Supplement to: The genomic landscapes of CHO cell lines as revealed by the *Cricetulus griseus* draft genome

Table of Contents

SUPPLEMENTARY NOTES

FILTERING OF RAW READS PRIOR TO ASSEMBLY: ................................. 2
ESTIMATION OF GENOME SIZE ............................................................... 2
REPEAT FEATURES IN THE CHINESE HAMSTER GENOME .................... 2
ENDOGENOUS RETROVIRAL SEQUENCES IN THE *C. GRISEUS* GENOME .... 3
PROPERTIES OF SCAFFOLD CHROMOSOME ASSIGNMENT .................... 3
DETAILS OF THE COMPARISON BETWEEN *C. GRISEUS* AND CHO-K1 ........ 4
CHROMOSOMAL LOCALIZATION OF GLYCOSYLATION ENZYMES IN *C. GRISEUS* AND CHO-K1 ... 4
COPY NUMBER VARIATIONS IN ANTI-APOPTOTIC GENES ......................... 4
MUTATIONS IN SUGAR NUCLEOTIDE SYNTHESIS AND GLYCOSYLATION .......... 5
VIRAL SUSCEPTIBILITY ........................................................................... 6

SUPPLEMENTARY FIGURES .................................................................... 7

SUPPLEMENTARY FIGURE 1 ........................................................................ 7
SUPPLEMENTARY FIGURE 2 ........................................................................ 8
SUPPLEMENTARY FIGURE 3 ........................................................................ 9
SUPPLEMENTARY FIGURE 4 ....................................................................... 10
SUPPLEMENTARY FIGURE 5 ....................................................................... 11
SUPPLEMENTARY FIGURE 6 ....................................................................... 11
SUPPLEMENTARY FIGURE 7 ....................................................................... 12
SUPPLEMENTARY FIGURE 8 ....................................................................... 12
SUPPLEMENTARY FIGURE 9 ....................................................................... 13

REFERENCES ......................................................................................... 13
Supplementary Notes

Filtering of raw reads prior to assembly:
We filtered the Illumina reads based on following criteria:

- Reads were filtered when more than 10% of the bases were degenerate (N) or poly-A’s.
- Reads were filtered if they were from large insert size libraries (2 kb, 5 kb, 10 kb and 20 kb) with 15 or more bases having phred quality score (given by Illumina sequencer) less than or equal to 7, or from short insert size libraries with 50 bases having quality score less than or equal to 7.
- Reads were filtered if more than 10 bp aligned to the adapter sequence.
- If read pairs in which read 1 and read 2 overlapped more than 10 bp (allowing 10% mismatch), they were filtered.
- Reads were filtered if they were PCR duplicates (i.e., two completely identical reads).

Using these criteria, 347.5 Gb of raw data was filtered to 240.5 Gb representing ~90X coverage of the hamster genome.

Estimation of genome size
Through k-mer estimation, we found the genome size to be 2.7 Gb (Supplementary Fig. 1), which is smaller than previous estimates using Feulgen densitometry (3.05 Gb ~ 3.99 Gb)1,2, but slightly larger than the k-mer estimate of Chinese hamster ovary (CHO)-K1 cell line, which was 2.6 Gb3 (Supplementary Fig. 1).

Repeat features in the Chinese hamster genome
Repeat elements in the genome were predicted using several methods. Tandem repeats were first predicted using RepeatMasker4 and TRF5. Then transposable elements (TEs) were identified using a combination of homology-based and de novo methods. For the homology-based method, we used databases of known repetitive sequences in Repbase6 and searched against the hamster draft genome using RepeatMasker. For the de novo method, we used three software packages, including LTR_FINDER7 (Version 1.0.3), PILER8 and RepeatScout9 (Version 1.05), to build a de novo repeat database of the Chinese hamster. We then used RepeatMasker (Version 3.2.7) to identify repeats and used RepeatProteinMask (http://www.repeatmasker.org/, Version 3.2.7) to search the protein database in Repbase against the genome to identify repeat-related proteins. Finally we combined the de novo prediction and the homolog prediction of TEs according to the position in the genome. We estimated that repeat elements account for 42.8% of the genome (Supplementary Table 6). Specifically, ~41% of the genome consists of transposable elements (TEs) (Supplementary Table 7), which is similar to that of the mouse (37.5%)10 and rat (40%)11. The most abundant TE, long interspersed repeated DNA (LINE), accounts for 27% of the genome. The distribution of sequence divergence rate for LINE elements was bimodal suggesting that two independent burst events of LINEs may have occurred within the hamster genome (Supplementary Fig. 4). Since other TEs, such as long terminal repeats (LTRs) and short interspersed repeated DNAs (SINEs), only have one peak, these might not have experienced the recent burst events as seen with LINEs. We further analyzed the detailed categories of different
TEs (Supplementary Table 8) and compared these to mouse, rat and human. The composition of TEs in hamster genome is similar to those of rat and mouse.

**Endogenous retroviral sequences in the *C. griseus* genome**

Consistent with other mammalian genomes, the *C. griseus* genome has many stretches of DNA with homology to viral genes. This is of particular interest in *C. griseus* since CHO cell lines have shown substantial resistance to many human viruses 12, and they do not seem to produce infectious retroviruses as seen in other rodent cell lines 13,14,15. However, numerous reports have observed viral particles budding off of CHO cells 16,14,15,17,18,19. Since these viral particles likely represent endogenous viral proteins as opposed to new viral infections, it would be of interest to assess the origin of these viral particles. Here we provide a preliminary identification of endogenous viral elements in the Chinese hamster genome.

Two classes of viral particles have been observed in CHO cell lines. Type A particles are immature intracellular particles that are derived from endogenous retroviral-like genes. Genes coding these particles often lack a functional env gene and therefore are unable to infect other cells, but instead behave like retrotransposons and readily spread through the host genome 20. Type A particles have been previously identified and their associated genes have been sequenced in Syrian hamster and mouse. These sequences were used to identify similar RNAs in a CHO-K1 derivative cell line 21,22. However, none of the identified sequences could encode functional proteins since they all contained premature stop codons or frameshift mutations. Similarly, budding type C particles have also been observed. When RNA was isolated and sequenced from these particles, it was found that these are also unable to encode functional proteins due to numerous mutations 23. Thus it is still unclear where these particles are encoded in the Chinese hamster and its derivative cell lines.

The existence of C-type particles suggests that full-length coding retroviral proteins should still exist. Thus, to gain a preliminary view of the landscape of endogenous retroviral elements, we compiled a list of 115 sequenced retroviral genomes and searched for genes that represent potentially intact viral proteins encoded by the hamster genome. This was done by conducting a blastp query of all retroviral proteins against all putative protein-coding ORFs in the hamster. With an E-value cutoff of 1x10^-6 we found 403 proteins in the hamster genome with homology to 174 proteins in the 115 retroviral genomes. These 174 viral proteins contained the domains common to retroviruses (Supplementary Fig. 5). Furthermore, we found that 80% of the 403 hamster proteins were at least as long as their retroviral homologs and 40% of their mRNAs were expressed, thereby suggesting that many these may be still be synthesizing retroviral components. Protein sequences for all of the ORFs with homology to retroviral proteins are provided in Supplementary Table 9.

**Properties of scaffold chromosome assignment**

To assign scaffolds to their respective chromosomes, our optical mapping data were used in conjunction with published BAC end sequencing and fluorescence in situ hybridization 24,25. Specifically, chromosomal assignments were obtained for each BAC, and then blastn was used to find scaffolds with the highest homology to BAC end sequences (E-value < 1x10^-5). Once chromosomal assignments based on the BAC library were obtained for scaffolds
(Supplementary Table 3), the chromosomal localization was inferred for all scaffolds that were joined to previously assigned scaffold by optical mapping (Supplementary Table 4). From this analysis, we were able reliably localize 26% of the genomic sequence to hamster chromosomes (Supplementary Fig. 6). Hamster scaffolds were aligned to the Mus musculus genome to assess the extent to which the species had diverged (Supplementary Fig. 7).

**Details of the comparison between C. griseus and CHO-K1**

We compared the hamster and CHO-K1 genome sequences and identified 25,711 structure variations including: 13,735 insertions and 11,976 deletions in CHO-K1, relative to the hamster. Most of these variations are shorter than 100bp (Supplementary Fig. 8). These variations were observed in 2,188 genes in the hamster genome and were enriched in genes with binding function, which might have changed during the cell line formation (Supplementary Table 12).

A comparison of the gene content of the two genomes showed that 99% of the hamster genes (24,021) had homologs in CHO-K1 with over 60% identity and 60% coverage. Similarly, 24,109 of the 24,383 (99%) CHO-K1 genes had homologs in the hamster genome with over 60% identity and 60% coverage.

**Chromosomal localization of glycosylation enzymes in C. griseus and CHO-K1**

One beneficial trait of Chinese hamster ovarian cell lines is their ability to glycosylate enzymes in a fashion that is compatible to humans. Thus, since one may desire to genetically modify CHO cell lines in order to improve product titer, it is desirable to know where the glycosylation enzymes reside, so that care can be taken to avoid unintended disruption of the glycosylation pathways. For each glycosylation enzyme (Supplementary Table 26), we identified the scaffold upon which it is located, allowing one to design primers to verify that desired glycosylation enzymes are not disturbed in cell engineering. Furthermore, using the BAC and optical mapping data, we associated 26% of the genomic sequence data to specific chromosomes. This allowed us to assign 49 glycosylation enzymes to specific hamster chromosomes. We note that the distribution of these glycosylation enzymes across all chromosomes roughly reflects the distribution of genes in general. Indeed, they were not significantly enriched or depleted from any one chromosome (p > 0.1; Supplementary Fig. 9).

**Copy number variations in anti-apoptotic genes**

Many proteins balance the pro-apoptotic activities, and when over-expressed, these proteins can inhibit apoptosis in CHO cells\(^{26,27,28,29}\). Different cell lines may naturally be less apoptotic, since some anti-apoptotic genes were duplicated in individual cell lines. For example, IAP genes inhibit caspases \(^{30}\), and we found one, BIRC7, is duplicated in all cell lines.

While several anti-apoptotic genes have increased copy numbers in CHO cell lines, we found that an upstream factor, phosphoinositide 3-kinase (PI3K) showed cell-type specific responses. The anti-apoptotic PI3K genes encode catalytic (PIK3C) and regulatory (PIK3R) subunits. Together, these proteins relay survival signals to proteins such as protein kinase B (Akt) \(^{31}\). We detected few
sequence variants in PI3K, but we found that CHO-S had a duplication of the catalytic subunit, and CHO-K1 had a deletion of the regulatory subunit. These differences in copy number may influence apoptotic activity in the cell lines with CNVs.

Mutations in sugar nucleotide synthesis and glycosylation
Protein glycosylation significantly influences biotherapeutic quality. Glycosylation can vary substantially between different parent organisms and cell lines, and is also influenced by mutations, culture conditions, and enzyme expression levels. Although there are fewer SNPs in glycosylation than expected by chance (mean p = 0.018), there are still many mutations. To gain a more detailed view of mutations in glycosylation, we conducted a bidirectional blastp between the human genome and CHO-K1 glycosylation genes. Among 256 enzymes associated with glycosylation in the hamster, 13%, 2% and 25% have non-synonymous SNPs, frame-shifting indels, or CNVs, respectively, in at least one cell line (Supplementary Fig. 3.a, Supplementary Tables 23-24, 26).

Sugar nucleotides are the building blocks for glycans. Transcripts for most sugar nucleotide synthesis enzymes are detected in the hamster and/or CHO-K1. However, between cell lines, there may be variations in sugar nucleotide abundance since most synthesis pathways have a mutation or CNV (Supplementary Fig. 3.b). Next in glycosylation, these sugars are sequentially added to growing oligosaccharide chains. To study these pathways, we determined human glycosylation reactions associated with the bidirectional hits using the human metabolic network, Recon. The reactions were then cross-referenced with glycosylation reactions required for producing N-glycans and O-glycans found on common IgG. We found that mutations and CNVs occur at various locations in the pathways synthesizing different glycans, including N- and O-glycans (Supplementary Fig. 3.b-d). While not all SNPs will have a measurable effect on enzyme activity or substrate preference, these glycans are produced through long, branching pathways. Thus, there is a considerable increase in the probability of having a mutation that changes the final glycoform in any given cell line. Furthermore, a few cell lines have CNVs in the FUCA1 fucosidase gene, α(2,3)-sialyltransferase (ST3GAL1), and the 3'-phosphoadenosine-5'-phosphosulfate transporter SLC35B3. These genes have important functions in removing fucose, adding sialic acid, and modulating sulfation, respectively. The processes to which these genes contribute also have been shown to affect the activity and longevity of recombinant proteins. The heterogeneity in glycosylation enzyme mutations across different cell lines could have dramatic effects on the protein glycoforms produced in different cell lines, yielding differences in in vivo product activity and clearance rates of recombinant therapeutic proteins. Thus, it will be important to assess the mutations, copy numbers and expression levels of these and other enzymes in production cell lines in order to optimize the modifications added to recombinant proteins.
**Viral susceptibility**

The resistance of CHO cell lines to viral infection is a beneficial trait from a bioprocessing perspective\(^\text{12}\); thus, we examined the hamster and cell line genomes for mutations or changes in gene expression that may influence viral susceptibility (Supplementary Table 27). Many viral susceptibility genes were previously shown in CHO-K1 to be not expressed, and many of these were key viral entry receptors \(^\text{3}\). We found that 70% of the viral susceptibility genes were expressed in the hamster transcriptome, which is more than expected by chance \((p = 3 \times 10^{-39})\), while only 55% of the viral susceptibility genes were expressed in CHO-K1, despite having more expressed transcripts across the entire CHO-K1 genome\(^\text{3}\). Non-expressed viral susceptibility genes in CHO-K1 were enriched in non-synonymous SNPs \((p = 0.02)\). Interestingly, 28 of these 33 SNPs in viral susceptibility genes were shared across all cell lines. These included plasma membrane receptors and viral entry receptors, such as integrins (ITAV, ITAX, ITAM), CD4, and CD86. While the hamster expresses many viral entry receptors, it seems that CHO cell lines have managed to down-regulate or mutate many viral susceptibility genes, thereby contributing to this favorable bioprocessing trait for CHO cell lines.
Supplementary Figures

**Supplementary Figure 1**

*Genome size of Chinese hamster and CHO-K1 cell line estimated by k-mer analysis.* The x-axis is depth (X); the y-axis represents the frequency at that depth. Without consideration of the sequence error rate, heterozygosis rate and repeat rate of the genome, the 17-mer of distribution should obey the Poisson theoretical distribution. From the actual data, due to the sequence error, the low depth of K-mer frequency will take up a large proportion. At the same time, for some specific genome, the certain heterozygosis rate can cause a sub peak at the position of the half of the main peak, while a certain repeat rate can cause a repeat peak at the position of the integer multiples of the main peak. The blue trace represents the 17-mer distribution for the Chinese hamster and the red trace for CHO-K1. For Chinese hamster genome, we estimated the genome to be 2.7Gb, while using the same amount of data (~50X), the CHO K1 genome size was estimated to be 2.6Gb.
Supplementary Figure 2

Predicted genes supported by different evidences. Venn diagram shows unique and shared gene number among different annotation method. Homolog support includes genes annotated by homolog method of CHO-K1 cell line. De novo support includes genes predicted by AUGUSTUS, GlimmerHMM and Genscan. RNA-seq support includes genes predicted by transcriptome data.
Supplementary Figure 3
Differences in mutations and copy number variations (CNVs) in cell lines may influence the glycoforms of recombinant proteins. 256 unique enzymes associated with glycosylation were identified in the C. griseus genome. (a) Many of these enzymes have one or more mutation or CNV in at least one cell line. These variations are associated with many aspects of glycosylation, such as (b) sugar nucleotide synthesis, (c) O-linked glycosylation, and (d) N-linked glycosylation. The combined effect of differential gene expression (two left columns of
grey/orange boxes), and differences in sequence and copy number variations between cell lines (two right columns of grey/orange boxes) can clearly influence the synthesis of O- and N-glycans that have been detected on native and recombinant IgG glycoforms (blue nodes). Numbers on the orange boxes refer to the number of cell lines with a SNP or CNV in that gene.

**Supplementary Figure 4**

**Divergence of different TE categories within the genome.** The divergence rate was calculated between the identified TE elements in the genome and the consensus sequence in the TE library used (Repbase or RepeatModeler).
Supplementary Figure 5

Number of hamster proteins showing homology to retroviral proteins with common retroviral protein domains. There are many hamster proteins that are homologous to retroviral gag and kinase proteins. Furthermore, transcripts for many of these were detected with RNA-seq. However, the low abundance of env proteins is consistent with the observation that endogenous retroviral elements lacking env genes tend to spread much more frequently.

Supplementary figure 6

Amount of sequence data associated with the 11 chromosomes of the female Chinese hamster. BACs were used to associate scaffolds with specific chromosomes, and in total 26% of sequenced genome could be associated with a specific chromosome.
**Supplementary Figure 7**

**Distribution of mouse chromosome with homology to Chinese hamster scaffolds.** Scaffolds associated with each hamster chromosome were aligned to the *mus musculus* genome to assess the extent to which the genomes have diverged. While each hamster chromosome demonstrated considerable rearrangement, similarities were seen between several hamster chromosomes and mouse chromosomes, such as hamster chromosomes 6, 7, 8, 10, and X, which showed considerable homology to mouse chromosomes 2, 11, 6, 15, and X.

**Supplementary Figure 8**

**Length distribution of Structural Variants (SVs).** The distribution of SVs frequency for different lengths. Most of these variations are shorter than 100bp.
Supplementary Figure 9
Distribution of genes among the 11 chromosomes in the hamster genome. BACs and optical mapping data were used to associate scaffolds with specific chromosomes, accounting for 26% of the genomic sequence. All genes in these scaffolds were identified and their distribution is shown here. We note that BAC coverage of chromosome 9 was considerably low and no gene-containing regions were found in the regions targeted. The distribution of glycosylation enzymes, mirrored that of all genes.

REFERENCES