Improved Single Cell Transcriptional Profiling with CRISPRclean™

Azeem Siddique,[1,5] Gaia Suckow,[1,5] Jay Carey,[3] Phillip Ordoukhanian,[1,5] Keith Brown[5] (keith@jumpcodegenomics.com), Steve Head,[1,5] Stan Nelson,[4,5] Doug Evans,[2] Ali Torkamani,[2] and Tom Alsaigh[2] | 1. The Scripps Research Institute, La Jolla, CA; 2. Scripps Research Translational Institute, Scripps Research Institute, La Jolla, CA; La Jolla, CA; 3. Fulcrum Genomics, Somerville, MA; 4. David Geffen School of Medicine, UCLA, Los Angeles, CA; 5. JumpCode Genomics, Carlsbad, CA

Introduction

Transcriptional profiling has been revolutionized with high throughput, massively parallel, single cell RNA sequencing. Microfluidic/microdroplet technologies recently introduced to the market produce complex data sets that allow investigators to deconvolute complex cell mixtures, identify cell types present in healthy and diseased tissues, and create cell type-specific transcriptional signatures. These technologies are dramatically enhancing our ability to identify transcriptional and cellular perturbations driving disease at the individual cell level. However, the techniques are expensive and enable only sparse sampling of RNAs from each cell, with many genes represented by only 1 or 2 sequencing reads. This limit is due partially to the fact that libraries are dominated by an abundance of housekeeping RNAs, which dominate sequencing reads and limit detection of even moderately expressed transcripts that often drive biological differences between cell types. To improve performance of single cell RNA sequencing (scRNA) sequencing, we have developed a CRISPR-based depletion system called CRISPRclean. CRISPRclean is a post-library depletion method that utilizes CRISPR-Cas technology to deplete uninformative molecules from a sequencing library, allowing sequencer capacity to be efficiently used for greater sensitivity and/or reduced cost. Here we apply CRISPRclean to remove 100 overexpressed, housekeeping transcripts that are abundantly expressed in all cell types. The genes consist of ribosomal and mitochondrial protein coding transcripts (mRNA with polyA tails). We treat fully prepared 10x Genomics single cell libraries with CRISPRclean and demonstrate depletion of sequencing reads associated with the target genes and enhanced sensitivity for other genes.

Methods:

We analyzed multiple, published single-cell studies that employed cells from a variety of tissues and different disease states (including neurological disorders, cardiovascular disease, and soft tissue cancer) to

Table 1: Genes targeted for depletion

Mitochondrial	Ribosomal								
MT-ATP6	RPL10	RPL18A	RPL27	RPL35A	RPL5	RPS11	RPS19BP1	RPS27L	RPS6KA3
MT-ATP6	RPL10A	RPL19	RPL27A	RPL36	RPL6	RPS12	RPS2	RPS28	RPS6KA4
MT-CO2	RPL11	RPL21	RPL28	RPL36A	RPL7	RPS13	RPS20	RPS29	RPS6KA5
MT-CO3	RPL12	RPL22	RPL29	RPL36AL	RPL7A	RPS14	RPS21	RPS3	RPS6KB1
MT-CYB	RPL13	RPL22L1	RPL3	RPL37	RPL7L1	RPS15	RPS23	RPS3A	RPS6KB2
MT-ND1	RPL13A	RPL23	RPL30	RPL37A	RPL8	RPS15A	RPS24	RPS4X	RPS6KC1
MT-ND2	RPL14	RPL23A	RPL31	RPL38	RPL9	RPS16	RPS25	RPS4Y1	RPS7
MT-ND3	RPL15	RPL24	RPL32	RPL39	RPLP1	RPS17	RPS26	RPS5	RPS8
MT-ND4	RPL17	RPL26	RPL34	RPL4	RPLP2	RPS18	RPS27	RPS6	RPS9
MT-ND5	RPL18	RPL26L1	RPL35	RPL41	RPS10	RPS19	RPS27A	RPS6KA1	RPSA

determine a candidate gene list for depletion. Our goal was to target genes of little biological interest that are consistently at high abundance in single cell sequence data. Transcript abundance was ranked first by read count and then filtered for a low coefficient of variation (between disease and normal states) of < 1.1. A set of 100 genes were identified as overabundant and invariant across all sample sets (Table 1, previous page). All of the genes targeted for depletion are mitochondrial and ribosomal protein-coding genes.

CRISPR guide RNAs were designed to the 100 target genes using the Jumpcode Genomics guide design pipeline. The guide RNAs were filtered for optimal sequence, secondary structure, *in vitro* cleavage efficiency and for minimal off-target effects. DNA templates containing a phage T7 RNA polymerase promoter upstream of the full length CRISPR single guide RNA sequence were synthesized on a DNA oligo array (Agilent, Santa Clara, CA), PCR amplified and transcribed *in vitro* to produce a pool of guide RNAs for use in the CRISPRclean protocol. Depletion was tested on single-cell RNA-Seq libraries generated via the 10x Genomics Chromium system at the Scripps Research Institute (San Diego, CA). The samples, for which data is presented here, comprise of matched diseased and normal atherosclerotic carotid artery tissue collected during carotid endarterectomy from three donors. The normal tissue samples contain approximately 3,500 cells and the diseased tissue samples have approximately 11,000 cells. Each 10x Genomics library was sequenced on an Illumina NextSeq instrument with approximately 250 million 2 x 150 PE reads corresponding to ~60K reads per cell for the normal and ~24K reads per cell for the diseased tissue samples (sequencing saturation of ~85% and ~65% for normal and diseased sample types, respectively). The prepared libraries were subsequently depleted with the CRISPRclean scRNA depletion product, as described in Figure 1. The samples were sequenced on the same NextSeq instrument again, for a similar read length and depth as the original (non-depleted) samples.

All sequence data was processed through Cell Ranger, subject to standard mitochondrial and doublet detection algorithms, and partitioned into cell-types using Monocle3. Differential expression was performed using a generalized linear model to control for patient-specific transcriptomic effects.

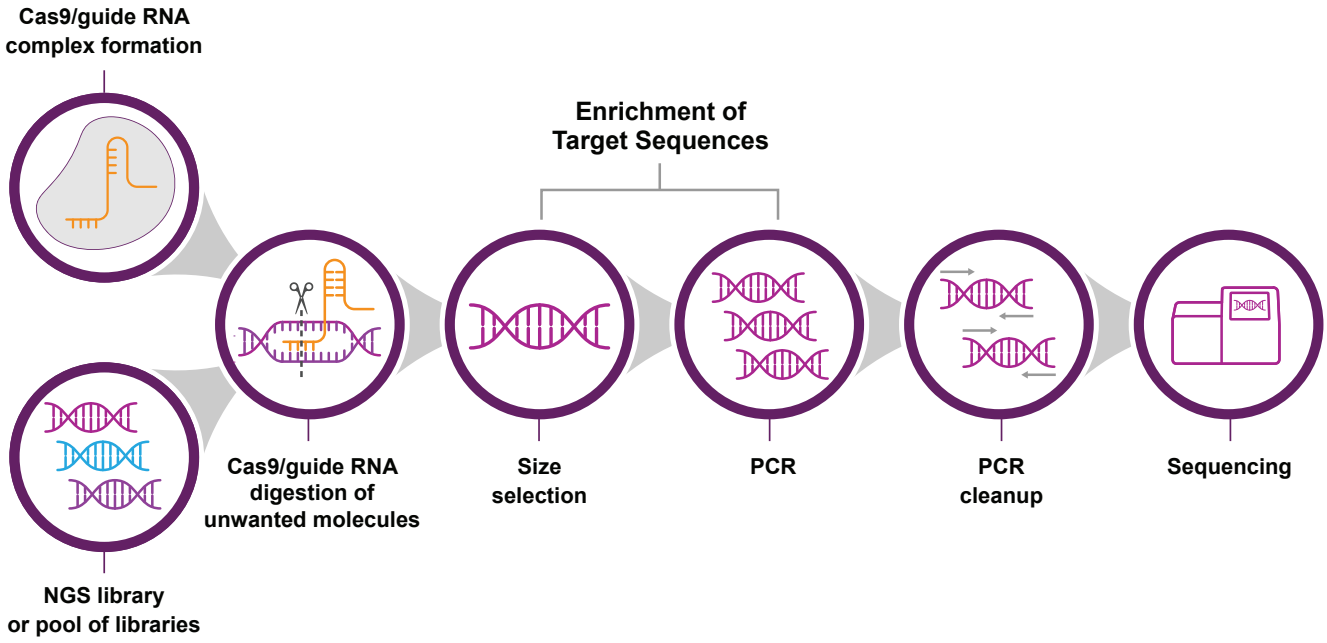

Figure 1: Schematic of CRISPRclean workflow.

Results:

Effect of abundant gene depletion on single cell gene expression

Differential gene expression results demonstrate that the vast majority of genes targeted for depletion display the greatest reduction in expression across all genes analyzed in both healthy and diseased samples.

Figure 2B shows that mitochondrial gene UMI counts are dramatically lower after depletion. As a result of mitochondrial and ribosomal gene depletion, total UMI counts per cell are lower as well (Figure 2A). Importantly, depletion results in an increase of 2,747 UMIs (12.6% of all genes in the library) derived from healthy tissue and 4,260 (18.4% of all genes in the library) derived from diseased tissue (Table 2) indicating that depletion has improved the ability to detect certain genes.

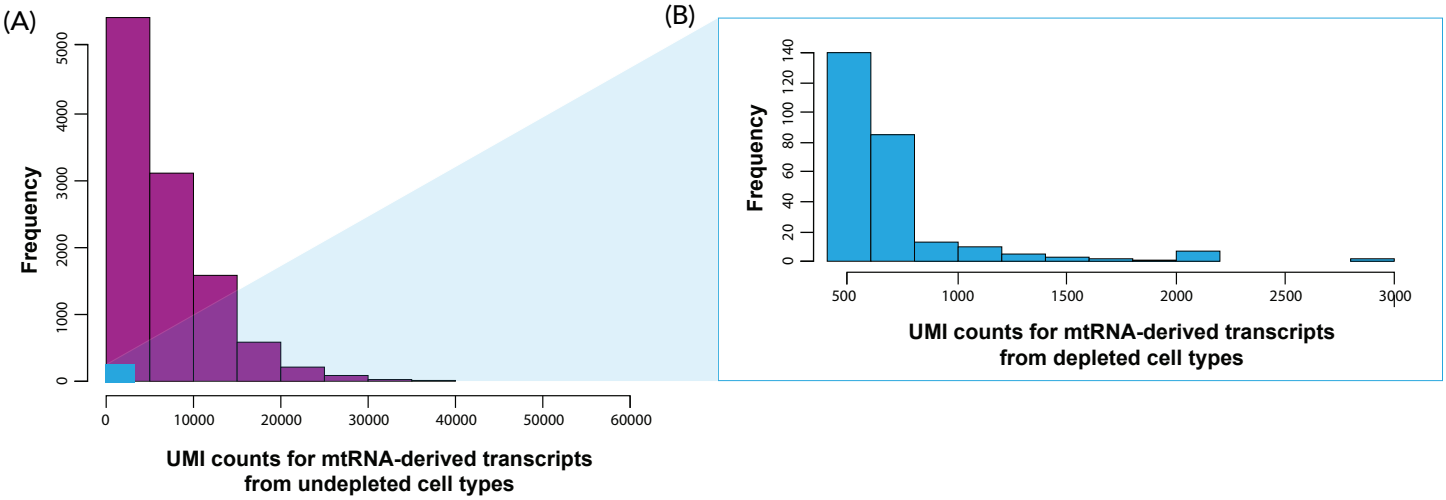

Figure 2: UMI counts for mitochondrial gene transcripts (mtRNA) targeted for CRISPRclean depletion from (A) undepleted and (B) depleted disease cell types. Results plot the number of unique UMIs on the X-axis and the frequency of the individual UMIs from mitochondrial-derived transcripts on the Y-axis. Results show > 95% reduction in total and unique UMI frequency of the depleted library targets.

Table 2: The number of genes (relative to total genes for which transcripts are observed in the sequencing data) for which UMI counts are increased, reduced or unaffected as a result of CRISPRclean-mediated depletion.

Genes with altered UMI abundance			
	Increased	**Decreased**	**Unaffected**
Healthy tissue	2747	142	18933
Diseased tissue	4260	228	18719

Since depleted libraries are derived from undepleted ones, 10x Genomics cell barcodes are identical between depleted and undepleted libraries of the same type. Consequently, cells that have the same barcodes in the depleted and undepleted libraries can be compared for alterations in UMI numbers per cell. The results of the comparison are shown in the form of linear regression plots for selected individual genes in Figure 3. The plots show a dramatic reduction in UMIs per cell for the genes selected for depletion (see UMI/cell plot for RPS12 as an example). The plots also show an increase in UMIs per cell for certain (non-targeted) genes as a consequence of depletion (see UMI/cell plot for CCL14 as an example).

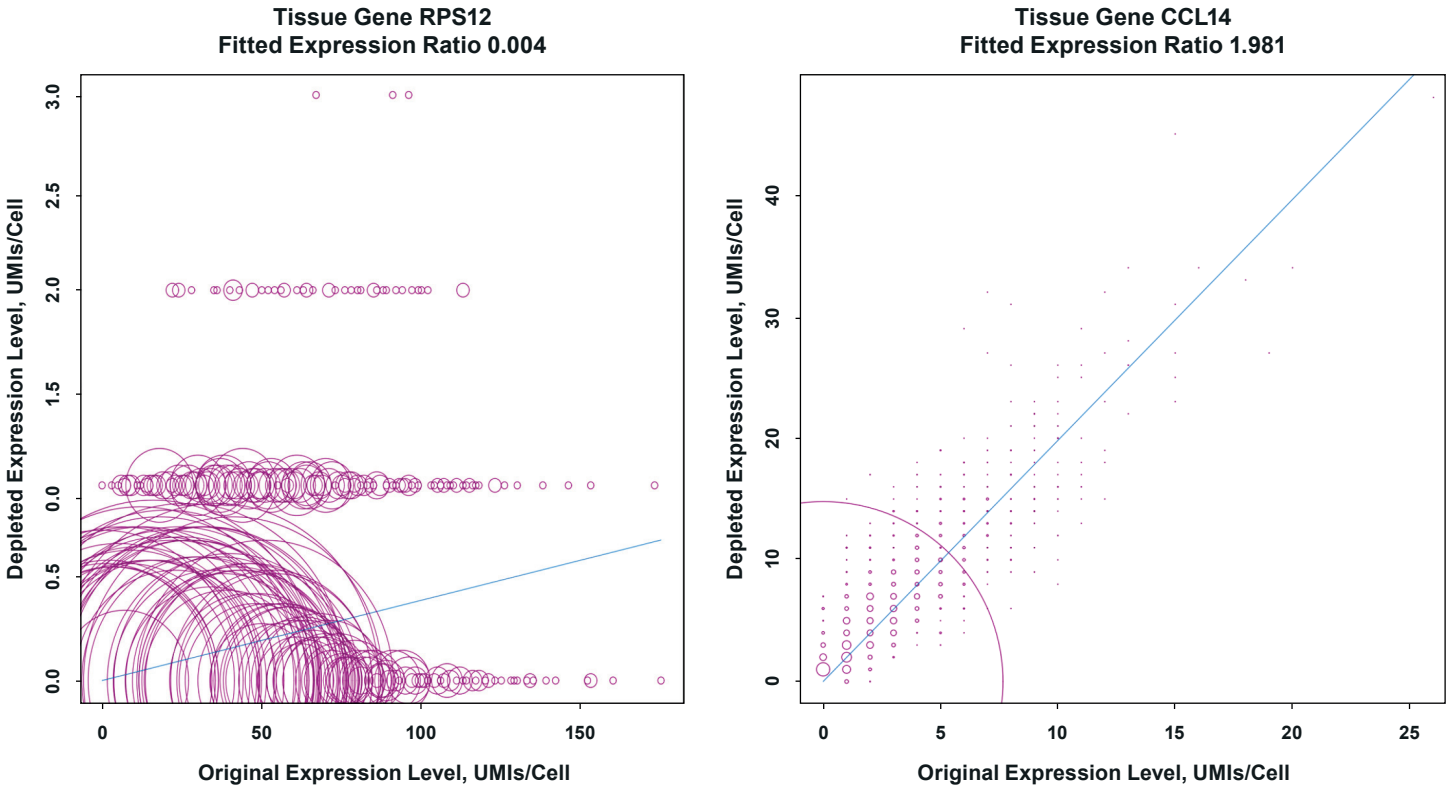

Figure 3: Linear regression plot comparing depleted and non-depleted states of UMIs per cell for RPS12 (gene targeted for depletion) and CCL14 (not targeted). Note: each purple circle corresponds to one or more cells. The area of the circle directly correlates with to the number of cells that have identical transcript-derived UMI counts.

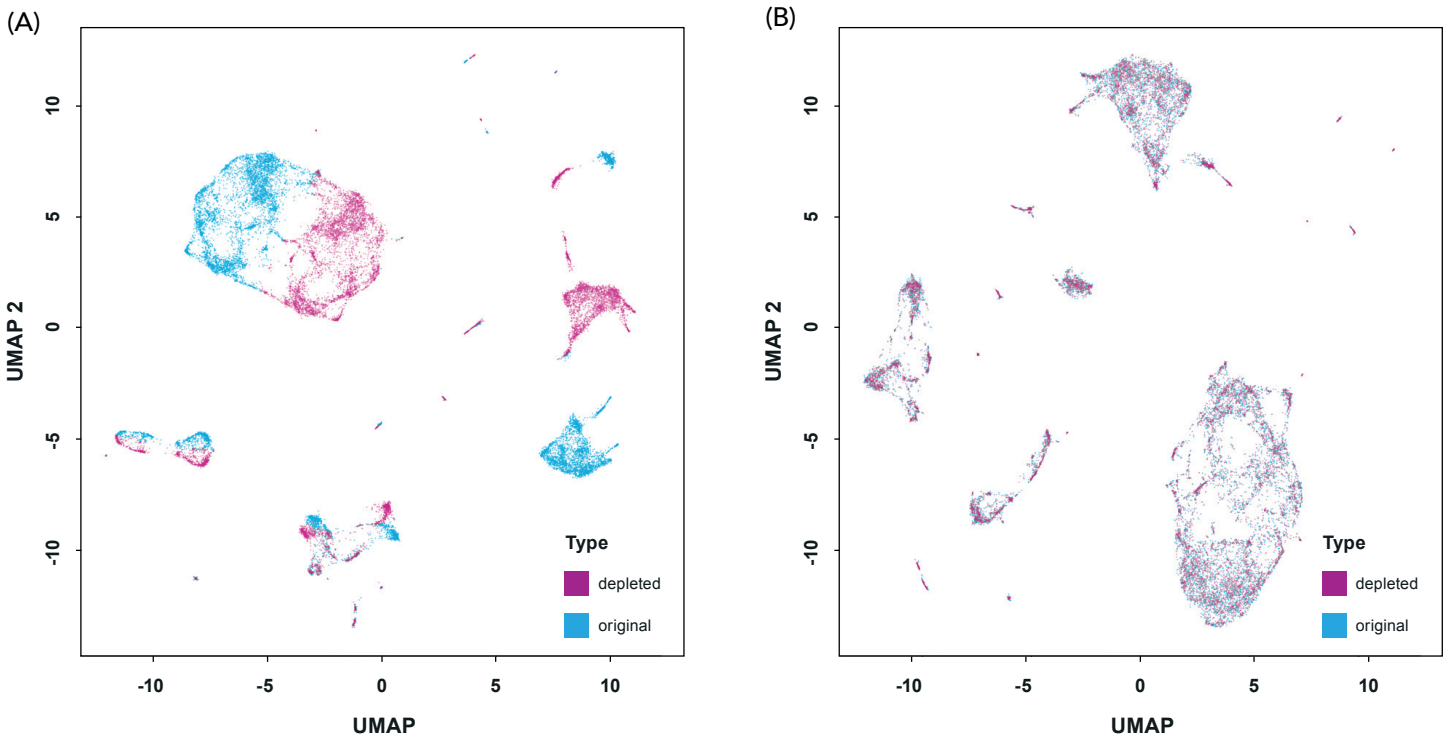

Figure 4: (A) UMAP plot comparing cell type resolution in the original sample (blue) with that of the depleted sample (purple). (B) The same plot adjusted to account for the depletion effect on the 100 targeted genes.

Effect of abundant gene depletion on cell type identification and resolution

Figure 4A shows a UMAP plot with an overlay of the original and depleted cell data demonstrating that cell type identification is highly preserved. Figure 4B shows the same plot adjusted to account for the depletion of the 100 targeted genes (i.e., depletion effect regressed out). The results indicate that depletion does not affect resolution of cell types within the cell population.

Conclusion:

CRISPRclean depletion of abundant and uninformative transcripts increases the ability to detect rare transcripts and adds confidence to differential gene expression data when applied to 10x Genomics single cell RNA sequencing libraries. It may allow the user to glean new biologically relevant information from scRNA-Seq data and, thus, gain novel insight into the genetic and cellular basis for disease.

To learn more, visit jumpcodegenomics.com