Sequencing the functional genome of plant species enhanced by the in-vitro use of CRISPR-CAS9 to remove repetitive elements from Lentil and Wheat genomes.

Keith Brown[1], Jon Armstrong[1], Azeem Siddique[1], Sridhar Ranganathan[1], Marzia Rossato[2], Jesse Poland[3] and Sandesh Shrestha[3]
1. Jumpcode Genomics, San Diego, CA; 2. University of Verona, Verona, IT, 3. Kansas State University, Wichita, KS

Introduction

Hybrid capture technologies have been used for decades to sequence unique protein coding regions across plant and animal species. However, the are flaws in this approach; allelic dropout, representation bias and limited specificity, are well documented. There is a need to expand beyond protein coding elements to other functional elements such as promoters, enhancers, UTRs, ATAC-seq accessible sites, CpG islands as well as introns while maintaining the cost benefits of genome wide targeted sequencing. Here we present CRISPRclean® depletion technology which reduces the sequence representation of repetitive elements from complex plant genomes including wheat and lentil. Nearly 600,000 single guide RNA:CAS 9 ribonucleoprotein (RNP) complexes were applied to target repetitive elements (transposable elements and simple repeats) that make up the majority of these genomes. By removing these abundant repetitive elements before sequencing, genotyping yields are dramatically increased while costs are significantly reduced.

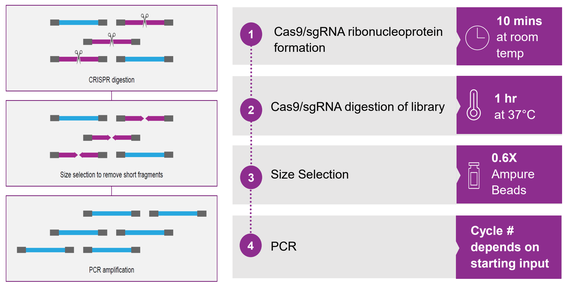

Figure 1. Overview of CRISPRclean technology. Double stranded cDNA library products are cleaved, following adapter ligation, using CAS9 and specifically designed guide RNAs. Cleaved fragments do not contain adapters on both ends and will not amplify in subsequent PCR reactions. The cleaved fragments are removed from the library pool during short fragment cleanup.

Conclusions

CRISPRclean depletion technology offers more comprehensive and lower cost coverage into the coding areas of complex and highly repetitive genomes. Ultimately, providing researchers with greater biological insight.

Lentil
- Decrease in repetitive content by 40%.
- Increase in coverage over coding regions by 2.8x.
- Genotyped bases equal to WGS at same amount of sequencing.
- 2.3-4x lower cost than WGS or exome genotyping, respectively.

Wheat
- Decrease in repetitive content by 53%.
- Increase in percent read alignment over coding regions by 14x.
- 32% of CDS bins at 5-fold or greater read enrichment in depleted samples.
- 2.3-4x lower cost than WGS or exome genotyping, respectively.

Lentil Methods

DNA was extracted from Lentil samples and libraries were generated using the iGenomx Riptide kit. 8 to 96 samples were pooled, at 50 – 100 ng of total DNA, for depletion. Single guide RNAs (sgRNAs) were designed against transposable elements and simple repeats (~85% of genome; REPEATS). Guides were selected for inclusion if they hit repeat areas of the genome > 25 times. Areas of the genome that were were to be retained included regulatory regions, derived from ATAC-seq data, coding and other non-repetitive regions making up a total of ~15% of the genome (CODING; Figure 1). The final guide design include 566,722 unique sgRNAs targeting 67.7% of the lentil genome. The sgRNAs were subdivided into pools based on cut frequencies and depletions were performed with all guides pooled, in groups of 3 pools or individual pools (n=11). Libraries were split and non-depleted and depleted replicates were generated. The replicates were sequenced on a Novaseq6000 at 150 base PE reads with 0.5 to 4x genome coverage. Sequencing data was analyzed for coverage of repeat and coding regions.

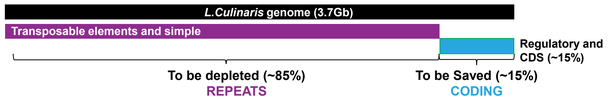

Figure 1. Target regions for CRISPR depletion from lentil genome.

Results

CRISPRclean enables genotyping equal to whole genome sequencing at a much lower cost.
Multiple genotyping methods were compared in Figure 2, including whole genome sequencing (WGS), iGenomx library prep with CRISPRclean depletion (iGx+CRISPRclean, purple dashed box), exome (Exome) and genotyping by sequencing (GBS). GBS shows sparse interspersed markers at a low cost, yet bases are lost for genotyping. WGS provides a much more complete view of the genome, however at a much higher cost. Exome sequencing concentrates data but at a higher cost and longer workflow. By reassigning sequencing reads to informative areas of the genome, such as coding regions, one is able in decrease the coverage over repetitive regions by 40% while increasing the coverage over coding and regulatory regions by nearly 3x (Figure 3). Only CRISPRclean provides a greater number of bases to genotype than exome sequencing, yet at a much lower cost than either WGS or exome sequencing (Figure 4).

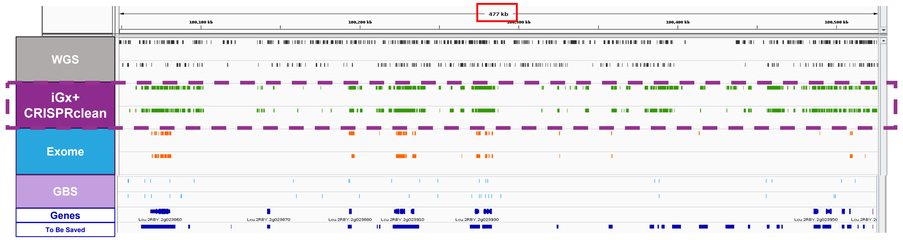

Figure 2. Tracks of genotypable positions for each method shown in a window of 477 kb of the lentil genome. Genes and to be saved annotations are shown at the bottom and the collapsed per base coverage of each method is shown above

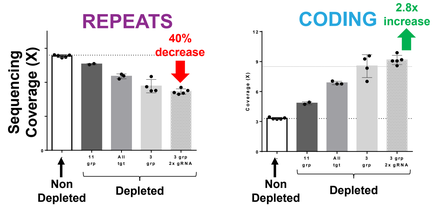

Figure 3. Repeat depletion increases coverage on meaningful regions

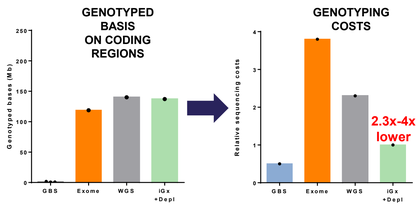

Figure 4. Lower relative sequencing costs to get same number of genotyped basis

Wheat Methods

DNA was extracted from Chinese Spring (CS) and Jagger wheat samples and libraries were generated using the NEB Ultra II RNA library prep. Single guide RNAs (sgRNAs) were designed using a k-mer approach where sgRNAs were designed against reads aligning to transposable elements and unmapped reads (~85% of genome). Coding sequences (CDS; CODING) were tagged to retain making up a total of ~15% of the genome (Figure 1). The final guide design included 30,00 unique sgRNAs targeting 78% of the wheat genome. The depletions were performed with all guides pooled. Libraries were generated with a single mock (no depletion) and 4 depleted. The samples were sequenced on a Hiseq2000 at 100 bp PE reads with ~1x genome coverage. Sequencing data was analyzed for coverage of repeat and coding regions.

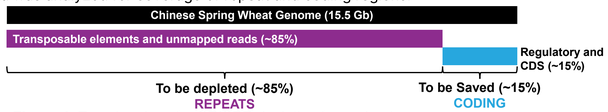

Figure 1. Target regions for CRISPR depletion from wheat genome

Results

CRISPRclean increases coverage over coding regions and enables genotyping at a lower cost than GBS.
CDS regions of the CS and Jagger genomes were split into 110,790 bins (average bin size = 3,065 bp) and Bedtools was used to count reads from the mock and depleted samples. A fold increase was realized if the depleted samples contained more reads in the bin than the mock (Figure 2 - Jagger, red dots). A fold decrease was shown if the mock sample bin contained more reads than the depleted (Figure 2 - Jagger, blue dots). Coverage analysis showed a 53% decrease in coverage over repeat areas targeted for removal and a concomitant increase of 14-fold coverage over coding bases (Figure 3). To obtain a more granular view of read coverage over CDS regions, The fold enrichment of reads (depleted / mock) was calculated for multiple levels of coverage and 32% of CDS bins showed 5-fold or greater read coverage in the depleted samples (Figure 4).

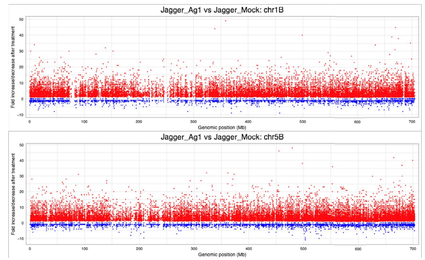

Figure 2. Fold increase/decrease of reads in CDS bins. Read fold changes were calculated by dividing the number of aligning reads from the the depleted sample by the mock across chromosome 1B and 5B in Jagger wheat strain.

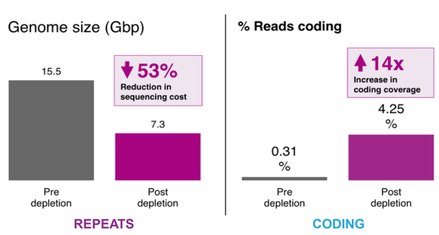

Figure 3. Repeat depletion increases coverage over meaningful regions translating to a reduction in sequencing cost and increased power for genotyping.

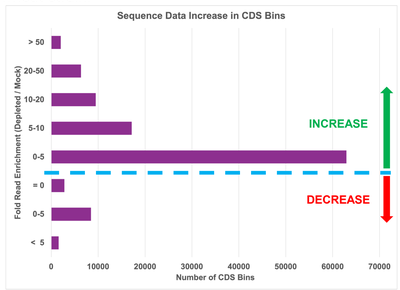

Figure 4. Large read enrichment over CDS regions.

To learn more visit jumpcodegenomics.com