Journal of Biotechnology 399 (2025) 91–98

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Investigating subpopulation dynamics in clonal CHO-K1 cells with single-cell RNA sequencing

Luke B. Morina [a], Haoyu Chris Cao [b], Siqi Chen [c], Swetha Kumar [b], Kevin S. McFarland [b], Natalia I. Majewska [b], Michael J. Betenbaugh [b], Winston Timp [a,c,d,*]

[a] Johns Hopkins Biomedical Engineering, USA
[b] Johns Hopkins Chemical and Biomolecular Engineering, USA
[c] Johns Hopkins University Department of Molecular Biology and Genetics, Baltimore, MD, USA
[d] Johns Hopkins University Department of Medicine, Division of Infectious Disease, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Single cell sequencing
CHO cells
Clonal heterogeneity
Transcriptomics

ABSTRACT

Chinese Hamster Ovary (CHO) cells produce monoclonal antibodies and other biotherapeutics at industrial scale. Despite their ubiquitous nature in the biopharmaceutical industry, little is known about the behaviors of individual transfected clonal CHO cells. Most CHO cells are assessed on their stability, their ability to produce the protein of interest over time. But CHO cells have primarily been studied in bulk, instead assuming that these bulk samples are homogenous because of presumed genetic clonality across the sample. This does not address cellular heterogeneity in these ostensibly clonal cells. These variable stability phenotypes may reflect heterogeneity within the clonal samples. In this study, we performed single-cell RNA sequencing on two clonal CHO-K1 cell populations with different stability phenotypes over a 90 day culture period. Our data showed that the instability of one of the clone's gene expression was due in part to the emergence of a low-producing subpopulation in the aged samples. This low-producing subpopulation did not exhibit markers of cellular stress which were expressed in the higher-producing populations. Further multiomic investigation should be performed to better characterize this heterogeneity.

1. Introduction

To understand the phenotype of cells, we typically rely on measuring RNA levels to give insight into active transcription hence translation. Cells express different genes depending on their age, differentiation status, environmental stimuli, physiological conditions, and even stochastic chance, all of which are reflected in a cell's transcriptome (Kim and Eberwine, 2010). These variations can impact cell behavior and productivity for recombinant cell lines used for industrial biotherapeutic production (Pilbrough et al., 2009). But when performing these methods on a population of cells, measurements can often obscure heterogeneity within the sample; changes within a subpopulation might be missed, or changes in proportions of cell type misconstrued as changes in bulk transcription. The advent of single-cell sequencing, either through split-recombine methods or droplet-based methods (Cao et al., 2017; Macosko et al., 2015), leverages individual cell oligonucleotide barcodes to allow for computational deconvolution and assignment of each RNA molecule to a cell. With these methods along with computational tools to support them, there has been an explosion in the investigation of heterogeneity within tissues (Cha and Lee, 2020; Choi and Kim, 2019; Deng et al., 2014; Haque et al., 2017; Shalek et al., 2014; Tirosh et al., 2016; Wang et al., 2021).

However such tools are rarely applied to ostensibly "clonal" cell populations. Clonal cell populations, generated from expansion from a single cell, are often assumed to be homogenous, though a cornucopia of data suggests this is not truly the case (Choi and Kim, 2019; Shalek et al., 2014). This is of particular interest in profiling Chinese hamster ovary (CHO) cells, the predominant mammalian expression platform for biotherapeutic production in the biopharmaceutical industry. CHO cells are responsible for producing over 70 % of recombinant therapeutic proteins, including monoclonal antibodies, hormones, and vaccines, due to their superior scalability and well-established regulatory acceptance (Kim et al., 2012). More importantly, their post-translational modifications and protein folding capabilities are crucial for producing highly compatible human therapeutics (Barnes et al., 2003).

CHO cells require optimal growth conditions to produce useful

* Corresponding author at: Johns Hopkins Biomedical Engineering, USA
E-mail address: wtimp@jhu.edu (W. Timp).

https://doi.org/10.1016/j.jbiotec.2025.01.010
Received 23 May 2024; Received in revised form 1 January 2025; Accepted 13 January 2025
Available online 15 January 2025

amounts of recombinant protein and are prone to genomic instability (Barnes et al., 2003; Dahodwala and Lee, 2019). But in production environments, it is hard to both predict and maintain high growth clones, even with selection pressure. The control mechanisms triggering shifts in response to environmental conditions in these cells are not yet fully understood. Previous transcriptome studies have mainly focused on comparative analyses between different environmental states or defined cell samples at the population level (Clarke et al., 2011; Doolan et al., 2013; Hsu et al., 2017).

The introduction of single-cell RNA sequencing techniques has provided new opportunities to investigate gene expression profiles at the single-cell resolution, offering insights into the sources of variation in cell lines and subclones. However, we argue that existing studies have not used enough cells to explore this given the expected proportions of cell populations or have not studied them over culture time. Tzani et al. performed bulk and single cell analysis on 1000 cells in CHO-K1 suspension culture but did not age them to measure changes over time. They did try to reconstruct heterogeneity in their populations with pseudo-temporal ordering, implicating stress-related divergence (Tzani et al., 2021). Ogata et al. analyzed single-cell transcriptomes of CHO-K1 suspension and adherent cultures, only using ~100 cells per condition and only aged for < 2 weeks (Ogata et al., 2021). Unlike Tzani et al., they observed a correlation between gene expression and cell cycle phase in adherent CHO cells (Ogata et al., 2021). They did not observe other clear substructures, though they did see significant variation in enolase expression within their clusters. Other recent work, such as Borsi et al., compared the transcriptional profiles of differently aged CHO-K1 vs HEK293FT cell lines, specifically looking for highly variable genes shared between the two cell lines (Borsi et al., 2023). Borsi et al. identified 53 genes that defined transcriptomic variability in HEK and CHO cell lines.

But the real concern is not just understanding a snapshot of heterogeneity (Tzani et al., 2021) nor how cells perform over a limited span (Borsi et al., 2023; Ogata et al., 2021) but rather *why production drops over time*. To accomplish this, we used 10X Chromium single-cell sequencing to examine the transcriptome of two CHO clones over 90 days (30 population doublings). We explored to see if the transcriptional heterogeneity and subpopulations could act to explain how titer/productivity drops over time.

2. Material and methods

2.1. Cell culture

Millipore Sigma (St. Louis, MO) provided two clonal cell lines producing the same monoclonal antibody and their host cell line (CHOZN-GS, SAFC) as cryo-preserved vials. CHOZN® ZFN-Modified GS-/- platform was used to generate the clones. This CHO strain has the endogenous genes encoding glutamine synthetase(*Glul*) knocked out. The host cells were transfected to introduce a human IgG light chain and heavy chain gene target along with an exogenous glutamine synthetase (*Glul*) gene. Clones were isolated from the pool generated by the above-described method.

These cells were cultured with imMEDIAte ADVANTAGE 87093 C (SAFC) media, a serum-free custom-made medium by Millipore Sigma, in 125 mL shake flasks (Fisher Scientific) with a working volume of 30 mL. Incubation was done in a humidified orbital shaking incubator set at 37 C, 80 % humidity, 5 % CO2 and 125 RPM shaking. The two clonal cell lines were thawed and seeded at 0.3×10^6 cells/mL.

The cells were passaged every 3 days while monitoring viable cell density (VCD) as determined by hemocytometer using trypan blue dye exclusion. The cells were then seeded in fresh media at 0.3×10^6 cells/mL. Cells were passaged for 30 passages over 90 days during the aging process. Clones were cultured in triplicate either with or without 6 mM L-Glutamine supplementation for the duration of the experiment. Culture samples at passage 0 (P0) (clone A, clone B) and passage 30 (P30) (clone A +Gln, clone A -Gln, clone B +Gln, clone B -Gln) were sampled for 10x cDNA library prep.

Cell banks were established from P30 cultures by freezing 10^7 cells/mL culture in media supplemented by 10 % DMSO. Media exchange was facilitated by pelleting cells at 500xg for 5 minutes and aspirating supernatants. Media exchanged cultures were then aliquoted into cryotubes and frozen in MrFrosty freezing containers (Nalgene) that gradually decreased content temperature in a -80C Freezer. Banked cells were then stored in a -80C freezer for short term storage and in a gas phase LN2 tank for long term storage. To thaw banked cells, we incubated the cryo-vial in a 37 C water bath, followed by adding 1 mL of thawed culture to 29 mL fresh media in a shaking flask.

2.2. Fed-batch experiment

Banked clones from the aging experiment were thawed to investigate fed-batch performance. Specifically, we revived clone A at P0 and clone A +Gln/-Gln at P30, as well as clone B at P0 and clone B +Gln/-Gln at P30. They were characterized in triplicate in a 14-day fed-batch process in 125 mL shake flasks (Fisher Scientific). Starting at a seeding density of 0.6×10^6 cells/mL, we cultured them with a working volume of 30 mL with glucose level set at 5.5 g/L. We used Sigma imMEDIAte ADVANTAGE 87093 C (SAFC) as media at D0 for all unaged samples as well as samples aged without glutamine(-Glu). The same basal media supplemented by 6 mM L-Glutamine was used for samples aged with glutamine (+Glu). Culture conditions were the same as the aging experiment(37 C, 5 % CO2 and 125 rpm shaking). Glucose was fed into the culture via bolus feed every other day after day 4 of the culture. VCD and viability were recorded daily via hemocytometer(Neubauer). Glucose and lactate measurements were also taken daily using a YSI 2900 Biochemistry Analyzer. Titer was measured from culture supernatant via HPLC run on a POROS Protein-A column. The calibration curve of the HPLC was determined with a manufacturer provided standard. Cell-specific productivity (Qp) was calculated as shown in (Clarke et al., 2011).

2.3. qPCR

qPCR was used to assess the variability of transgene copy number between samples. Genomic DNA was sampled from Host, Clone A and Clone B at early and late passages at day 3 after seeding in suspension batch culture. Quick-DNA™ Microprep Plus Kit (Zymo Research) was used to extract total DNA from the samples. We designed primers targeting the IgG LC as well as 2 housekeeping genes: *Gapdh* and *Actb* (Table S1). No template control and the host cell line were used as controls, and each condition was run in triplicate. We performed qPCR on the QuantStudio3 using Forget-Me-Not™ EvaGreen® qPCR Master Mix (Biotium). Thermal cycling recipe was 95 C for 3 minutes, followed by 40 cycles of 1) 95 C 20 s, 2) 66 C 20 s, 3) 72 C 10 s. Fluorescence signals were collected during the extension at 72 C. Following qPCR, a melt curve was performed from 60 C to 95 C to validate the products. The data was analyzed with Quantstudio3 Design&Analysis Software version 2.8.0 with default settings for Ct value in each well. Select amplicons were loaded on 1 % agarose gel for quality control and verification of amplicon size.

2.4. Library prep

We targeted 8000–10,000 cells per condition (measured via hemocytometer) to generate single cell libraries using a Chromium Next GEM Single Cell 3' v3.1 dual index chip (10X Genomics). In brief, poly(A) mRNA transcripts were captured with poly(dT) oligo gel beads, and reverse transcribed into barcoded full-length cDNA. Then, scRNA-seq libraries were prepared with the Chromium Single Cell 3' library construction kit (10X Genomics), where each sample is amplified with primers containing unique i5 and i7 sample indexes and common P5 and P7 sequencing adaptors via PCR. Libraries were subsequently pooled

based on their molar concentration. Six pooled libraries were then loaded at 4 nM and sequenced on a NovaSeq SP flow cell (Illumina) with 28 bases for read1, 91 bases for read2, and 10 bases for i5 and i7 index respectively.

2.5. Cell ranger

A Cell Ranger reference was made from the CHOK1GS genome and transcriptome from Ensembl (GCA_900186095.1) using Cell Ranger version 6.0.2. We manually added heavy chain IgG (IgH), light chain IgG (IgL), and the CHO *Glul* gene (NM_001246770) sequences to our reference. We then generated an alignment index from the reference using the *cellranger mkref* command with default settings. This reference was then used with cellranger count, also run using default settings, for all six samples. Barcodes with over 500 UMIs were considered cells (Figure S1).

2.6. QC & filtering

Quality control was done using Scanpy version 1.8.2 (Wolf et al., 2018). Cells with less than 500 genes and genes identified in less than 10 cells were removed. We retained cells with mitochondrial gene counts between 0 % and 18 %, total transcript counts between 5000–100,000, and total number of genes between 500 and 10,000 for subsequent analysis (Figure S2). A mitochondrial gene list was obtained from Quiros et al. (2017). We normalized the remaining cells to 10,000 transcript counts per cell, log1p transformed with base 2, and merged into one single-cell object. Cell cycle regression was then performed using S and G2M phase gene lists to remove cell cycle phase gene variability using the built-in Scanpy regression function (Figures S3–S4). Human S and G2M gene lists were lifted over onto the K1GS CHO genome (Tirosh et al., 2016).

2.7. Highly-variable gene detection

Highly variable genes were selected per sample using the default settings in Scanpy *highly_variable_genes* function. Genes identified as highly variable in 5 or more samples were kept for clustering, resulting in 78 genes (Table S2).

2.8. Clustering and UMAP

Uniform manifold approximation and projection (UMAP) was calculated using 25 principal components on the selected highly variable genes. Clustering was performed using the Louvain algorithm in Scanpy at a resolution of 0.25. Quality control metrics did not demonstrate any cluster specific nor cell cycle biases (Figure S5).

2.9. Differential gene expression

The Wilcoxon Rank Sum test was used to identify differentially expressed genes between Louvain clusters. The raw log-transformed counts were used over the scaled counts, otherwise the default settings were used in the Scanpy rank_genes_groups function.

For the Clone B cell filtering analysis, the Wilcoxon Rank Sum test was again used to identify differentially expressed genes between the two IgG expression groups for both clone B P30 samples. Genes with an absolute log2foldchange greater than 1, Bonferroni corrected p-value less than 0.05, average expression greater than 0.2, and non-zero expression percentage greater than 15 % were considered significantly differentially expressed. The minimum non-zero expression percentage of 15 % was included to remove genes that might appear as differentially expressed due to sparsity of the single cell counts data (Figure S6).

2.10. Gene set enrichment analysis

Gene set enrichment analysis was performed using the GSEApy package version 0.14.0 and the Enrichr module (Fang et al., 2023). For each Louvain cluster, differentially expressed genes with a log2fold-change greater than 0.15 and adjusted p-value less than 0.05 were passed into Enrichr with mouse as the set organism and GO_Biological_Process_2023, GO_Molecular_Function_2023, and KEGG_2019_Mouse libraries loaded (Ashburner et al., 2000; Gene Ontology Consortium et al., 2023; Kanehisa and Goto, 2000). The top 5 terms per gene set library as ranked by adjusted p-value were kept.

3. Results

3.1. Cell samples

To characterize changes in cell populations over time, we used two CHO clones provided by our collaborators at Millipore-Sigma, based on the CHOZN® ZFN-Modified GS-/- platform. In these CHO cells, the endogenous glutamine synthetase (*Glul*) is knocked out, requiring that L-glutamine be supplemented in their media. These host cells are then transfected to introduce a human IgG with lambda light chain (IgL) and heavy chain (IgH) along with glutamine synthetase (*Glul*). *Glul* allows for selection pressure, as only cells with a successful transfection will be able to grow in the absence of supplemented glutamine. Individual cells were isolated from the pool of transfected cells to form founding clonal populations. We cultured these cells in orbital shaking incubators in a 90 day aging campaign, passing every 3 days for 30 passages either in the presence (no selection) or absence (selection) of 6 mM glutamine. Samples from passage 0 and passage 30 were banked in cryogenic storage for subsequent analysis.

We then revived banked cell samples and cultured them in triplicate in a fed-batch system for a maximum of 15 days (Fig. 1A). Throughout the fed-batch culture period, mAb titer, viable cell density (VCD), and cell viability were measured daily (Figures S7–S11). Plotting cell-specific productivity (Qp) vs VCD (Fig. 1B), we found that samples with high productivity typically have a lower VCD and samples with a high VCD low productivity - as might be expected due to the metabolic cost of IgG production. Notably, clone B has a large increase in VCD after aging, independent of selection pressure. At P0, clone A and B both showed high titer/productivity, though clone B was noticeably lower. However, by P30 clone A under selection pressure (-Gln) is the only sample that maintained a high productivity. The other conditions all have significantly reduced titer (<1000 mg/L) and productivity (<20 pg/cell/day).

3.2. Single cell RNA-sequencing

From day 6 of the fed-batch samples, we generated 10X Single Cell Gene Expression profiling data. We targeted 10,000 cells for each condition to generate barcoded cDNA. The cDNA was prepared for Illumina sequencing following established protocols (Methods) and sequenced on a NovaSeq, generating an average of ~636 million reads per sample. The data generated is summarized in Table S3. This data was then used to generate a count matrix using 10X software as summarized in Methods.

From this data, we first wanted to measure target gene expression levels. We plotted the measured aggregated RNA-sequencing reads for IgG, both heavy chain (IgH) and light chain (IgL) versus the cell-specific productivity at day 6 (Fig. 1C). We found these values had Pearson correlations of R= 0.98 and 0.67 for the heavy and light chains, respectively. The high correlation for IgH reflects the nature of the titer assay - it measures IgH levels. Notably the samples with higher productivity have lower growth rates - as shown in Figure S10, the Clone B P30 cultures (both +/- Gln) maintain positive growth rates for longer during the fed-batch culture compared to the high-producing Clone A cultures. This growth advantage of the low-producing populations is

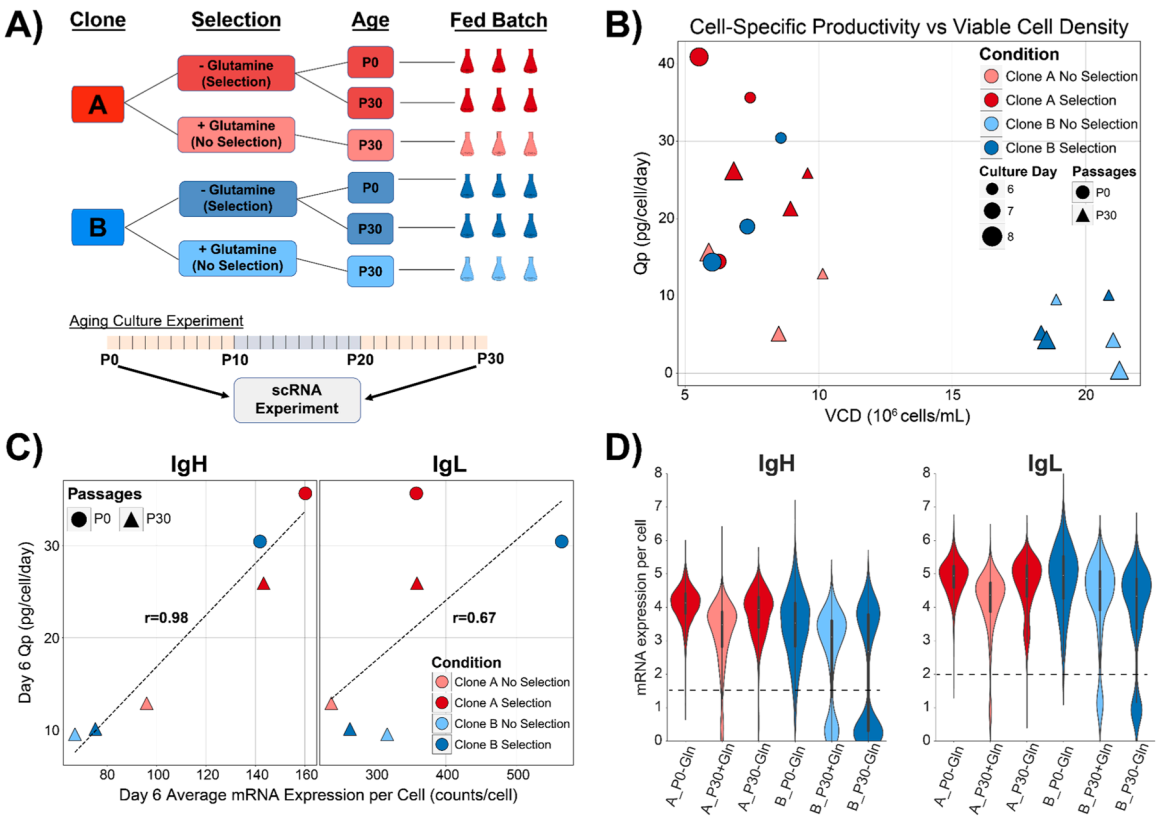

Fig. 1. Experimental design. A) Two CHO clones with the same gene target were cultured with and without glutamine for 90 days with passages every 3 days. To measure productivity, these six cell samples were cultured in triplicate for measuring mAb titer and cell viability. B) Cell-specific productivity (Qp) and viable cell density (VCD) were measured each day of the cell culture period. Days 6–8 of the fed-batch experiment have VCD plotted vs. the same day's productivity measurements. C) The antibody titer and single cell experiments were performed on day 6 of the fed-batch experiment. For each sample, the cell-specific productivity is plotted against the average mRNA transcripts per cell of heavy and light chain IgG. The Pearson correlation coefficient was calculated for both IgG chains. D) Violin plots of IgG gene expression per cell per sample. The y-axis is the log2 of the cDNA counts + 1 for each gene per cell. We observed high expression levels per cell with distinct bimodality in some Day 30 samples. Dotted lines indicate thresholds dividing high vs low IgG expression.

particularly evident on Day 6 when we collected samples for scRNA-seq - the Clone B P30 cultures show positive growth rates while Clone A cultures exhibit negative growth rates.

Moving from the bulk RNA expression, we examined the single cell log-transformed expression of the three inserted genes (*IgH*, *IgL*, *Glul*) (Fig. 1D, Figure S10A). We observe a clear bimodality of IgG in clone B in both P30 conditions, with a *slight* bimodality of clone A at P30 without selection pressure. Setting a threshold of 1.5 for IgH and 2 for IgL log-transformed expression, we split cells into "high" or "low" expressors. To be a high expressor cells have to pass both thresholds, otherwise, they are classified as low. Notably, clone B P30 samples had 25.7 % in P30 + Gln and 47.4 % in P30-Gln samples exhibiting a low-expressing phenotype (Table S4). This suggests that the observed lower IgG protein titer, echoed by the lower bulk RNA level in the B clone is partially due to a subpopulation of cells with lower expression, explaining the variance in viability within the low-titer samples. Curiously, we did not observe the same variation in Glul transcript expression (Figure S10), with the caveat that its low levels of expression make it difficult to conclusively rule out variation. The apparent bimodality in the violin plot for Glul is due to a failure to detect *Glul* transcripts due to low expression.

To inspect gene copy number variation, we performed qPCR for the inserted genes with Clone A and Clone B at early and late passages. Primers were designed against the inserted construct (IgG) and housekeeping genes *Actb* and *Gapdh*. However, melt-curve and subsequent electrophoresis revealed that any signals from control samples (NTC and Host) had a higher Ct, and different melting temperature and size from those in experimental groups These bands were not detected in the experimental group (Figure S12). After normalization against *Actb*, we found Clone A at P2 and P31 did not display a large variation in transgene copy number. By contrast, Clone B at P2 had a much higher transgene copy number compared to Clone B at P31(~35-fold change) (Figure S12).

After cell cycle regression, highly variable gene selection, and dimensionality reduction (Methods), we performed Louvain clustering (Fig. 2A). Three distinct clusters emerged from the data, each with a unique IgG transcript expression profile (Fig. 2B-C). This clustering captures the subpopulations observed in the IgG transcript violin plots. Selection pressure acts to maintain a high proportion of productive cells in clone A but seems to have the inverse effect on clone B (Fig. 2D), though we observe no difference in Glul expression between the clusters (Figure S14B-C). Curiously, the IgH and IgL expression are not correlated in Cluster 1 which has low IgG heavy chain but high IgG light chain expression.

We then found differentially expressed genes between Louvain clusters (Figure S15) and performed gene set enrichment analysis (GSEA) using Enrichr in the GSEApy wrapper (Figure S16). GSEA showed an enrichment for protein production and transport pathways in Cluster 0 whereas Clusters 1 and 2 were enriched for Proteasome and RNA processing pathways. This suggests that these low-producing cells are utilizing cellular mechanisms to avoid the selection pressure applied, possibly by ribosomal or proteasomal methods.

3.3. *Digging into cluster analysis*

To look for biomarkers beyond the target genes different between

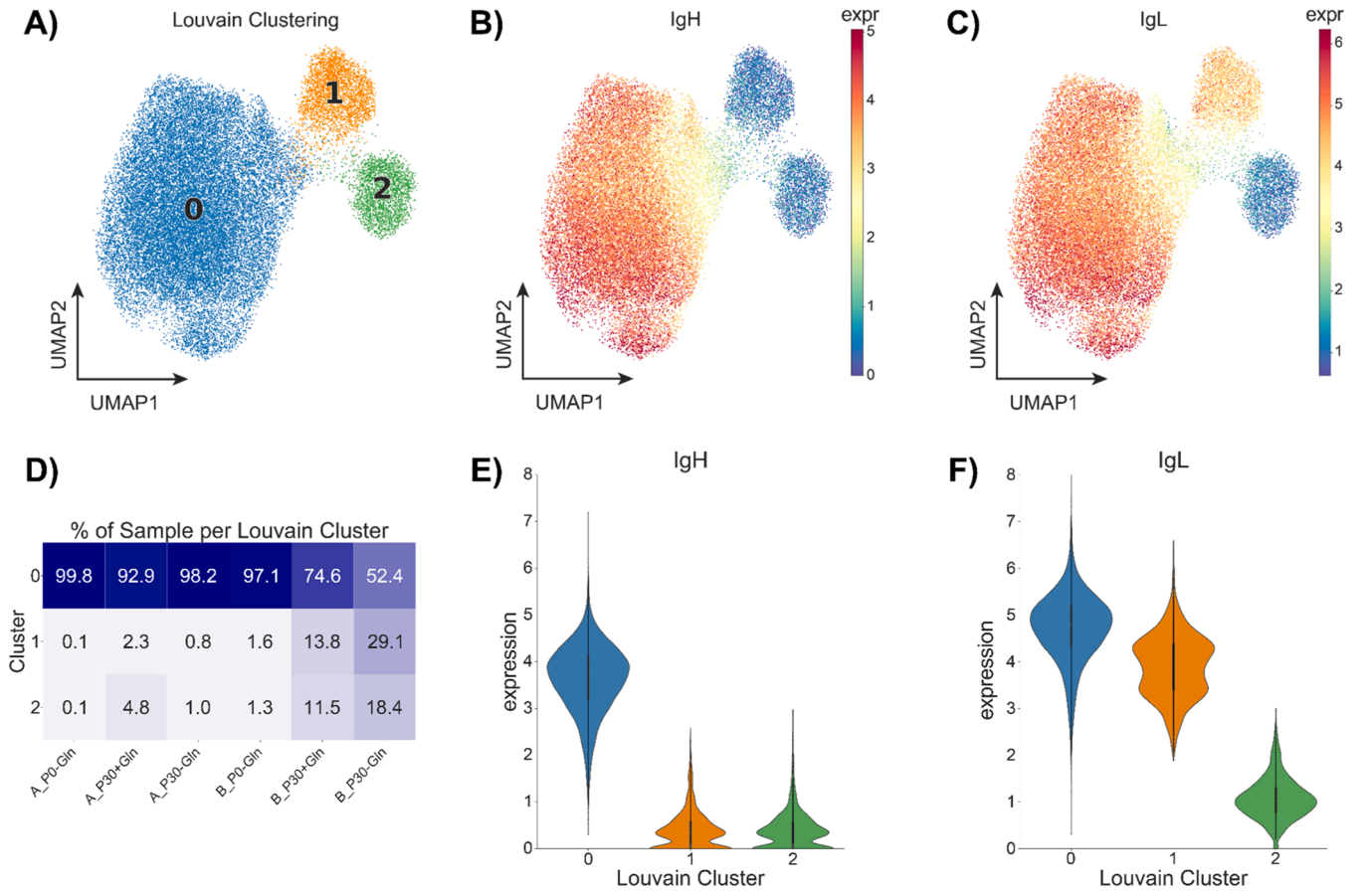

Fig. 2. UMAP projections and Louvain clustering of cells from all six samples. A) UMAP of cells merged from all six samples colored by Louvain clustering. B) UMAP with cells colored by log-transformed IgG heavy chain mRNA expression. Expression is higher in the larger cluster 0 and lower in the two smaller clusters 1 & 2. C) UMAP colored by log-transformed IgG light chain mRNA expression; highest in the large cluster, middling in one of the smaller clusters, and low in the other small cluster. D) Percent cells from each sample in each Louvain cluster. The clone B P30 samples had a higher presence in clusters 1 & 2, the clusters associated with lower IgG chain expression. UMAP of this data colored by sample can be seen in Figure S13E-F) Log-transformed mRNA expression violin plots of the IgG genes by Louvain cluster. Cluster 0 has high IgG heavy chain, and IgG light chain expression. Cluster 1 has low IgG heavy chain and middling IgG light chain. Cluster 2 has a low IgG heavy chain and low IgG light chain.

these clusters, we measured differential gene expression using the Wilcoxon Rank-Sum test between the high IgG producing cluster 0 and the low IgG producing clusters 1 and 2. We restricted our analysis to clone B P30 samples to remove other clone or age specific artifacts from our analysis. To identify truly differentially expressed genes versus noise, we performed multiple filtering steps. Genes where more than 5 % of cells had zero expression were removed from differential gene expression analysis. These often have false inflation of either the p-value or $\log_2$(-foldchange) when there is a marginal difference in non-zero expression between groups (Jiang et al., 2022). Genes were further filtered to remove genes with average expression < 0.20 (Fig. 3B). Finally, genes with magnitude $\log_2$(foldchange) > 1, p-value < 0.05, and non-zero expression > 15 % were considered as significantly differentially expressed genes (Fig. 3A).

This resulted in 5 genes: *IgH*, *IgL*, *Scd2*, *Ccnd1*, and *Pdlim2*. As shown in Fig. 3C and F, we found that *Scd2* and *Ccnd1* were both highly expressed in Cluster 0 (high production cluster) but lowly expressed or absent from Clusters 1 and 2 (low production cluster). Violin plots of these genes in the clone B P30 samples can be found in Figure S18.

We then compared these two genes across all six samples (Figure S19). We found that *Ccnd1* was highly expressed in the clone B P0 sample alone, with little difference across the three Louvain clusters. However, *Scd2* was more highly expressed in the four conditions with higher titer and no bimodality in IgG expression (clone A samples, clone B P0). Given this trend, it follows that *Scd2* could be used as a marker to filter low IgG producing cells in Clone B populations.

We wanted to see if these differentially expressed genes could act as marker genes to sort out high versus low expression populations. Through manual inspection we chose a minimum *Scd2* expression threshold of 1.19 transcripts per 10,000 (0.25 in log2 space) for both of the clone B P30 samples. Filtering out *Scd2* expression below this threshold, we plotted the distributions of *IgH* and *IgL* in Fig. 4 with (purple) and without (grey) filtering.This filtering strategy works best on the samples under selection pressure even though Scd2 was identified as a differentially expressed gene for both clone B P30 samples. In the clone B P30-Gln sample, roughly half of the cells (6539 down to 3294, 50.4 %) are retained after filtering, but this results in a noticeable increase in average $\log_2$ mRNA expression from 2.09 to 3.04 for IgH and 3.81–4.51 for IgL. By comparison, the clone B P30 + Gln sample retains less cells (7145–1708, 24.0 %) in filtering and has a less noticeable change in average expression, including no change for IgL and 2.58–2.89 for IgH.

To further investigate this strategy we also applied the filtering to the clone B P0 sample (Figure S20). We found that using this *Scd2* threshold most cells were retained (4965–4939, 99.5 %) so there was minimal change in average expression. We also performed this same filtering strategy with the three clone A samples; however, there was minimal filtering across all samples (P0: 9474–9082, 95.9 %; P30 no selection: 9898–9562, 96.6 %; P30 selection: 6846–6742, 98.5 %).

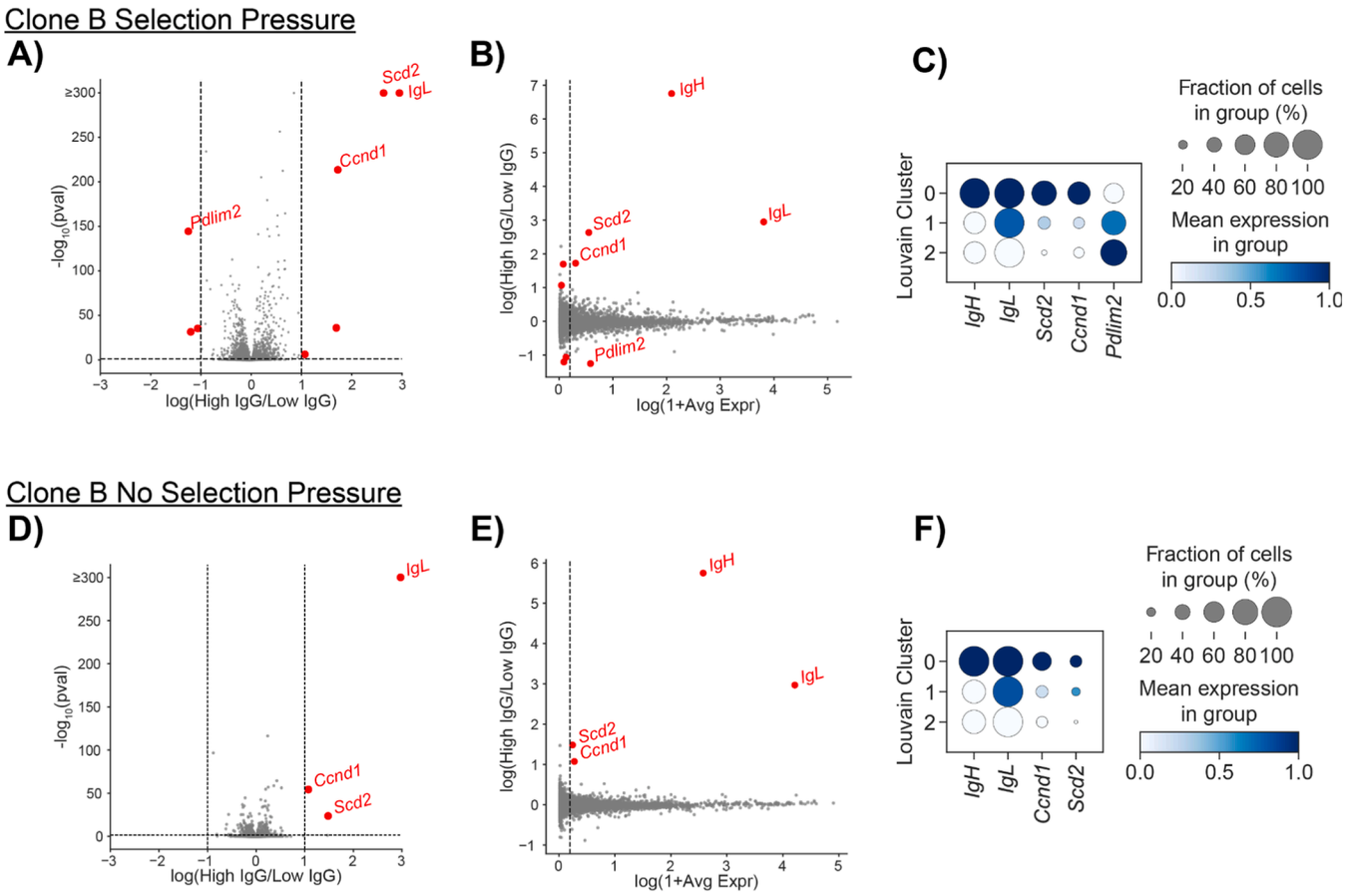

Fig. 3. Differentially Expressed Genes between IgG Expression Groups in aged clone B samples. A) Volcano plot of genes with p-value generated by Wilcoxon test between the High IgG (Cluster 0) and Low IgG (Clusters 1 and 2) for the clone B P30-Gln sample. Genes expressed in < 15 % of cells were not plotted. Genes with magnitude $\log_2$(foldchange) > 1, p-value < 0.05, and non-zero expression > 15 % are highlighted in red as significantly differentially expressed genes. Log p-values were capped at 300 for visualization. X-axis scaling was set from −3–3, removing IgH from the plot; full range figure available in Figure S17. B) Magnitude-Amplitude (MA) plot for all genes in the clone B P30-Gln sample. Genes with average expression greater than 0.2 are labeled red. Five genes that passed all filtering are labeled with red text. C) Dot plots of the five significantly differentially expressed genes between IgG groups displayed across clusters in the clone B P30-Gln sample. The expression is normalized per gene; size of each dot represents the percentage of cells with non-zero expression for each gene in each cluster. D-F) Corresponding plots for clone B P30 + Gln.

4. Discussion

Through our single-cell RNA expression analysis of two clonal IgG-transfected CHO-K1 cell lines, we investigated how population heterogeneity evolved over time and correlated with declining productivity. The clonal nature of our samples, combined with their foreign gene insertions and limited biological replication, presented unique analytical challenges. Standard scRNA analysis tools, typically designed for highly variable cell types like PBMCs, proved inadequate for detecting subtle differences in these more homogeneous populations. Therefore, we implemented stringent criteria for identifying variable genes, focusing our analysis on the most prominent transcriptional differences while acknowledging we might miss more subtle variations.

Using Louvain clustering, we identified three distinct subpopulations in our clonal samples (Fig. 2A). Two of these were "low expressing" subpopulations with distinct IgG expression phenotypes: one maintained IgG light chain expression while showing minimal heavy chain expression, while the other expressed neither chain. These low-producing subpopulations became increasingly dominant as the cells aged, particularly in the P30 samples. Most notably, in clone B P30 without Gln, the low-producing cells comprised 47.6 % of the population, compared to just 2.9 % in P0.

This shift in population composition could be driven by relative fitness. The low IgG expressing samples exhibited a higher cell viability than cells with high IgG expression (Figure S8). This suggests that there is an inverse relationship between cell viability and inserted IgG production. This finding is consistent with the titer and VCD measurements taken from the fed-batch experiment of these same cell lines, as the three samples with the highest titer had lower and more consistent VCD than the three samples with lower titer across replicates. It also follows that these three low titer samples had a larger fraction of the low IgG expressing cell subpopulations. Additionally, qPCR revealed that the lower expressing IgG samples (Clone B P30 Gln-/+) had reduced target gene copy numbers, though single cell analysis showed the genes were still expressed universally, albeit at varying levels.

A striking finding was that selection pressure in Clone B appeared to have the opposite effect from what would be expected - the B P30-Gln sample had a larger proportion of low-IgG-producing cells (47.6 %) compared to B P30 + Gln (25.4 %). This counterintuitive result suggests that sustained selection pressure may actually accelerate the emergence of resistant subpopulations. Our GSEA analysis provides some mechanistic insight, showing enrichment of proteasomal and RNA processing pathways in the low-producing clusters. This indicates these cells may be utilizing post-transcriptional mechanisms to maintain sufficient Glul activity for survival while downregulating the metabolically costly IgG production. The qPCR data showing reduced transgene copy number in aged Clone B samples further supports this adaptation model - cells that shed copies of the transgene cassette while retaining minimal Glul

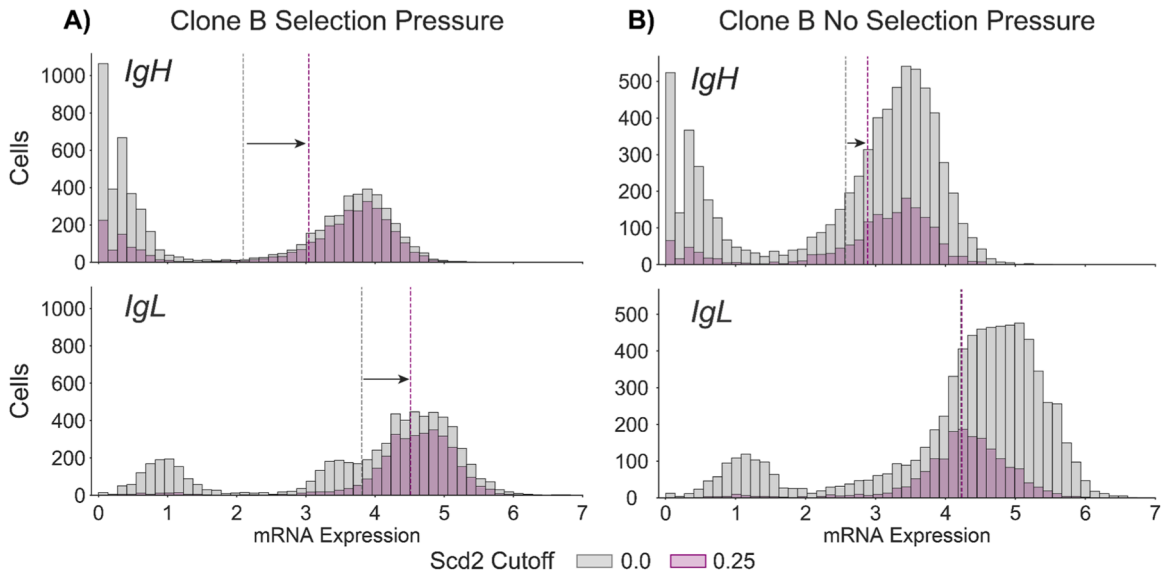

Fig. 4. *In silico* cell filtering by coexpressed marker gene *Scd2* in clone B P30 samples. A) mRNA expression distributions (grey) of IgH and IgL per cell (n = 6539) are plotted for the clone B P30 sample under selection pressure. Cells were then filtered for a minimum *Scd2* log2 expression threshold of 0.25, with n = 3294 cells passing. IgG expression was then plotted only for these filtered cells (purple). B) As in (A) but for clone B P30 without selection pressure (n = 7145) and filtered by Scd2 (n = 1708 cells passed threshold). Average mRNA expression per filter condition is displayed as colored vertical dotted lines.

expression would have a fitness advantage under selection.

We then focused our analysis on Clone B P30 samples, which showed the most prominent low-expressing subpopulation among the unstable cells (Fig. 2D-E). Specifically, we examined the Clone B P30 samples, where a substantial percentage of cells showed low production: 25.4 % in P30 + Gln and 47.6 % in P30-Gln. Using strict filtering criteria (p-value cutoff, minimum fold-change, and minimum percent non-zero expression to control for sparse data), our analysis identified five differentially expressed genes, four of them shared between both Clone B P30 samples. As expected, two of the shared genes were the IgG heavy and light chain transcripts, validating our classification of high and low-producing populations.

The other two shared genes were acyl-CoA desaturase 2 (*Scd2*) and cyclin D1 (*Ccnd1*), both showing coexpression patterns with the IgG chains. *Ccnd1*, which functions in cell cycle regulation and DNA damage response, can localize to either the nucleus or cytoplasm. Its cytoplasmic form has been implicated in metabolic regulation, potentially contributing to adaptation to the glutamine synthetase selection system (Tchakarska and Sola, 2020). Importantly, this *Ccnd1* association is not merely a cell cycle artifact, as we had regressed out cell cycle effects and found no cluster-specific cell cycle components (Figures S3–S4, 5D-E). *Scd2*, located in the endoplasmic reticulum, participates in fatty acid synthesis and influences plasma membrane flexibility through its role in membrane content regulation (Sun et al., 2003). It also affects mitochondrial metabolism and mTOR activity, typically supporting protein production (Zhou et al., 2021). Both genes may serve as cellular stress indicators associated with protein production under selection pressure.

Given the consistent association of *Scd2* with high-producing cells, we tested its potential as a marker gene through *in silico* cell filtering experiments. Using *Scd2* expression levels effectively identified high-producing cells, particularly in aged B P30 samples under selection pressure. This filtering approach proved specific to Clone B - when applied to Clone A samples, it had minimal effect, though this may simply reflect the already low proportion of low-producing cells in Clone A (Figure S21). While *Scd2* appears to be a Clone B-specific marker, the approach of identifying such markers could be valuable for other clones. Additional studies across multiple CHO cell lines could determine whether *Scd2* or similar genes could serve as general markers of productivity loss across different CHO cell clones.

5. Conclusions

Our findings reinforce the notion that high-IgG-producing cells are under higher stress than the low-producing subpopulation. Given the higher survivability, lower IgG titer, and lack of defining stress response genes, the low producing subpopulations seem to have subverted the selection pressure and are trending toward cellular stability. We identified *Scd2* as a marker that can distinguish between high and low-producing subpopulations within Clone B. This suggests that identified clone-specific markers could be valuable in cell sorting or other forms of continued selection. However, Clone A's loss in titer seems to derive from a more uniform population, without clear transcriptional heterogeneity. Different mechanisms than sample heterogeneity measurable with single cell RNA sequencing could cause loss of productivity. Further investigation should be performed to validate and ultimately apply the differential genes and biomarkers identified.

CRediT authorship contribution statement

Winston Timp: Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Michael Betenbaugh:** Writing – review & editing, Funding acquisition, Conceptualization. **Siqi Alice Chen:** Methodology, Investigation. **Haoyu Chris Cao:** Writing – original draft, Data curation. **Luke B. Morina:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Natalia I. Majewska:** Investigation. **Kevin S. McFarland:** Investigation. **Swetha Kumar:** Methodology, Investigation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Winston Timp reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thanks to all AMBIC member universities (Johns Hopkins University, Clemson University, University of Delaware, University of Maryland College Park, and the University of Massachusetts Lowell) and companies for their mentorship and financial support. Thanks to MilliporeSigma for sharing CHO cells. This work was supported by IUCRC NSF Grant 1624684.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jbiotec.2025.01.010.

Data availability

Data is available through GEO GSE275930 - the dataset is public: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE275930

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Barnes, L.M., Bentley, C.M., Dickson, A.J., 2003. Stability of protein production from recombinant mammalian cells. Biotechnol. Bioeng. 81, 631–639.

Borsi, G., Motheramgari, K., Dhiman, H., Baumann, M., Sinkala, E., Sauerland, M., Riba, J., Borth, N., 2023. Single-cell RNA sequencing reveals homogeneous transcriptome patterns and low variance in a suspension CHO-K1 and an adherent HEK293FT cell line in culture conditions. J. Biotechnol. 364, 13–22.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., Adey, A., Waterston, R.H., Trapnell, C., Shendure, J., 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 357, 661–667.

Cha, J., Lee, I., 2020. Single-cell network biology for resolving cellular heterogeneity in human diseases. Exp. Mol. Med. 52, 1798–1808.

Choi, Y.H., Kim, J.K., 2019. Dissecting cellular heterogeneity using single-cell RNA sequencing. Mol. Cells 42, 189–199.

Clarke, C., Doolan, P., Barron, N., Meleady, P., O'Sullivan, F., Gammell, P., Melville, M., Leonard, M., Clynes, M., 2011. Predicting cell-specific productivity from CHO gene expression. J. Biotechnol. 151, 159–165.

Dahodwala, H., Lee, K.H., 2019. The fickle CHO: a review of the causes, implications, and potential alleviation of the CHO cell line instability problem. Curr. Opin. Biotechnol. 60, 128–137.

Deng, Q., Ramsköld, D., Reinius, B., Sandberg, R., 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343, 193–196.

Doolan, P., Clarke, C., Kinsella, P., Breen, L., Meleady, P., Leonard, M., Zhang, L., Clynes, M., Aherne, S.T., Barron, N., 2013. Transcriptomic analysis of clonal growth rate variation during CHO cell line development. J. Biotechnol. 166, 105–113.

Fang, Z., Liu, X., Peltz, G., 2023. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics 39, btac757.

Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., Hill, D.P., Lee, R., Mi, H., Moxon, S., Mungall, C.J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P.D., Van Auken, K., Ramsey, J., Siegele, D.A., Chisholm, R.L., Fey, P., Aspromonte, M.C., Nugnes, M.V., Quaglia, F., Tosatto, S., Giglio, M., Nadendla, S., Antonazzo, G., Attrill, H., Dos Santos, G., Marygold, S., Strelets, V., Tabone, C.J., Thurmond, J., Zhou, P., Ahmed, S.H., Asanitthong, P., Luna Buitrago, D., Erdol, M. N., Gage, M.C., Ali Kadhum, M., Li, K.Y.C., Long, M., Michalak, A., Pesala, A., Pritazahra, A., Saverimuttu, S.C.C., Su, R., Thurlow, K.E., Lovering, R.C., Logie, C., Oliferenko, S., Blake, J., Christie, K., Corbani, L., Dolan, M.E., Drabkin, H.J., Hill, D. P., Ni, L., Sitnikov, D., Smith, C., Cuzick, A., Seager, J., Cooper, L., Elser, J., Jaiswal, P., Gupta, P., Jaiswal, P., Naithani, S., Lera-Ramirez, M., Rutherford, K., Wood, V., De Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Lauledeerkind, S.J.F., Tutaj, M.A., Vedi, M., Wang, S.-J., D'Eustachio, P., Aimo, L., Axelsen, K., Bridge, A., Hyka-Nouspikel, N., Morgat, A., Aleksander, S.A., Cherry, J. M., Engel, S.R., Karra, K., Miyasato, S.R., Nash, R.S., Skrzypek, M.S., Weng, S., Wong, E.D., Bakker, E., Berardini, T.Z., Reiser, L., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Blatter, M.-C., Boutet, E., Breuza, L., Bridge, A., Casals-Casas, C., Coudert, E., Estreicher, A., Livia Famiglietti, M., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Pedruzzi, I., Pourcel, L., Poux, S., Rivoire, C., Sundaram, S., Bateman, A., Bowler-Barnett, E., Bye-A-Jee, H., Denny, P., Ignatchenko, A., Ishtiaq, R., Lock, A., Lussi, Y., Magrane, M., Martin, M.J., Orchard, S., Raposo, P., Speretta, E., Tyagi, N., Warner, K., Zaru, R., Diehl, A.D., Lee, R., Chan, J., Diamantakis, S., Raciti, D., Zarowiecki, M., Fisher, M., James-Zorn, C., Ponferrada, V., Zorn, A., Ramachandran, S., Ruzicka, L., Westerfield, M., 2023. The Gene Ontology knowledgebase in 2023. Genetics 224.

Haque, A., Engel, J., Teichmann, S.A., Lönnberg, T., 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 9, 75.

Hsu, H.-H., Araki, M., Mochizuki, M., Hori, Y., Murata, M., Kahar, P., Yoshida, T., Hasunuma, T., Kondo, A., 2017. A systematic approach to time-series metabolite profiling and RNA-seq analysis of Chinese Hamster Ovary Cell Culture. Sci. Rep. 7, 43518.

Jiang, R., Sun, T., Song, D., Li, J.J., 2022. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. 23, 31.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27–30.

Kim, J.Y., Kim, Y.-G., Lee, G.M., 2012. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl. Microbiol. Biotechnol. 93, 917–930.

Kim, T.K., Eberwine, J.H., 2010. Mammalian cell transfection: the present and the future. Anal. Bioanal. Chem. 397, 3173–3178.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J. R., Shalek, A.K., Regev, A., McCarroll, S.A., 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214.

Ogata, N., Nishimura, A., Matsuda, T., Kubota, M., Omasa, T., 2021. Single-cell transcriptome analyses reveal heterogeneity in suspension cultures and clonal markers of CHO-K1 cells. Biotechnol. Bioeng. 118, 944–951.

Pilbrough, W., Munro, T.P., Gray, P., 2009. Intraclonal protein expression heterogeneity in recombinant CHO cells. PLoS One 4, e8432.

Quiros, P.M., Goyal, A., Jha, P., Auwerx, J., 2017. Analysis of mtDNA/nDNA Ratio in Mice. Curr. Protoc. Mouse Biol. 7, 47–54.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A.P., Regev, A., 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature 510, 363–369.

Sun, Y., Hao, M., Luo, Y., Liang, C.-P., Silver, D.L., Cheng, C., Maxfield, F.R., Tall, A.R., 2003. Stearoyl-CoA desaturase inhibits ATP-binding Cassette Transporter A1-mediated cholesterol efflux and modulates membrane domain structure. J. Biol. Chem. 278, 5813–5820.

Tchakarska, G., Sola, B., 2020. The double dealing of cyclin D1. Cell Cycle 19, 163–178.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer, S.W., Gaillard, A., Kolb, K.E., Villani, A.-C., Johannessen, C. M., Andreev, A.Y., Van Allen, E.M., Bertagnolli, M., Sorger, P.K., Sullivan, R.J., Flaherty, K.T., Frederick, D.T., Jané-Valbuena, J., Yoon, C.H., Rozenblatt-Rosen, O., Shalek, A.K., Regev, A., Garraway, L.A., 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–196.

Tzani, I., Herrmann, N., Carillo, S., Spargo, C.A., Hagan, R., Barron, N., Bones, J., Shannon Dillmore, W., Clarke, C., 2021. Tracing production instability in a clonally derived CHO cell line using single-cell transcriptomics. Biotechnol. Bioeng. 118, 2016–2030.

Wang, Y., Tian, X., Ai, D., 2021. Cell heterogeneity analysis in single-cell RNA-seq data using mixture exponential graph and markov random field model. Biomed. Res. Int. 2021, 9919080.

Wolf, F.A., Angerer, P., Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15.

Zhou, X., Zhu, X., Li, C., Li, Y., Ye, Z., Shapiro, V.S., Copland, J.A., 3rd, Hitosugi, T., Bernlohr, D.A., Sun, J., Zeng, H., 2021. Stearoyl-CoA desaturase-mediated monounsaturated fatty acid availability supports humoral immunity. Cell Rep. 34, 108601.