Method

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

Marzia Rossato,[1,2,6] Luca Marcolungo,[1,6] 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]

[1]*Department of Biotechnology, University of Verona, 37134 Verona, Italy;* [2]*Genartis s.r.l., 37126 Verona, Italy;* [3]*Department of Agricultural, Food and Environmental Sciences, Polytechnic University of Marche, 60131 Ancona, Italy;* [4]*Department of Plant Sciences, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5A8, Canada;* [5]*National Research Council Canada, Saskatoon, Ontario S7N 0W9, Canada*

Q1 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 because of 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 nondepleted 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 because of 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 showed 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.

[Supplemental material is available for this article.]

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, whereas 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 (5–10×) whole-genome sequencing (lcWGS), allowing the characterization of several million variants (Friel et al. 2021; Tanaka et al. 2021). To reduce costs enough to make WGS affordable even in large germplasm collections, very low coverage (0.5–2×) WGS (ultra-lcWGS) can be combined with imputation to infer positions that are not sequenced or genotyped (Wang et al. 2016; Zan et al. 2019; Deng et al. 2022). Alternatively,

[6]**These authors contributed equally to this work.**
Corresponding authors: Marzia.rossato@univr.it, massimo.delledonne@univr.it, r.papa@staff.univpm.it

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277628.122. Freely available online through the *Genome Research* Open Access option.

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) (Baird et al. 2008; Davey et al. 2011) and double-digest RAD-Seq (ddRADseq) (Peterson et al. 2012; Truong 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 analyzed 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 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 nonrandom sampling of millions of genomic positions (Siddique et al. 2019).

Despite the advantages of sequencing-based genotyping over SNP arrays, one common disadvantage is that sequencing methods generally do not distinguish between repetitive (low-complexity) and single-copy (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 (Yan et al. 2022). 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, in which 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 onion (*Allium cepa*) (Feuillet et al. 2011). One strategy to address this issue is whole exome sequencing (WES), which selects coding regions for preferential sequencing (Hodges et al. 2007) as shown in lentil, wheat and barley (Ogutcen et al. 2018; He et al. 2019). However, WES only focuses on coding sequences and thus overlooks regulatory elements, which are equally important as sources of genetic diversity (Ricci et al. 2019; Wang et al. 2019; Tian et al. 2020).

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 (Zhulidov et al. 2004). This method has been used in RNA-seq analysis to remove abundant transcripts (Miller et al. 2013; Zhao et al. 2014) and, just occasionally, also to delete repetitive elements in DNA-seq libraries generated from plant genomes (Matvienko et al. 2013; Ichida and Abe 2019). However (Matvienko et al. 2013), DSN can also remove 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 (Matvienko et al. 2013). More recently, the CRISPR-Cas9 (clustered regularly interspaced short palindromic repeats and CRISPR-associated nuclease 9) system has been used for the selective depletion of unwanted genome fractions from sequencing libraries (Gu et al. 2016). 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 to detect rare pathogens in metagenomic samples (Gu et al. 2016). A similar technology has been recently commercialized under the name "CRISPRclean" by Jumpcode Genomics (Jumpcode Genomics 2021).

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 (*Lens 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 (Ramsay et al. 2021), targeting transposable elements (totaling 3.1 Gbp of sequence across the whole genome, corresponding to 82.5% of its size), simple sequence repeats (58 Mbp, 1.5%) and tandem repeats (13 Mbp, 0.4%) in the nuclear genome, as well as the entire mitochondrial genome (mtDNA, 489 kbp) and chloroplast genome (cpDNA, 118 kbp) (Supplemental 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%) (Supplemental Table S1). All nuclear DNA outside the gRNA target regions is hereafter defined as single copy. The final design comprised 566,766 gRNAs with at least 25 recognition sites, potentially targeting 2.9 Gbp (77%) of the *L. culinaris* nuclear genome and 93.5% of its repetitive regions when using a sequencing library with 500-bp inserts (Supplemental Table S2; Supplemental File S1). An additional 2366 gRNAs targeted the mitochondrial and chloroplast genomes (Supplemental Table S2). The gRNAs were assigned to 11 pools based on their cutting frequency in the *L. culinaris* genome (Supplemental Table S2).

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 8-plex library comprised replicates of three distinct *L. culinaris* samples (cv. Castelluccio) (Supplemental Table S3). Cas9/gRNAs ribonucleoprotein (RNP) complexes (1:2.5 protein/gRNA ratio) were generated, and gRNAs with more target sites in the genome were used at higher relative concentrations in the final reaction (Supplemental 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 (Supplemental Table S2) and using the groups sequentially, starting with the lowest cutting frequency. Depleted and nondepleted libraries were sequenced, generating

91 million fragments on average (Supplemental Table S4). The sequencing data were normalized at ~50 million fragments per library to compare the proportion of reads mapping on repetitive and single-copy regions of the nuclear genome and on the organelle genomes (Fig. 1). The number of reads mapping to repetitive regions (total repeats, nuclear repeats, mtDNA, and cpDNA) 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 (Fig. 1A). The results were similar when considering only the nuclear repetitive regions (37.2% depletion) (Fig. 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 (Fig. 1C). In contrast, no significant depletion was observed in the mitochondrial genome (Fig. 1D). In parallel, the sequential depletion strategy achieved a 130% increase in the number of reads mapping to single-copy regions, from 17.5 to 40.4 million (Fig. 1E). Given that the concentration of Cas9 RNPs influences the cutting efficiency (Gu et al. 2016), we repeated the sequential depletion strategy using double amount of Cas9 and gRNAs. This modified the read distribution further, achieving 41.2% depletion of nuclear repeats and a 160% increase in reads mapped to single-copy regions. The sequential depletion strategy with double RNPs was therefore the most efficient, and was applied in all subsequent experiments. Overall, our results showed that the custom gRNA set and Cas9 effectively targeted fragments containing repetitive DNA sequences and depleted them in the resulting sequencing libraries. Figure 2A and B shows the alignments of reads at two representative genomic regions, confirming that less sequencing data was assigned to regions of repetitive DNA and more reads were mapped to single-copy parts of the genome.

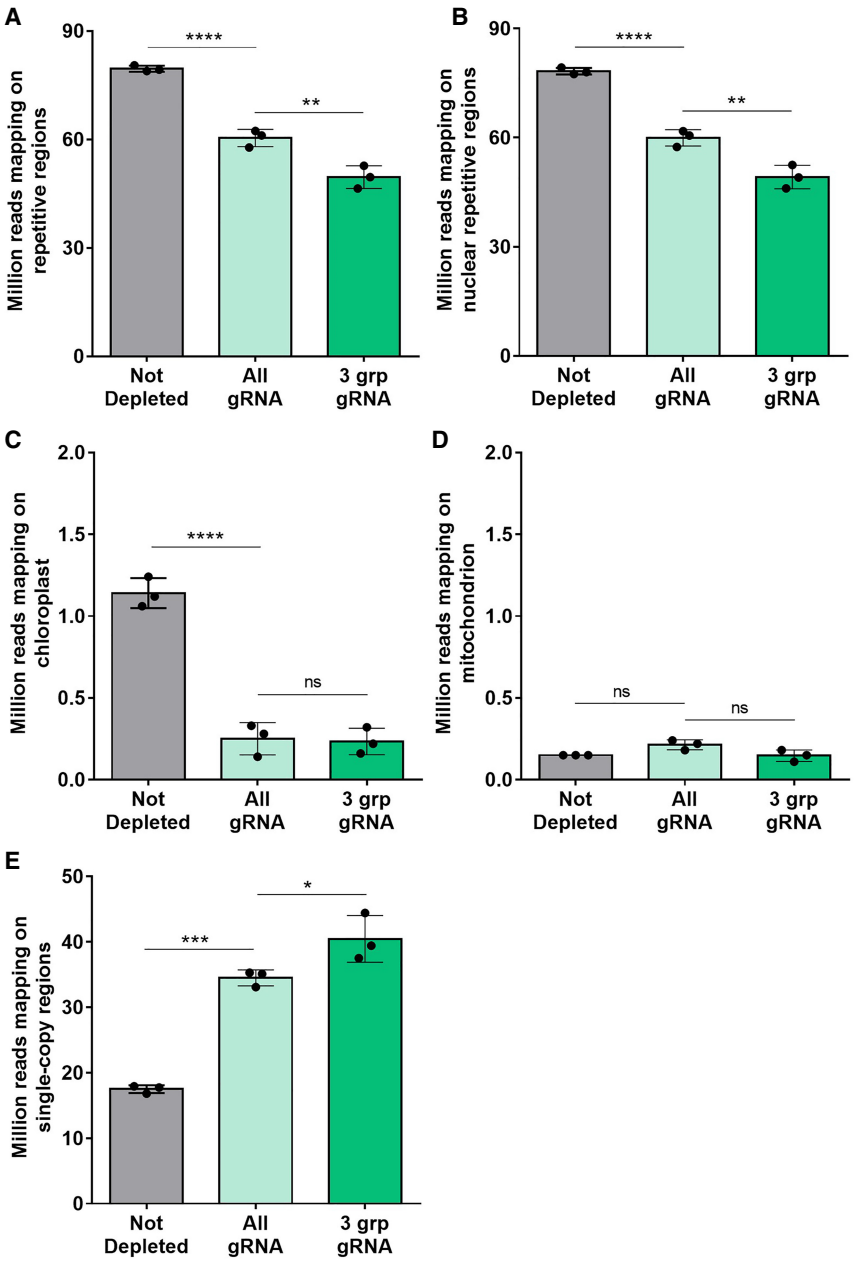

Figure 1. Distribution of mapped reads after CRISPR-Cas9-mediated repeat depletion. Libraries of *L. culinaris* cv. Castelluccio DNA (8-plex) were depleted using the custom gRNA set and Cas9. The gRNAs were used simultaneously (All) or were split into three groups that were used sequentially in order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

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 (Fig. 3A; 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) (Fig. 3A). We observed a

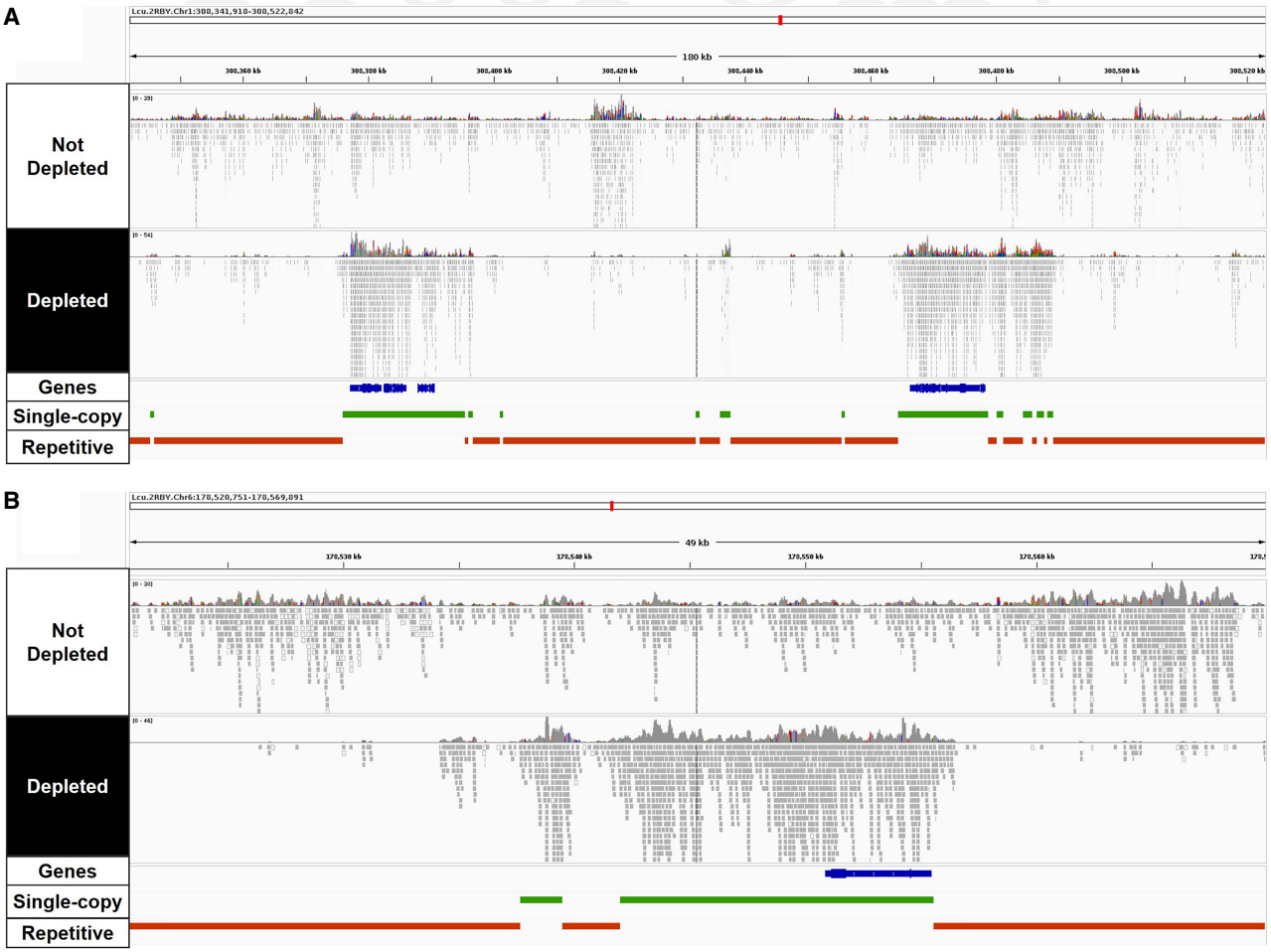

Figure 2. Reads mapped to repetitive or single-copy regions of the lentil genome with or without CRISPR-Cas9-mediated repeat depletion. Integrative Genome Browser Visualization (IGV) (Thorvaldsdóttir et al. 2013) of Illumina sequencing data mapped to two representative genomic sites of ~180 and ~50 kbp. Tracks in blue, green, and red represent annotated genes, single-copy regions, and repetitive regions, respectively.

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 (Fig. 3B,C; Supplemental 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 (Fig. 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 (Supplemental Fig. S1). In particular, target regions with a density >8 gRNAs/kbp showed a read reduction in 85% of cases (Supplemental Fig. S1) whereas regions targeted by <8 gRNAs/kbp usually showed limited or no depletion (Supplemental Fig. S1). We, therefore, concluded that the depletion efficiency across different repeat classes was dependent on the density of gRNA targets.

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 single-copy regions where a base can be reliably genotyped (PASS at a depth of ≥5 reads). For this analysis, sequencing data generated from depleted and nondepleted *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 single-copy regions of the depleted samples over the whole range considered (from ~3.5- to ~13-fold), with the highest gains at the lowest amounts of sequencing data (6 to 25 million fragments) (Fig. 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 single-copy regions increased significantly from ~4.5- to ~18-fold, with a delta of ~25,000 to ~1 million more variants identified in the depleted sample (Fig. 4B). Also the number of heterozygous variants identified was increased, although these constituted a minor fraction of total variants, as lentil is an autogamous species (Fig. 4C). This allowed us to identify from ~650 up to ~50,000 variant positions that would be erroneously classified as reference without the depletion (Fig. 4D). These false negative variants in the nondepleted sample were not called because of allelic imbalance caused by the low coverage (Supplemental Fig. S2). Consistently, most of these false negative variants rescued by depletion were heterozygous (~97%, Fig. 4D).

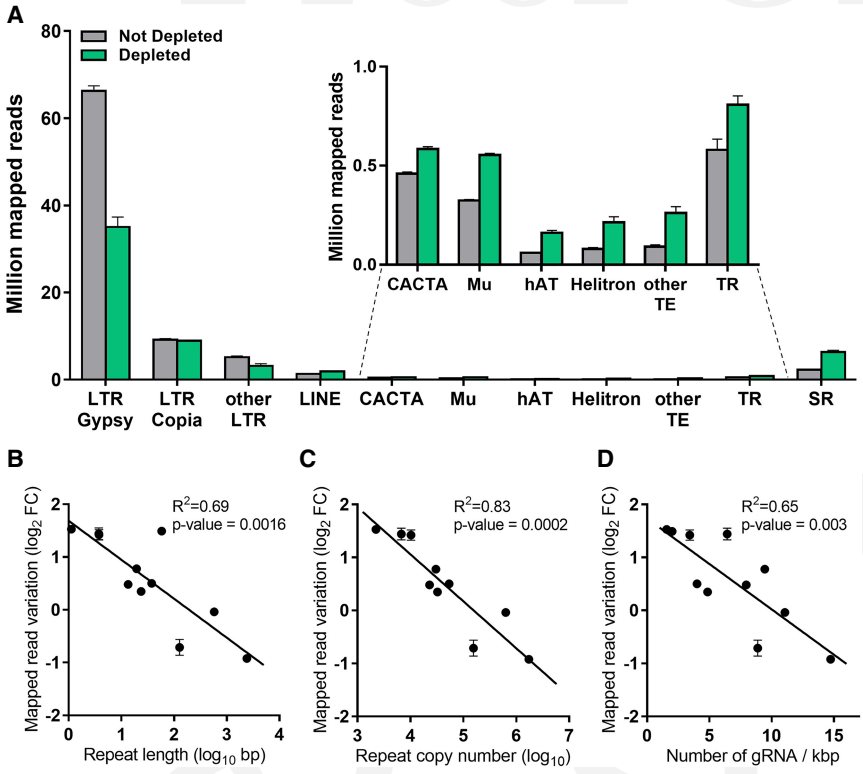

Figure 3. Sequencing data distribution after CRISPR-Cas9-mediated repeat depletion according to the class of nuclear repeat. (*A*) Number of reads (in millions) mapping to different nuclear repeat classes with or without CRISPR-Cas9-mediated repeat depletion: Ty3-Gypsy (LTR Gypsy), Ty1-Copia (LTR Copia), 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.

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 (Supplemental Tables S3, S4). Similar variations in the coverage of repetitive/single-copy regions and the number of mapped reads were observed when depleting multiplex libraries generated from a different cultivar (RB, Redberry) or from the **F5** closely related species *Lens orientalis* (Fig. 5A,B) when compared to the original Castelluccio cultivar (Fig. 1). To assess the impact of depletion when comparing different samples, these data were downsampled as described above, genotyped positions were identified and intersected with those of the Castelluccio samples. Depletion improved the genotyping reproducibility, as the number of genotyped positions in common between all analyzed samples, within the single-copy regions, was consistently higher than in the condition without depletion (Fig. 5C). Finally, there was no significant difference in the performance of CRISPR-Cas9-mediated repeat depletion when library multiplexing was increased from 8-plex to 96-plex while maintaining the 1:2.5 Cas9:gRNA ratio and **F6** 1 ng of treated library per sample (Fig. 6A,B), or when treating standard singleplex WGS libraries (Fig. 6C,D). Overall, these results showed that CRISPR-Cas9-mediated repeat depletion using the same gRNA set is at least equally effective when applied to a group of two lentil cultivars and one close wild relative, providing an improved genotyping reproducibility and the possibility to process individual samples or multiple samples simultaneously.

Discussion

In a typical genotyping experiment based on sequencing, the data derived from repetitive DNA is directly proportional to the repeat content of the genome, 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 (Prezza et al. 2020), metatrascriptomics (Gu et al. 2016), pathogen detection (Gu et al. 2016; Le et al. 2021) and single-cell analysis (Le et al. 2021; Homberger et al. 2023). Here, the main challenge is that 84% of the 3.7-Gb *L. culinaris* genome is repetitive DNA. The design of the gRNA array, therefore, required stringent multistep filtering 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 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 (41.2% depletion) and enriched the single-copy regions (160% increase), thus allowing the generation of more meaningful sequencing data. Equivalent results have been achieved not only in *L. culinaris* cv. Redberry (source of the reference genome for gRNA design) but also in another *L. culinaris* cultivar (Castelluccio) and in the closely related species *L. orientalis*, using the same gRNA set. The approach was shown to be useful also for other species beyond lentil, namely in bread wheat (*Triticum aestivum*), in which a dedicated repeat-specific gRNA set showed similar depletion performances, both in Chinese Spring and Jagger cultivars (Jumpcode Genomics, pers. comm.).

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 (Bennetzen and Wang 2014). 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

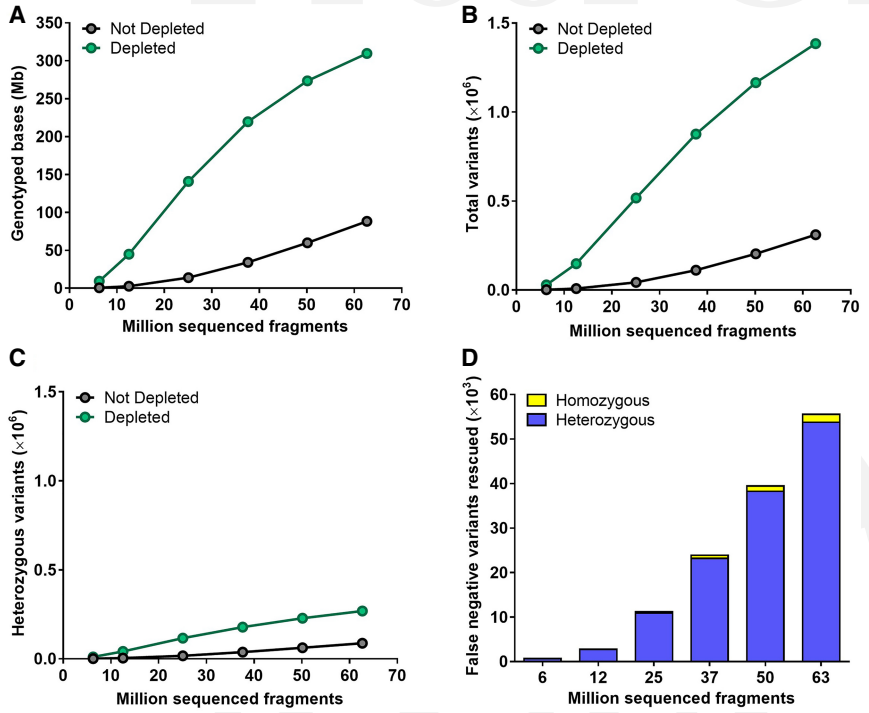

Figure 4. Genotyping performance on single-copy regions with or without CRISPR-Cas9-mediated repeat depletion starting from different amounts of sequencing data. (*A*) Number of genotyped positions. (*B*) Number of total variants identified. (*C*) Number of heterozygous variants identified. (*D*) Number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

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. Given that some repetitive elements were not efficiently depleted, it is likely that some gRNAs in the current lentil design are not yet optimal. Another factor influencing the depletion performances was probably 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 usually featured <8 gRNA/kbp, namely a density associated with a low depletion rate. A similar observation was reported for RNA-seq libraries, where a gRNA every 50–100 nucleotides (10–20 gRNAs/kbp) achieves excellent depletion results (Gu et al. 2016). 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 (Müller Paul et al. 2022). 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 (Bennetzen and Wang 2014; Lee and Kim 2014), 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 their dose indeed improved repeat depletion by ~10%, albeit with a slight increase in overall costs. To maximize depletion, the concentration of most efficient gRNAs could be also increased by excluding gRNA-Cas9 complexes with low cut performances and preoccupying limited Cas9 molecules, while maintaining the final amount of gRNA and Cas9 enzyme. Finally, also the order of gRNA addition was important, as higher depletion efficiency was achieved when splitting gRNAs into three groups based on cutting frequency and using the groups sequentially. A possible explanation of this phenomenon is that most abundant gRNAs, with the highest number of target sites in the genome, could interfere with the "genome patrolling" of other RNPs targeting less abundant elements.

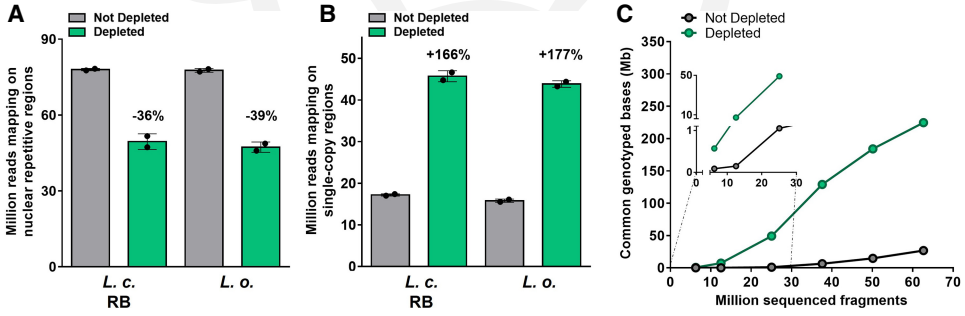

Figure 5. Performances of CRISPR-Cas9-mediated repeat depletion in different lentil samples. (*A*,*B*) Sequencing reads mapping to the repetitive or single-copy regions with or without CRISPR-Cas9-mediated repeat depletion in multiplex libraries generated from different lentil samples, namely *L. culinaris* 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). Variation percentages observed following CRISPR-Cas9-mediated repeat depletion are reported *above* each condition. (*C*) Number of genotyped positions in common between all samples analyzed in the study (three distinct samples of *L. culinaris* cv. Castelluccio, one sample of *L. culinaris* cv. Redberry, and one sample 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.

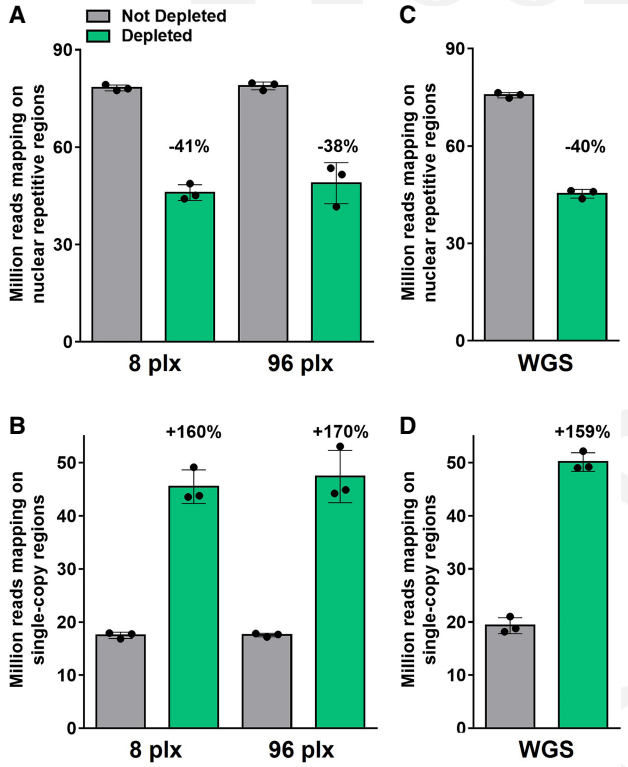

Figure 6. Performances of CRISPR-Cas9-mediated repeat depletion at different multiplexing level and library types. Sequencing reads mapping to the repetitive or single-copy regions with or without CRISPR-Cas9-mediated repeat depletion in (*A*,*B*) multiplex libraries containing eight or 96 samples (plx)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

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

Organelle genomes often constitute a large fraction of DNA derived from plants (Sakamoto and Takami 2018). 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 (Gargiulo et al. 2021; Ren et al. 2021). CRISPR-Cas9-mediated repeat depletion has been shown to reduce the fraction of sequencing libraries derived from organelle genomes in ATAC-seq experiments (Montefiori et al. 2017). Our method was efficient for the depletion of chloroplast DNA (by 67%), whereas the depletion of mitochondrial DNA was only marginal, possibly reflecting the different abundance of the two organelle genomes in the starting genomic sample. Still, in lentil, the fraction of sequencing data attributable to organelles was largely because of 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 versus nuclear DNA, such as *Cypripedium calceolus* (Gargiulo et al. 2021) and *Haematococcus pluvialis* (Ren et al. 2021). Although this has not been investigated in the present work, it is plausible that gRNAs designed for organelle's genomes could also target nuclear integrants of plastid/mitochondrial DNA (NUPTs and NUMTs), that are homologous to the cpDNA and mtDNA (Sloan et al. 2018; Zhang et al. 2020).Although the efficiency of CRISPR-Cas9-mediated repeat depletion could be improved, the current set of gRNAs allowed us to genotype consistently more bases on single-copy as compared to not-depleted libraries, and consequently to identify more genetic variants. We observed that coupling the depletion with 25 million sequencing fragments provided the best balance between costs and fraction of genotyped bases. In this condition, depleted samples reached an average coverage of approximately fivefold on single-copy regions, with gains of 10- and 12-fold in the number of genotyped positions and identified variants, respectively, as compared to not-depleted libraries. We estimated that one would need 3.5–4 times more sequencing data to achieve the same performances when using not-depleted libraries. By targeting the majority of genomic length (84%), the depletion allowed to pour a large fraction of sequencing data on the single-copy regions, that are instead just a small fraction (16%). For this reason, the CRISPR-Cas9-repeat depletion was more effective to improve genotyping performances than increasing the overall sequencing coverage, being also beneficial for the identification of heterozygous variants that would otherwise be missed because of unbalanced allelic sampling. Although this involved only ~3% of total variants identified in lentil, as this is an autogamous diploid species, allelic imbalance is a well-known cause of errors in genotyping experiments based on sequencing (Cooke et al. 2016). CRISPR-Cas9-mediated repeat depletion can therefore improve the accuracy of genotyping experiments, especially in plants with highly heterozygous genomes and/or in polyploids.

Given that CRISPR-Cas9 repeat depletion allows to concentrate the sequencing data on the desired regions, the amount of positions that were genotyped in common between multiple samples was consistently higher in the depleted ones. In population studies, the shift in the distribution of sequencing data may contribute to reduce the number of missing data, thereby detecting a larger number of differences between samples. Furthermore, 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 accessions that includes both cultivated varieties and local landraces (Guerra-García et al. 2021). Further experiments could determine whether the gRNAs designed in this study are also suitable for the depletion of repeats in other closely related leguminous species (Fabaceae) with very large and repetitive genomes, such as pea (*Pisum sativum*, 3.92 Gb, 83% repetitive) (Kreplak et al. 2019) and faba bean (*Vicia faba* L., 12 Gb, 79% transposon-derived repeats) (Jayakodi et al. 2023).

The cost of NGS library preparation for a genotyping 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 genotyping in general because random priming reduces uniform genome coverage but allows more reproducible sampling of the same sites across multiple samples (Siddique et al. 2019). 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). To design, filter, and synthesize the gRNA set targeting *L. culinaris* repeats required 10 weeks and cost ~20,000 U.S. dollars (USD). The latter comprised gRNAs and depletion reagents sufficient for 30 reactions, each for a maximum of 96 samples treated in multiplex, corresponding to approximately 10 USD per sample. We estimated that the net cost to achieve approximately fivefold average coverage of the single-copy regions of the lentil genome by combining CRISPR-Cas9-mediated repeat depletion with a 96-plex Twist library is approximately USD75, made up of USD15 for library preparation, USD10 for depletion and USD40 for sequencing on a NovaSeq 6000 S4 flowcell, generating 25 million fragments per sample. As such, our results showed that depleted libraries are more informative than standard ones when normalized for the amount of sequencing data. 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 single-copy regions can be detected with much less sequencing data. CRISPR-Cas9-mediated repeat depletion combined with Twist multiplex libraries is therefore an effective strategy for genotyping projects involving hundreds or thousands of samples. Dealing with less-repetitive data sets can also reduce the complexity of the genotyping analysis and the computational resources required.

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

Methods

Multiplex-library preparation and sequencing

We prepared 8-plex and 96-plex multiplex libraries according to the Twist 96-Plex Library Preparation Kit protocol (Twist Bioscience) with the following modifications. For each sample, we denatured 100 ng of genomic DNA (25 ng/µL) at 98°C for 1 min. Ultralow (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 (0.65× 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) and size distributions were assessed using a Tape Station System (Agilent Technologies). Nondepleted libraries were pooled at equimolar concentrations and sequenced on a NovaSeq 6000 instrument (Illumina) to generate 150-bp paired-end reads.

WGS library preparation and sequencing

Genomic DNA samples were fragmented using a Covaris sonicator to achieve an average size of 400 bp, and Illumina PCR-free libraries were prepared from 700 ng DNA using the KAPA Hyper prep kit and unique dual-indexed adapters (5 µL of a 15 µM stock) according to the supplier's protocol (Roche). The library concentration and size distribution were assessed on a Bioanalyzer (Agilent Technologies). Nondepleted WGS libraries were pooled at equimolar concentrations and sequenced on a NovaSeq 6000 instrument (Illumina) to generate 150-bp paired-end reads.

Design of gRNAs

The gRNA set was designed by Jumpcode Genomics against the repetitive regions of the *L. culinaris* CDC Redberry v2.0 reference genome (Ramsay et al. 2021) (https://knowpulse.usask.ca/genome-assembly/Lcu.2RBY). 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 (Supplemental 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) (Doench et al. 2016). Third, the resulting guides were filtered to minimize off-target cleavage in single-copy 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 (PRJNA912311) (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 fewer than 25 cleavage sites each and to maintain an interguide 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 containing the target-specific 20 nt gRNA sequence and invariant single gRNA sequence were synthesized, after which pools of oligonucleotides were amplified by PCR and converted to RNA by in vitro transcription. 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 (Supplemental Table S2; Supplemental File S1). The number of gRNAs targeting each repeat class are reported in Supplemental Table S5.

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) with the following modifications. The input was 10 and 100 ng for the 8-plex and 96-plex libraries, respectively. Depletion was performed 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 (Supplemental 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

Supplemental Table S2 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 1× 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 0.6× volume of AMPure XP Beads (Beckman Coulter). Libraries were amplified with 10 and 6 PCR cycles for the 8-plex and 96-plex libraries, respectively before final purification with 60 µL (0.6× 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 NovaSeq 6000 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 (http://fulcrumgenomics.github.io/fgbio/), 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 (https://github.com/vsbuffalo/scythe) and sickle v1.33 (https://github.com/najoshi/sickle), respectively. Filtered reads were aligned to the *L. culinaris* v2.0 reference genome using BWA-MEM v2.2.1 (Li 2013; Md et al. 2019) and the resulting alignments were converted to BAM files and sorted using SAMtools v1.13 (Li et al. 2009; Danecek et al. 2021). PCR-derived duplicates were removed using the GATK MarkDuplicates tool v4.1.7.0 (Van der Auwera and O'Connor 2020) and overlapping portions of the paired-end reads were clipped using the fgbio v1.3.0 ClipBam tool (http://fulcrumgenomics.github.io/fgbio/). The resulting BAM files were used to calculate coverage depth, breadth, and fraction of PASS bases (at ≥5×) using BEDTools v2.30.0 genomecov (Quinlan and Hall 2010) and GATK v3.8 CallableLoci, respectively (Van der Auwera and O'Connor 2020). The number of reads aligning to the reference genome and to different regions of interest was calculated using SAMtools v1.13 (Li et al. 2009; Danecek et al. 2021) with option -c to discard reads with a 2308 sam flag to consider only the primary alignment, thus omitting repetitive counts of the same multimapping reads. When necessary, sequencing data were normalized to a predefined number of input fragments using seqtk sample v1.3 (https://github.com/lh3/seqtk).

The variation of mapped reads in depleted versus not-depleted libraries was calculated using the formula:

$$Variation\ of\ mapped\ reads = \frac{Mapped\ reads\ (depleted) - Mapped\ reads\ (not\ depleted)}{Mapped\ reads\ (not\ depleted)}.$$

To achieve a normal distribution of variation, in Supplemental Figure S1 the variation of mapped read coverage between the depleted and not-depleted libraries was calculated using the following formula:

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

Regions with zero coverage in either depleted or not-depleted conditions (6097 and 6319, respectively, over a total of 237,136 repetitive segments, of which 4662 in common) were excluded from the calculation, as they represented a negligible fraction of total (3.3%) and showed a minimal coverage also in the opposite condition (<0.2-fold). The plot in Supplemental Figure S1 was generated using the ggplot package in R (Hadley 2016; R Core Team 2020).

Genomic variants were identified using GATK HaplotypeCaller v4.1.7.0 with the parameters "--min-base-quality-score 20 -ERC GVCF" (Van der Auwera and O'Connor 2020). 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 (Van der Auwera and O'Connor 2020). Variant filtration was achieved using GATK hard filters (https://gatk.broadinstitute.org/hc/en-us/articles/360037499012?id=3225).

Data access

The sequencing data sets generated and analyzed in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA915594 and PRJNA912311.

Competing interest statement

Authors M.R. and M.D. are partners of Genartis s.r.l. 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.

Acknowledgments

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

Author contributions: Conceptualization: M.R., M.D., and R.P.; methodology: M.R., L.M., and M.D.; software: L.M. and G.L.; investigation: L.D.A., E.Be., G.C., and F.L.; formal analysis: M.R., L.M., and L.D.A.; validation: G.F., L.N., and L.V.; resources: K.E.B., L.R., and D.J.K.; data curation: L.M. and G.L.; writing—original draft: M.R.; writing—review and editing: L.M., L.D.A., and M.D.; visualization: M.R. and L.M.; supervision: M.R. and M.D.; project administration: M.R.; funding acquisition: E.Bi., M.D., and R.P.

References

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3:** e3376. doi:10.1371/journal.pone.0003376

Bellucci E, Mario Aguilar O, Alseekh S, Bett K, Brezeanu C, Cook D, De la Rosa L, Delledonne M, Dostatny DF, Ferreira JJ, et al. 2021. The INCREASE project: intelligent collections of food-legume genetic resources for European agrofood systems. *Plant J* **108:** 646–660. doi:10.1111/tpj.15472

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol* **65:** 505–530. doi:10.1146/annurev-arplant-050213-035811

Cooke TF, Yee MC, Muzzio M, Sockell A, Bell R, Cornejo OE, Kelley JL, Bailliet G, Bravi CM, Bustamante CD, et al. 2016. GBStools: a statistical method for estimating allelic dropout in reduced representation

sequencing data. *PLoS Genet* **12:** e1005631. doi:10.1371/journal.pgen.1005631

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12:** 499–510. doi:10.1038/nrg3012

Deng T, Zhang P, Garrick D, Gao H, Wang L, Zhao F. 2022. Comparison of genotype imputation for SNP array and low-coverage whole-genome sequencing data. *Front Genet* **12:** 704118. doi:10.3389/fgene.2021.704118

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, et al. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34:** 184–191. doi:10.1038/nbt.3437

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6:** e19379. doi:10.1371/journal.pone.0019379

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. 2011. Crop genome sequencing: lessons and rationales. *Trends Plant Sci* **16:** 77–88. doi:10.1016/j.tplants.2010.10.005

Friel J, Bombarely A, Fornell CD, Luque F, Fernández-Ocaña AM. 2021. Comparative analysis of genotyping by sequencing and whole-genome sequencing methods in diversity studies of *Olea europaea* L. *Plants* **10:** 2514. doi:10.3390/plants10112514

Gargiulo R, Kull T, Fay MF. 2021. Effective double-digest RAD sequencing and genotyping despite large genome size. *Mol Ecol Resour* **21:** 1037–1055. doi:10.1111/1755-0998.13314

Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL. 2016. Depletion of abundant sequences by hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol* **17:** 41. doi:10.1186/s13059-016-0904-5

Guerra-García A, Gioia T, von Wettberg E, Logozzo G, Papa R, Bitocchi E, Bett KE. 2021. Intelligent characterization of lentil genetic resources: evolutionary history, genetic diversity of germplasm, and the need for well-represented collections. *Curr Protoc* **1:** e134. doi:10.1002/cpz1.134

Hadley W. 2016. *ggplot2: elegant graphics for data analysis*. Springer-Verlag, New York. https://ggplot2.tidyverse.org.

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 sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat Genet* **51:** 896–904. doi:10.1038/s41588-019-0382-2

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* **39:** 1522–1527. doi:10.1038/ng.2007.42

Homberger C, Hayward RJ, Barquist L, Vogel J. 2023. Improved bacterial single-cell RNA-seq through automated MATQ-seq and Cas9-based removal of rRNA reads. *mBio* **14:** e0355722. doi:10.1128/mbio.03557-22

Ichida H, Abe T. 2019. An improved and robust method to efficiently deplete repetitive elements from complex plant genomes. *Plant Sci* **280:** 455–460. doi:10.1016/j.plantsci.2018.10.021

Jayakodi M, Golicz AA, Kreplak J, Fechete LI, Angra D, Bednář P, Bornhofen E, Zhang H, Boussageon R, Kaur S, et al. 2023. The giant diploid faba genome unlocks variation in a global protein crop. *Nature* **615:** 652–659. doi:10.1038/s41586-023-05791-5

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

Kreplak J, Madoui MA, Cápal P, Novák P, Labadie K, Aubert G, Bayer PE, Gali KK, Syme RA, Main D, et al. 2019. A reference genome for pea provides insight into legume genome evolution. *Nat Genet* **51:** 1411–1422. doi:10.1038/s41588-019-0480-1

Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* **35:** 780–786. doi:10.1002/bies.201300014

Le C, Liu Y, López-Orozco J, Joyce MA, Le XC, Tyrrell DL. 2021. CRISPR technique incorporated with single-cell RNA sequencing for studying hepatitis B infection. *Anal Chem* **93:** 10756–10761. doi:10.1021/acs.analchem.1c02227

Lee S-I, Kim N-S. 2014. Transposable elements and genome size variations in plants. *Genomics Inform* **12:** 87. doi:10.5808/GI.2014.12.3.87

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997 [q-bio.GN].

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Matvienko M, Kozik A, Froenicke L, Lavelle D, Martineau B, Perroud B, Michelmore R. 2013. Consequences of normalizing transcriptomic and genomic libraries of plant genomes using a duplex-specific nuclease and tetramethylammonium chloride. *PLoS One* **8:** e55913. doi:10.1371/journal.pone.0055913

Md V, Misra S, Li H, Aluru S. 2019. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *Proceedings - 2019 IEEE 33rd international parallel and distributed processing symposium, IPDPS 2019*, pp. 314–324. Institute of Electrical and Electronics Engineers Inc.

Miller DFB, Yan PS, Buechlein A, Rodriguez BA, Yilmaz AS, Goel S, Lin H, Collins-Burow B, Rhodes LV, Braun C, et al. 2013. A new method for stranded whole transcriptome RNA-seq. *Methods* **63:** 126–134. doi:10.1016/j.ymeth.2013.03.023

Montefiori L, Hernandez L, Zhang Z, Gilad Y, Ober C, Crawford G, Nobrega M, Sakabe NJ. 2017. Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Sci Rep* **7:** 2451. doi:10.1038/s41598-017-02547-w

Müller Paul H, Istanto DD, Heldenbrand J, Hudson ME. 2022. CROPSR: an automated platform for complex genome-wide CRISPR gRNA design and validation. *BMC Bioinformatics* **23:** 74. doi:10.1186/s12859-022-04593-2

Ogutcen E, Ramsay L, von Wettberg EB, Bett KE. 2018. Capturing variation in *Lens* (Fabaceae): development and utility of an exome capture array for lentil. *Appl Plant Sci* **6:** e01165. doi:10.1002/aps3.1165

Pavan S, Delvento C, Ricciardi L, Lotti C, Ciani E, D'Agostino N. 2020. Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. *Front Genet* **11:** 447. doi:10.3389/fgene.2020.00447

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* **7:** e37135. doi:10.1371/journal.pone.0037135

Prezza G, Heckel T, Dietrich S, Homberger C, Westermann AJ, Vogel J. 2020. Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. *RNA* **26:** 1069–1078. doi:10.1261/rna.075945.120

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

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 rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. *bioRxiv* doi:10.1101/2021.07.23.453237

R Core Team. 2020. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Ren Q, Wang YC, Lin Y, Zhen Z, Cui Y, Qin S. 2021. The extremely large chloroplast genome of the green alga *Haematococcus pluvialis*: genome structure, and comparative analysis. *Algal Res* **56:** 102308. doi:10.1016/j.algal.2021.102308

Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, Noshay JM, Galli M, Mejía-Guerra MK, Colomé-Tatché M, et al. 2019. Widespread long-range *cis*-regulatory elements in the maize genome. *Nat Plants* **5:** 1237–1249. doi:10.1038/s41477-019-0547-0

Sakamoto W, Takami T. 2018. Chloroplast DNA dynamics: copy number, quality control and degradation. *Plant Cell Physiol* **59:** 1120–1127. doi:10.1093/pcp/pcy084

Siddique A, Suckow G, Ordoukhanian P, Head S, Homer N, Hernandez A, Brown K, Glick L, Baruch K, Doran P, et al. 2019. Riptide high throughput NGS library prep for genotyping in populations. *J Biomol Tech* **30 (Suppl):** S35–S36.

Sloan DB, Warren JM, Williams AM, Wu Z, Abdel-Ghany SE, Chicco AJ, Havird JC. 2018. Cytonuclear integration and co-evolution. *Nat Rev Genet* **19:** 635–648. doi:10.1038/s41576-018-0035-9

Tanaka N, Shenton M, Kawahara Y, Kumagai M, Sakai H, Kanamori H, Yonemaru JI, Fukuoka S, Sugimoto K, Ishimoto M, et al. 2021. Investigation of the genetic diversity of a rice core collection of Japanese landraces using whole-genome sequencing. *Plant Cell Physiol* **61:** 2087–2096. doi:10.1093/pcp/pcaa125

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178–192. doi:10.1093/bib/bbs017

Tian F, Yang DC, Meng YQ, Jin J, Gao G. 2020. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res* **48:** D1104–D1113. doi:10.1093/nar/gkz1020

Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJA, Huvenaars KHJ, Hogers RCJ, van Enckevort LJG, Janssen A, van Orsouw NJ, et al. 2012. Sequence-based genotyping for marker discovery and co-

dominant scoring in germplasm and populations. *PLoS One* **7:** e37565. doi:10.1371/journal.pone.0037565
Van der Auwera GA, O'Connor BD. 2020. *Genomics in the cloud: using docker, GATK, and WDL in terra*, 1st ed. O'Reilly Media.
Wang J, Sun G, Ren X, Li C, Liu L, Wang Q, Du B, Sun D. 2016. QTL underlying some agronomic traits in barley detected by SNP markers. *BMC Genet* **17:** 103. doi:10.1186/s12863-016-0409-y
Wang P, Xiong Y, Gong R, Yang Y, Fan K, Yu S. 2019. A key variant in the *cis*-regulatory element of flowering gene *Ghd8* associated with cold tolerance in rice. *Sci Rep* **9:** 9603. doi:10.1038/s41598-019-45794-9
Yan H, Haak DC, Li S, Huang L, Bombarely A. 2022. Exploring transposable element-based markers to identify allelic variations underlying agronomic traits in rice. *Plant Commun* **3:** 100270. doi:10.1016/j.xplc.2021.100270
Zan Y, Payen T, Lillie M, Honaker CF, Siegel PB, Carlborg Ö. 2019. Genotyping by low-coverage whole-genome sequencing in intercross pedigrees from outbred founders: a cost-efficient approach. *Genet Sel Evol* **51:** 44. doi:10.1186/s12711-019-0487-1
Zhang GJ, Dong R, Lan LN, Li SF, Gao WJ, Niu HX. 2020. Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. *Int J Mol Sci* **21:** 707. doi:10.3390/ijms21030707
Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15:** 419. doi:10.1186/1471-2164-15-419
Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, et al. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res* **32:** e37. doi:10.1093/nar/gnh031

Received December 27, 2022; accepted in revised form April 26, 2023.

GR277628Ros

Queries

Marzia Rossato et al.

Q1 As outlined in our Instructions to Authors, you must use approved nomenclature for gene and protein names and symbols, as it applies for each organism, in text, tables, and figures. In addition, it is the journal's style to set gene symbols, alleles, and loci in italic, and proteins in Roman type. Please verify that all have been properly set throughout the manuscript.

Q2 Would you like to include author V.D.V. among the "Author contributions"?

Q3 Reference entry for "Baird et al. 2008" was updated to match details for this article record; please confirm accuracy of updated entry.

Q4 Reference entry for "Cooke et al. 2016" was updated to match details for this article record; please confirm accuracy of updated entry.

Q5 Reference entry for "Danecek et al. 2021" was updated to match details for this article record; please confirm accuracy of updated entry.

Q6 Reference entry for "Deng et al. 2022" was updated to match details for this article record; please confirm accuracy of updated entry.

Q7 Reference entry for "Elshire et al. 2011" was updated to match details for this article record; please confirm accuracy of updated entry.

Q8 Reference entry for "Friel et al. 2021" was updated to match details for this article record; please confirm accuracy of updated entry.

Q9 Reference entry for "Gu et al. 2016" was updated to match details for this article record; please confirm accuracy of updated entry.

Q10 Reference entry for "Guerra-García et al. 2021" was updated to match details for this article record; please confirm accuracy of updated entry.

Q11 Reference entry for "Homberger et al. 2023" was updated to match details for this article record; please confirm accuracy of updated entry.

Q12 Reference entry for "Jayakodi et al. 2023" was updated to match details for this article record; please confirm accuracy of updated entry.

Q13 Please check that "Li 2013" is the correct reference to cite for BWA-MEM.

Q14 Please check that "Li et al. 2009" is the correct reference to cite for SAMtools.

Q15 Reference entry for "Matvienko et al. 2013" was updated to match details for this article record; please confirm accuracy of updated entry.

Q16 Reference entry for "Montefiori et al. 2017" was updated to match details for this article record; please confirm accuracy of updated entry.

Q17 Reference entry for "Müller Paul et al. 2022" was updated to match details for this article record; please confirm accuracy of updated entry.

Q18 Reference entry for "Ogutcen et al. 2018" was updated to match details for this article record; please confirm accuracy of updated entry.

Q19 Reference entry for "Pavan et al. 2020" was updated to match details for this article record; please confirm accuracy of updated entry.

Q20 Reference entry for "Peterson et al. 2012" was updated to match details for this article record; please confirm accuracy of updated entry.

Q21 Reference entry for "Prezza et al. 2020" was updated to match details for this article record; please confirm accuracy of updated entry.

Q22 If this bioRxiv preprint (Ramsay et al. 2021) has now been published, please update the citation and reference with the journal details.

Q23 Reference entry for "Ren et al. 2021" was updated to match details for this article record; please confirm accuracy of updated entry.

Q24 Please check that "Thorvaldsdottir et al. 2013" is the correct reference to cite for IGV.

Q25 Reference entry for "Truong et al. 2012" was updated to match details for this article record; please confirm accuracy of updated entry.

Q26 Please provide publisher location for "Van der Auwera and O'Connor 2020".

Q27 Reference entry for "Wang et al. 2016" was updated to match details for this article record; please confirm accuracy of updated entry.

Q28 Reference entry for "Wang et al. 2019" was updated to match details for this article record; please confirm accuracy of updated entry.

Q29 Reference entry for "Yan et al. 2022" was updated to match details for this article record; please confirm accuracy of updated entry.

Q30 Reference entry for "Zan et al. 2019" was updated to match details for this article record; please confirm accuracy of updated entry.

Q31 Reference entry for "Zhang et al. 2020" was updated to match details for this article record; please confirm accuracy of updated entry.

Q32 Reference entry for "Zhao et al. 2014" was updated to match details for this article record; please confirm accuracy of updated entry.

Q33 Reference entry for "Zhulidov et al. 2004" was updated to match details for this article record; please confirm accuracy of updated entry.