INVESTIGATING THE SOURCES OF VARIATION IN SINGLE-CELL GENE EXPRESSION STUDIES

Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles*, Jonathan E Burnett,
Jonathan K Pritchard* and Yoav Gilad
Department of Human Genetics, University of Chicago, *Department of Genetics, Stanford University

Introduction

Single-cell RNA sequencing (scRNA-seq) can be used to characterize variation in gene expression levels at high resolution. However, the sources of experimental noise in scRNA-seq are not yet well understood. We investigated the technical variation associated with sample processing using the single-cell Fluidigm C1 platform. To do so, we processed three C1 replicates from three human induced pluripotent stem cell (iPSC) lines. We added unique molecular identifiers (UMIs) to account for amplification bias. With these data, we were able to elucidate technical variability both within and between C1 batches.

In addition to the variation associated with the C1 Fluidigm platform, we currently know little about the contribution of cell cycle to the variation of single-cell gene expression. We thus propose to perform scRNA-seq using iPSCs that express the Fluorescence Ubiquitin Cell Cycle Indicator (FUCCI) - which can provide gold-standard measurements of individual cell cycle status. To our knowledge, this will be the first to quantify and measure the different sources of variations in scRNA-seq from both cell cycle status and C1 batches. The data will be used to develop generally applicable statistical methods for the analysis of cell cycle effect in single-cell gene expression studies.

UMI Barcoding of Cellular Transcriptome

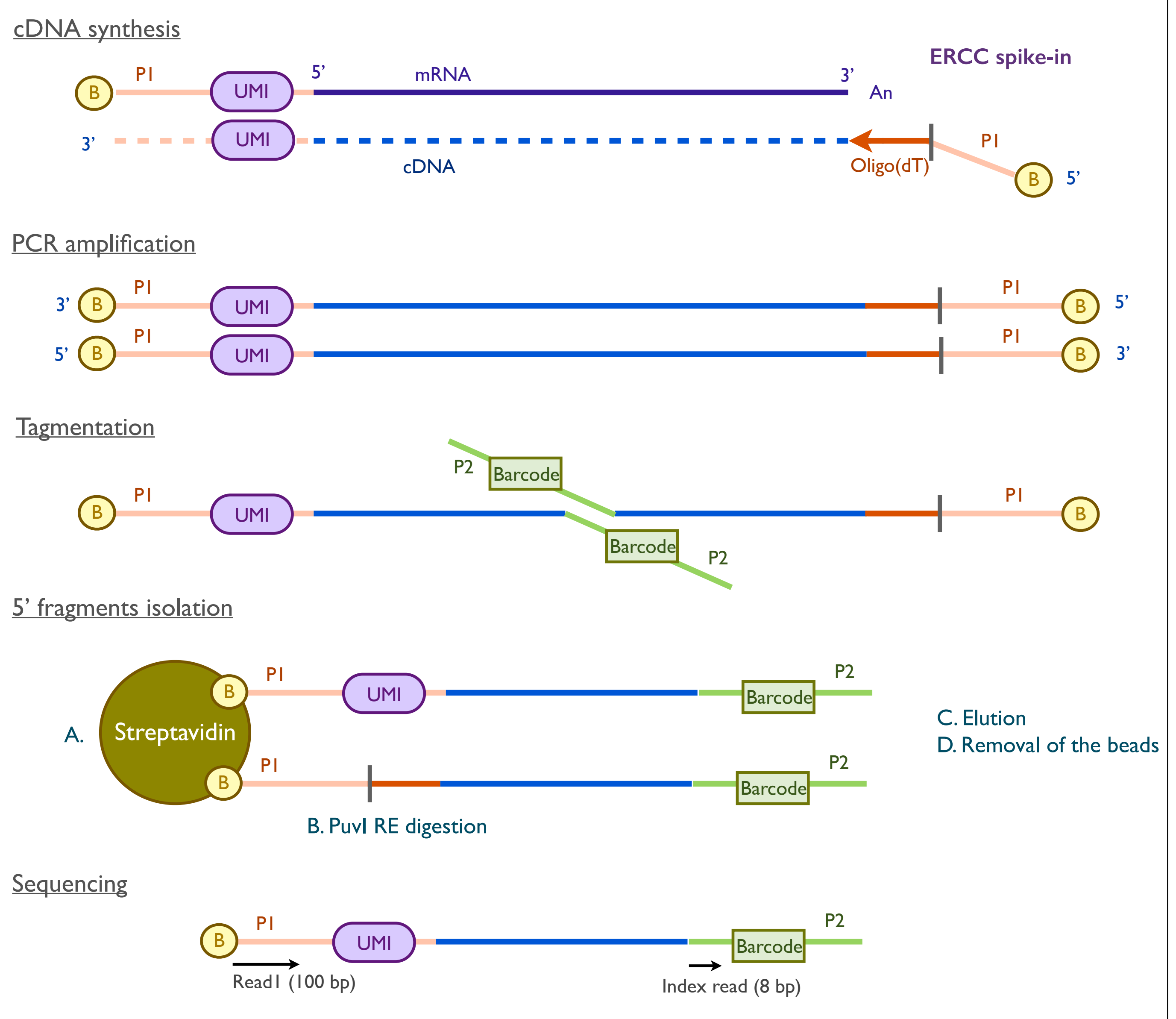

Figure 1. Schematic of single-cell library preparation. The unique molecular identifiers (UMIs) were added to each transcript during reverse transcription prior to amplification to account for the PCR bias. The addition of ERCC spike-ins in cell lysis buffer allows for the estimation of technical variation and also better quantitation.

Data Collection

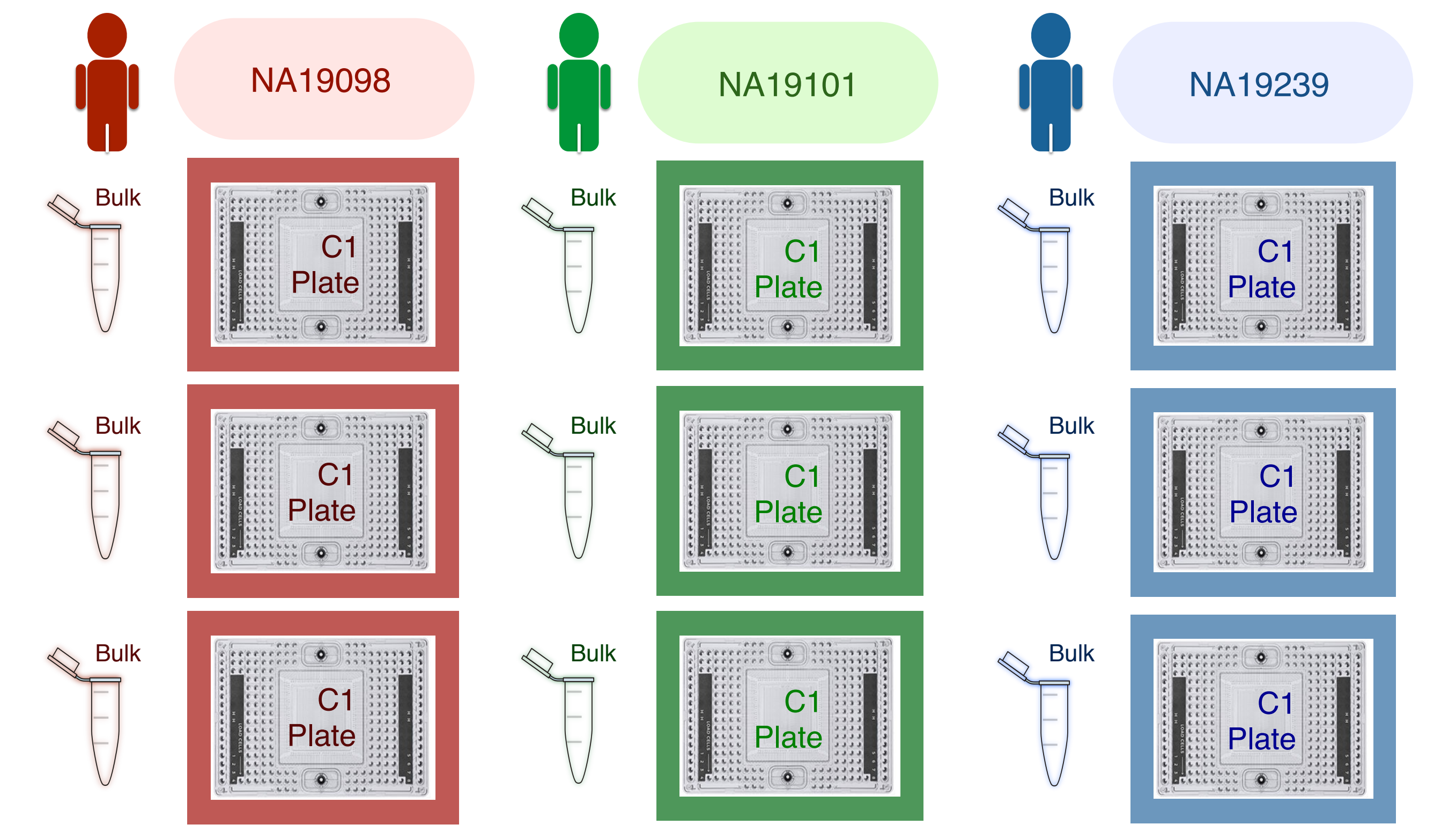

Figure 2. Study Design. Three C1 96 well-integrated fluidic circuit (IFC) replicates were collected from three Yoruba individuals. A bulk sample was included in each batch using the exact same chemicals as the C1 IFC.

Batch Effect Associated with UMI-Based Single-Cell Data

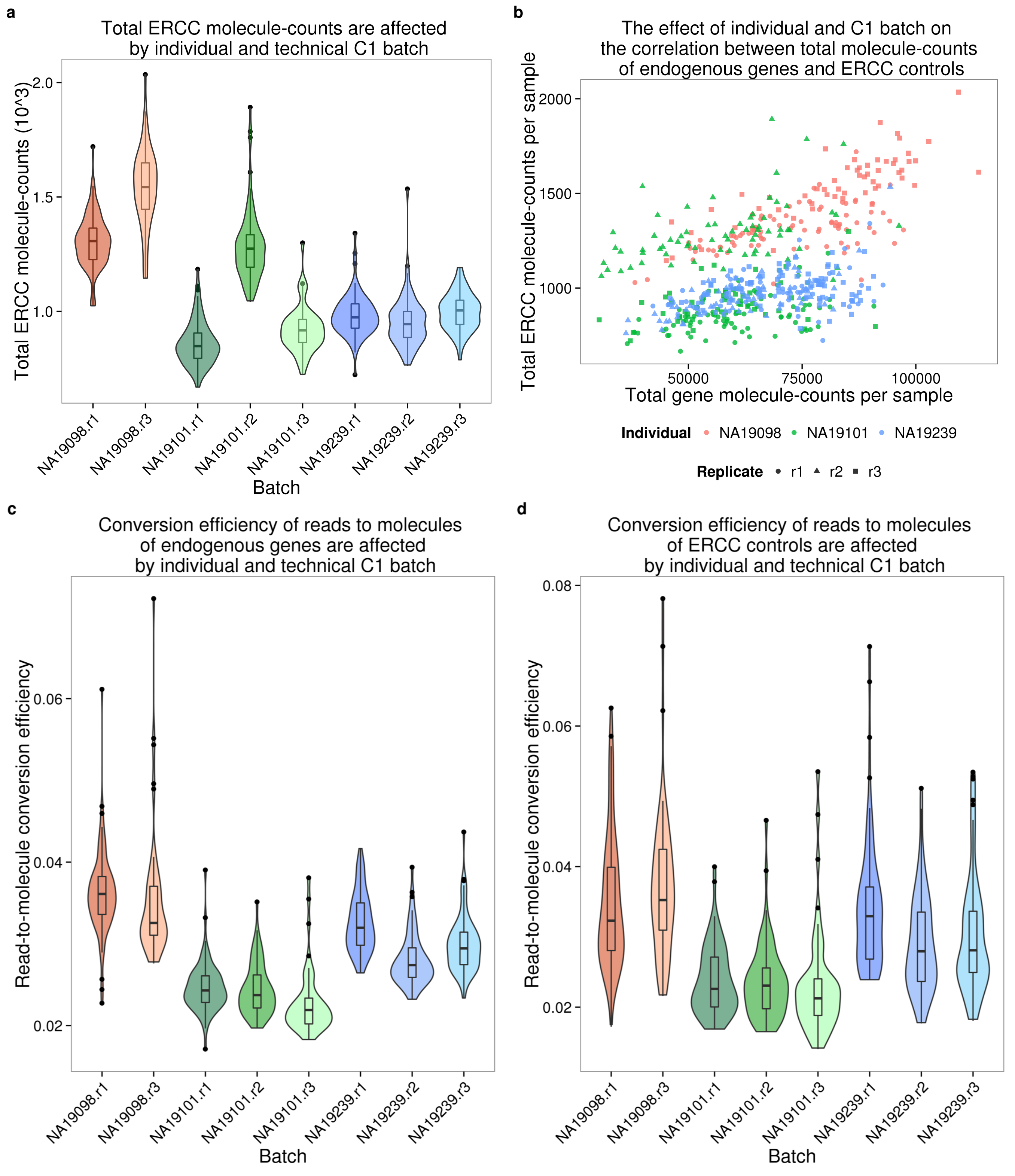

Figure 3. Batch effect of scRNA-seq data using the C1 platform. (a) Violin plots of the number of total ERCC spike-in molecule-counts in single cell samples per C1 replicate. (**b**) Scatterplot of the total ERCC molecule-counts and total gene molecule-counts. (**c** and **d**) Violin plots of the reads to molecule conversion efficiency (total molecule-counts divided by total read-counts per single cells) by C1 replicate. There is significant difference across individuals of both endogenous genes (P < 0.001) and ERCC spike-ins (P < 0.05). The differences across C1 replicates per individual of endogenous genes and ERCC spike-ins were also evaluated (both P < 0.01).

Variation in scRNA-seq from C1 Batch

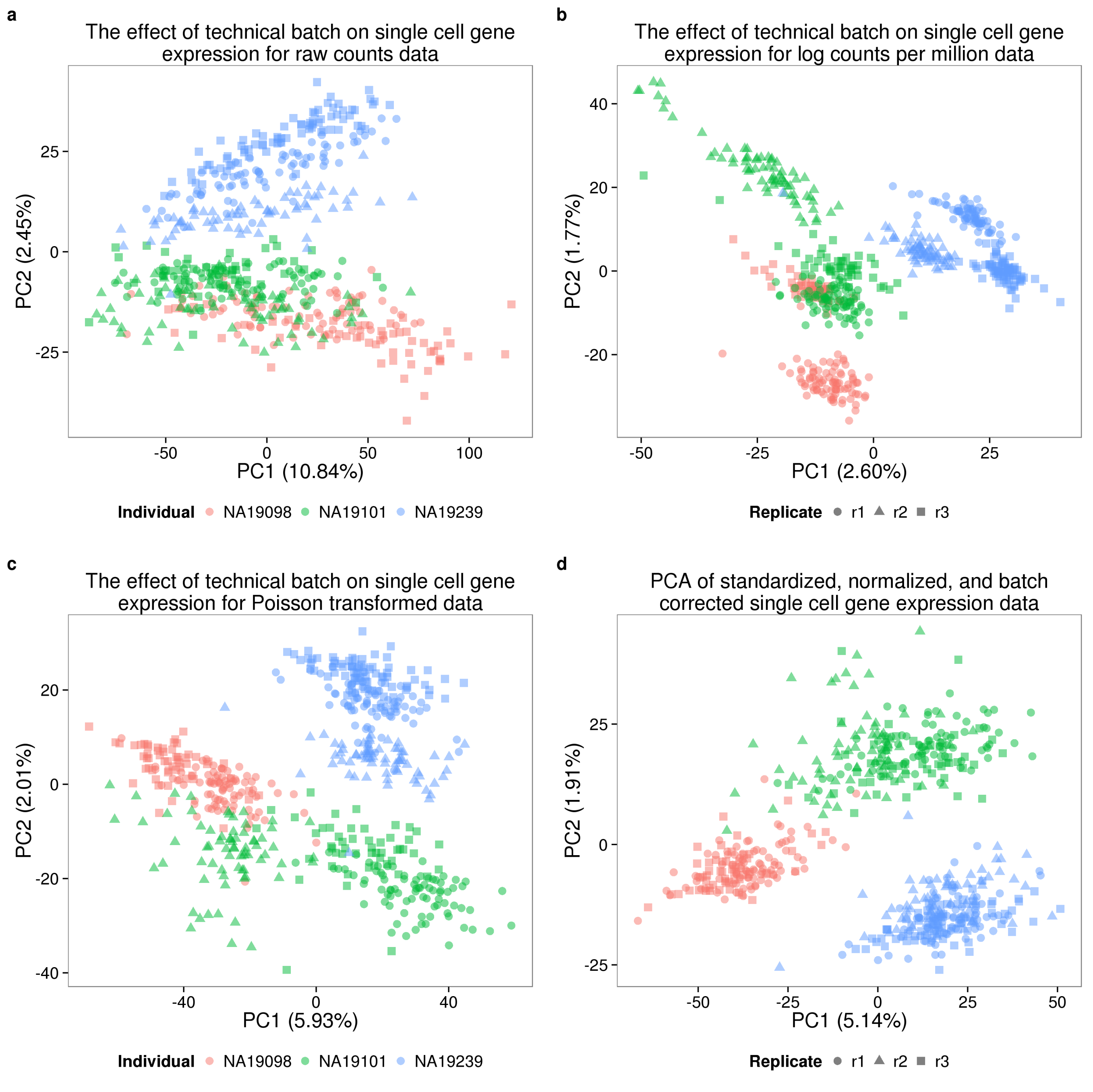

Figure 4. Normalization and removal of technical variability. Principal component (PC) 1 versus PC2 of the (**a**) raw molecule counts, (**b**) log2 counts per million (cpm), (**c**) Poisson transformed expression levels (accounting for technical variability modeled by the ERCC spike-ins), and (**d**) batch-corrected expression levels.

Variation Associated with Cell Cycle State

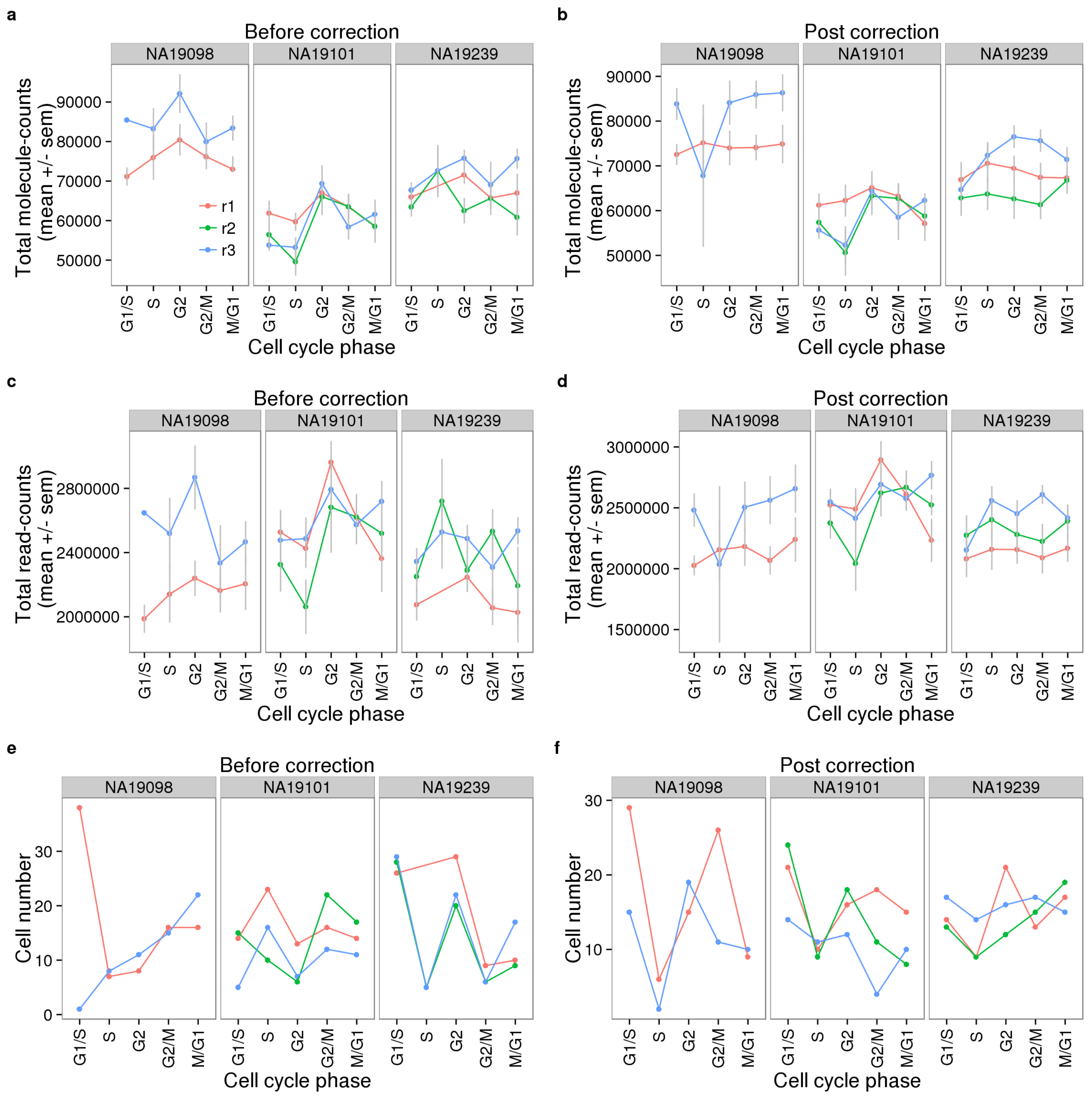

Figure 5. Cell cycle status and cellular mRNA content. (a and b) Single cells were assigned a cell cycle phases (G1/S, S, G2, G2/M, and M/G1) according to the expressions of a subset of genes based on molecule counts before (a) or after (b) applying the correction (standardization, normalization, and batch correction). (c and d) Similarly, the mean of total read-counts per sample were shown using the same classification of cell cycle phases as in (a) and (b). (e and f) The total cell numbers of each cell cycle phase were shown before correction in (e) or after correction in (f). We found that in general G2 and G2/M phases (larger cells) have higher total molecule-counts, but we also observed variation of this pattern from different individuals as well as different C1 replicates from the same individual.

Investigating The Cell Cycle Effect in scRNA-seq Data

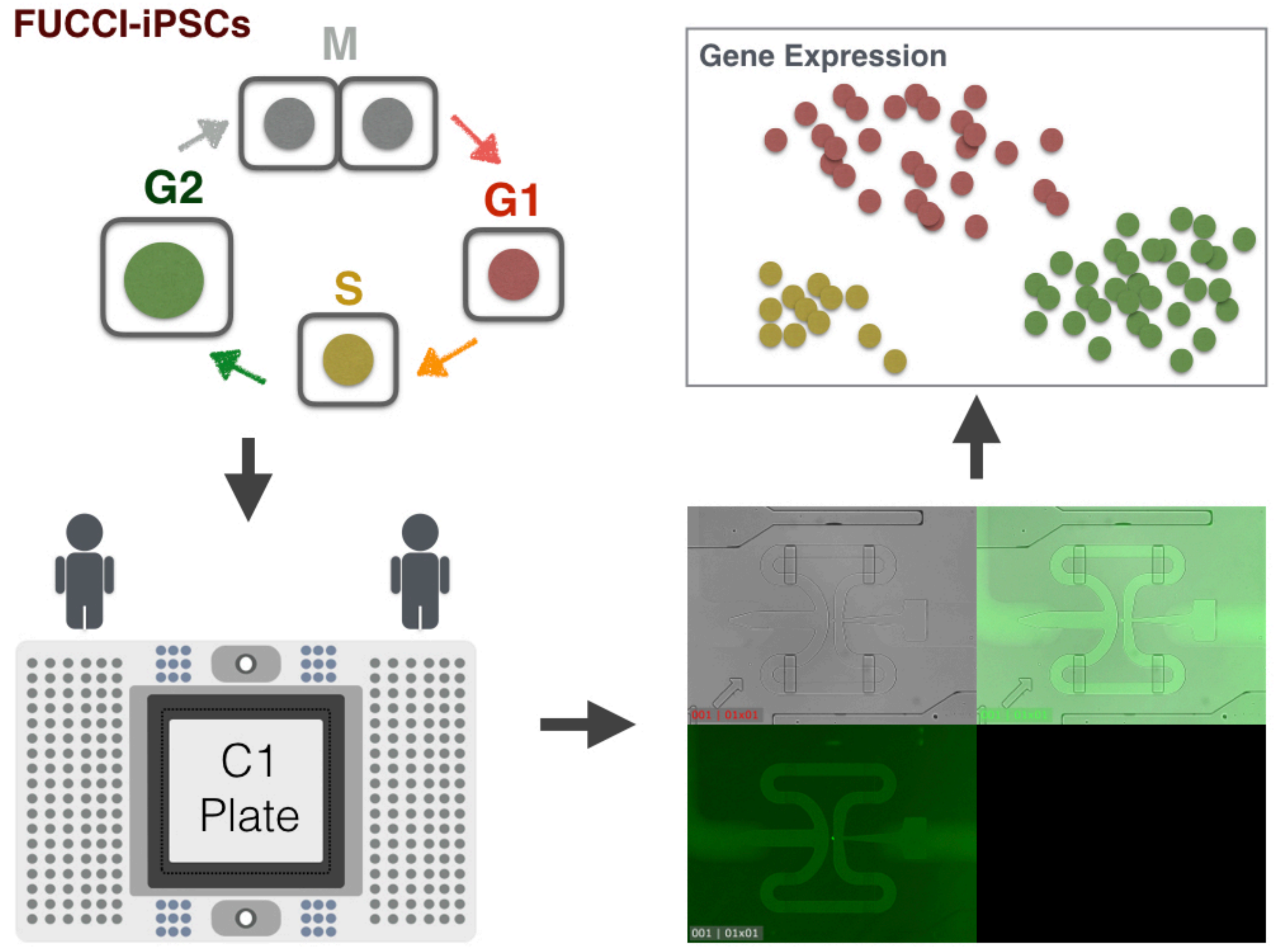

Figure 6. Schematic of FUCCI-iPSC scRNA-seq. Six FUCCI-iPSC lines will be generated using the Yoruba lines. Two individuals will be collected on the same C1 plate to account for C1 batch effect. Images will be taken before the cell lysis step on the C1 plate to determine the cell cycle status of each captured cell.

Conclusions and Reference

We found that the major source of variation in the gene expression data was driven by genotype, but we also observed substantial variation between the technical replicates. Importantly, we also observed that the conversion of reads to molecules using the UMIs was impacted by both biological and technical variation, indicating that UMI counts are not an unbiased estimator of gene expression levels. Based on our results, we suggest a framework for effective scRNA-seq studies. *Tung et al., Sci Rep. 2017 Jan 3;7:39921. doi: 10.1038/srep39921.*