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.

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), PIK3CA (cancer syndromes), and STRC (spinal muscular dystrophy), TYR (oculocutaneous albinism), and VWF (von Willebrand disease). Among them, CYP21A2, GBA, SOD1, STRC, 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 depletion reduces the number of reads aligning to homologous regions.

CRISPR-Cas9 depletion reduces the number of reads aligning to homologous genomic 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.

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.