Improved Transcript Discovery in Single Cell RNA-Seq with CRISPRclean®

Jon Armstrong, Jon Bezney, Sridhar Ranganathan and Azeem Siddique

Introduction

There are vast opportunities to apply single cell genomics to the unanswered in biology. With its unique ability to study the individuality of cells, single cell genomics has become an increasingly common, widely adopted approach. This research will not only have important implications for disease but can revolutionize biology. In order to profile large amounts of individual cells, transcriptional profiling with single cell RNA Seq provides the most thorough analysis for large amounts of individual cells. This empowers researchers with the knowledge of what genes are expressed, in what quantity, and how they differ across the cells within a sample at the single cell level.

Single cell studies require a significant amount of sequencing to understand transcript levels within individual cells. It can take up anywhere from 50,000 to 150,000 reads per cell in order to get the informative, low abundant genes you want for calling differences between cells—with reads for a single sample often exceeding 150 million reads.

Traditionally, single cell data processing incorporates certain filtering and normalization steps prior to cell clustering and downstream interpretation. Instead of removing those reads in-silico, CRISPRclean removes those reads in-vitro ahead of sequencing, redistributing ~50% sequencing reads to unique biologically relevant transcripts —allowing you to maximize gene and UMI sensitivity.

CRISPRclean leverages Cas9 and a specifically designed guide set to remove reads filtered by secondary analysis. CRISPRclean Single Cell RNA Boost Kit gives you the ability to cut through the noise with minimal impact on your workflow, and maximum confidence in your results.

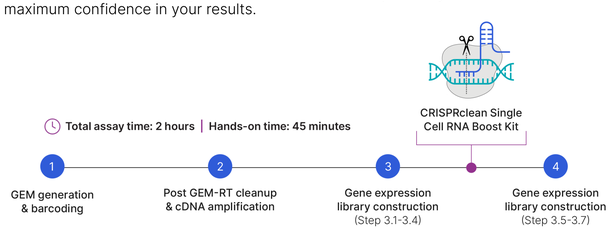

Figure 1: CRISPRclean is a simple 3-step protocol integrated into the 10x Genomics® Chromium™ Next GEM Single Cell 3' workflow prior to the final PCR amplification.

Methods

Content for depletion was designed by analyzing a cohort of 14 publicly available single cell 10x Genomics® data sets, roughly 30-50% of reads aligned to the genome but not the transcriptome, and thus, were ignored. Guides were designed to deplete these genomic intervals in addition to the highest expressed protein coding ribosomal and mitochondrial genes and non-variable genes removed by Pegasus software (Table 1). In-silico depletions were performed across these 14 data sets and showed an average read redistribution of ~50% (data not shown).

PBMC samples were isolated from a single donor and a single control library and triplicate depleted libraries prepared using 10x Genomics® Chromium™ Next GEM Single Cell 3' Reagent Kit (v3.1) protocol. 150 ng of cDNA product, collected at the end of Step 3.4 in the Next GEM Single Cell 3' Reagent Kit protocol, was used as starting material for CRISPRclean Single Cell RNA Boost Kit protocol. The 10x protocol was resumed at step 3.5 with depleted libraries. Four libraries were loaded onto a P3-100 cycle flow cell and sequenced on a NextSeq™ 2000. We recovered ~11,000 cells per sample, each with >25,000 reads per cell.

Methods, cont.

Raw sequencing reads were processed through Cell Ranger and then used the filtered feature barcode matrix. Secondary analysis was performed using Pegasus, the python based single cell toolbox. This included cell filtering, normalization as well as the generation of UMAP plots and single cell trajectory plots.

Designed to deplete	Description
Ribosomal, mitochondrial	Poly A ribosomal, mitochondrial genes. Abundant ribosomal and mitochondrial rRNA
Unaligned reads	Reads that do not align to the transcriptome
Non-variable genes	155 genes commonly expressed across 14 different sample types

Table 1: CRISPRclean Single Cell RNA Boost Kit redistributed ~50% of reads through in-silico depletion by depleting unaligned intervals and abundant protein coding ribosomal and mitochondrial genes.

Equivalent performance with ~50% less reads

	Undepleted	CRISPRclean Depleted
Reads per cell	27,000	13,500
Number of cells	11,479	11,479
Genes per cell	1,353	1,340
UMIs per cell	1,976	1,905

Table 2: Detected the same number of genes and UMIs with reduced sequencing.

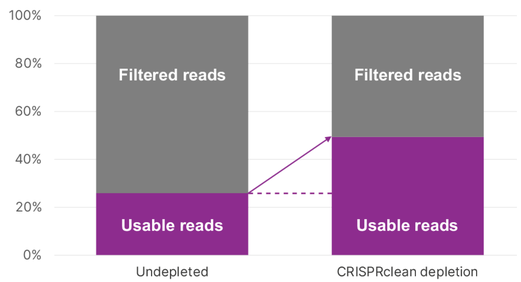

Figure 2: 100% increase in reads mapped to the transcriptome with CRISPRclean. Filtered reads were not used for secondary analysis.

1.5x improvement in detection sensitivity with depletion

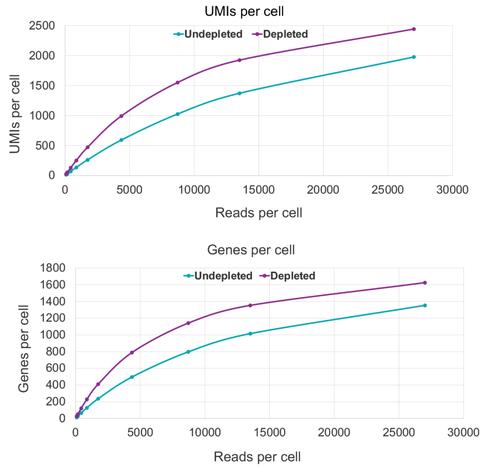

Figure 3: 1.5x improvement in detection sensitivity of genes and UMIs per cell for depleted PBMC samples compared to undepleted sample.

Results

We were able to reserve flow cell space for fragments that would generate informative sequencing reads by performing depletion between adapter ligation and final PCR and cleanup. The number of informative reads or reads aligning to the transcriptome and utilized by primary and secondary analysis, increased by 2-fold (Figure 2). This 2-fold increase led to a concomitant increase of approximately 1.5-fold in unique molecular identifiers (UMIs) and genes identified per cell, as shown in Figure 3.

To elucidate the relationship between increased UMIs and genes per cell on cell identification, UMAP plots were generated for depleted and non-depleted samples (Figure 4). Identical cell types were identified in both non-depleted and depleted samples, showing depletion does not perturb cell type identification. However, the UMAP plots for depleted samples showed one additional minor cell type that was not present in the non-depleted UMAP plots. Since 2-fold more reads aligned to the transcriptome in the depleted sample, we investigated the effect of removing half of the sequencing data from the depleted samples by random down-sampling. Table 2 shows when half of the sequencing data is removed from the depleted samples, the cell counts and frequencies are comparable, with the exception of the two additional cell types identified in the depleted samples.

Gain a deeper view of expression profiles

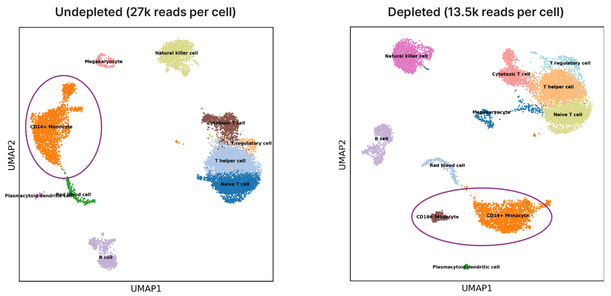

Figure 4: UMAP plots of cell clusters with and without depletion using PBMC samples showed that depletion did not perturb cell type calls. One additional cell type was identified in depleted samples.

	Undepleted (27k reads /cell)		Depleted (13.5k reads/cell)	
Cell Type	**Cell Number**	**Cell Frequency**	**Cell Number**	**Cell Frequency**
Naïve T Cell	2,414	22%	2,216	21%
CD14+ Monocyte	2,260	21%	1,949	18%
T Helper Cell	2,254	21%	2,372	22%
Natural Killer Cell	1,468	14%	1,444	13%
Cytotoxic T Cell	994	9%	982	9%
B Cell	693	6%	683	6%
Red Blood Cell	246	2%	272	3%
T Regulatory Cell	203	2%	295	3%
Megakaryocyte	165	2%	288	3%
Plasmacytoid Dendritic Cell	67	1%	61	1%
CD16+ Monocyte	**0**	**0%**	**209**	**2%**
	10,764		10,771	

Table 3: Improvement in cell frequency for CRISPRclean depleted PBMC samples compared to undepleted samples using Pegasus, a secondary analysis single cell toolbox.

Conclusions

- Simple 3-step protocol integrated into the 10x Genomics® Chromium™ 3' workflow.
- Depletes sequences not used for secondary analysis including unaligned reads, ribosomal, mitochondria and non-variable genes.
- Gain a deeper view of expression profiles.
- 1.5-fold increase in UMIs and genes per cell.
- 1 additional PBMC cell type identified in depleted samples.
- Genes / cell and UMIs / cell are comparable between full sequencing read numbers without depletion and ½ sequencing read numbers with CRISPRclean depletion.

To learn more visit jumpcodegenomics.com