further
278 optimization of the *L. culinaris* design should maximize the gRNA number on these most abundant elements, instead of
279 dispersing the effort across the remaining repetitive fraction (< 10%). Another factor influencing the efficiency of
280 CRISPR/Cas9-mediated repeat depletion was the dose of Cas9 and gRNAs; doubling their dose indeed improved repeat
281 depletion by ~10%, albeit with a slight increase in overall costs. Finally, also the order of gRNA addition was important, as
282 higher depletion efficiency was achieved when splitting gRNAs into three groups based on cutting frequency and using
283 the groups sequentially. A possible explanation of this phenomenon is that most abundant gRNAs, with the highest
284 number of target sites in the genome, could interfere with the "genome patrolling" of other RNPs targeting less abundant

285 elements. Therefore, the gRNA target density, RNP concentration and prioritization of gRNAs with less abundant targets
286 are factors that can improve depletion efficiency.

287 Organelle genomes often constitute a large fraction of DNA derived from plants (Sakamoto and Takami 2018). Although
288 the mitochondrial and chloroplast genomes are smaller than the nuclear genome, they are present in multiple copies per
289 cell, and they can represent > 20% of the total sequence data (Gargiulo et al. 2021; Ren et al. 2021). CRISPR/Cas9-
290 mediated repeat depletion has been shown to reduce the fraction of sequencing libraries derived from organelle
291 genomes in ATAC-Seq experiments (Montefiori et al. 2017). Our method was efficient for the depletion of chloroplast
292 DNA (by 67%) while the depletion of mitochondrial DNA was only marginal, possibly reflecting the different abundance
293 of the two organelle genomes in the starting genomic sample. Still, in lentil, the fraction of sequencing data attributable
294 to organelles was largely due to chloroplasts (88%), whose depletion was therefore sufficient to decrease the data
295 mapping on organelles after Cas9 treatment (-65% overall). Although in the case of lentil the total sequencing data
296 attributable to organelles was rather low (~1.3% in the not depleted libraries), the depletion of organelle DNA from
297 sequencing libraries will be highly beneficial for organisms with a strongly unbalanced ratio of organelle vs nuclear DNA,
298 such as *Cypripedium calceolus* (Gargiulo et al. 2021) and *Haematococcus pluvialis* (Ren et al. 2021). Although this has not
299 been investigated in the present work, it is plausible that gRNAs designed for organelle's genomes could also target
300 nuclear integrants of plastid/mitochondrial DNA (NUPTs and NUMTs), that are homologous to the cpDNA and mtDNA
301 (Zhang et al. 2020; Sloan et al. 2018).Although the efficiency of CRISPR/Cas9-mediated repeat depletion could be
302 improved, the current set of gRNAs allowed us to genotype consistently more bases on single-copy as compared to not
303 depleted libraries, and consequently to identify more genetic variants. We observed that coupling the depletion with 25
304 million sequencing fragments provided the best balance between costs and fraction of genotyped bases. In this condition,
305 depleted samples reached an average coverage of ~5 fold on single-copy regions, with gains of 10- and 12-fold in the
306 number of genotyped positions and identified variants, respectively, as compared to not depleted libraries. We estimated
307 that one would need 3.5-4 times more sequencing data to achieve the same performances when using not depleted
308 libraries. By targeting the majority of genomic length (84%), the depletion allowed to pour a large fraction of sequencing
309 data on the single-copy regions, that are instead just a small fraction (16%). For this reason, the CRISPR/Cas9-repeat
310 depletion was more effective to improve genotyping performances than increasing the overall sequencing coverage,
311 being also beneficial for the identification of heterozygous variants that would otherwise be missed due to unbalanced
312 allelic sampling. Although this involved only ~3% of total variants identified in lentil, as this is an autogamous diploid
313 species, allelic imbalance is a well-known cause of errors in genotyping experiments based on sequencing (Cooke et al.

314 2016). CRISPR/Cas9-mediated repeat depletion can therefore improve the accuracy of genotyping experiments,
315 especially in plants with highly heterozygous genomes and/or in polyploids.

316 Given that CRISPR/Cas9 repeat depletion allows to concentrate the sequencing data on the desired regions, the amount
317 of positions that were genotyped in common between multiple samples was consistently higher in the depleted ones. In
318 population studies, the shift in the distribution of sequencing data may contribute to reduce the number of missing data,
319 thereby detecting a larger number of differences between samples. Furthermore, this approach was successful in
320 different cultivars of *L. culinaris*, and also in the closely-related species *L. orientalis.* This is important because genotyping
321 experiments typically include distant/wild relatives and related species, from which it is possible to develop evolutionary
322 studies and plan breeding experiments, including the introgression of characters of interest. For example, the INCREASE
323 project features a collection of 2000 lentil accessions that includes both cultivated varieties and local landraces (Guerra-
324 García et al. 2021). Further experiments could determine whether the gRNAs designed in this study are also suitable for
325 the depletion of repeats in other closely related leguminous species (Fabaceae) with very large and repetitive genomes,
326 such as pea (*P. sativum*, 3.92 Gb, 83% repetitive)(Kreplak et al. 2019) and faba bean (*Vicia faba* L., 12 Gb, 79% transposon-
327 derived repeats)(Jayakodi et al. 2023).

328

329 The cost of NGS library preparation for a genotyping project can easily exceed the cost of sequencing in the case of small
330 genomes and/or ultra-lcWGS, especially given the steadily falling price of sequencing. More recent library-preparation
331 kits circumvent several lengthy steps that require expensive reagents, and allow large sample sets to be processed in
332 multiplex reactions. We used the Twist 96-Plex Library Prep Kit (formerly iGenomX Riptide) that constructs Illumina NGS
333 libraries by polymerase-mediated extension of barcoded random primers. This type of library is beneficial for genotyping
334 in general because random priming reduces uniform genome coverage but allows more reproducible sampling of the
335 same sites across multiple samples (Siddique et al. 2019). Most importantly, the kit is designed to process large numbers
336 of samples (up to 96 simultaneously) at low costs and without advanced equipment (just a multichannel pipette). To
337 design, filter and synthesize the gRNA set targeting *L. culinaris* repeats required 10 weeks and cost ~20,000 USD. The
338 latter comprised gRNAs and depletion reagents sufficient for 30 reactions, each for a maximum of 96 samples treated in
339 multiplex, corresponding to approximately 10 USD per sample. We estimated that the net cost to achieve ~5-fold average
340 coverage of the single-copy regions of the lentil genome by combining CRISPR/Cas9-mediated repeat depletion with a
341 96-plex Twist library is approximately US$75, made up of US$15 for library preparation, US$10 for depletion and US$40

342 for sequencing on a NovaSeq6000 S4 flowcell, generating 25 million fragments per sample. As such, our results
343 demonstrated that depleted libraries are more informative than standard ones when normalized for the amount of
344 sequencing data. Alternatively, CRISPR/Cas9-mediated repeat depletion can be used to reduce sequencing costs (by ~75%
345 in lentil), because the same number of genotyped bases (or detected variants) in single-copy regions can be detected
346 with much less sequencing data. CRISPR/Cas9-mediated repeat depletion combined with Twist multiplex libraries is
347 therefore an effective strategy for genotyping projects involving hundreds or thousands of samples. Dealing with less-
348 repetitive datasets can also reduce the complexity of the genotyping analysis and the computational resources required.

349 The method therefore has the potential to increase our genetic knowledge of plant species that are currently difficult to
350 analyze without a significant economic investment due to the large genome size and high proportion of repetitive DNA.
351 Population studies, eQTL analysis, GWAS and pre-breeding programs are just some of the approaches that can benefit
352 from CRISPR/Cas9-mediated repeat depletion.

353

354 METHODS

355 **Multiplex-library preparation and sequencing.** We prepared 8-plex and 96-plex multiplex libraries according to the Twist
356 96-Plex Library Preparation Kit protocol (Twist Bioscience, South San Francisco, CA, USA) with the following modifications.
357 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
358 set A was used for the extension and termination reaction (Reaction A) followed by 8 and 9 cycles of PCR amplification
359 for the 96-plex and 8-plex libraries, respectively. Final libraries were purified using Twist DNA Purification Beads (0.65x
360 volume) and a second round of purification was applied to the supernatant using 10 µl of beads to achieve a median
361 insert size of 500 bp. Libraries were quantified using the Qubit BR DNA kit and a Qubit device (Thermo Fisher Scientific,
362 Waltham, MA, USA) and size distributions were assessed using a Tape Station System (Agilent Technologies, Santa Clara,
363 CA, USA). Non-depleted libraries were pooled at equimolar concentrations and sequenced on a NovaSeq6000 instrument
364 (Illumina, San Diego, CA, USA) to generate 150-bp paired-end reads.

365 **WGS library preparation and sequencing.** Genomic DNA samples were fragmented using a Covaris sonicator to achieve
366 an average size of 400 bp, and Illumina PCR-free libraries were prepared from 700 ng DNA using the KAPA Hyper prep kit
367 and unique dual-indexed adapters (5 µL of a 15 µM stock) according to the supplier's protocol (Roche, Basel, Switzerland).
368 The library concentration and size distribution were assessed on a Bioanalyzer (Agilent Technologies, Santa Clara, CA,

369 USA). Non-depleted WGS libraries were pooled at equimolar concentrations and sequenced on a NovaSeq6000
370 instrument (Illumina, San Diego, CA, USA) to generate 150-bp paired-end reads.

371 **Design of gRNAs.** The gRNA set was designed by JumpCode Genomics (San Diego, CA, USA) against the repetitive regions
372 of the *L. culinaris* CDC Redberry v2.0 reference genome (Ramsay et al. 2021) (https://knowpulse.usask.ca/genome-
373 assembly/Lcu.2RBY). The available repeat annotation (transposable elements and tandem repeats) was integrated with
374 the annotation of simple and tandem repeats identified by RepeatMasker v4.0.6 and Tandem Repeat Finder v4.9 using 2
375 7 7 80 10 50 2000 -d -h parameters to identify intervals for gRNA design (**Supplemental Table S1**). Adjacent or overlapping
376 intervals were collapsed into single intervals before design. As a first step, all 20 nt sequences with adjacent PAM sites
377 for Cas9 (NGG) were identified in the target intervals. Second, the guides were filtered to exclude secondary structure,
378 high and low GC content, homopolymers, dinucleotide repeats and low *in vitro* cleavage efficiency prediction scores
379 (Azimuth algorithm; (Doench et al. 2016)). Third, the resulting guides were filtered to minimize off-target cleavage in
380 single-copy regions of the genome by excluding guides that have complementary sites in genomic regions corresponding
381 to genes and open-chromatin regions identified by ATAC-Seq (PRJNA912311) (allowing for up to 3 mismatches). As a final
382 step, and to reduce the number of guides in the set, guides were selected to have no fewer than 25 cleavage sites each
383 and to maintain an inter-guide spacing of at least 500 bp. The final guide set, comprising 569,088 unique guides, was split
384 into 11 pools for the purpose of synthesis. The number of copies of each guide varied and reflected the number of on-
385 target cleavage sites for each guide. DNA oligonucleotides containing the target-specific 20 nt gRNA sequence and
386 invariant single gRNA sequence were synthesized, after which pools of oligonucleotides were amplified by PCR and
387 converted to RNA by *in vitro* transcription. The products of transcription were treated with DNase I and column purified
388 to generate the final gRNA material. Pools 1–3, 5–8 and 10 contain only gRNAs targeting the nuclear genome. Pools 9 and
389 11 contain both nuclear and chloroplast genome gRNAs, and pool 4 is the only pool containing gRNAs that target the
390 nuclear, chloroplast and mitochondrial genomes (**Supplemental Table S2** and **Supplemental File S1**). The number of gRNAs
391 targeting each repeat class are reported in **Supplemental Table S5**.

392 **Repeat depletion with JumpCode CRISPRclean.** Repetitive regions were depleted using the Cas9 protein and the custom
393 gRNA set described above, according to the Jumpcode CRISPRclean Ribosomal RNA Depletion from Human RNA-Seq
394 Libraries for Illumina Sequencing protocol (Jumpcode Genomics, San Diego, CA, USA) with the following modifications.
395 The input was 10 and 100 ng for the 8-plex and 96-plex libraries, respectively. Depletion was carried out either using all
396 gRNA simultaneously or by splitting the gRNA pools into three groups based on cutting frequency, which were used

397 sequentially in order of increasing cutting frequency (**Supplemental Table S2**). The sequential depletion strategy was also
398 conducted using the double amounts of gRNAs and Cas9. The reaction volume was 20 µl when using all gRNAs
399 simultaneously or 26 µl for the sequential and double sequential protocols. The quantity of each gRNA pool per reaction
400 is shown in **Supplemental Table S2** and amounted to 620 ng in the simultaneous and sequential depletion reactions or
401 1240 ng in the double sequential depletion reaction. The Cas9 enzyme was diluted 1:5 in 1x Cas9 Buffer and 0.0029 µl
402 was used per ng gRNA. The reactions were incubated at 37 °C and libraries were treated in the presence of gRNAs for a
403 total of 1 h (simultaneous depletion protocol) or 3 h (sequential and double sequential depletion protocols, with the
404 sequential gRNA pools added at 1-h intervals). The depleted samples were then size selected using 0.6x volume of
405 AMPure XP Beads (Beckman Coulter, Brea, CA, USA). Libraries were amplified with 10 and 6 PCR cycles for the 8-plex and
406 96-plex libraries, respectively before final purification with 60µl (0.6x volume) AMPure XP Beads. The concentrations of
407 depleted libraries were measured using the Qubit system and size distributions were assessed on a Tape Station System
408 as described above. Depleted libraries were pooled at equimolar concentrations and sequenced on a NovaSeq6000
409 instrument to generate 150-bp paired-end reads.

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

423 The variation of mapped reads in depleted *vs* not depleted libraries was calculated using the formula:

424 $Variation\ of\ mapped\ reads = \frac{Mapped\ reads\ (depleted) - Mapped\ reads\ (not\ depleted)}{Mapped\ reads\ (not\ depleted)}$ To achieve a normal distribution
425 of variation, in Supplemental Figure S1 the variation of mapped read coverage between the depleted and not depleted
426 libraries was calculated using the following formula:

427 $$Variation\ of\ mapped\ read\ coverage = \log_2 \frac{Mean\ coverage\ (depleted)}{Mean\ coverage\ (not\ depleted)}$$

428 Regions with zero coverage in either depleted or not depleted conditions (6,097 and 6,319, respectively, over a total of
429 237,136 repetitive segments, of which 4,662 in common) were excluded from the calculation, as they represented a
430 negligible fraction of total (3.3%) and showed a minimal coverage also in the opposite condition (< 0.2 fold). The plot in
431 Supplemental Figure S1 was generated using the ggplot package in R.

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

436

437 **DATA ACCESS.** The sequencing datasets generated and analysed in this study have been submitted to the NCBI GenBank
438 repository (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA915594 and PRJNA912311.

439 **COMPETING INTEREST STATEMENT.** Authors MR and MD are partners of Genartis srl. The remaining authors declare that
440 the research was conducted in the absence of any commercial or financial relationships that could be construed as a
441 potential conflict of interest.

442 **ACKNOWLEDGMENTS.** This research was supported by the European Union's Horizon 2020 research and innovation
443 program, through the project INCREASE (www.pulsesincrease.eu) (grant agreement No. 862862). The founding sponsors
444 had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the
445 manuscript; and in the decision to publish the results. We acknowledge Matteo De Biasi for the support in bioinformatic
446 analysis, the Twist Bioscience and Jumpcode Genomics teams for the excellent technical support.

447 **Authors' contributions.** Conceptualization MR, MD and RP; Methodology MR, LM and MD; Software LM and GL;
448 Investigation LDA, EBE, GC and FL; Formal analysis MR, LM and LDA; Validation GF, LN and LV; Resources KB, LR and DJK;

449 Data Curation LM and GL; Writing - Original Draft MR; Writing - Review & Editing LM, LDA and MD; Visualization MR and
450 LM; Supervision MR and MD; Project administration MR; Funding acquisition EBI, MD and RP.

451

452 REFERENCES

453 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008.
454 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**.

455 Bellucci E, Mario Aguilar O, Alseekh S, Bett K, Brezeanu C, Cook D, De la Rosa L, Delledonne M, Dostatny DF,
456 Ferreira JJ, et al. 2021. The INCREASE project: Intelligent Collections of food-legume genetic resources
457 for European agrofood systems. *Plant Journal* **108**: 646–660.

458 Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and
459 evolution of plant genomes. *Annu Rev Plant Biol* **65**: 505–530.

460 Cooke TF, Yee MC, Muzzio M, Sockell A, Bell R, Cornejo OE, Kelley JL, Bailliet G, Bravi CM, Bustamante CD,
461 et al. 2016. GBStools: A Statistical Method for Estimating Allelic Dropout in Reduced Representation
462 Sequencing Data. *PLoS Genet* **12**.

463 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker
464 discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510.

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

467 Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard
468 R, et al. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-
469 Cas9. *Nat Biotechnol* **34**: 184–191.

470 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple
471 genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**.

472 Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. 2011. Crop genome sequencing: Lessons and
473 rationales. *Trends Plant Sci* **16**: 77–88.

474 Friel J, Bombarely A, Fornell CD, Luque F, Fernández-Ocaña AM. 2021. Comparative analysis of genotyping
475 by sequencing and whole-genome sequencing methods in diversity studies of olea europaea l. *Plants*
476 **10**.

477 Gargiulo R, Kull T, Fay MF. 2021. Effective double-digest RAD sequencing and genotyping despite large
478 genome size. *Mol Ecol Resour* **21**: 1037–1055.

479 Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL. 2016. Depletion of
480 Abundant Sequences by Hybridization (DASH): Using Cas9 to remove unwanted high-abundance
481 species in sequencing libraries and molecular counting applications. *Genome Biol* **17**.

482 Guerra-García A, Gioia T, von Wettberg E, Logozzo G, Papa R, Bitocchi E, Bett KE. 2021. Intelligent
483 Characterization of Lentil Genetic Resources: Evolutionary History, Genetic Diversity of Germplasm,
484 and the Need for Well-Represented Collections. *Curr Protoc* **1**.

485 He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, Forrest K, Fritz A, Hucl P, Wiebe K, et al. 2019. Exome
486 sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the
487 wheat genome. *Nat Genet* **51**: 896–904.

488 Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ,
489 et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**: 1522–1527.

490 Homberger C, Hayward RJ, Barquist L, Vogel J. 2023. Improved Bacterial Single-Cell RNA-Seq through
491 Automated MATQ-Seq and Cas9-Based Removal of rRNA Reads. *mBio*.

492 Ichida H, Abe T. 2019. An improved and robust method to efficiently deplete repetitive elements from
493 complex plant genomes. *Plant Science* **280**: 455–460.

494 Jayakodi M, Golicz AA, Kreplak J, Fechete LI, Angra D, Bednář P, Bornhofen E, Zhang H, Boussageon R, Kaur
495 S, et al. 2023. The giant diploid faba genome unlocks variation in a global protein crop. *Nature*.

496 JumpCode Genomics. 2021. Technology Overview Version 1.2 Harnessing CRISPR to boost NGS sensitivity
497 with CRISPRclean ™. https://www.jumpcodegenomics.com/wp-content/uploads/2021/07/jumpcode-
498 technical-overview-20210521_v1-1_F.pdf (Accessed December 24, 2022).

499 Kreplak J, Madoui MA, Cápal P, Novák P, Labadie K, Aubert G, Bayer PE, Gali KK, Syme RA, Main D, et al.
500 2019. A reference genome for pea provides insight into legume genome evolution. *Nat Genet* **51**:
501 1411–1422.

502 Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: Why it is important,
503 and how to correct it. *BioEssays* **35**: 780–786.

504 Le C, Liu Y, López-Orozco J, Joyce MA, Le XC, Tyrrell DL. 2021. CRISPR Technique Incorporated with Single-
505 Cell RNA Sequencing for Studying Hepatitis B Infection. *Anal Chem* **93**: 10756–10761.

506 Lee S-I, Kim N-S. 2014. Transposable Elements and Genome Size Variations in Plants. *Genomics Inform* **12**:
507 87.

508 Matvienko M, Kozik A, Froenicke L, Lavelle D, Martineau B, Perroud B, Michelmore R. 2013. Consequences
509 of Normalizing Transcriptomic and Genomic Libraries of Plant Genomes Using a Duplex-Specific
510 Nuclease and Tetramethylammonium Chloride. *PLoS One* **8**.

511 Miller DFB, Yan PS, Buechlein A, Rodriguez BA, Yilmaz AS, Goel S, Lin H, Collins-Burow B, Rhodes L V., Braun
512 C, et al. 2013. A new method for stranded whole transcriptome RNA-seq. *Methods* **63**: 126–134.

513 Montefiori L, Hernandez L, Zhang Z, Gilad Y, Ober C, Crawford G, Nobrega M, Sakabe NJ. 2017. Reducing
514 mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Sci Rep* **7**.

515 Müller Paul H, Istanto DD, Heldenbrand J, Hudson ME. 2022. CROPSR: an automated platform for complex
516 genome-wide CRISPR gRNA design and validation. *BMC Bioinformatics* **23**.

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

519 Pavan S, Delvento C, Ricciardi L, Lotti C, Ciani E, D'Agostino N. 2020. Recommendations for Choosing the
520 Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies.
521 *Front Genet* **11**.

522 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: An inexpensive
523 method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7**.

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

526 Ramsay L, Koh CS, Kagale S, Gao D, Kaur S, Haile T, Gela TS, Chen L-A, Cao Z, Konkin DJ, et al. 2021. Genomic
527 rearrangements have consequences for introgression breeding as revealed by genome assemblies of
528 wild and cultivated lentil species. https://doi.org/10.1101/2021.07.23.453237.

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

531 Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, Noshay JM, Galli M, Mejía-Guerra MK, Colomé-
532 Tatché M, et al. 2019. Widespread long-range cis-regulatory elements in the maize genome. *Nat*
533 *Plants* **5**: 1237–1249.

534 Sakamoto W, Takami T. 2018. Chloroplast DNA dynamics: Copy number, quality control and degradation.
535 *Plant Cell Physiol* **59**: 1120–1127.

536 Siddique A, Suckow G, Ordoukhanian P, Head S, Homer N, Hernandez A, Brown K, Glick L, Baruch K, Doran
537 P, et al. 2019. RipTide High Throughput NGS Library Prep for Genotyping in Populations. *J Biomol*
538 *Tech.*; **30**(Suppl):S35-S36.

539 Sloan DB, Warren JM, Williams AM, Wu Z, Abdel-Ghany SE, Chicco AJ, Havird JC. 2018. Cytonuclear
540 integration and co-evolution. *Nat Rev Genet* **19**: 635–648.

541 Tanaka N, Shenton M, Kawahara Y, Kumagai M, Sakai H, Kanamori H, Yonemaru JI, Fukuoka S, Sugimoto K,
542 Ishimoto M, et al. 2021. Investigation of the Genetic Diversity of a Rice Core Collection of Japanese
543 Landraces using Whole-Genome Sequencing. *Plant Cell Physiol* **61**: 2087–2096.

544 Tian F, Yang DC, Meng YQ, Jin J, Gao G. 2020. PlantRegMap: Charting functional regulatory maps in plants.
545 *Nucleic Acids Res* **48**: D1104–D1113.

546 Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJA, Huvenaars KHJ, Hogers RCJ, van Enckevort
547 LJG, Janssen A, van Orsouw NJ, et al. 2012. Sequence-based genotyping for marker discovery and co-
548 dominant scoring in germplasm and populations. *PLoS One* **7**.

549 Wang J, Sun G, Ren X, Li C, Liu L, Wang Q, Du B, Sun D. 2016. QTL underlying some agronomic traits in
550 barley detected by SNP markers. *BMC Genet* **17**.

551 Wang P, Xiong Y, Gong R, Yang Y, Fan K, Yu S. 2019. A key variant in the cis-regulatory element of flowering
552 gene Ghd8 associated with cold tolerance in rice. *Sci Rep* **9**.

553 Yan H, Haak DC, Li S, Huang L, Bombarely A. 2022. Exploring transposable element-based markers to
554 identify allelic variations underlying agronomic traits in rice. *Plant Commun* **3**.

555 Zan Y, Payen T, Lillie M, Honaker CF, Siegel PB, Carlborg Ö. 2019. Genotyping by low-coverage whole-
556 genome sequencing in intercross pedigrees from outbred founders: A cost-efficient approach.
557 *Genetics Selection Evolution* **51**.

558 Zhang GJ, Dong R, Lan LN, Li SF, Gao WJ, Niu HX. 2020. Nuclear integrants of organellar DNA contribute to
559 genome structure and evolution in plants. *Int J Mol Sci* **21**.

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

563 Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz M V.,
564 Meleshkevitch E, Moroz LL, Lukyanov SA, et al. 2004. Simple cDNA normalization using kamchatka
565 crab duplex-specific nuclease. *Nucleic Acids Res* **32**.

566