Received: 16 March 2022 | Revised: 20 July 2022 | Accepted: 28 August 2022

DOI: 10.1002/bit.28226

ARTICLE

BIOTECHNOLOGY AND BIOENGINEERING WILEY

De novo assembly and annotation of the CHOZN® GS$^{-/-}$ genome supports high-throughput genome-scale screening

Corey Kretzmer[1] | Rajagopalan Lakshmi Narasimhan[2] | Rahul Deva Lal[2] | Vincent Balassi[1] | James Ravellette[1] | Ajaya Kumar Kotekar Manjunath[2] | Jesvin Joy Koshy[2] | Marta Viano[3] | Serena Torre[3] | Valeria M. Zanda[3] | Mausam Kumravat[2] | Keith Metelo Raul Saldanha[2] | Harikrishnan Chandranpillai[2] | Ifra Nihad[2] | Fei Zhong[4] | Yi Sun[5] | Jason Gustin[1] | Trissa Borgschulte[1] | Jiajian Liu[4] | David Razafsky[1]

[1]Upstream Research and Development, MilliporeSigma, St. Louis, Missouri, USA

[2]Bioinformatics, IT R&D Applications, Merck (Sigma-Aldrich Chemicals Pvt. Ltd., A subsidiary of Merck KGaA, Darmstadt, Germany), Bangalore, India

[3]Istituto di Ricerche Biomediche "A. Marxer" RBM S.p.A., Ivrea, Italy

[4]Life Science Bioinformatics, IT, MilliporeSigma, St. Louis, Missouri, USA

[5]Bioinformatics, IT R&D Applications, MilliporeSigma, St. Louis, Missouri, USA

Correspondence

David Razafsky, Upstream Research and Development, MilliporeSigma, St. Louis, MO, USA.
Email: David.Razafsky@MilliporeSigma.com

Jiajian Liu, Life Science Bioinformatics, IT, MilliporeSigma, St. Louis, Missouri, USA.
Email: Jiajian.Liu@MilliporeSigma.com

Abstract

Chinese hamster ovary (CHO) cells have been used as the industry standard for the production of therapeutic monoclonal antibodies for several decades. Despite significant improvements in commercial-scale production processes and media, the CHO cell has remained largely unchanged. Due to the cost and complexity of whole-genome sequencing and gene-editing it has been difficult to obtain the tools necessary to improve the CHO cell line. With the advent of next-generation sequencing and the discovery of the CRISPR/Cas9 system it has become more cost effective to sequence and manipulate the CHO genome. Here, we provide a comprehensive de novo assembly and annotation of the CHO-K1 based CHOZN® GS$^{-/-}$ genome. Using this platform, we designed, built, and confirmed the functionality of a whole genome CRISPR guide RNA library that will allow the bioprocessing community to design a more robust CHO cell line leading to the production of life saving medications in a more cost-effective manner.

KEYWORDS

cell line engineering, Chinese hamster ovary, CRISPR, genome-wide pooled screens, Bioprocessing

1 | INTRODUCTION

Chinese hamster ovary (CHO) cells are the predominant cell line used to produce recombinant therapeutic proteins in the biopharmaceutical industry. The decision to utilize CHO cells for the production of recombinant therapeutics stems from several advantageous characteristics, including their adaptability to suspension culture conditions, vigorous growth characteristics, ability to grow in chemically defined media and capacity to secrete properly folded, post-translationally modified biotherapeutic proteins (Bandaranayake & Almo, 2014; Stolfa et al., 2018). As a testament to the utility of CHO cells, the biotechnology and pharmaceutical communities have made substantial investments over the past several decades to improve every facet of the production process; leading to higher

Corey Kretzmer, Rajagopalan Lakshmi Narasimhan, and Rahul Deva Lal contributed equally to this study.

protein titers, better control of protein quality attributes and ultimately safer products for patients (Stolfa et al., 2018; Walsh, 2006). Despite these advances, there continues to be increasing pressure to reduce the cost of goods and timelines associated with the development and large-scale manufacturing of these life-saving medications (Tihanyi & Nyitray, 2021).

While improvements to the manufacturing process have already contributed to reduced drug production costs, enhancements to the CHO cell lines themselves have lagged-behind; likely as a result of inadequate genomic resources, as well as the cost and complexity of designing gene-editing reagents. With the enhanced efficiency of next-generation sequencing technologies, as well as the development of new chromosome conformation capture techniques, such as Hi-C, it is now more financially feasible to sequence the genome and transcriptome of CHO cell lines which could provide a greater understanding of the genetic basis of favorable manufacturing phenotypes (Hilliard et al., 2020; Rupp et al., 2018). Moreover, with the discovery of the CRISPR/Cas9 system we can now more effectively identify and modulate the expression of genes that lead to favorable manufacturing phenotypes, including enhanced biotherapeutic protein productivity (Cox et al., 2015; H. Li et al., 2020).

Although several groups have published genomic assemblies of CHO cell genomes, these efforts have suffered from the well-documented short-comings of short-read sequencing technologies, namely the inability to place repetitive sequences within the overall framework of the assembly (Hilliard et al., 2020; Kaas et al., 2015; Lewis et al., 2013; Rupp et al., 2018; Xu et al., 2011). Furthermore, those assemblies that have utilized long-read and chromosome conformation capture sequencing techniques were initiated using liver tissue from a Chinese hamster, which has been previously shown to have significantly less genomic plasticity and be less representative of the aneuploid CHO cell lines by both sequencing and karyotyping (Hilliard et al., 2020; Kaas et al., 2015; Rupp et al., 2018; Vcelar et al., 2018). To improve the tools available to study CHO cells, we have utilized a combination of short- and long-read sequencing technologies along with chromosome conformation capture techniques to sequence the genome and transcriptome of the CHO-K1 derived, industry-relevant, CHOZN® GS$^{-/-}$ host cell line. Our efforts have resulted in an assembly that encompasses >90% of the predicted genome size as well as the annotation of >20,000 genes.

The CHOZN® GS$^{-/-}$ assembly has allowed us to design complex genetic engineering screens for CHO cells, similar to screens that have been available for scientific communities working in human or mouse systems for nearly a decade. These tools have been instrumental in completing genetic screens that have elucidated the cellular networks that play a role in regulating cancer cell growth rates, altering sensitivity of cells to selective pressures (such as drugs or toxins) and identifying genes that are essential for cell survival (Joung et al., 2017; Koike-Yusa et al., 2014; Peng et al., 2015; Shalem et al., 2014; Wang et al., 2014, 2015; Xiong et al., 2021; Zhou et al., 2014; Zhu, et al., 2016). Using the CHOZN® GS$^{-/-}$ assembly we designed and built a whole-genome CRISPR/Cas9 gRNA (guide RNA) library for CHO cells. To test the efficacy of the pooled whole-genome gRNA library, we developed CHOZN$^{®Cas9}$ helper subclones which constitutively, and stably, express Cas9 as well as an industry relevant therapeutic IgG_1 molecule. To confirm proper functionality of the genomic tools, as well as the CHOZN$^{®Cas9}$ helper subclones, we performed a screen to identify genes that, in CHO cells, confer resistance to the toxic nucleotide analog 6-thioguanine (6-TG). Upon treatment with 6-TG, we observed a substantial enrichment of gRNAs targeting the gene *HPRT1*, which encodes the protein hypoxanthine phosphoribosyltransferase 1. This enrichment in *HPRT1* gRNAs is consistent with previous screens (Koike-Yusa et al., 2014; Peng et al., 2015; Wang et al., 2014). To validate this observation, we developed CHOZN® GS$^{-/-}$ clones that contained frameshift mutations in the *HPRT1* gene and treated the resultant clones with 6-TG. As expected, when 6-TG was supplemented in the media of clones harboring a frameshift mutation in HPRT1 the cells survived, while cells with no modification to the *HPRT1* coding sequence did not. Together this suggests that the scientific tools and processes described here are fully functional and can be utilized to modify CHO cells to be more resilient and productive in bioproduction processes. We believe the combination of the tools described here, as well as the continued advancement of both the upstream- and downstream-bioprocessing units, will play a pivotal role in providing patients with access to the medications they depend on in a more timely and financially responsible manner.

2 | RESULTS

2.1 | De novo assembly and annotation of the CHOZN® GS$^{-/-}$ genome

To provide a robust CHO cell genomic platform and lay the foundation for further ‘omics studies, the CHOZN® GS$^{-/-}$ genome and transcriptome were sequenced using a variety of second- and third-generation sequencing technologies as well as chromosome conformation capture techniques (Supporting Information: Supplemental Tables 1–4). The CHOZN® GS$^{-/-}$ genome was constructed in two stages using a hybrid approach. First, SOAPdenovo 2.04 (Luo et al., 2012) was utilized to generate the CHOZN® GS$^{-/-}$ genome assembly from paired-end and mate-paired Illumina reads from libraries with insert sizes ranging from 430 bp to 10 kb (Supporting Information: Supplemental Table 1), while the gaps within the assembled scaffolds were filled with error-corrected PacBio long-reads (Supporting Information: Supplemental Table 2). Next, the HiRise pipeline was implemented using the assembled CHOZN® GS$^{-/-}$ genome, obtained above, as an input, along with CHiCAGO and Hi-C sequencing libraries (Supporting Information: Supplemental Table 3), permitting the development of a more robust and contiguous assembly, referred to as the CHOZN® GS$^{-/-}$ genome version 2.3 (v2.3). The genome size and scaffold N50 for CHOZN® GS$^{-/-}$ v2.3 are 2.4 Gbp and 43.52 Mbp, respectively. The CHOZN® GS$^{-/-}$ genome v2.3 provides the single most continuous assembly of a

CHO cell line and the second most continuous genome assembly, falling short of only the PICRH assembly, of any *Cricetulus griseus* derived cell line or tissue assembly currently available (Table 1). While most CHOZN® $GS^{-/-}$ v2.3 assembly metrics, including the N50, L50, and total number of scaffolds ≥2 kb, represent a significant improvement over historical CHO cell data we do note that there are two metrics, Total number of scaffolds and % Gaps, which could be improved in future CHO cell assemblies (Table 1). Several factors may affect gaps in de novo genome assemblies, including read depth and the extent of repeats in the genome. A large fraction of our mate-paired reads were removed due to their redundancy which lead to a large reduction of qualified mate-paired reads in this assembly (Supporting Information: Supplemental Table 1). We attribute the high number of Total scaffolds and % Gaps to the relative lack of diversity in the length of the Illumina mate-pair libraries, which could benefit from the incorporation of sequencing data from a more diverse array of longer insert mate-pair libraries in the future, as mate-pair libraries provide the most information towards the resolution of repetitive genome sequences.

After testing several *ab initio* and evidence-based gene prediction methods, AUGUSTUS (version 3.2.2) (Stanke et al., 2006) was selected. Before gene prediction with AUGUSTUS, we collected CHO cDNA sequences and performed RNA transcript assembly from RNA-seq data (Supporting Information: Supplemental Table 4) that was used to train the AUGUSTUS gene models. This refined AUGUSTUS (version 3.2.2) gene model was then used to predict genes at a genome scale in the CHOZN® $GS^{-/-}$ assembly v2.3, by integrating extrinsic evidence from RNA-seq data. The final gene set includes 20,414 genes. The predicted genes were then annotated using tools including InterPro, pfam, gene ontology (GO), and KEGG pathway. To examine how well our predicted genes are conserved in the mouse genome, we employed OrthoMCL (L. Li et al., 2003) and mutual best hits to identify orthologous genes in the mouse genome with stringent criteria. Of the 20,414 genes in CHOZN® $GS^{-/-}$ v2.3, 17,338 genes (84.9%) were found conserved in the Ensemble (Ensemble 90) mouse genome version GRCm38.p5, indicating that the annotated genes in the CHOZN® $GS^{-/-}$ genome are highly homologous to those in the mouse genome. Similarly, we found ~70%–79% of genes in the CHOZN v2.3 assembly had orthologs in previously published *C. griseus* and CHO cell genomes (Table 2) (Brinkrolf et al., 2013; Hilliard et al., 2020; Lewis et al., 2013; Xu et al., 2011). Comparatively, about 60-65% of genes in previous *C. griseus* and CHO cell genomes had been shown to be orthologous (Brinkrolf et al., 2013; Xu et al., 2011; Lewis et al., 2013).

2.2 | Construction and transduction of a CHO-K1 genome-wide lentiviral gRNA library

Using the CHOZN® $GS^{-/-}$ genome assembly version 2.3, all possible gRNAs targeting the forward- or reverse-strand of coding genes, with an adjacent 5′-NGG-3′ PAM sequence, were identified. Any gene for which we could not identify at least five gRNAs was excluded from the library. A maximum of six gRNAs per gene that exhibited the desired characteristics were included in the library (summarized in Figure 1a, general rules). If more than six gRNAs were identified then those gRNAs with the most stringent criteria were given priority. If more than six gRNAs were identified in a single round of gRNA design then the gRNAs were ranked by specificity score and the six with the highest specificity score were used. All gRNAs were designed within the first two-thirds of the coding region to prevent truncated, yet functional, protein from complicating interpretations of the data. Likewise, gRNA designs within the first exon were excluded to reduce the probability that an alternative downstream start codon would result in production of smaller, yet functional, protein isoforms. To maximize library diversity and minimize overlap between gRNAs targeting the same gene, we required a minimum of 20 bp between adjacent gRNAs. Finally, a minimum of 3 bp mismatches were required in the 5'-N19 region of gRNAs, to minimize potential off-target effects. According to these specifications, if we were unable to identify six unique gRNAs per gene then criteria, such as the minimum sequence between adjacent gRNAs

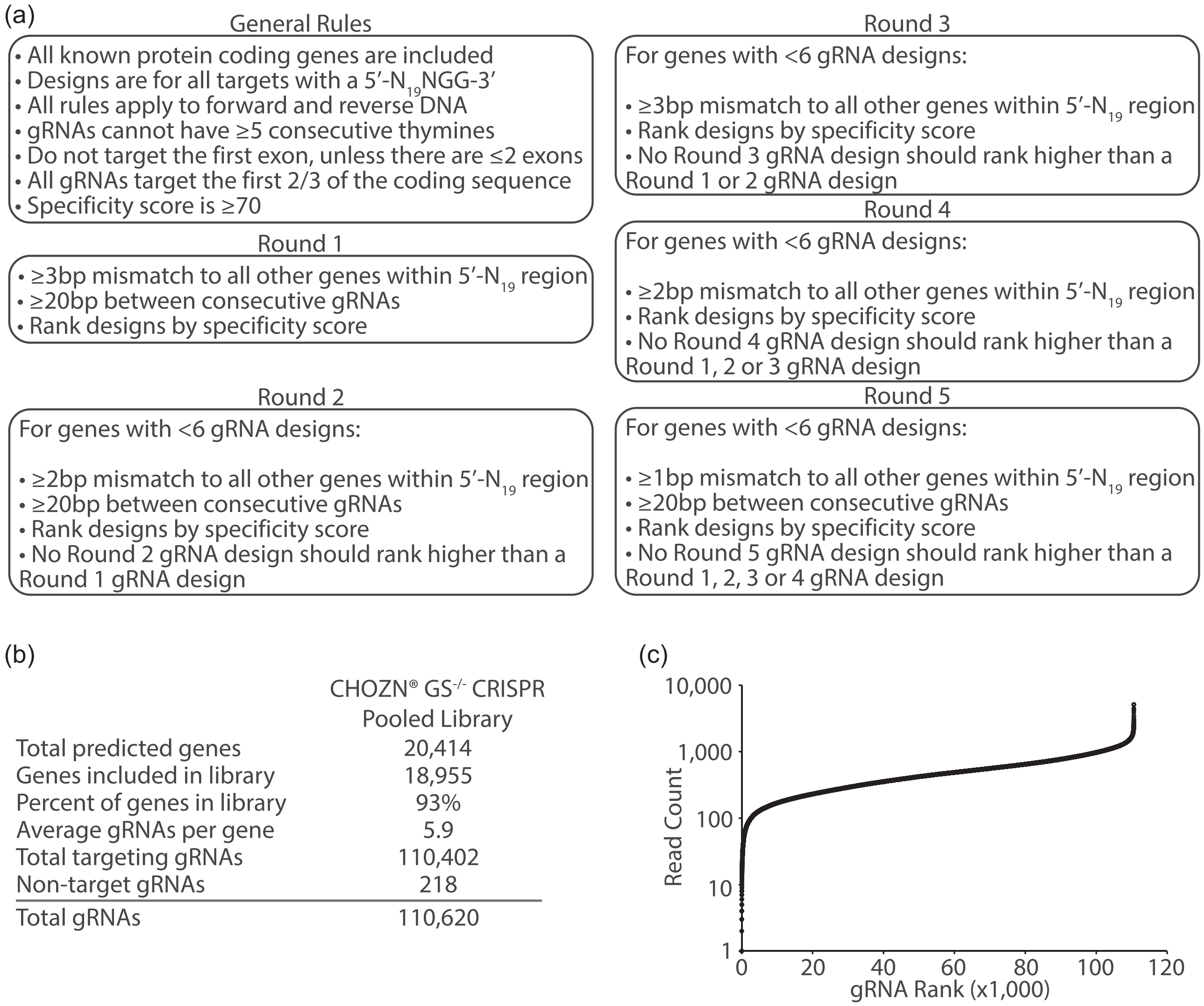
TABLE 1 Comparison of *Cricetulus griseus* and Chinese hamster ovary (CHO) cell genome assembly metrics

	CHOZN v2.3	CriGri_1.0 (2011)	C-griseus_v1.0 (2013)	Cgr1.0 (2013)	PICR (2018)	PICRH (2020)
DNA source	CHO-K1 cells	CHO-K1 cells	*C. griseus* tissues	*C. griseus* tissues	*C. griseus* tissues	*C. griseus* tissues
Assembly length (Gb)	2.4	2.4	2.4	2.3	2.4	2.4
Scaffold N50 (bp)	43,523,667	1,115,615	1,544,832	1,245,000	20,188,720	274,391,693
Scaffold N90 (bp)	549	102,441	346,540	180,686	4,400,570	127,255,434
Scaffold L50	17	547	450	501	32	4
Total number of scaffolds	1,634,314	109,151	52,710	28,749	1,829	647
Total number of scaffolds ≥2 kb	6,285	14,128	6,747	14,081	X	643
% Gaps	27.40	3.30	2.49	10.45	0.12	0.10

Note: Assembly statistics are provided from the referenced publications, no new computational analysis was performed except in the case of CHOZN® $GS^{-/-}$ v2.3. CriGri_1.0 (2011) (Xu et al., 2011), C-griseus_v1.0 (2013) (Lewis et al., 2013), Cgr1.0 (2013) (Brinkrolf et al., 2013), PICR (2018) (Rupp et al., 2018), PICRH (2020) (Hilliard et al., 2020).

TABLE 2 Comparison of *Cricetulus griseus* and Chinese hamster ovary (CHO) cell genome annotations

	CHOZN v2.3	CriGri_1.0 (2011)	C-griseus_v1.0 (2013)	Cgr1.0 (2013)	PICR (2018)	PICRH (2020)
Total predicted genes	20,414	24,383	24,044	X	24,686	X
Protein coding genes	20,414	21,278	20,350	21,779	21,394	21,776
Orthologs to CHOZN v2.3 (protein-coding genes)	X	15,560	15,312	14,322	15,881	16,056
% Orthologous to CHOZN v2.3 (protein-coding genes)	X	76	75	70	74	79



(a)

General Rules
- All known protein coding genes are included
- Designs are for all targets with a 5′-N_{19}NGG-3′
- All rules apply to forward and reverse DNA
- gRNAs cannot have ≥5 consecutive thymines
- Do not target the first exon, unless there are ≤2 exons
- All gRNAs target the first 2/3 of the coding sequence
- Specificity score is ≥70

Round 1
- ≥3bp mismatch to all other genes within 5′-N_{19} region
- ≥20bp between consecutive gRNAs
- Rank designs by specificity score

Round 2
For genes with <6 gRNA designs:
- ≥2bp mismatch to all other genes within 5′-N_{19} region
- ≥20bp between consecutive gRNAs
- Rank designs by specificity score
- No Round 2 gRNA design should rank higher than a Round 1 gRNA design

Round 3
For genes with <6 gRNA designs:
- ≥3bp mismatch to all other genes within 5′-N_{19} region
- Rank designs by specificity score
- No Round 3 gRNA design should rank higher than a Round 1 or 2 gRNA design

Round 4
For genes with <6 gRNA designs:
- ≥2bp mismatch to all other genes within 5′-N_{19} region
- Rank designs by specificity score
- No Round 4 gRNA design should rank higher than a Round 1, 2 or 3 gRNA design

Round 5
For genes with <6 gRNA designs:
- ≥1bp mismatch to all other genes within 5′-N_{19} region
- ≥20bp between consecutive gRNAs
- Rank designs by specificity score
- No Round 5 gRNA design should rank higher than a Round 1, 2, 3 or 4 gRNA design

(b)

	CHOZN® GS$^{-/-}$ CRISPR Pooled Library
Total predicted genes	20,414
Genes included in library	18,955
Percent of genes in library	93%
Average gRNAs per gene	5.9
Total targeting gRNAs	110,402
Non-target gRNAs	218
Total gRNAs	110,620

FIGURE 1 Design and analysis of a genome-wide CHO-K1 gRNA library. (a) Workflow for the design of the genome-wide gRNA library. (b) The number of gRNAs designed in the genome-wide library. (c) Deep-sequencing analysis of the gRNAs in the lentiviral plasmid DNA library.

and/or the number of off-target mismatches with other coding genes, were relaxed to permit additional gRNA designs (Figure 1a). Utilizing these gRNA design criteria, we built a library that consists of 110,402 gRNAs which target 93% (18,955 out of 20,414) of the predicted genes in the CHOZN® GS$^{-/-}$ assembly. As expected, all 18,955 gRNAs targeted the appropriate gene in the CHOZN® GS$^{-/-}$ v2.3 assembly with 0 bp mismatches to the target sequence. However, when the gRNAs were mapped to other publicly available CHO and *C. griseus* genomes we found that ~20% of the gRNAs targeted the appropriate gene with up to a 3 bp mismatch (Table 3). These mismatches are likely a result of mutations that have occurred over time in different cell line lineages, single base-pair sequencing errors

TABLE 3 Alignment of gRNA library with *Cricetulus griseus* and Chinese hamster ovary (CHO) cell genome assemblies

	CHOZN v2.3	CriGri_1.0 (2011)	C-griseus_v1.0 (2013)	PICRH (2020)
Total genes targeted (% orthologs targeted)	18,955 (100)	15,168 (97)	14,976 (98)	15,800 (98)
Genes with 0 bp gRNA mismatch (% of genes)	18,955 (100)	11,782 (78.0)	11,801 (78.8)	12,550 (79.4)
Genes with 1 bp gRNA mismatch (% of genes)	0 (0)	423 (3.0)	408 (2.7)	520 (3.3)
Genes with 2 bp gRNA mismatch (% of genes)	0 (0)	623 (4.0)	582 (3.9)	656 (4.2)
Genes with 3 bp gRNA mismatch (% of genes)	0 (0)	2,340 (15.0)	2,185 (14.6)	2,074 (13.1%)

Note: Percentage of total genes targeted is based on the total number of orthologous genes for each specific genome as presented in Table 2. The percentage of genes with gRNAs targeting with 0, 1, 2, or 3 bp mismatches is calculated based on the number of total genes targeted.

that may be present in different assemblies, assembly errors or incomplete annotations across different genome assemblies. An additional 218 nontargeting gRNAs were included as controls, bringing the total synthesized library to 110,620 gRNAs (Figure 1b). These gRNAs were cloned into a lentiviral vector that contains a puromycin resistance cassette for selection of transduced cells as well as a BFP coding sequence to monitor transduction efficiency. The gRNA library was deep sequenced at a depth of >500x and 99.99% of the gRNAs were detected in the library. Only six gRNAs, each representing a unique gene, were undetectable by deep sequencing, however, all six of these genes were still targeted by a minimum of five gRNAs. As would be anticipated in a library of this size, a small fraction of the gRNAs were over- or underrepresented in the plasmid pool, however, we observed that 99.5% of gRNAs had read counts within 1-log of the median read count (Figure 1c), furthermore we observed only an 8.6-fold difference in the read count of the top and bottom gRNAs within the middle 90% of the gRNA population (Supporting Information: Supplemental Tables 5 and 6).

2.3 | Development and transduction of CHOZN® GS$^{-/-}$ Cas9 helper cell lines

It has been presumed that expression levels of Cas9 can affect the mutation efficiency mediated by specific gRNA sequences (Peng et al., 2015). Therefore, to enhance the statistical power of the screening process, we developed and fully characterized two clonal CHOZN® GS$^{-/-}$ helper cell lines that constitutively express Cas9 as well as an industry-relevant IgG_1 molecule (referred to as CHOZN$^{®Cas9}$). The IgG_1 heavy chain (HC), light chain (LC), and Glutamine Synthetase (GS) genes were introduced into the CHOZN® GS$^{-/-}$ cell line via electroporation of a single plasmid. Top IgG_1 producing mini pools and clones were generated using standard protocols, based on growth, viability, IgG_1 productivity, long-term stability of IgG_1 expression, as well as protein quality attributes (data not shown). Based on clone performance, clone 18-15 was selected to develop the CHOZN$^{®Cas9}$ helper subclones. Parental cell line 18-15 was transduced at two multiplicities of infection (MOIs) with lentivirus harboring the Cas9 coding sequence. Transduced cells were selected with blasticidin and subpools were plated. CHOZN$^{®Cas9}$ subpools were screened via droplet digital PCR (ddPCR) to determine the number of Cas9 integration sites and Cas9 activity was evaluated via the Surveyor Nuclease Detection Assay utilizing a single gRNA targeting Caspase 3. Finally, the performance of both CHOZN$^{®Cas9}$ subpools was evaluated in a 7-day batch assay. Performance of the two CHOZN$^{®Cas9}$ subpools is summarized in Supporting Information: Supplemental Table 7. Given the equivalent cutting activity observed in both of the CHOZN$^{®Cas9}$ subpools, cloning was initiated via limiting dilution from the 1x MOI pool for two important reasons. First, as expected when cells are transduced at a lower MOI, fewer Cas9 transgene integration events were detected which decreases the likelihood of genomic disruptions that could impact cell performance with respect to growth and IgG_1 productivity. Second, the pool transduced at a lower MOI exhibited a more consistent performance in the 7-day batch assay, relative to the 18-15 parental cell line.

Genomic DNA (gDNA) was isolated from all 47 subclones and primers targeting the Cas9 coding sequence were utilized to screen for the stable integration of Cas9 in each subclone. Approximately 95% of the screened subclones contained a Cas9 amplicon of the predicted size (Figure 2a). As previously described, to minimize disruption to the CHOZN® GS$^{-/-}$ genome and increase the probability that CHOZN$^{®Cas9}$ subclones would perform in a manner consistent with the 18-15 parental cell line, we sought subclones that had integrated a single, or at most two, copies of Cas9. ddPCR suggests that 18 of the 47 CHOZN$^{®Cas9}$ subclones contained ≤2 copies of the Cas9 coding sequence (Figure 2b). To assure that the stably integrated copies of Cas9 are functional and to identify subclones with maximal on-target cutting activity, we electroporated the 18 CHOZN$^{®Cas9}$ subclones with a gRNA targeting Caspase 3 and performed a Surveyor Nuclease Detection Assay. Densitometry measurements of the digested PCR products showed consistent- and high- cutting activity across all of the CHOZN$^{®Cas9}$ helper subclones tested (Figure 2c). In fact, all 18 of the subclones displayed cutting efficiencies in the range of 34%–39%, which is comparable to the 27% average cutting efficiency reported previously in human cells (Metzakopian et al., 2017).

Finally, to assure that the introduction of Cas9 had no significant effect on the growth, viability, productivity, and protein quality attributes associated with production or secretion of the therapeutic IgG_1, we performed a fed-batch assay on the CHOZN$^{®Cas9}$ helper

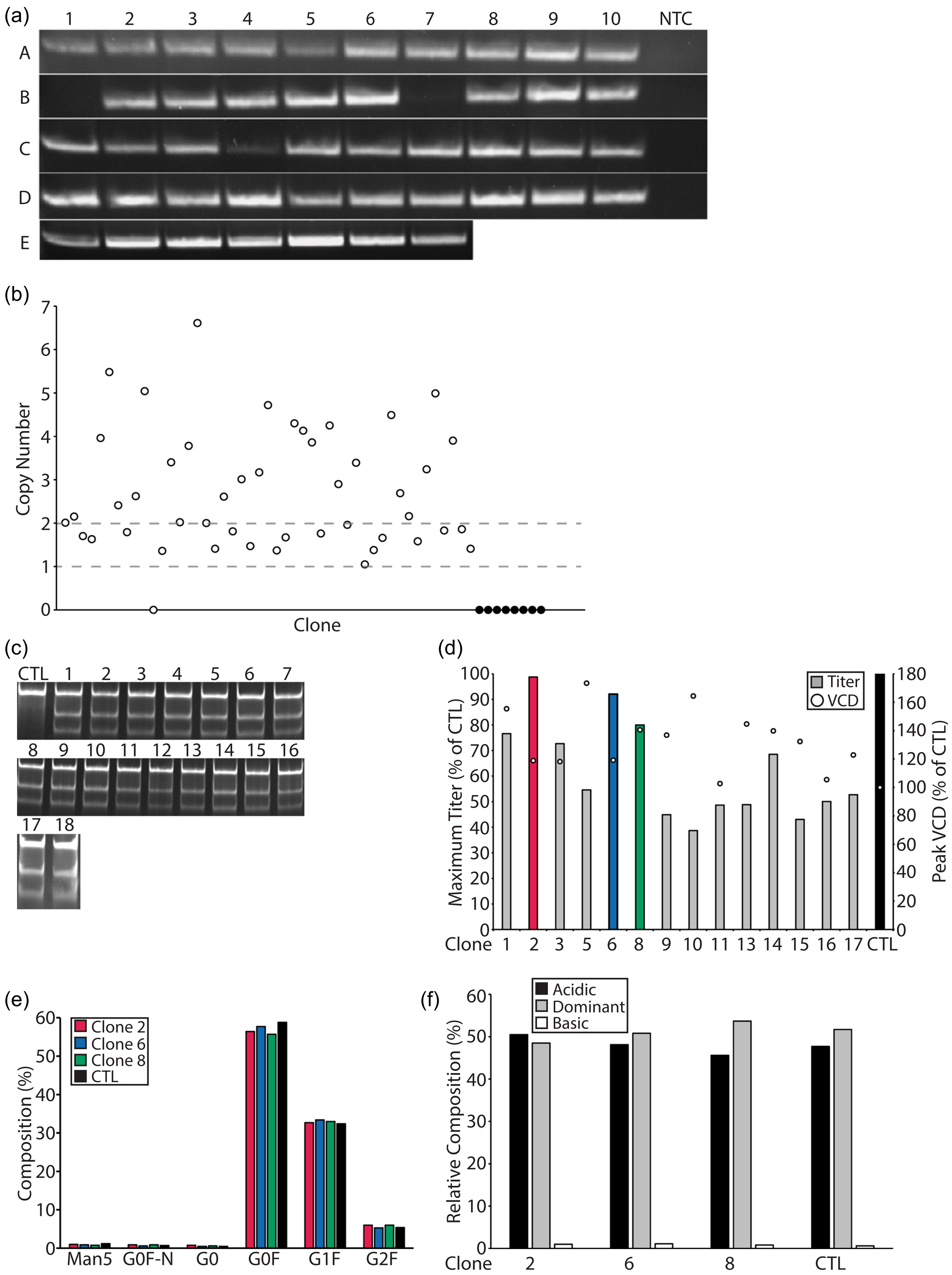


FIGURE 2 (See caption on next page)

subclones. Peak viable cell densities ranged from about 9.5×10^6 cells/ml to 16×10^6 cells/ml, corresponding to 103%–173% of the peak viable cell density of the parental cell line. Similarly, volumetric titers ranged from 39% to 99% of the maximum titer of the parental cell line (Figure 2d). Charge heterogeneity and variation in glycosylation properties of IgG molecules have been of particular concern for the biotechnology and pharmaceutical industry because alterations in these protein quality attributes could significantly alter the biological activity, stability, solubility, bioavailability, pharmacokinetics, and immunogenicity of recombinant protein therapeutics (Khawli et al., 2010; Yehuda & Padler-Karavani, 2020). Having observed no differences in the growth and productivity of clones 2, 6, and 8, relative to the parental cell line, supernatants from these clones were utilized to measure the charge variant and glycosylation pattern of the secreted IgG_1. No significant differences in the charge variant or glycosylation patterns, relative to the parental cell line, were observed in clones 2, 6, and 8 (Figure 2e,f). As a result of the nearly identical performance of these three clones, clones 2 and 8 were chosen for all further studies described here.

The cloned gRNA plasmid library was used to generate a lentivirus pool and transduce CHOZN$^{®Cas9}$ helper subclones 2 and 8. As previously described, transductions were performed at a low MOI to decrease the probability that any given cell would harbor more than a single gRNA (Joung et al., 2017; Zhu, et al., 2016). Four days after transduction, gDNA was isolated from the transduced cells and ddPCR was performed using primers and probes targeting the puromycin selection cassette, as well as a single copy housekeeping gene, Slc35a1 (Kaas et al., 2015). ddPCR indicates <0.4 copies of the lentiviral transgene integrated per CHO cell genome, suggesting that ~35%–40% of cells in the pool were transduced (Supporting Information: Supplemental Figure 1). Poisson statistics indicate that this would result in ~5% of the population of cells harboring ≥2 gRNAs. Transduced cells were selected with puromycin then the gRNAs in the CHOZN$^{®Cas9}$ libraries were deep sequenced at a depth of ~500x and we found that 90.4% and 94.4% of the gRNAs were represented in CHOZN$^{®Cas9}$ clones 2 and 8, respectively. A small portion of gRNAs were underrepresented in the CHOZN$^{®Cas9}$ helper subclone libraries, however 87.1% and 89.0% of gRNAs were represented within 1-log of the median read count for clones 2 and 8, respectively. As previously suggested, we attribute the reduction in the representation of gRNAs in the cell population to the inactivation of essential genes (Wang et al., 2014, 2015; Xiong et al., 2021). Interestingly, our observations correspond with previous estimations that suggest ~10% of genes in the human genome are essential for cell survival (Supporting Information: Supplemental Table 8) (Wang et al., 2015).

2.4 | Validation of the CHO-K1 lentiviral gRNA library

In mammalian cells, purine nucleotides are synthesized via two pathways, the energetically expensive, de novo purine biosynthesis pathway and the purine salvage pathway, the latter of which fulfills the majority of the cells' purine needs by recycling degraded bases via two key enzymes, adenine phosphoribosyltransferase (*APRTase*) and *HPRT1* (Yin et al., 2018). HPRT1 is a nonredundant enzyme that catalyzes the transfer of a phosphoribosyl group from phosphoribosylpyrophosphate (PRPP) to generate inosine monophosphate (IMP) and guanine monophosphate (GMP). Cells harboring functional *HPRT1* will incorporate 6-TG into the DNA via the Purine Salvage Pathway, which hampers DNA replication and ultimately leads to cell death. On the other hand, when *HPRT1* is functionally inactivated, the cell is forced to shift resources to the *de novo* purine biosynthesis pathway, preventing the incorporation of 6-TG into the DNA and allowing the cell to survive (Golan et al., 2012; Yin et al., 2018).

To test the mutant libraries in the CHOZN$^{®Cas9}$ helper subclones we performed a screen to identify target genes that confer resistance to treatment with the toxic guanine analog, 6-TG. Briefly, cells in the mutant libraries were scaled-up, seeded in fresh media with or without an optimized concentration of 6-TG at a total cell count that provides >550x gRNA representation. After five weeks in culture, gDNA was isolated from enough cells to maintain >500x coverage of the whole gRNA library and gRNA abundance was evaluated via deep sequencing (Figure 3a). Reads were mapped to each gRNA in the library and only counted if they perfectly matched the gRNA sequence. Read counts for all six of the *HPRT1* gRNAs were significantly higher in the 6-TG treated samples than in the untreated samples when the screen was performed with both CHOZN$^{®Cas9}$ helper subclones 2 and 8 (Figure 3b,c and Supporting Information: Supplemental Table 8). We then ranked gRNAs by computing the difference in abundance of each gRNA in the treated- and untreated-samples (Figure 3d,e and Supporting Information: Supplemental Table 8). Importantly, all six gRNAs targeting *HPRT1* were within the top 25 most abundant when either clone was treated with 6-TG. However, only a single gRNA targeting *HPRT1* was ranked in the top

FIGURE 2 Generation and characterization of CHOZN$^{®Cas9}$ helper clones. (a) End-point PCR screen for integration of the Cas9 cDNA into CHOZN$^{®Cas9}$ helper subclones. (b) ddPCR copy number analysis of Cas9 integrations into CHOZN$^{®Cas9}$ helper subclones. Black dots represent a control cell line that does not express the Cas9 coding sequence (c) Agarose gel electrophoresis of Surveyor nuclease digestion in CHOZN$^{®Cas9}$ helper subclones transfected with a Caspase3 gRNA. (d) Maximum titer and peak viable cell density (VCD) of CHOZN$^{®Cas9}$ helper subclones, as a percentage of the control parental clones, in a fed-batch assay. (e) Comparison of glycosylation patterns of the secreted IgG_1 in CHOZN$^{®Cas9}$ helper subclones from a fed-batch assay. No significant differences in glycosylation patterns were identified in clones 2, 6, and 8 relative to the control parental clone (CTL). (f) Charge variant analysis of the secreted IgG_1 in CHOZN$^{®Cas9}$ helper subclones from a fed-batch assay. No significant differences were observed in clones 2, 6, and 8 relative to the CTL.

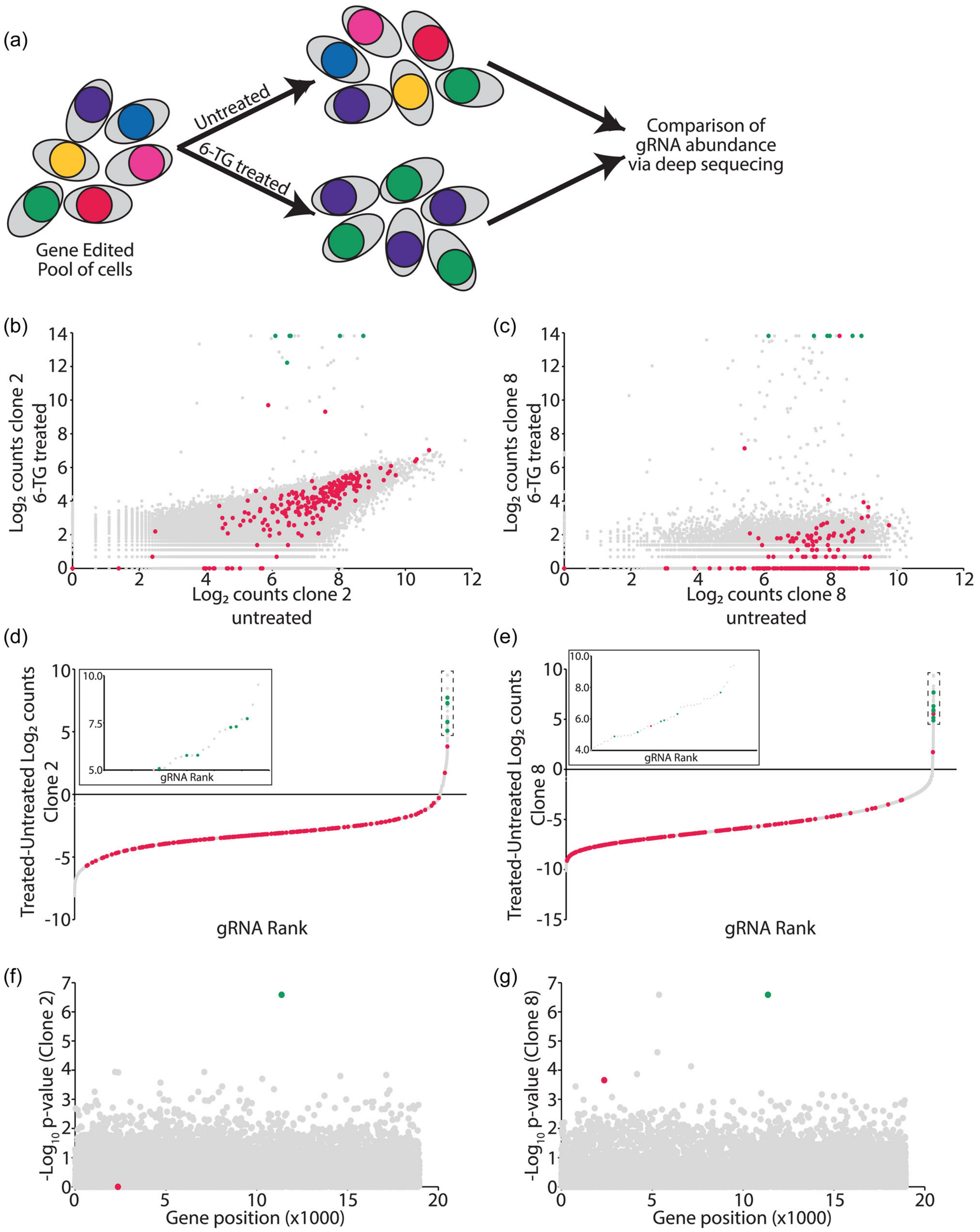


FIGURE 3 (See caption on next page)

1,000 most abundant in the untreated controls. Likewise, *HPRT1* was the only gene that had more than a single gRNA present in the top 100 most abundant when either subclone was treated with 6-TG (Supporting Information: Supplemental Table 8). In addition, we used the MAGeCK algorithm, as previously described, to further confirm our results (Zhu et al., 2016). The output of MAGeCK is a set of positively selected gRNAs, which are capable of functionally inactivating genes that confer cells with resistance to treatment with 6-TG. Through this analysis, *HPRT1* was the highest ranked gene target in screens completed with both CHOZN$^{®Cas9}$ helper subclones (Figure 3f,g and Supporting Information: Supplemental Table 9). Only a single target in either screen, annotated as g13276 in the CHOZN$^{®}$ GS$^{-/-}$ genome, and containing substantial homology to *C. griseus* Washc5, had a significance score comparable to *HPRT1* (Figure 3g). In the CRISPR gRNA library there are six gRNAs targeting *g13276* and only one of these six gRNAs was significantly enriched and only in the screen performed with CHOZN$^{®Cas9}$ helper subclone 8 (Supporting Information: Supplemental Table 8). For comparison, we observed > 100-fold enrichment of all six HPRT1 gRNAs with both CHOZN$^{®Cas9}$ helper subclones 2 and 8 (Supporting Information: Supplemental Table 8), suggesting that g13276 is a false positive, likely related to off-target cutting.

2.5 | Confirmation of CRISPR screen results via targeted disruption of HPRT1

To validate the results from our CRISPR/Cas9 whole-genome screen we opted to utilize an alternative gene-editing technology to discount any possibility that the observations are related to deficiencies in the Cas9 approach, including the miniscule probability that all six *HPRT1* gRNAs are modulating sensitivity to 6-TG through off-target genetic modifications. As a result, we utilized zinc finger nucleases (ZFNs) to obtain isogenic clones in which *HPRT1* is functionally inactive (Chen et al., 2011). Since CHO cells are known to be a complex aneuploid cell line, we used ddPCR to confirm that *HPRT1* is a single copy gene in the CHOZN$^{®}$ GS$^{-/-}$ host cell line (Supporting Information: Supplemental Figure 2) (Kaas et al., 2015). Accordingly, we have never identified more than a single mutant allele in any *HPRT1* knockout clone that has been generated. cDNA from clones 4 and 6 was sequenced and each clone was found to contain a deletion of the entirety of exon 2 (Figure 4a). To ensure these mutations resulted in cell lines with no detectable HPRT1 protein, we performed a western blot using an antibody targeting HPRT1. While HPRT1 protein was detectable at the predicted molecular weight in the CHOZN$^{®}$ GS$^{-/-}$ parental cell line, we were unable to detect HPRT1 protein in both knockout clones 4 and 6 while the housekeeping gene β-actin was detected in the parental cell line as well as both *HPRT1* knockout clones (Figure 4b). Finally, to confirm that the inactivation of HPRT1 confers resistance to 6-TG exposure, we treated clones 4, 6 and the unmodified CHOZN$^{®}$ GS$^{-/-}$ parental cell line with 6-TG for 7 days and measured the viability and viable cell density. As expected, the CHOZN$^{®}$ GS$^{-/-}$ parental cells, which contain a functional *HPRT1* gene, were susceptible to 6-TG treatment, indicated by a drop in cell viability and the absence of cell growth, which were both detectable as early as Day 3 of treatment. Conversely, when the CHOZN$^{®}$ GS$^{-/-}$ parental cell line was grown in the absence of 6-TG we observed high viability and typical growth rates, indicating there are no media, process, or cellular insufficiencies. On the other hand, clones 4 and 6 displayed a viability of ~95% and consistent cell growth throughout the duration of the assay in both the presence and absence of 6-TG (Figure 4c,d). Consistent with previous reports in mouse and human cell lines, these results confirm our findings from the genome-wide CRISPR-screen, that the genetic inactivation of *HPRT1* confers resistance to 6-TG treatment in CHO-K1 cells (Koike-Yusa et al., 2014; Peng et al., 2015; Wang et al., 2014). More importantly, these results indicate that the tools described here, including the CHOZN$^{®Cas9}$ helper subclones as well as the genome-wide CHO-K1 CRISPR library are fully functional and will permit more efficient phenotypic screening to improve CHO cell bioprocessing.

Using high-throughput genetic screens in CHO cells has already proven to be advantageous in mitigating the effects of chemically induced ER-stress, which led to cell lines with more robust cell growth and prolonged culture viability (Xiong et al., 2021). Tools like those reported here, along with whole-genome CRISPR-activator and CRISPR-interference screens, will allow for the unbiased and empirical discovery of genes and cellular pathways that could be genetically engineered to enhance the productivity of CHO cells thus permitting more recombinant protein to be produced on a per cell or per volume basis. Likewise, these tools could be used in the discovery of genes whose modification could increase the rate of cell growth and thus minimize the seed train timeline, defined as the amount of

FIGURE 3 6-TG resistance screen using a genome-wide CHO-K1 gRNA library. (a) Outline of the genetic screening strategy using a genome-wide lentiviral gRNA library. (b, c) Primary 6-TG screening data. HPRT1 gRNAs (green) in subclone 2 (b) and subclone 8 (c) have significantly more mapped reads in 6-TG treated cultures relative to all other gRNAs in both 6-TG and untreated samples. Green represents HPRT1 gRNAs, red represents nontargeting gRNAs, gray represents all other gRNAs. (d, e) gRNAs from the 6-TG screens in subclone 2 (d) and subclone 8 (e) were ranked by differential abundance between treated and untreated conditions. gRNAs with no change in abundance were removed for clarity. Inset indicates that all six gRNAs targeting HPRT were among the most enriched gRNAs in both screens. Green represents HPRT1 gRNAs, red represents nontargeting gRNAs, gray represents all other gRNAs. Higher differential expression indicates a larger enrichment in the 6-TG treated sample relative to the untreated sample. (f, g) Gene hit identification comparing differential abundances of all gRNAs targeting a gene in the 6-TG screens calculated by MAGeCK. HPRT1 is the most significant hit identified using both subclone 2 (f) and subclone 8 (g). Green represents the aggregate score of HPRT1, red represents the aggregate score of all nontargeting controls, and gray represents the aggregate score of each remaining gene. Higher -Log$_{10}$ *p*-values indicate a stronger selection of the corresponding gene.

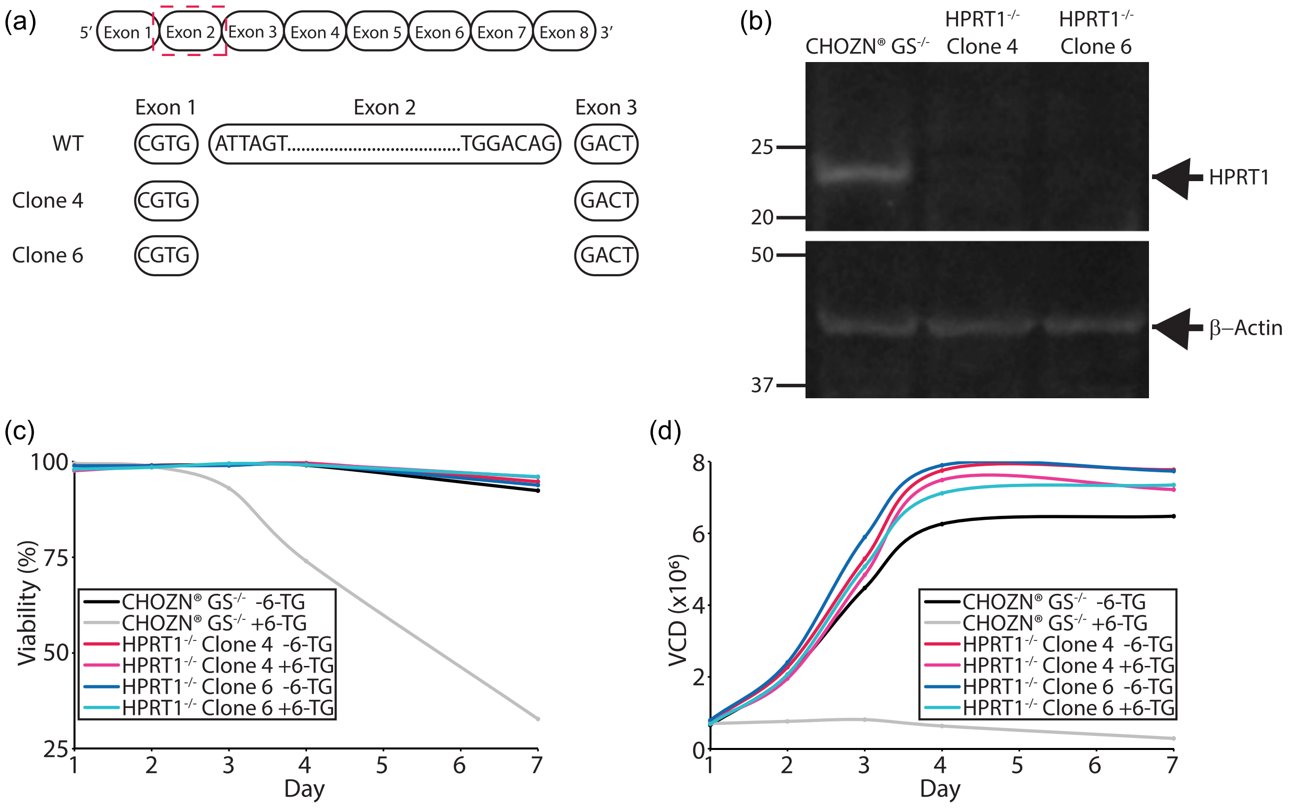


FIGURE 4 Validation of HPRT1 in conferring 6-TG resistance to CHO-K1 cells. (a) Depiction of the deletions identified in the cDNA of HPRT1. Both clones 4 and 6 contain deletions of the entirety of exon 2 while exons 1, 3, and all other downstream exons remain completely intact. (b) Western blot of HPRT1 and β-Actin. HPRT1 protein was detectable at the expected molecular weight in CHOZN® GS$^{-/-}$, but not HPRT1$^{-/-}$ clones 4 and 6 while the loading control β-Actin was identified in all three samples. (c, d) Viability and cell growth plots of control CHOZN® GS$^{-/-}$ cells as well as HPRT1$^{-/-}$ clones 4 and 6. Both HPRT1$^{-/-}$ clones survive and grow in media containing 6-TG while CHOZN® GS$^{-/-}$ cells are only able to survive and grown in media lacking 6-TG.

time required to gain sufficient biomass to go from cryopreserved cells to a production scale bioreactor which can often take ≥30 days, and save substantial time in the scale-up process. In summary, while the enhancement of CHO cells themselves has historically lagged-behind process engineering improvements, we now have the genomic- and genetic engineering-tools available to reduce the time required to identify genes that could lead to better performance of CHO cells and substantially reduce the cost and time associated with manufacturing the life-saving medications that patients rely on every day.

3 | DISCUSSION

For decades CHO cells have been the predominant cell line used by the biomanufacturing community to produce recombinant therapeutic proteins. During this time, we have seen substantial improvements in many facets of the production process, including improved media formulations, improved process parameters and improved hardware designs. Despite these advances, improvements to the CHO cell line have been underwhelming. This is likely a reflection of the historically high cost of curating high-quality genomic and transcriptomic libraries as well as uncertainty about the similarity of *the C. griseus* and CHO cell genomes. Several pieces of evidence suggest that there are significant differences between the genome of *C. griseus* and CHO cells as well as important differences between the diverse CHO host cell lines. For example, at the chromosome level, spectral karyotyping suggests there are significant variations in the chromosome content of *C. griseus*, CHO-S and CHO-K1 host cell lines (Vcelar et al., 2018). Likewise, whole-genome sequencing indicates that while *C. griseus* is largely a diploid organism, there is significant gene copy number variability amongst different CHO host cell lines (Kaas et al., 2015). As a result of this genomic uncertainty, it has been more difficult to develop large-scale ‘omics tools that will be useful across the entire bioprocessing industry, including accurate genomic databases as well as large scale genome editing resources.

We opted to start from a fresh foundation by sequencing the genome of the commercially available and industry-relevant CHOZN® GS$^{-/-}$ host cell line. Our efforts have resulted in substantial improvements to the assembly and annotation of a CHO-K1 derived host cell genome. Through this study, we have improved the genome continuity, as represented by the N50, of the original CHO-K1 sequencing efforts by ~40x and reduced the number of scaffolds ≥2 kb by more than 50% (Xu et al., 2011). These improvements, combined with a fully annotated genome in which we observe ~70%–80% gene orthology with previously published CHO-K1 and *C. griseus* annotations, allowed us to build a whole-genome CRISPR/Cas9 gRNA library. Our analysis of the CHOZN® $GS^{-/-}$ v2.3 gRNA library suggests that using a gRNA library derived from a cell line specific genome assures the highest gRNA targeting efficiency. However, one could anticipate that nearly 80% of the gRNAs would also target orthologous genes in similarly derived CHO cell lines or primary cells from *C. griseus* with 0 bp mismatches. In a whole genome gRNA library as expansive as that reported here, this would reduce the statistical power of having up to six gRNA target each gene, however, whole genome CRISPR screens have been performed in CHO cells using gRNA libraries with fewer gRNAs targeting each coding gene (Xiong et al., 2021). On the other hand, if less stringent gRNA targeting criteria are implemented by allowing up to 3 bp mismatches between gRNAs and their target sequence, then we anticipate that the gRNA library provided here would perform equally well in any CHO-K1 cell line or primary cells from *C. griseus*. Importantly, the CHOZN® $GS^{-/-}$ v2.3 assembly does not account for splice variants of any gene and as a result the gRNA library may not target all transcript isoforms. However, we anticipate that future improvements to the assembly and annotation will include incorporation of splice variants allowing for a more careful analysis and, if needed, a new gRNA library with modified designs.

Finally, we show here that the tools derived from this study, including the assembly, annotation, and CHOZN®^{Cas9} cells are fully functional and were able to recapitulate CRISPR screens previously conducted in other mammalian cell lines (Koike-Yusa et al., 2014; Peng et al., 2015; Wang et al., 2014). While the 6-TG screen was essential to validate the performance of the tools described here, the full benefit of this study will only be realized as novel screens are conducted to understand processes more pertinent to the biomanufacturing community. As an example, some cell culture media raw materials are difficult to procure or validate while other components can be more susceptible to degradation during storage (Neutsch et al., 2018). Using the tools available through this study one could conduct a genome-wide CRISPR screen in a media deficient in one or more of these challenging media components to determine if there are genetic modifications that could be incorporated into the cell line that might permit the cells to be cultured in media lacking this component(s). As a second example, there are several FACS-based assays that can be utilized to identify and isolate CHO-K1 clones with high recombinant protein productivity (Chakrabarti et al., 2019). Using the tools described here, one could introduce a whole genome gRNA library into CHO cells, use these FACS assays to isolate pools of cells that have favorable phenotypes that are correlated with higher protein productivity and identify genes that more consistently provide high protein production. Once these genes are identified it would simply require routine cell line engineering to obtain a new genetically modified isogenic host cell line with enhanced protein productivity. In summary, the tools developed here lay the foundation to systematically improve the CHO-K1 cell line and usher in the next generation of bioprocessing efficiency improvements.

4 | MATERIALS AND METHODS

4.1 | Cell culture and genome sequencing

CHOZN® $GS^{-/-}$ cells were maintained in suspension culture conditions at 37°C, 5% CO_2 and 80% relative humidity in EX-CELL® CHO CD Fusion media supplemented with 6 mM ʟ-glutamine. Cells were thawed, passaged one time, grown to the exponential phase of the growth cycle then pelleted, rinsed with PBS several times and the rinsed cell pellets were frozen at −80°C. Genomic DNA isolation, library preparation and Illumina® MiSeq/HiSeq2500, PacBio® Sequel™, and Hi-C/Chicago® Library sequencing were performed by the Istituto Di Ricerche Biomediche "Antoine Marxer" RBM S.p.A. (affiliate of Merck KGaA), Ivrea [TO], GeneWiz (South Plainfield) and Dovetail Genomics® (Santa Cruz) respectively.

4.2 | RNA isolation and sequencing

A research cell bank of CHOZN® $GS^{-/-}$ was thawed and passaged three times in EX-CELL® CHO CD Fusion media supplemented with 6mM ʟ-glutamine. RNA was isolated from 1×10^7 cells in the exponential phase of growth using an RNeasy Mini kit (Qiagen) according to the manufacturer's instructions. RNA quantity and integrity (350 ng/µl and RIN 9.8) was measured using a Qubit fluorometer (Thermo Fisher) and Bioanalyzer (Agilent Technologies). Ribosomal RNA was removed with a Ribo-Zero Globin depletion kit (Illumina®) and library prepared by Truseq Stranded Total RNA protocol (Illumina®) according to the manufacturer's instructions. Library was sequenced on an Illumina® NextSeq500 (PE75 reads with Q30 > 80%).

4.3 | De novo assembly of the CHOZN® $GS^{-/-}$ genome

The CHOZN® $GS^{-/-}$ genome assembly was performed using a hybrid approach with Illumina short-read libraries, PacBio long reads, Hi-C and CHiCAGO read libraries. The Illumina short-read libraries consist of six mate-paired libraries with insert sizes from 5,600 to 10,100 nt and three paired-end libraries with two libraries of insert sizes 430nt each and one library with 1,200nt (Supporting Information: Supplemental Table 1). PacBio long reads, Hi-C- and CHiCAGO- libraries were used for the genome assembly and improvement.

SOAPDenovo short-read assembler (version 2.04) was used to assemble the Illumina® paired-end and mate-paired libraries (Luo et al., 2012). The PacBio error-corrected reads were applied to fill gaps in the Illumina® assembly using GMCloser (Kosugi et al., 2015) resulting in v1.0 of the CHOZN® $GS^{-/-}$ genome. HiRise assembler (https://github.com/DovetailGenomics/HiRise_July2015_GR) was used to improve assembly v1.0 by integration with the reads from Chicago and Hi-C libraries to obtain the CHOZN® $GS^{-/-}$ genome version 2.3.

4.4 | CHOZN® GS$^{-/-}$ RNA-seq transcriptome assembly

A de novo CHOZN® GS$^{-/-}$ RNA-seq assembly was performed using Trinity RNA-seq de novo assembler based on a single Illumina paired-end sequencing reads (Grabherr et al., 2011). The raw reads were first subjected to a quality check using FastQC to estimate the overall read quality. Next read correction was performed using Rcorrector, which is a k-mer-based error correction method for Illumina RNA-seq reads (Song & Florea, 2015). Once the reads were corrected, a custom python script, FilterUncorrectablePEfastq. py (https://github.com/harvardinformatics/TranscriptomeAssemblyTools) was used to further remove reads which were labeled as 'unfixable' from Rcorrector. In addition, the sequence header format was corrected to avoid any issues in downstream analysis. FASTX-Toolkit was used for quality trimming at the 3′ end of the reads, in which, the bases were removed if the quality was less than a phred score of 30. Based on the corrected and quality trimmed reads, the coverage was estimated around 55x. These corrected and quality trimmed reads were then used for de novo transcriptome assembly using Trinity with CuffFly option. The transcriptome assembly quality was further assessed using rnaQUAST (version 1.4.0) (Bushmanova et al., 2016).

4.5 | Genome-scale gene prediction and annotation

The genes in the assembled CHOZN® GS$^{-/-}$ genome v2.3 were predicted using AUGUSTUS (version 3.2.2) eukaryotic gene finding tool using two independent training and prediction steps (Stanke et al., 2006). During AUGUSTUS training, a gene training set was created using PASA by aligning the assembled transcriptome assembly with the assembled genomic scaffolds, and a collection of training sets with defined gene structures. The training set was used to learn gene HMM model parameters. The refined gene models were then used to predict CHOZN® GS$^{-/-}$ genes, in which assembled RNA transcripts serves as extrinsic evidence to improve dictions.

InterPro, pfam, KEGG, and GO were utilized to perform genome-wide annotations. Two approaches were used to identify orthologs between CHOZN® GS$^{-/-}$ and mouse genes including OrthoMCL (version 2.0.9) and mutual best hit (L. Li et al., 2003). In mutual best hit approach, three stringent criteria were applied to identify mutual best hits including, (a) mutual top hit, (b) *e*-value 1e-10, and (c) overlap fraction >70%.

4.6 | Genome-wide lentiviral gRNA design

The genome-wide lentiviral gRNA library was designed using our proprietary custom CRISPR design pipeline (Bradford & Perrin, 2019). The designs were specified to target annotated coding genes in the CHOZN® GS$^{-/-}$ genome. For each gene, coding sequences were extracted based on the annotation and gRNAs were designed. For all target sequences, adjacent 5′-NGG-3′ PAM sequences on both, were identified. Each gRNA was designed within the first two-thirds of the coding region to minimize the possibility of truncated, yet functional protein from complicating interpretations of the screens. Likewise, gRNA designs within the first exon were excluded to reduce the probability that an alternative downstream start codon would result in production of smaller, yet functional, protein isoforms. In accordance with standard gRNA design criteria, a stretch of TTTTT was avoided as it reduces the efficiency of gRNA transcription and decreases the efficiency of genome editing (Hiranniramol et al., 2020).

All gRNAs were selected to include a minimum of 20 bp between adjacent gRNAs. Our standard pipeline aligns the gRNAs against the whole soft masked genome and reports aligned locations with mismatch position and the sequence. To minimize potential off-target effects, a minimum of 3 bp mismatches were required for selected gRNAs. For each gene, a maximum of six gRNA designs were included in the library. For the genes where six gRNAs were not obtained, criteria such as the minimum sequence between adjacent gRNAs and/or the number of off-target mismatches with other coding genes were relaxed to acquire a minimum of four gRNA designs. Genes for which at least four gRNA designs could not be obtained were excluded from the library.

4.7 | Development of Cas9 helper subclones

An IgG$_1$ expressing CHOZN® GS$^{-/-}$ clone was transduced with lentivirus (MilliporeSigma) harboring the Cas9 coding sequence as well as a blasticidin antibiotic selection cassette. At the time of virus addition, 8 µg/ml of Polybrene was added to the media to enhance transduction efficiency. Cells were incubated at 37°C, 5% CO$_2$ and 80% relative humidity for 24 h. Cells were then pelleted, and the media was fully exchanged with EX-CELL® CHO CD Fusion media and incubated for an additional 24 h before the initiation of selection with 5 µg/ml of blasticidin. After selection, Cas9 activity was assessed via a Surveyor™ enzyme assay (IDT) following the manufacturer's instructions and a 7-day batch assay was conducted in TPP® TubeSpin® bioreactors. Briefly, cells were cultured in EX-CELL® Advanced CHO Fed-batch media (MilliporeSigma) and fed glucose as needed. Culture viability and cell density were measured using Vi Cells (Beckman Coulter) and IgG$_1$ productivity was assessed from supernatants after 7 days on an Octet Platform (ForteBio) using Protein A biosensors. Single-cell cloning was performed via limiting dilution at 0.5 cells/well in 96-well tissue culture plates (Corning) as previously described (Sealover et al., 2013). Clones were scaled-up, genomic DNA was isolated and ddPCR was performed according to the manufacturer's instructions (Bio-Rad) to determine Cas9 copy number. Cas9 cutting activity of clones was assessed exactly as described for the selected pools and cell growth, viability and IgG$_1$ production was measured via a fed-batch assay essentially as previously described (Lin et al., 2011). Protein quality attributes were determined for the antibody product as follows. The IgG samples were purified using immobilized Protein-A resin, washed with 20 mM citrate, 150 mM NaCl, pH 7 buffer, and eluted with 25 mM citrate buffer, pH 3. N-glycan analysis was performed using UHPLC-FLR-MS based detection where glycans are removed using PNGase F and labeled with procainamide. A BIOshell Glycan column was used for separation (15 x 2.1 x 2.7 μm), fluorescence signal collected using excitation at 308 nm with emission at 359 nm, and MS and MS2 data collected using a Thermo Q-Exactive Plus Mass Spectrometer. Distribution of charge variants was determined by imaged capillary isoelectric focusing (ICIEF) using the ICE3 system by (ProteinSimple) according to the manufacturer's instructions.

4.8 | Whole-genome CRISPR screen and analysis

The 6-TG CRISPR screen was performed essentially as previously described (Joung et al., 2017). Briefly, 2.65×10^8 CHOZN$^{®Cas9}$ cells were transduced with the gRNA library at an infectious MOI of ~0.3, as determined by the percentage of BFP positive cells from seven small-scale preliminary transductions performed in triplicate (Supporting Information: Supplemental Table 10), in EX-CELL® CHO CD Fusion media supplemented with 5 μg/ml blasticidin and 8 μg/ml polybrene. Approximately 24 h later the cells were pelleted and the media was exchanged with fresh EX-CELL® CHO CD Fusion media supplemented with 5 μg/ml blasticidin. Cultures were passaged routinely for 6 days to maintain the appropriate number of cells and gRNA representation. After 6 days, the media was exchanged with fresh EX-CELL® CHO CD Fusion media, supplemented with 5 μg/ml blasticidin and 10 μg/ml puromycin. Upon recovery to ≥95% viability the cells were split into two flasks and a minimum of 6.5×10^7 viable cells were maintained in logarithmic growth in either EX-CELL® CHO CD Fusion media supplemented with 5 μg/ml blasticidin and 10 μg/ml puromycin or EX-CELL® CHO CD Fusion media supplemented with 5 μg/ml blasticidin, 10 μg/ml puromycin and 15 μM 6-TG (MilliporeSigma). Cultures were pelleted, spent media was removed and the pellet was resuspended in fresh media as needed for about 4 weeks. Upon completion, gDNA was isolated from 6.5×10^7 cells (to maintain >500x gRNA coverage) in each condition using the DNeasy miniprep kit (Qiagen) and a QiaCube automated nucleic acid extractor according to the manufacturer's instructions. PCR was performed using JumpStart REDTaq Ready Mix (MilliporeSigma), with 2 μg of gDNA per reaction for 35 cycles. The number of PCR reactions was scaled up such that all harvested gDNA (minimum of 200 μg) was amplified and sequencing was performed on a NexSeq500. Reads were then demultiplexed and trimmed to remove barcodes and adapters. The trimmed reads were aligned with Bowtie2 and ambiguous hits were filtered out. SAMtools was applied to the alignment results to generate depth data. Final analysis was performed using the raw counts as well as the MAGeCK pipeline (W. Li et al., 2014).

4.9 | Generation of HPRT1$^{-/-}$ clones and 6-TG assay

ZFN expression plasmids targeting *HPRT1* genomic sequences were designed using a proprietary algorithm as previously described (Mascarenhas et al., 2017). Gene-editing reagents were electroporated into cells using a Gene Pulser (Bio-Rad) at 140 V and 950 μF in a 0.2 cm cuvette. Electroporated cells were placed in 2 ml of media in a 6-well plate at 30°C for 48-h then returned to 37°C. Single cell cloning was performed via limiting dilution at 0.5 cells/well in 96-well tissue culture plates (Corning) as previously described (Sealover et al., 2013). Modifications to the *HPRT1* coding sequence were confirmed via Sanger sequencing of full-length RT-PCR products. Briefly, RNA was isolated using the RNeasy miniprep kit (Qiagen) and a QiaCube automated nucleic acid extractor according to the manufacturer's instructions. RT-PCR reactions were carried out using Superscript IV Reverse-transcriptase (Thermo Fisher) and Platinum Taq DNA Polymerase High-Fidelity (Thermo Fisher) following the manufacturer's recommendations. PCR clean-ups and Sanger sequencing were performed by Elim Biopharmaceuticals.

4.10 | Western blots

Cells (1×10^7 cells) were pelleted, resuspended, and immediately heated in Laemmli buffer/5% β-mercaptoethanol. Protein lysates were separated by SDS-PAGE, transferred to nitrocellulose membranes (Trans-Blot Turbo Transfer Pack, ThermoFisher), blocked with 5% milk in TBST for 1 h at room temperature and incubated overnight at 4°C with HPRT1 (Abcam #ab109021) and β-actin primary antibodies (Cell Signaling #4967). After three washes with TBST, membranes were incubated with the appropriate Alexa Fluor Plus 555-conjugated secondary antibodies for 1 h at room temperature and images were captured on a ChemiDoc MP imaging system (Bio-Rad).

AUTHOR CONTRIBUTIONS

Corey Kretzmer, Vincent Balassi, Marta Viano, Serena Torre, Valeria M. Zanda, Jason Gustin, Trissa Borgschulte, Jiajian Liu, and David Razafsky designed and conceived the experiments. Corey Kretzmer, Rajagopalan Lakshmi Narasimhan, Rahul Deva Lal, Vincent Balassi, James Ravellette, Ajaya Kumar Kotekar Manjunath, Jesvin Joy Koshy, Marta Viano, Serena Torre, Valeria M. Zanda, Mausam Kumravat, Keith Metelo Raul Saldanha, Harikrishnan Chandranpillai, Ifra Nihad, Fei Zhong, Yi Sun, Jiajian Liu, and David Razafsky acquired, analyzed or interpreted the data. Rajagopalan Lakshmi Narasimhan, Rahul Deva Lal, Ajaya Kumar Kotekar Manjunath, Jason Gustin, Trissa Borgschulte, Jiajian Liu, and David Razafsky drafted or revised the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Robert Grady, Casey Jowdy, Jeremy Lehmann and Shawn Shafer for providing computational assistance and invaluable advice in the design and analysis of the CRISPR screens. We would also like to thank Jasna Despot, Brian Verstraete, David Cutter, and Mathew Schmatz for assistance in the production and sequencing of the CRISPR library.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

David Razafsky http://orcid.org/0000-0001-6292-0169

REFERENCES

Bandaranayake, A. D., & Almo, S. C. (2014). Recent advances in mammalian protein production. *FEBS Letters*, *588*(2), 253–260. https://doi.org/10.1016/j.febslet.2013.11.035

Bradford, J., & Perrin, D. (2019). A benchmark of computational CRISPR-Cas9 guide design methods. *PLoS Computational Biology*, *15*(8), 1007274. https://doi.org/10.1371/journal.pcbi.1007274

Brinkrolf, K., Rupp, O., Laux, H., Kollin, F., Ernst, W., Linke, B., Kofler, R., Romand, S., Hesse, F., Budach, W. E., Galosy, S., Müller, D., Noll, T., Wienberg, J., Jostock, T., Leonard, M., Grillari, J., Tauch, A., Goesmann, A., … Borth, N. (2013). Chinese hamster genome sequenced from sorted chromosomes. *Nature Biotechnology*, *31*, 694–695. https://doi.org/10.1038/nbt.2645

Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V., & Prjibelski, A. D. (2016). rnaQUAST: A quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*, *32*(14), 2210–2212. https://doi.org/10.1093/bioinformatics/btw218

Chakrabarti, L., Mathew, A., Li, L., Han, S., Klover, J., Albanetti, T., & Hawley-Nelson, P. (2019). Mitochondrial membrane potential identifies cells with high recombinant protein productivity. *Journal of Ummunological Methods*, *464*, 31–39. https://doi.org/10.1016/j.jim.2018.10.007

Chen, F., Pruett-Miller, S. M., Huang, Y., Gjoka, M., Duda, K., Taunton, J., Collingwood, T. N., Frodin, M., & Davis, G. D. (2011). High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nature Methods*, *8*(9), 753–755. https://doi.org/10.1038/nmeth.1653

Cox, D., Platt, R. J., & Zhang, F. (2015). Therapeutic genome editing: Prospects and challenges. *Nature Medicine*, *21*(2), 121–131. https://doi.org/10.1038/nm.3793

Golan, D. E., Tashjian, Jr., A. H., Armstrong, E. J., & Armstrong, A. W. (2012). *Principles of pharmacology: The pathophysiologic basis of drug therapy*. Wolters Kluwer.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Hilliard, W., MacDonald, M. L., & Lee, K. H. (2020). Chromosome-scale scaffolds for the Chinese hamster reference genome assembly to facilitate the study of the CHO epigenome. *Biotechnology and Bioengineering*, *117*(8), 2331–2339. https://doi.org/10.1002/bit.27432

Hiranniramol, K., Chen, Y., & Wang, X. (2020). CRISPR/Cas9 Guide RNA Design Rules for Predicting Activity. In M. Sioud (Ed.), *Methods in molecular biology: RNA interference and CRISPR technologies* (Vol. 2115, pp. 351–364). Humana. https://doi.org/10.1007/978-1-0716-0290-4_19

Joung, J., Konermann, S., Gootenberg, J. S., Abudayyeh, O. O., Platt, R. J., Brigham, M. D., Sanjana, N. E., & Zhang, F. (2017). Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nature Protocols*, *12*(4), 828–863. https://doi.org/10.1038/nprot.2017.016

Kaas, C. S., Kristensen, C., Betenbaugh, M. J., & Andersen, M. R. (2015). Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy. *BMC Genomics*, *16*(160), 160. https://doi.org/10.1186/s12864-015-1391-x

Khawli, L. A., Goswami, S., Hutchinson, R., Kwong, Z. W., Yang, J., Wang, X., Yao, Z., Sreedhara, A., Cano, T., Tesar, D., Nijem, I., Allison, D. E., Wong, P. Y., Kao, Y. H., Quan, C., Joshi, A., Harris, R. J., & Motchnik, P. (2010). Charge variants in IgG1: Isolation, characterization, in vitro binding properties and pharmacokinetics in rats. *mAbs*, *2*(6), 613–624. https://doi.org/10.4161/mabs.2.6.13333

Koike-Yusa, H., Li, Y., Tan, E. -P., Velasco-Herrera, M. D., & Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, *32*(3), 267–273. https://doi.org/10.1038/nbt.2800

Kosugi, S., Hirakawa, H., & Tabata, S. (2015). GMcloser: Closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics*, *31*(23), 3733–3741. https://doi.org/10.1093/bioinformatics/btv465

Lewis, N. E., Liu, X., Li, Y., Nagarajan, H., Yerganian, G., O'Brien, E., Bordbar, A., Roth, A. M., Rosenbloom, J., Bian, C., Xie, M., Chen, W., Li, N., Baycin-Hizal, D., Latif, H., Forster, J., Betenbaugh, M. J., Famili, I., Xu, X., … Palsson, B. O. (2013). Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. *Nature Biotechnology*, *31*(8), 759–765. https://doi.org/10.1038/nbt.2624

Li, H., Yang, Y., Hong, W., Huang, M., Wu, M., & Zhao, X. (2020). Applications of genome editing technology in the targeted therapy of human diseases: Mechanisms, advances and prospects. *Signal Transduction and Targeted Therapy*, *5*(1), 1. https://doi.org/10.1038/s41392-019-0089-y

Li, L., Stoeckert, Jr., C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., & Liu, X. S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology*, *15*(554), 554. https://doi.org/10.1186/s13059-014-0554-4

Lin, N., Davis, A., Bahr, S., Borgschulte, T., Achtien, K., & Kayser, K. (2011). Profiling highly conserved mirorna expression in recombinant IgG-producing and parental Chinese hamster ovary cells. *Biotechnology Progress*, *27*(4), 1163–1171. https://doi.org/10.1002/btpr.556

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: An empirically imporved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Mascarenhas, J. X., Korokhov, N., Burger, L., Kassim, A., Tuter, J., Miller, D., Borgschulte, T., George, H. J., Chang, A., Pintel, D. J., Onions, D., & Kayser, K. J. (2017). Genetic engineering of CHO cells for viral resistance to minute virus of mice. *Biotechnology and Bioengineering*, *114*(3), 576–588. https://doi.org/10.1002/bit.26186

Metzakopian, E., Strong, A., Iyer, V., Hodgkins, A., Tzelepis, K., Antunes, L., Friedrich, M. J., Kang, Q., Davidson, T., Lamberth, J., Hoffmann, C., Davis, G. D., Vassiliou, G. S., Skarnes, W. C., & Bradley, A. (2017). Enhancing the genome editing toolbox: Genome wide CRISPR arrayed libraries. *Scientific Reports*, *7*(2244), 2244. https://doi.org/10.1038/s41598-017-01766-5

Neutsch, L., Kroll, P., Brunner, M., Pansy, A., Kovar, M., Herwig, C., & Klein, T. (2018). Media photo-degradation in pharmaceutical biotechnology-impact of ambient light on media quality, cell physiology and IgG productin in CHO cultures. *Journal of Chemical Technology and Biotechnology*, *93*(8), 2141–2151. https://doi.org/10.1002/jctb.5643

Peng, J., Zhou, Y., Zhu, S., & Wei, W. (2015). High-throughput screens in mammalian cells using the CRISPR-Cas9 system. *The FEBS Journal*, *282*(11), 2089–2096. https://doi.org/10.1111/febs.13251

Rupp, O., MacDonald, M. L., Li, S., Dhiman, H., Polson, S., Griep, S., Heffner, K., Hernandez, I., Brinkrolf, K., Jadhav, V., Samoudi, M., Hao, H., Kingham, B., Goesmann, A., Betenbaugh, M. J., Lewis, N. E., Borth, N., & Lee, K. H. (2018). A reference genome of the Chinese hamster based on a hybrid assembly strategy. *Biotechnology and Bioengineering*, *115*(8), 2087–2100. https://doi.org/10.1002/bit.26722

Sealover, N. R., David, A. M., Brooks, J. K., George, H. J., Kayser, K. J., & Lin, N. (2013). Engineering Chinese hamster ovary (CHO) cells for producing recombinant proteins with simple glycoforms by zinc-finger nuclease (ZFN)-mediated geneknockout of mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase (Mgat1). *Journal of Biotechnology*, *167*(1), 24–32. https://doi.org/10.1016/j.jbiotec.2013.06.006

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., & Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, *343*(6166), 84–87. https://doi.org/10.1126/science.1247005

Song, L., & Florea, L. (2015). Rcorrector: Efficient and accurate error correction for illumina RNA-seq reads. *GigaScience*, *4*(48), 48. https://doi.org/10.1186/s13742-015-0089-y

Stanke, M., Schoffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *9*(7), 62. https://doi.org/10.1186/1471-2105-7-62

Stolfa, G., Smonskey, M. T., Boniface, R., Hachmann, A.-B., Gulde, P., Joshi, A. D., Pierce, A. P., Jacobia, S. J., & Campbell, A. (2018). CHO-omics review: The impact of current and emerging technologies on Chinese hamster ovary based bioproduction. *Biotechnology Journal*, *13*(3), 1700227. https://doi.org/10.1002/biot.201700227

Tihanyi, B., & Nyitray, L. (2021). Recent advances in CHO cell line development for recombinant protein production. *Drug Discovery Today: Technologies*, *38*, 25–34. https://doi.org/10.1016/j.ddtec.2021.02.003

Vcelar, S., Jadhav, V., Melcher, M., Auer, N., Hrdina, A., Sagmeister, R., Heffner, K., Puklowski, A., Betenbaugh, M., Wenger, T., Leisch, F., Baumann, M., & Borth, N. (2018). Karyotype variation of CHO host cell lines over time in culture characterized by chromosome counting and chromosome painting. *Biotechnology and Bioengineering*, *115*(1), 165–173. https://doi.org/10.1002/bit.26453

Walsh, G. (2006). Biopharmaceutical benchmarks 2006. *Nature Biotechnology*, *24*(7), 769–776. https://doi.org/10.1038/nbt0706-769

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., & Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, *350*(6264), 1096–1101. https://doi.org/10.1126/science.aac7041

Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, *343*(6166), 80–84. https://doi.org/10.1126/science.1246981

Xiong, K., Karottki, K., Hefzi, H., Li, S., Grav, L. M., Li, S.,Spahn, P., Lee, J. S., Ventina, I., Lee, G. M., Lewis, N. E., Kildegaard, H. F., & Pedersen, L. E. (2021). An optimized genome-wide, virus-free CRISPR screen for mammalian cells. *Cell Reports Methods*, *1*(4), 100062. https://doi.org/10.1016/j.crmeth.2021.100062

Xu, X., Nagarajan, H., Lewis, N. E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., Andersen, M. R., Neff, N., Passarelli, B., Koh, W., Fan, H. C., Wang, J., Gui, Y., Lee, K. H., Betenbaugh, M. J., … Wang, J. (2011). The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nature Biotechnology*, *29*, 735–741. https://doi.org/10.1038/nbt.1932

Yehuda, S., & Padler-Karavani, V. (2020). Glycosylated biotherapeutics: Immunological effects of N-glycolylneuraminic acid. *Frontiers in Immunology*, *11*(21), 21. https://doi.org/10.3389/fimmu.2020.00021

Yin, J., Ren, W., Huang, X., Deng, J., Li, T., & Yin, Y. (2018). Potential mechanisms connecting purine metabolism and cancer therapy. *Frontiers in Immunology*, *9*(1697), 1697. https://doi.org/10.3389/fimmu.2018.01697

Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., & Wei, W. (2014). High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, *509*(7501), 487–491. https://doi.org/10.1038/nature13166

Zhu, S., Li, W., Liu, J., Chen, C. -H., Liao, Q., Xu, P., Xu, H., Xiao, T., Cao, Z., Peng, J., Yuan, P., Brown, M., Liu, X. S., & Wei, W. (2016). Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nature Biotechnology*, *34*(12), 1279–1286. https://doi.org/10.1038/nbt.3715

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kretzmer, C., Narasimhan, R. L., Lal, R. D., Balassi, V., Ravellette, J., Kotekar Manjunath, A. K., Koshy, J. J., Viano, M., Torre, S., Zanda, V. M., Kumravat, M., Saldanha, K. M. R., Chandranpillai, H., Nihad, I., Zhong, F., Sun, Y., Gustin, J., Borgschulte, T., Liu, J., & Razafsky, D. (2022). De novo assembly and annotation of the CHOZN® GS$^{-/-}$ genome supports high-throughput genome-scale screening. *Biotechnology and Bioengineering*, 119, 3632–3646. https://doi.org/10.1002/bit.28226