Application of CRISPR-Cas9 and next-generation sequencing to resolve highly homologous genes in the human genome

Hegde MR[1], Ma Z[1], Kothandaraman A[1], Balciuniene J[1], Guo F[1], Brown K[2], De La Toba D[2], Choudhary S[2], Siddique A[2], Ranganathan S[2], Armstrong JR[2]

[1]PerkinElmer, Waltham, MA
[2]Jumpcode Genomics, San Diego, CA

Introduction

Next-generation sequencing (NGS) based targeted gene panels, exome sequencing (ES), and genome sequencing (GS) are now routinely used to interrogate large sets of genes for diagnostic use. Irrespective of the chosen assay, genes with high sequence homology continue to be a major challenge for short- read technologies and can lead to false-positive and false-negative diagnostic errors. Long-read sequencing has the potential to resolve this issue for many genes, but analysis of the homologous sequences typically necessitates advanced bioinformatics pipelines that have not been validated for clinical use. Traditionally, laboratories have used targeted Sanger sequencing and/or long-range PCR techniques to resolve these genes. These methods are gene-specific, difficult to design and expensive to perform in a clinical setting. The target regions with homology are complex and can be divided by varying degrees of homology, medical relevance, and type of homology (functional homolog, known pseudogene, partial or within-gene homology, uncharacterized noncoding region). To address this unmet need, we have described a proof-of-concept for using the CRISPRclean® technology which harnesses the specificity of the CRISPR-Cas9 system to degrade abundant, uninformative sequences in next-generation sequencing libraries.

Figure 1

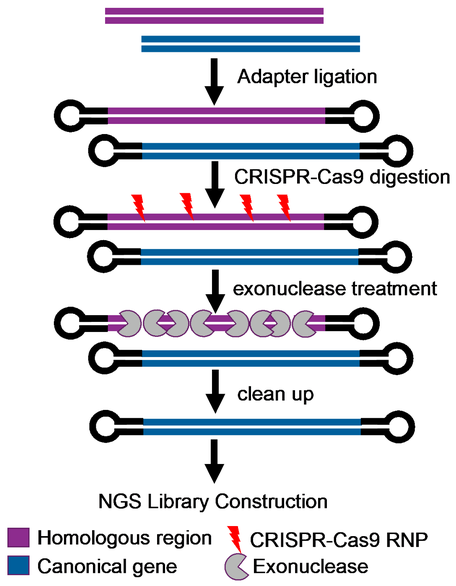


Figure 1. Overview of CRISPRclean technology. Double stranded genomic fragments are cleaved, following adapter ligation, using CAS9 and guide RNAs designed to regions homologous to canonical genes. Cleaved fragments are digested with exonuclease and removed following cleanup. A next-generation sequencing library is generated from the remaining genomics fragments.

Methods

Twenty clinically relevant genes with low mapability score, due to the presence of highly homologous sequences, were identified. We developed a two-step approach to pinpoint genomic regions for depletion. First, we enumerated all potential homologous regions by mapping the DNA sequence of the parent gene to the reference genome. Subsequently, we quantified sequence similarity by performing pairwise alignments between the parent gene and homologous regions. Finally, we identified all available CRISPR target sites (recognized by SpCas9) on these homologous regions and guides were stringently filtered for off targets to the canonical genes. High molecular weight DNA was obtained from NA12878 and NA24631, nick repaired, and ligated to short hairpin adapters on the 3' and 5' ends to produce a library of exonuclease deficient "protected" DNA. Protected libraries were incubated with the pool of pre-formed ribonucleoprotein complexes (RNPs) producing double-stranded cuts in undesired library fragments. Cleaved fragments with exposed phosphodiester bonds are then degraded by exonuclease. Remaining DNA fragments were converted into standard NGS libraries preparation followed by 2x150 paired-end reads sequencing (Figure 1).

Clinically relevant genes include: ABCD1 (adrenoleukodystrophy), ALG1 (disorder of glycosylation), CHEK2 and PMS2 (cancer syndromes), CYP21A2 (adrenal hyperplasia), FANCD2 (Fanconi anemia), GBA (Gaucher disease), GUSB and IDS (mucopolysaccharidosis), NF1 (Neurofibromatosis), OTOA and STRC (hearing loss), PIK3CA (overgrowth syndrome), PKD1 (polycystic kidney disease), SBDS (Shwachman-Diamond syndrome), SDHA (pheochromocytoma-paraganglioma), SLC6A8 (creatine transporter deficiency), SMN1 (spinal muscular dystrophy), TYR (oculocutaneous albinism), and VWF (von Willebrand disease). Among them, CYP21A2, GBA, SBDS, SDHA, SLC6A8, SMN1, and STRC are known to be "blind spots" for short-read NGS due to the presence of highly homologous sequences spanning across the full length of the gene.

Results

CRISPR-Cas9 treatment is effective at depleted homology intervals.
Figure 2a-b show the depletion rate of genomic regions homologous to the 20 canonical genes in NA12878 and NA24631 for condition 2 (2-hour incubation). The length of the homologous regions is shown on the x-axis and depletion rate on the y-axis. The depletion rate for separate homologous regions of the same length were averaged. Longer regions of homology are generally depleted at a higher rate because they contain a larger number of guide sites. On average, CRISPR-Cas9 treatment resulted in nearly 50% reduction in reads across all homology regions for both NA12878 and NA24631.

Figure 2a

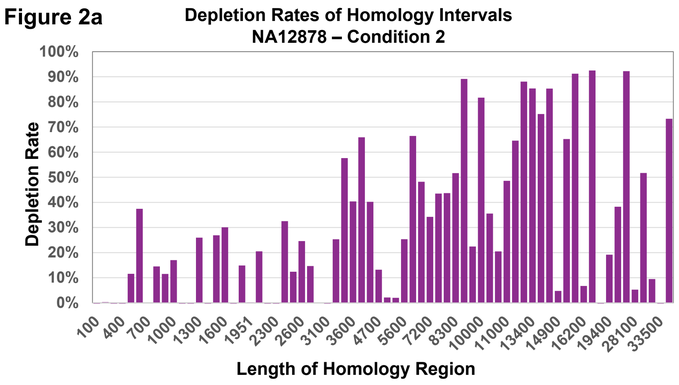
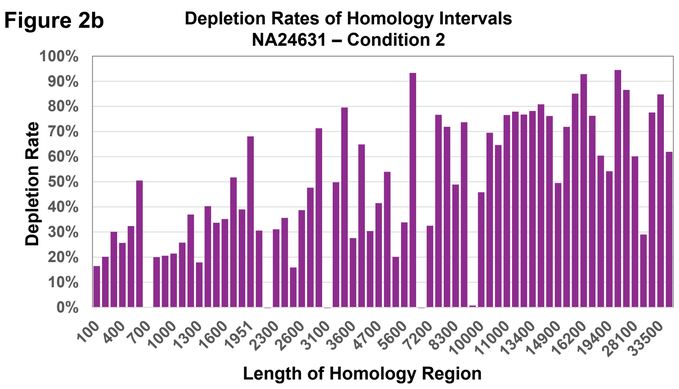
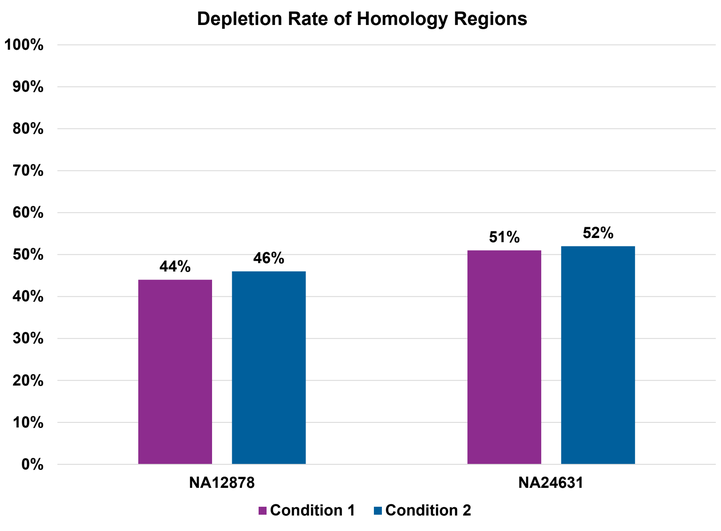


Figure 2b



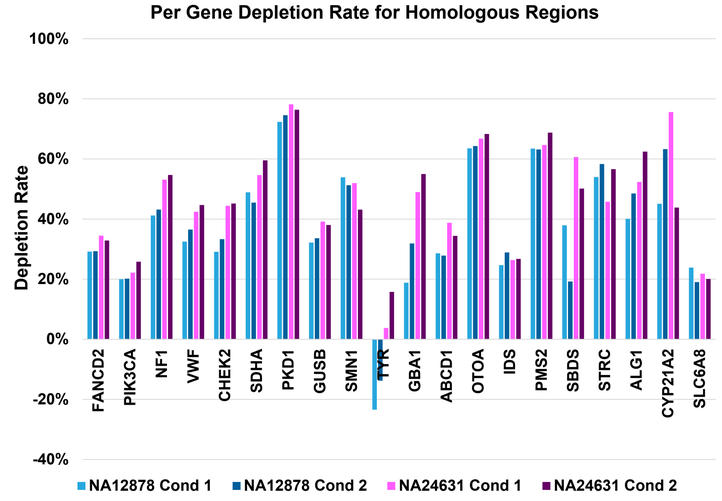
The average depletion rate across all homology regions is ~50%.
Figure 3 shows ~50% depletion rate across all targeted genomic homology regions in samples NA12878 and NA24631, for condition 1 (1-hour incubation; purple bars) and 2 (2-hour incubation; blue bars). Treatment with CRISPR-Cas9 removes fragments originating from homology regions and reduces their read representation in sequencing libraries.

Figure 3



CRISPR-Cas9 depletion reduces the number of reads aligning to genomic intervals homologous to canonical genes.
Figure 4 shows the depletion rate for genomic intervals homologous to the 20 canonical genes in samples NA12878 and NA24631. The read coverage for each homology reagion was calculated, averaged on a per gene basis and compared between control and depleted samples. Condition 2 provides slightly better results across most genes and NA24631 shows slightly higher depletion percentages as compared to NA12878. Depletion with CRISPR-Cas9 is consistent across samples and conditions within each gene providing evidence that depletion rates are affected by the parameters of the genomic regions homologous to canonical genes and not the activity of CRISPR-Cas9 within a sample. The regions homologous to the TYR gene shows an enrichment of reads in NA12878, likely due to inefficient cutting of the homologous genomic intervals by CRISPR-Cas9, and/or the redistribution of sequencing reads to this gene.

Figure 4



CRISPR-Cas9 depletion reduces read alignment to homologous genomics regions.
Figures 5a-b show IGV plots of reads aligning to pseudogenes of the ALG1 (CARD11) and GBA (GBAP1) genes. The pseudogene features are shown in blue (Homologous gene), the region of homology, identified with our bioinformatics pipelines, is shown in red (Homology region), the individual guides targeted the homology regions are shown in purple (Guide targets) and the read alignment plots are shown for untreated and depleted samples. Depleted samples show a large reduction in the number of reads aligning to homologous regions, thus contributing fewer confounding read alignments to canonical genes.

Figure 5a

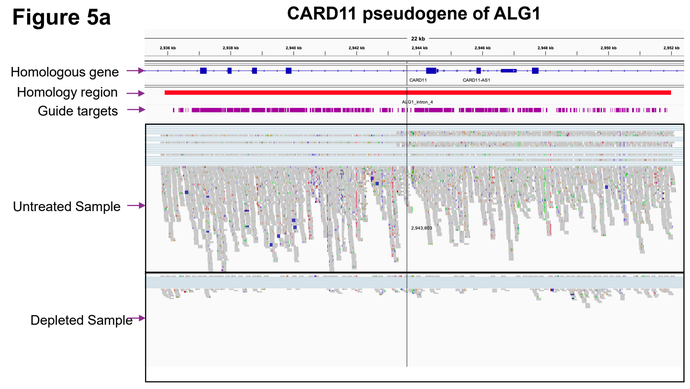
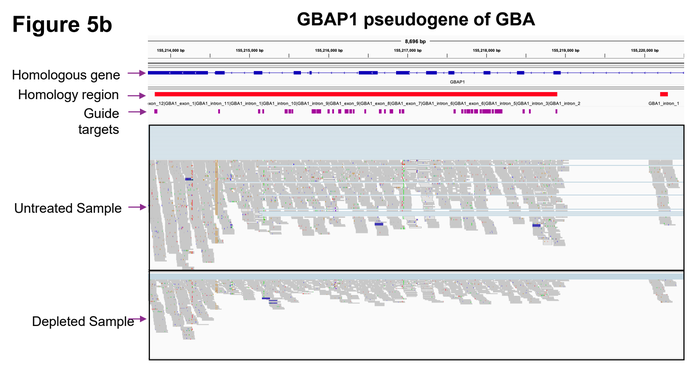


Figure 5b



Conclusions

- CRISPR-Cas9 depletion enables the removal of genomic fragments homologous to pseudogenes, prior to library generation for next generation sequencing.
- Depletion is effective in removing an average ~50% of the reads emanating from homologous genomic regions.
- The depletion rate generated by CRISPR-Cas9 depletion is consistent across sample types.
- Jumpcode's CRISPR-Cas9 system is highly programmable and can be configured for any gene of interest.

Future Work

- Test different conditions and guide design methods to further increase the percent of fragments and reads removed from homology regions.
- Investigate the addition or subtraction of guides to help balance the resulting read coverage for a sample.
- Explore the effect that size selection of genomic DNA has on the percent of fragments and reads removed from homology regions.