nature biotechnology

OPEN

Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome

Nathan E Lewis[1,14], Xin Liu[2,3,14], Yuxiang Li[2,14], Harish Nagarajan[1,14], George Yerganian[4,5], Edward O'Brien[1], Aarash Bordbar[1], Anne M Roth[6,13], Jeffrey Rosenbloom[6,13], Chao Bian[2], Min Xie[2], Wenbin Chen[2], Ning Li[2,3,7], Deniz Baycin-Hizal[8], Haythem Latif[1], Jochen Forster[9], Michael J Betenbaugh[8,9], Iman Famili[6,13], Xun Xu[2,3], Jun Wang[2,10–12] & Bernhard O Palsson[1,9]

Chinese hamster ovary (CHO) cells, first isolated in 1957, are the preferred production host for many therapeutic proteins. Although genetic heterogeneity among CHO cell lines has been well documented, a systematic, nucleotide-resolution characterization of their genotypic differences has been stymied by the lack of a unifying genomic resource for CHO cells. Here we report a 2.4-Gb draft genome sequence of a female Chinese hamster, *Cricetulus griseus*, harboring 24,044 genes. We also resequenced and analyzed the genomes of six CHO cell lines from the CHO-K1, DG44 and CHO-S lineages. This analysis identified hamster genes missing in different CHO cell lines, and detected >3.7 million single-nucleotide polymorphisms (SNPs), 551,240 indels and 7,063 copy number variations. Many mutations are located in genes with functions relevant to bioprocessing, such as apoptosis. The details of this genetic diversity highlight the value of the hamster genome as the reference upon which CHO cells can be studied and engineered for protein production.

Recombinant therapeutic proteins are increasingly important to the pharmaceutical industry. Global spending on biologics, such as antibodies, hormones and blood factors, reached $138 billion dollars in 2010 (ref. 1). CHO cell lines are the preferred host expression system for many therapeutic proteins[2], and the cells have been repeatedly approved by regulatory agencies. Moreover, they can be easily cultured in suspension and can produce high titers of human-compatible therapeutic proteins[3].

Most improvements in CHO-based recombinant protein titer and quality have been achieved by random cell-line mutagenesis and media optimization[4]. Meanwhile, efforts to engineer mouse cells have greatly benefited from numerous genomic tools and technologies, owing in large part to the availability of the *Mus musculus* reference genome sequence. Genomic resources are also becoming available for CHO cells, such as the CHO-K1 genome[5], expressed sequence tag[6,7] and bacterial artificial chromosome (BAC) libraries[8], and compendia of proteomic[9–11] and transcriptomic data[7,12–16]. However, much like how murine cell line data are routinely studied in the context of the *Mus musculus* reference genome, there is a need for a standard reference for all CHO cell lines to contextualize all of these valuable genomic resources.

Many recombinant protein–producing CHO cell lines were derived from the CHO-K1, CHO-S and DG44 lineages. Each has undergone extensive mutagenesis and clonal selection[17]. Hence, a standard reference genome that is representative of the genomic sequence of all native CHO genes and regulatory elements would be advantageous for the successful implementation of genomic resources in CHO-based bioprocessing[4,17]. To address this need, we present a draft genome sequence of the *C. griseus* (Chinese hamster) colony from which the CHO cell lines have been derived. This reference sequence is used to analyze the genomic composition and mutational diversity among seven CHO cell lines, and to study how sequence variations may affect cellular processes that are of bioprocessing relevance. The *C. griseus* genome may serve along with the previously published CHO-K1 genome as primary reference resources in future analyses of omics data sets derived from CHO cells. This will also aid in bioprocessing systems analysis and in cell line engineering studies.

RESULTS

Genome assembly

Female Chinese hamster DNA was acquired from various tissues and sequenced using the Illumina HiSeq 2000 platform, yielding 347.5 Gb of raw data (**Supplementary Tables 1** and **2**). Using SOAPdenovo, we assembled 2.4 Gb of the genome with a contig N50 (the shortest length of sequence contributing more than half of assembled sequences) of 26.5 kb and scaffold N50 of 1.54 Mb (**Table 1**). The genome was

[1]CHOmics, Inc., San Diego, California, USA. [2]BGI-Shenzhen, Shenzhen, People's Republic of China. [3]BGI Europe, BGI-Shenzhen, Copenhagen Bio Science Park, Copenhagen, Denmark. [4]Cytogen Research and Development, Inc., West Roxbury, Massachusetts, USA. [5]Foster Biomedical Research Laboratory, Brandeis University, Waltham, Massachusetts, USA. [6]GT Life Sciences, San Diego, California, USA. [7]BGI Europe Institute, BGI-Shenzhen, Copenhagen Bio Science Park, Copenhagen, Denmark. [8]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [9]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark. [10]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [11]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [12]King Abdulaziz University, Jeddah, Saudi Arabia. [13]Present addresses: Life Technologies, Carlsbad, California, USA (A.M.R.), and Cell Engineering Unit, Intrexon Corporation, San Diego, California, USA (J.R. and I.F.). [14]These authors contributed equally to this work. Correspondence should be addressed to B.O.P. (bpals@biosustain.dtu.dk) or J.W. (wangj@genomics.org.cn).

Received 7 August 2012; accepted 3 June 2013; published online 21 July 2013; doi:10.1038/nbt.2624

Table 1 Assembly statistics

	Contig		Scaffold		Super-scaffolds	
	Size (bp)	Number	Size (bp)	Number	Size (bp)	Number
N90	6,390	91,476	346,540	1,637	443,523	1,091
N80	11,724	65,207	656,362	1,156	939,760	723
N70	16,531	48,549	950,835	853	1,417,091	519
N60	21,461	36,190	1,249,430	634	1,994,221	378
N50	26,761	26,456	1,544,832	461	2,491,721	271
Longest (bp)	219,443	–	8,324,132	–	10,797,402	–
Total size (bp)	2,332,459,831	–	2,393,115,851	–	2,400,585,184	–
Total number (≥100 bp)	–	458,620	–	287,210	–	286,619
Total number (≥2 kb)	–	128,107	–	6,947	–	6,356

Nx contig (scaffold) size is the length of the smallest contig (scaffold) S in the sorted list of all contigs (scaffolds) where the cumulative length from the largest contig to contig S is at least x% of the total assembly length.

further assembled into super-scaffolds with optical mapping, yielding an N50 of 2.49 Mb. Ninety percent of the genome assembly was included in the 1,091 longest super-scaffolds (**Table 1**). The overall size of the hamster genome was estimated to be 2.7 Gb using the k-mer estimation method (**Supplementary Fig. 1**). Optical mapping data were further combined with published BAC-based fluorescence *in situ* hybridization data[8] to successfully associate 26% of the genome sequence data to specific hamster chromosomes (**Supplementary Tables 3** and **4**).

To assess the coverage of the hamster transcripts in the assembly, we sequenced mRNA from a pool of hamster tissues and assembled the transcriptome *de novo* into 98,116 contigs (Online Methods). Mapping RNA-seq contigs to the genome assembly demonstrated that >90% of the assembled transcripts could be associated with annotated genes (**Supplementary Table 5**).

Genome annotation

We annotated repeat features and identified endogenous retroviral elements (**Supplementary Notes** and **Supplementary Tables 6–9**). We next predicted genes using homology-based approaches, *de novo* gene prediction algorithms and transcriptome-based methods (**Supplementary Table 10** and **Supplementary Fig. 2**). The final gene set consisted of 24,044 genes in the hamster genome, which is similar to that of the CHO-K1 cell line[5]. Of these predicted genes, 23,473 clustered into 21,628 gene families (**Fig. 1a**), and 3,052 (14.1%) gene families contained more than one gene in the hamster. Only 20 gene families were unique to the hamster, when compared to the rat, mouse and CHO-K1 genomes (**Fig. 1b**). We functionally annotated 82% (19,775) of the predicted genes using InterPro, Swiss-Prot, TrEMBL, Gene Ontology (GO) and KEGG (**Supplementary Table 11**).

Comparison between hamster and CHO-K1 genomes

Mutations and structural variations are common in mammalian cell line genomes[17–19]. Although large chromosomal rearrangements have been shown in CHO cell lines previously[8], the extent of these changes at the sequence level remains unknown. Thus, we compared the structure and gene content of the Chinese hamster genome and the published genome of CHO-K1 cells from the American Type Culture Collection (ATCC)[5]. To facilitate this comparison, we aligned all large hamster and CHO-K1 scaffolds to the mouse chromosomes. Numerous chromosomal translocations have occurred through evolution since the mouse and hamster diverged (**Fig. 2a**). However, no large sections of the mouse chromosomes were missing in the hamster (**Fig. 2b**). On the other hand, CHO-K1 scaffolds failed to align to portions of mouse chromosomes 5, 7, 15 and 19 (**Fig. 2b**). Meanwhile, Illumina sequencing reads from CHO-K1 (ref. 5) aligned to the hamster scaffolds corresponding to these regions. This result suggests the possibility that these regions are in CHO-K1, albeit considerably mutated or rearranged. We next directly assessed the scope of mutations by comparing the CHO-K1 genome to the hamster genome. CHO-K1 contained 25,711 structure variations, including 13,735 insertions and 11,976 deletions (**Supplementary Notes** and **Supplementary Table 12**). Despite the large number of structural variations in CHO-K1, the set of annotated genes in the hamster and CHO-K1 were highly similar. Specifically, there was a 99% overlap in gene content between the two genomes, and an assessment of GOslim terms for these genes confirmed the similarity in gene content (**Fig. 2c**).

Variation between different CHO cell lines

Despite the similarity in gene content, numerous genomic variations were detected in CHO-K1 relative to the hamster. To elucidate

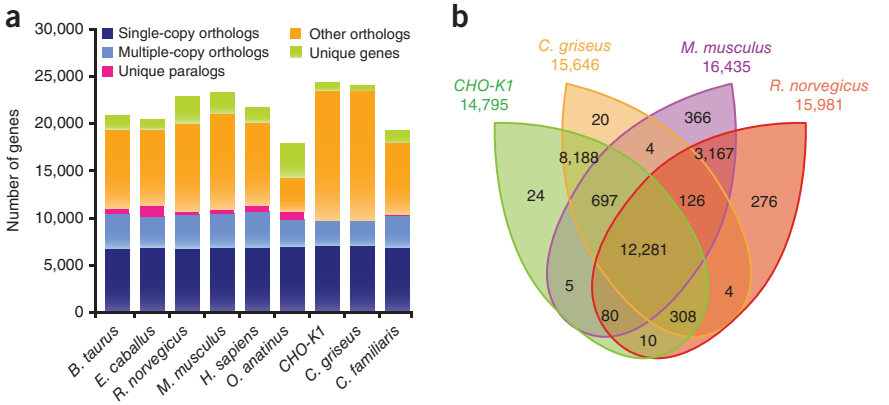

Figure 1 Gene families across *C. griseus* and several mammalian genomes. (**a**) The majority of mammalian genes are orthologous, with more than 5,000 preserved as single copies in each species (dark blue). A few thousand have species-specific duplications (light blue), whereas other orthologs were shared by only some of the nine mammals studied here (orange). A small fraction of genes were unique to just one species (green), and occasionally had paralogs in that one species (pink). (**b**) The overlap of orthologous gene clusters is shown among the CHO-K1, *C. griseus*, *M. musculus* and *Rattus norvegicus* genomes. ENSEMBL (v58) annotated genes were used for the CHO-K1, *M. musculus* and *R. norvegicus* genomes.

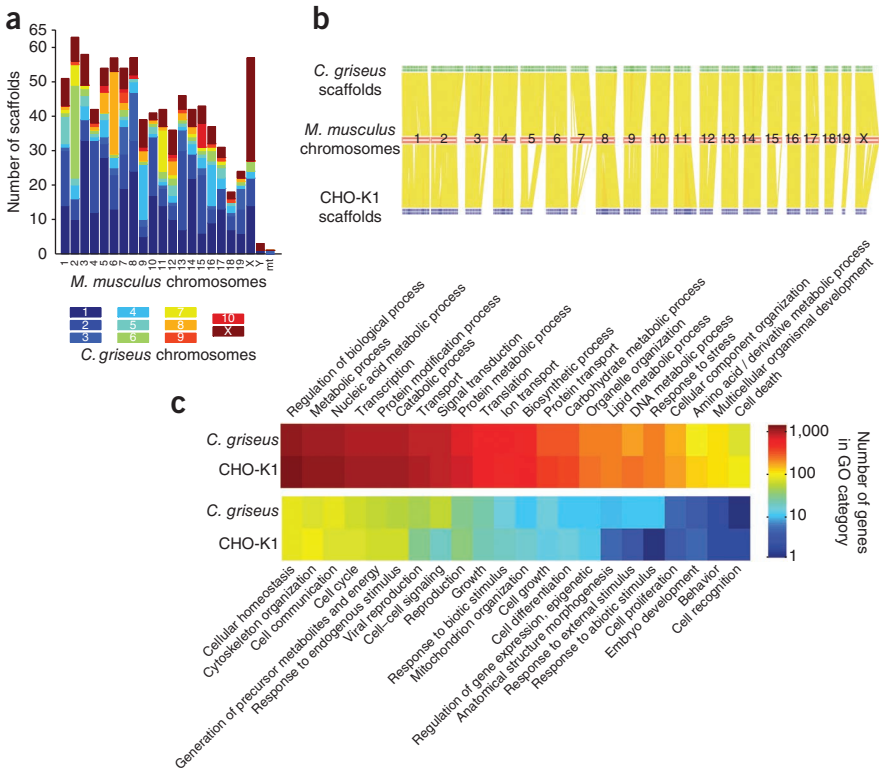

Figure 2 Genome comparison between mouse, Chinese hamster and CHO-K1. Conserved sequences among the mouse, CHO-K1 and *C. griseus* genomes were determined by aligning their scaffolds (larger than 1 Mb) to the mouse genome. (**a**) Assignment of *C. griseus* scaffolds to *M. musculus* chromosomes. The *C. griseus* scaffolds with chromosomal assignment (accounting for more than a quarter of the 2.4 Gb of genomic sequence) were compared to mouse chromosomes to assess the scale of chromosomal rearrangement. (**b**) Alignment of CHO-K1 and *C. griseus* genomes. Few large DNA stretches are missing in the hamster, whereas there are more regions to which CHO-K1 scaffolds could not align. (**c**) Gene annotation. The number of genes was determined for each "Biological Process" GO slim category in both the *C. griseus* and CHO-K1 genomes.

the extent of genomic heterogeneity across other cell lines, we sequenced six additional CHO cell lines (**Fig. 3a**) to >9× depth, covering ~95% of each genome. Including the previously sequenced CHO-K1 genome, the seven cell lines accounted for three different lineages and several different phenotypic features, for example, cells adapted to different media, suspension-grown cells and antibody-producing cells (**Supplementary Table 13**).

To initially validate our cell line resequencing data, we inspected the genotype related to an important phenotypic marker for CHO cell lines. Certain cell lines lack dihydrofolate reductase (DHFR) activity[20], and cannot grow without glycine, hypoxanthine and thymidine (GHT). However, when an exogenous DHFR gene is coupled to a gene encoding a desired protein product on the same plasmid, the GHT media and methotrexate can be used to select for clones that overproduce DHFR and the recombinant protein of interest. Among the cell lines sequenced here, only the DG44 cell line is known to carry the DHFR-negative phenotype[20]. Consistent with this characteristic,

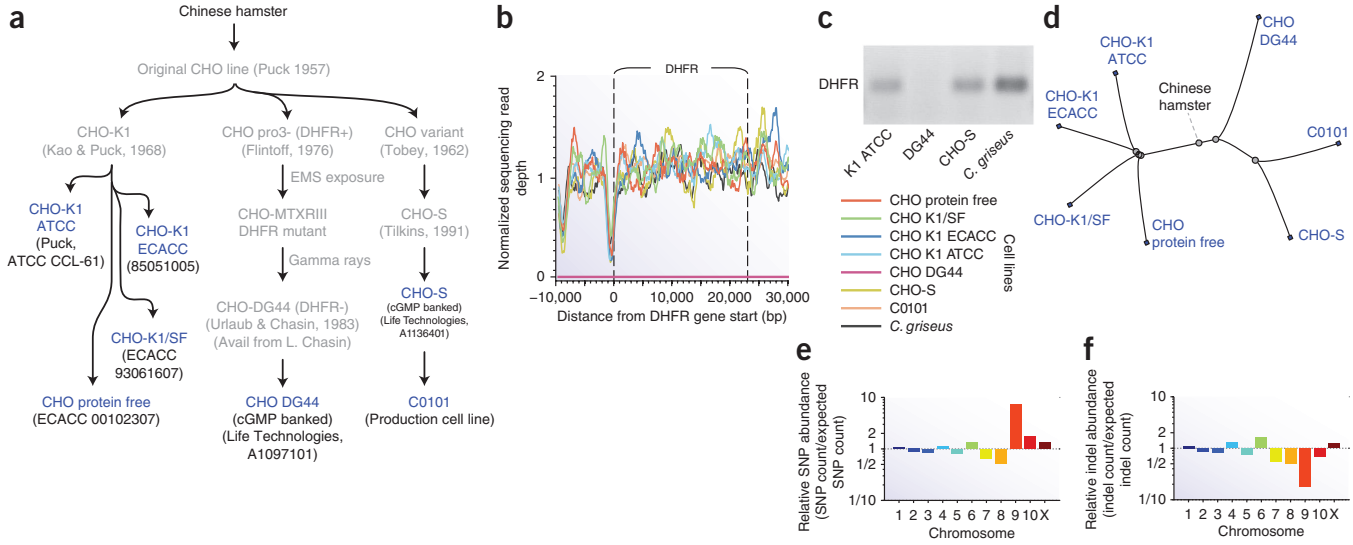

Figure 3 Mutation landscape of CHO cell lines. CHO cell lines have diverged over time due to numerous iterations of mutation, selection and clonal isolation. (**a**) The family tree of a few cell lines are shown here, with the sequenced lines highlighted in blue. Where known, the name of those who isolated the strain and the year it was done are given in parentheses. (**b**) Sequencing read depth (normalized by the average read depth for the cell line, and averaged over 100 bp bins) was assessed for the DHFR gene, a selectable marker for some CHO cell lines. The DHFR gene was clearly deleted in the DG44 cell line, as no DG44 reads aligned to this region and (**c**) no PCR product was obtained for the gene. Mutations were further analyzed on a genomic-wide scale. (**d**) A phylogenetic reconstruction based on the diversity of SNPs recapitulate the known historical divergence of these CHO cell lines from inferred ancestral cell lines (gray parent nodes). (**e**,**f**) Furthermore, the abundance of SNPs (**e**) and indels (**f**) varied between the hamster chromosomes, as determined using all scaffolds that could be assigned to specific chromosomes (~26% of the sequence data). ECACC, European Collection of Cell Cultures; ATCC, American Type Culture Collection.

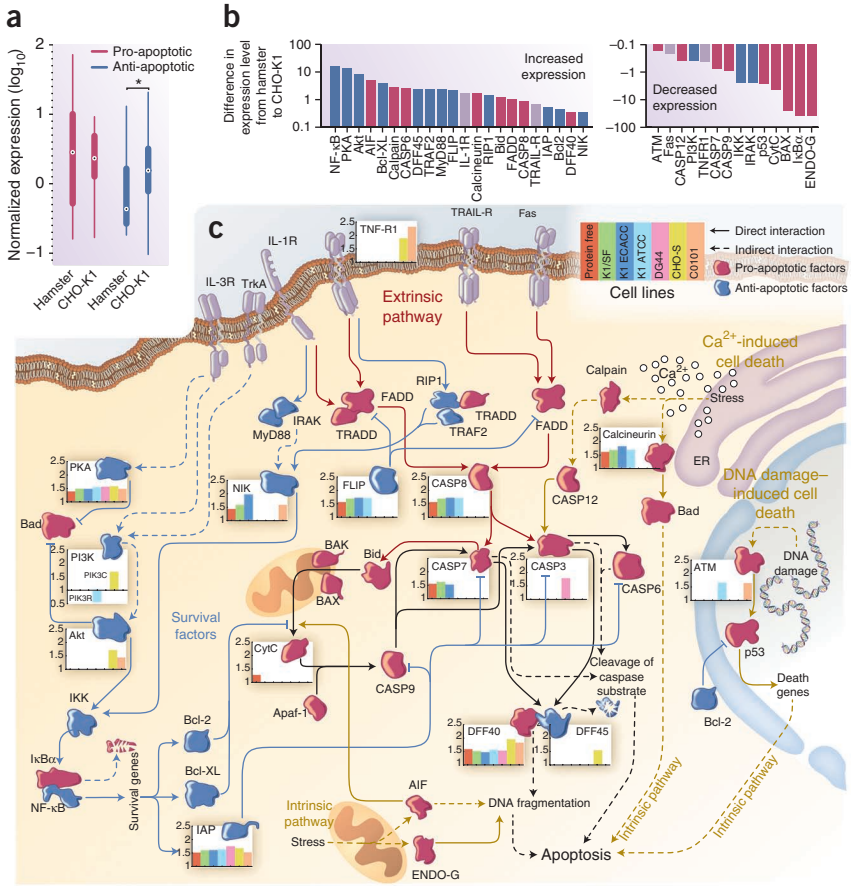

Figure 4 Expression changes and CNVs of key members of the apoptotic pathways. Apoptosis is a complex network of proteins that integrates several external and internal signals to make decisions about programmed cell death. (**a**) On average, gene expression levels of pro-apoptotic genes are only slightly lower in CHO-K1, in comparison to the Chinese hamster. However, anti-apoptotic gene expression is significantly higher in CHO-K1 (*: $P < 0.02$, Wilcoxon rank-sum test). (**b**) When assessing expression of individual genes, pro-apoptotic genes (red) tend to more frequently decrease mRNA expression, whereas anti-apoptotic genes (blue) more frequently increase expression. (**c**) Many major pro-apoptotic (red) and anti-apoptotic (blue) proteins are represented here in the context of the extrinsic (brown), intrinsic (red) or survival (blue) pathways. Proteins that have CNVs are plotted in bar graphs with each bar representing a unique cell line as detailed in the legend, and copy numbers are normalized to the copy number in hamster. Thus, a value less than one suggests a loss of a gene copy, whereas a value greater than one suggests duplication. Details on each gene abbreviation are included in **Supplementary Table 21**.

all cell lines had genomic sequence data for the DHFR gene, except for DG44 (**Fig. 3b**). This DG44-specific deletion was further confirmed by PCR (**Fig. 3c**).

To assess the genome-wide differences between these CHO cell lines, we used the hamster genome assembly as the reference sequence. This reference sequence allowed us to determine SNPs, short insertions and deletions (indels) and gene copy number variations (CNVs) (**Supplementary Table 14**). Across the cell lines, we identified 3,715,639 SNPs, and a phylogenetic reconstruction based on these SNPs accurately recapitulated the cell line history (**Fig. 3d**). We also identified 551,240 indels shorter than 5 bp, 319 of which are predicted to be frame-shifting indels in coding regions. SNPs and indels did not occur uniformly, and some hamster chromosomes were more affected than others (**Fig. 3e**,**f**).

We also found 3,383 nonredundant duplicated regions in at least one cell line and 177 duplicated regions in all seven cell lines (**Supplementary Table 15**). In total, 4,241 genes resided entirely within these 3,383 duplicated regions. Moreover, 113 genes were found to have a reduced copy number in one or more cell lines. In addition, 17 hamster genes were completely missing in at least one cell line, and the missing genes often differed between the lineages (**Supplementary Table 16**).

A variety of genes are associated with mutations and CNVs (**Supplementary Tables 17–20**). Of the SNPs, 5,487 (0.15%) were nonsynonymous and significantly enriched in many GO classes (false discovery rate < 0.01), such as olfactory genes and G protein–coupled receptors ($P < 2 \times 10^{-25}$ and 6×10^{-21}, respectively; hypergeometric test), whereas genes in these same classes were rarely duplicated ($P < 1 \times 10^{-5}$ and 0.02, respectively; hypergeometric test). In addition, proteins involved in cell adhesion were also enriched in SNPs ($P < 0.004$; hypergeometric test). It is possible that these mutations influence the ability of CHO cells to grow in suspension cultures without adhesion factors.

Other genes were protected from SNPs, such as genes associated with DNA binding transcription factor activity and metabolism ($P < 0.006$ and $P < 9 \times 10^{-5}$, respectively; hypergeometric test). Notably, some signaling pathways were insulated from SNPs, such as the WNT and mTOR signaling pathways ($P < 0.02$ and $P < 0.002$, respectively; hypergeometric test) and autophagy ($P < 0.01$). These pathways all contribute to the proliferative and immortalized phenotypes in cancer cells[21–23] and likely play a similar role in CHO cell lines. Protein glycosylation was also significantly insulated from SNPs in all cell lines (mean hypergeometric $P = 0.018$). Thus, the distribution of mutations and CNVs seems consistent with traits that make CHO cell lines desirable protein production hosts (that is, high proliferation rate, suspension growth and protected protein glycosylation).

Using the genome to study the apoptosis pathway

CHO production strains can be grown to high cell densities in fed-batch cultures with serum-free media. Bioprocessing limitations in nutrients in these environments can lead to apoptosis, thereby limiting viable cell density and volumetric productivity. To improve bioprocessing efficiencies, many researchers have sought to improve cell-line longevity by suppressing apoptosis in CHO cells. These efforts involve modulating protein activity by overexpressing anti-apoptotic pathways[24] and blocking pro-apoptotic pathways with chemicals[25], short interfering RNA (siRNA)[26] and gene deletions[27]. However, the complex nature of apoptosis has made it nontrivial to optimize in CHO cells. Thus, a more complete view of gene expression and mutations in the apoptosis system could facilitate bioprocessing and cell engineering efforts to control cell death.

To assess changes in apoptosis in CHO cells, we first identified homologs for anti- and pro-apoptotic proteins in the *C. griseus* genome (**Supplementary Table 21**). Of the 62 KEGG orthologous

gene identifiers in apoptosis, 92% were in the hamster genome. Consistent with observations in mouse, caspase-10 was missing[28]. Other missing genes included interleukin-3, interleukin-3 receptor alpha and interleukin-1 alpha. Although these genes were undetected, apoptosis utilizes redundant pathways, and the lack of these genes should not hinder the system.

In the CHO-K1 cell line, no additional genes for anti- and pro-apoptotic proteins were lost relative to the hamster. Instead, apoptotic gene expression significantly changed. Pro-apoptotic genes exhibited slightly lower gene expression in CHO-K1 in comparison to *C. griseus*, although this was not statistically significant. However, anti-apoptotic genes in CHO-K1 exhibited significantly higher median expression ($P < 0.02$; Wilcoxon rank-sum test; **Fig. 4a**). Apoptotic genes with the greatest increase in expression tended to be anti-apoptotic (e.g., NF-κB, protein kinase A, Akt and Bcl-XL), whereas repressed genes tended to be pro-apoptotic (e.g., endonuclease G, IκBα, BAX, and p53) (**Fig. 4b**). Thus, CHO-K1 suppresses apoptosis, and we anticipate that similar gene expression changes occur in other CHO cell lines.

In addition to changes in apoptotic gene expression, CNVs also frequently occur in apoptotic genes in mammalian cell lines[29]. As CNVs can complicate efforts to engineer cell lines, we also analyzed CHO CNVs in the context of the apoptosis pathways.

The apoptotic network is stimulated by external signals through the extrinsic pathway, or internal stress signals (e.g., increases in cytosolic Ca^{2+} or DNA damage) through the intrinsic pathway. The diverse signals transmitted by each pathway converge upon the caspase proteases, which cleave protein targets and lead to cell death[28]. As a strategy to increase CHO cell longevity, caspase activation has been targeted with chemical inhibitors[30] and caspase-inhibiting proteins[24,31–33]. We found that several cell lines contained extra copies of various caspases (**Fig. 4c**). Thus, efforts to remove pro-apoptotic genes, such as caspases, should account for potential CNVs for those genes. Some anti-apoptotic genes were duplicated only in individual cell lines, which may lead to these lines being more resilient against apoptosis activation. For example, the inhibitors of apoptosis (IAP) family of proteins inhibit caspases[34], and we found that one IAP gene, *BIRC7*, is duplicated in all cell lines. In addition, another anti-apoptotic factor, phosphoinositide 3-kinase (PI3K), also showed cell type–specific CNVs.

In general, CNVs occur in various pathways, such as apoptosis and glycosylation (**Supplementary Fig. 3**) and can differ between cell lines (**Supplementary Tables 22** and **23**). Knowledge of CNVs can help researchers avoid unexpected genomic changes[35–37] when using nucleases in duplicated regions. CNVs can be clone-specific as gene copy numbers in a single cell line vary considerably during growth media adaptation or after several cell passages[29,38]. Thus, clone-specific genomic data may indicate which cell line modifications will be effective in developing a particular production cell line.

DISCUSSION

Genomic resources have provided a wealth of tools in biotechnology[4], ranging from phenotyping tools, such as transcriptomics, to genome editing technologies. These resources have transformed our ability to study and modify the functions of human cells (e.g., cancer and human embryonic kidney cells) and other model organisms. Similar tools are becoming available for CHO cells[16,39,40], but maximizing their potential requires a clear picture of the genomic landscape of CHO cells. Here, we demonstrated how the *C. griseus* genome can provide a sequence-level view of genomic heterogeneity between cell lines and yield a more comprehensive picture of the variants in a cell line of choice.

Numerous studies have shown large chromosomal rearrangements in CHO cells, using banding techniques[41–43] and fluorescence *in situ* hybridization[8,44–48]. These approaches identified large translocations in CHO cells, providing a coarse-grained view of genomic variations in these unstable genomes. We present, for the first time to our knowledge, a whole-genome, sequence-level view of the heterogeneity between CHO cell lines. We showed that each cell line harbors a unique set of mutations, including SNPs, indels, CNVs and missing genes. CNVs were particularly heterogeneous, with 48% (mostly duplications) being unique to one cell line (**Supplementary Table 16**). We also found that mutations rapidly accumulate during development of production cell lines. For example, during the development of the C0101 antibody-producing cell line from CHO-S, 301,753 new SNPs arose, representing 9% of the SNPs in that cell line.

The nonuniform distribution of mutations in each cell line seemed to have some phenotypic relevance. Indeed, several processes associated with proliferation and immortalized phenotypes were more insulated from mutation. These included the WNT and mTOR signaling pathways and processes such as autophagy. Mutations in other pathways such as glycosylation and viral susceptibility (**Supplementary Notes**) varied between cell lines and might influence desired phenotypic properties, although careful biochemical studies are needed. Duplications were also seen for many apoptotic genes. Notably, many of the sequence variations were shared between members of the same family of CHO cells (that is, CHO-K1, DG44 or CHO-S), but these were frequently not shared across CHO cell families (**Supplementary Tables 16** and **22**–**24**). A detailed knowledge of mutations in each cell line may be valuable for cell line selection, characterization and engineering, as well as bioprocess and media optimization. This knowledge for each cell line may further improve the success of siRNAs, zinc finger nucleases and other cell-line engineering tools. Additionally, as more sequence variation data are collected on diverse cell lines, it may be possible to associate cell phenotypes with different mutations (as is commonly done in model organisms[49]).

To fully detail the sequence variations, it is necessary to have a well-defined reference genome with relevance to all CHO cell lines. The reference genome should exhibit several properties. First, it must contain the genomic sequence of all native CHO genes and their regulatory elements. We found that CHO-K1 seems to be missing certain hamster genes, and that cell lines from other lineages are missing other genes (**Supplementary Table 16**). Although we focused on genes that are entirely missing, many more truncated genes and disrupted promoter elements may be found in each cell line as gene models are improved and as regulatory elements are discovered.

Second, it is often desirable to identify all variants in a cell line, and not just the genomic differences between two cell lines. There are clear ultrastructural differences between the hamster and CHO cells. Some chromosomal translocations are conserved among cell lines[8]. These structural variations are likely conserved because CHO cells from the CHO-K1, DG44 and CHO-S lineages share a common highly mutated ancestor. Indeed, we found that 67% of SNPs (~2.5 million) were shared among all CHO cell lines. These shared variants would be missed if the CHO-K1 genome were used as the sole reference. Mutations with deleterious effects on expression and/or activity can be more comprehensively cataloged using the hamster genome as the reference. Thus, endemic loss-of-function mutations in CHO could be identified and remedied as needed for a desired phenotype.

Third, a reference genome must be amenable to improvement over time. The chromosomes of CHO cell lines are unstable, with nonnegligible karyotypic differences even in the same culture[17,43]. Thus, it will be much easier to develop and maintain a gold standard reference sequence of the more stable Chinese hamster genome. This resource will be valuable for characterizing CHO cell lines and using omic technologies, akin to how the *M. musculus* genome is used for studying murine cell lines. Furthermore, although regulatory challenges remain for cell line engineering, whole-genome resequencing against a reference genome will provide transparency as regulatory agencies assess products from engineered cell lines for approval.

There are important differences in genomic content among CHO cell lines that can influence cell line traits. These are likely to be further influenced by differences in gene expression levels. As a result, genome-scale viewpoints will likely become increasingly relevant for CHO-based bioprocessing, as they have for microbe-based manufacturing over the past decade. Although these approaches can require expensive phenotyping and omic technologies, costs are rapidly decreasing. Thus, genome-scale analyses may enhance our ability to understand the production characteristics of CHO cell lines and aid in the production of therapeutic proteins in the coming decades.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GenBank: AMDS00000000; the version described in this study is AMDS00000000.1. Accession codes for the sequencing data for the cell lines and the hamster transcriptome are listed in **Supplementary Table 25**.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to thank K.C. Hayes, at Brandeis University, for enabling the continued housing of the Chinese hamster colony from which CHO cells were derived. B. Monroe, L. Chasin and S. Gorfien kindly aided in delineating the history of the cell lines used in this study, and T. Omasa gave guidance in the chromosomal assignments of scaffolds. This work was funded in part by the China National GeneBank-Shenzhen, the Shenzhen Engineering Laboratory for Genomics-Assisted Animal Breeding, and the Shenzhen Key Laboratory of Transomics Biotechnologies (NO.CXB201108250096A). This work and the Center for Biosustainability at the Danish Technical University were also funded with generous support from the Novo Nordisk Foundation. Female Chinese hamsters were kindly provided by G. Yerganian. The authors would also like to thank L. Donahue-Hjelle for kindly providing their cell lines.

AUTHOR CONTRIBUTIONS

B.O.P., I.F., X.X., J.W., M.J.B. and J.F. conceived, designed and guided the study. N.E.L. and H.N. wrote the manuscript. G.Y., A.M.R., J.R., D.B.-H. and N.L. prepared tissue and cells for sequencing. X.X., X.L., Y.L., C.B., W.C. and M.X. performed the genome assembly, optical mapping and annotation. N.E.L., X.L., H.N., Y.L., E.O'B., A.B. and H.L. analyzed the variant and transcriptomic data. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. IMS Institute for Healthcare Informatics. *The Global Use of Medicines: Outlook Through 2015* ⟨http://www.imshealth.com/ims/Global/Content/Insights/IMS%20Institute%20for%20Healthcare%20Informatics/Documents/The_Global_Use_of_Medicines_Report.pdf⟩ (IMS, 2011).
2. Walsh, G. Biopharmaceutical benchmarks 2010. *Nat. Biotechnol.* **28**, 917–924 (2010).
3. De Jesus, M. & Wurm, F.M. Manufacturing recombinant proteins in kg-ton quantities using animal cells in bioreactors. *Eur. J. Pharm. Biopharm.* **78**, 184–188 (2011).
4. Wuest, D.M., Harcum, S.W. & Lee, K.H. Genomics in mammalian cell culture bioprocessing. *Biotechnol. Adv.* **30**, 629–638 (2012).
5. Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* **29**, 735–741 (2011).
6. Wlaschin, K.F. *et al.* EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotechnol. Bioeng.* **91**, 592–606 (2005).
7. Kantardjieff, A. *et al.* Developing genomic platforms for Chinese hamster ovary cells. *Biotechnol. Adv.* **27**, 1028–1035 (2009).
8. Cao, Y. *et al.* Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. *Biotechnol. Bioeng.* **109**, 1357–1367 (2012).
9. Baycin-Hizal, D. *et al.* Proteomic analysis of Chinese hamster ovary cells. *J. Proteome Res.* **11**, 5265–5276 (2012).
10. Meleady, P. *et al.* Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. *Biotechnol. Bioeng.* **109**, 1386–1394 (2012).
11. Wei, Y.Y. *et al.* Proteomics analysis of chinese hamster ovary cells undergoing apoptosis during prolonged cultivation. *Cytotechnology* **63**, 663–677 (2011).
12. Becker, J. *et al.* Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. *J. Biotechnol.* **156**, 227–235 (2011).
13. Birzele, F. *et al.* Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic Acids Res.* **38**, 3999–4010 (2010).
14. Clarke, C. *et al.* CGCDB: A web-based resource for the investigation of gene coexpression in CHO cell culture. *Biotechnol. Bioeng.* **109**, 1368–1370 (2011).
15. Hackl, M. *et al.* Computational identification of microRNA gene loci and precursor microRNA sequences in CHO cell lines. *J. Biotechnol.* **158**, 151–155 (2012).
16. Kildegaard, H.F., Baycin-Hizal, D., Lewis, N.E. & Betenbaugh, M.J. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Curr. Opin. Biotechnol.* doi:10.1016/j.copbio.2013.02.007 (20 March 2013).
17. Wurm, F.M. & Hacker, D. First CHO genome. *Nat. Biotechnol.* **29**, 718–720 (2011).
18. Mayshar, Y. *et al.* Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7**, 521–531 (2010).
19. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
20. Urlaub, G., Kas, E., Carothers, A.M. & Chasin, L.A. Deletion of the diploid dihydrofolate reductase locus from cultured mammalian cells. *Cell* **33**, 405–412 (1983).
21. Rosenfeldt, M.T. & Ryan, K.M. The multiple roles of autophagy in cancer. *Carcinogenesis* **32**, 955–963 (2011).
22. Sabatini, D.M. mTOR and cancer: insights into a complex relationship. *Nat. Rev. Cancer* **6**, 729–734 (2006).
23. Yu, M. *et al.* RNA sequencing of pancreatic circulating tumour cells implicates WNT signalling in metastasis. *Nature* **487**, 510–513 (2012).
24. Becker, E., Florin, L., Pfizenmaier, K. & Kaufmann, H. Evaluation of a combinatorial cell engineering approach to overcome apoptotic effects in XBP-1(s) expressing cells. *J. Biotechnol.* **146**, 198–206 (2010).
25. Fussenegger, M., Schlatter, S., Datwyler, D., Mazur, X. & Bailey, J.E. Controlled proliferation by multigene metabolic engineering enhances the productivity of Chinese hamster ovary cells. *Nat. Biotechnol.* **16**, 468–472 (1998).
26. Kim, S.H. & Lee, G.M. Down-regulation of lactate dehydrogenase-A by siRNAs for reduced lactic acid formation of Chinese hamster ovary cells producing thrombopoietin. *Appl. Microbiol. Biotechnol.* **74**, 152–159 (2007).
27. Cost, G.J. *et al.* BAK and BAX deletion using zinc-finger nucleases yields apoptosis-resistant CHO cells. *Biotechnol. Bioeng.* **105**, 330–340 (2010).
28. Ghavami, S. *et al.* Apoptosis and cancer: mutations within caspase genes. *J. Med. Genet.* **46**, 497–510 (2009).
29. Laurent, L.C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8**, 106–118 (2011).
30. Arden, N. *et al.* Chemical caspase inhibitors enhance cell culture viabilities and protein titer. *Biotechnol. Prog.* **23**, 506–511 (2007).
31. Dorai, H. *et al.* Combining high-throughput screening of caspase activity with anti-apoptosis genes for development of robust CHO production cell lines. *Biotechnol. Prog.* **26**, 1367–1381 (2010).
32. Kim, Y.G., Kim, J.Y. & Lee, G.M. Effect of XIAP overexpression on sodium butyrate-induced apoptosis in recombinant Chinese hamster ovary cells producing erythropoietin. *J. Biotechnol.* **144**, 299–303 (2009).
33. Wang, Z., Park, J.H., Park, H.H., Tan, W. & Park, T.H. Enhancement of therapeutic monoclonal antibody production in CHO cells using 30Kc6 gene. *Process Biochem.* **45**, 1852–1856 (2010).
34. Dasgupta, A., Alvarado, C.S., Xu, Z. & Findley, H.W. Expression and functional role of inhibitor-of-apoptosis protein livin (BIRC7) in neuroblastoma. *Biochem. Biophys. Res. Commun.* **400**, 53–59 (2010).

35. Bueno, C. *et al.* Etoposide induces MLL rearrangements and other chromosomal abnormalities in human embryonic stem cells. *Carcinogenesis* **30**, 1628–1637 (2009).
36. Lee, H.J., Kweon, J., Kim, E., Kim, S. & Kim, J.S. Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases. *Genome Res.* **22**, 539–548 (2012).
37. Piganeau, M. *et al.* Cancer translocations in human cells induced by zinc finger and TALE nucleases. *Genome Res.* **7**, 1182–1193 (2013).
38. Narva, E. *et al.* High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nat. Biotechnol.* **28**, 371–377 (2010).
39. Griffin, T.J., Seth, G., Xie, H., Bandhakavi, S. & Hu, W.S. Advancing mammalian cell culture engineering using genome-scale technologies. *Trends Biotechnol.* **25**, 401–408 (2007).
40. Datta, P., Linhardt, R.J. & Sharfstein, S.T. An 'omics approach towards CHO cell engineering. *Biotechnol. Bioeng.* **110**, 1255–1271 (2013).
41. Deaven, L.L. & Petersen, D.F. The chromosomes of CHO, an aneuploid Chinese hamster cell line: G-band, C-band, and autoradiographic analyses. *Chromosoma* **41**, 129–144 (1973).
42. Derouazi, M. *et al.* Genetic characterization of CHO production host DG44 and derivative recombinant cell lines. *Biochem. Biophys. Res. Commun.* **340**, 1069–1077 (2006).
43. Worton, R.G., Ho, C.C. & Duff, C. Chromosome stability in CHO cells. *Somatic Cell Genet.* **3**, 27–45 (1977).
44. Balajee, A.S., Dominguez, I. & Natarajan, A.T. Construction of Chinese hamster chromosome specific DNA libraries and their use in the analysis of spontaneous chromosome rearrangements in different cell lines. *Cytogenet. Cell Genet.* **70**, 95–101 (1995).
45. Davies, J. & Reff, M. Chromosome localization and gene-copy-number quantification of three random integrations in Chinese-hamster ovary cells and their amplified cell lines using fluorescence *in situ* hybridization. *Biotechnol. Appl. Biochem.* **33**, 99–105 (2001).
46. Simi, S., Xiao, Y., Campagna, M., Doehmer, J. & Darroudi, F. Dual-colour FISH analysis to characterize a marker chromosome in cytochrome P450 2B1 recombinant V79 Chinese hamster cells. *Mutagenesis* **14**, 57–61 (1999).
47. Xiao, Y., Slijepcevic, P., Arkesteijn, G., Darroudi, F. & Natarajan, A.T. Development of DNA libraries specific for Chinese hamster chromosomes 3, 4, 9, 10, X, and Y by DOP-PCR. *Cytogenet. Cell Genet.* **75**, 57–62 (1996).
48. Cao, Y. *et al.* Fluorescence in situ hybridization using bacterial artificial chromosome (BAC) clones for the analysis of chromosome rearrangement in Chinese hamster ovary cells. *Methods* **56**, 418–423 (2012).
49. Rosenberg, N.A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).

ONLINE METHODS

Sample preparation and DNA sequencing. Female Chinese hamsters were kindly provided by G. Yerganian. Genomic DNA was isolated from multiple tissues using a modified SDS method[50]. Seven different paired-end libraries were constructed with 170 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb insert sizes, using the standard protocol provided by Illumina (San Diego). The sequencing was done using Illumina HiSeq 2000 according to the manufacturer's standard protocol. The raw data were filtered to remove low-quality reads, reads with adaptor sequences, and duplicated reads before *de novo* genome assembly (**Supplementary Notes**).

Optical mapping. High molecular weight DNA was obtained from Chinese hamster tissues. Whole genome shotgun, single-molecule restriction maps were generated using the automated Argus system (OpGen Inc., Maryland, USA), based on the optical mapping technology[51,52]. Individual DNA molecules were deposited onto silane-derivatized glass surfaces in MapCards (OpGen Inc., MD, USA) and digested by BamHI enzyme. DNA was subsequently stained with JOJO fluorescence dye (Invitrogen, CA, USA) and imaged within the Argus system. A total of 28 MapCards were processed. The DNA molecules were marked up and restriction fragment size was determined by image processing in parallel with image acquisition. This yielded ~26× optical data.

Genome assembly. Similar to the assembly of the CHO-K1 genome, SOAPdenovo v.1.06 (ref. 53) was used to assemble the hamster genome into contigs and scaffolds as well as for gap closure. The final genome assembly was 2.4 Gb in length, which is about 89% of the estimated genome. The contig N50 (the shortest length of sequence contributing more than half of assembled sequences) was 26.5 kb and the scaffold N50 was 1.54 Mb (**Table 1** for statistics on genome assembly). Optical mapping data were used to further assemble the genome into super-scaffolds. The scaffolds were extended according to the optical maps to determine overlapping regions between scaffolds and their relative location and orientation. First, the sequence scaffolds were converted into restriction maps by *in silico* restriction enzyme digestion by BamHI. These *in silico* restriction maps were used as seeds to identify single-molecule restriction maps of DNA from the corresponding genomic regions by map-to-map alignment. These single-molecule maps were then assembled together by using the *in silico* maps, to produce elongated consensus maps (extended scaffolds). The low coverage regions near the ends of the extended scaffolds were trimmed off to maintain high extension quality. To generate sufficient extension length, we repeated the alignment-assembly process 4–5 times, using the extended scaffolds as seeds for each subsequent iteration. All of the extended scaffolds were then aligned to each other. Any pair-wise alignments above an empirically decided confidence threshold were considered as initial candidates for scaffold connection. Alignments that overlapped substantially with the initial scaffolds were excluded from the candidates. Among the remaining alignments, those with the highest score were considered. The relative location and orientation of each pair of connected scaffolds were used to generate super-scaffolds. This resulted in 6,356 super-scaffolds (>2 kb) with N50 of 2.49 Mb (**Table 1**).

Chromosomal assignment of scaffolds. 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[8]. Specifically, chromosomal assignments were obtained for each BAC, and then blastn was used to find scaffolds with the highest homology to the BAC end-sequences (E-value < 1×10^{-5}). Scaffolds aligned to BACs from more than one chromosome were filtered from the analysis. Once chromosomal assignments were obtained for scaffolds (**Supplementary Table 3**), they were extended to super-scaffolds based on optical mapping data (**Supplementary Table 4**). From this analysis, we were able to reliably localize 26% of the genomic sequence to specific hamster chromosomes.

RNA sequencing and assembly. RNA was isolated from eight tissues from several Chinese hamsters. Total RNA was extracted using Trizol (Invitrogen, USA). The isolated RNA was then treated by RNase-free DNase. The RNA was subsequently mixed and treated using the Illumina mRNA-Seq Prep Kit following the manufacturer's instructions. The insert size of the RNA libraries was about 170 bp, and the sequencing was done using Illumina HiSeq 2000. Raw reads were filtered out if they contained contamination or were of low quality (more than 10% of the bases with unknown quality). The resulting 5 Gb of RNA-seq data were assembled into transcriptional fragments by Trinity[54] (version: r2011-08-20). We then assessed the coverage of the transcripts in the genome assembly by mapping the assembled transcriptional fragments to the genome assembly using BLAT[55].

Gene annotation. We predicted gene models using *de novo*, homology-based and transcriptome-aided prediction approaches. For *de novo* gene prediction, we used a repeat-masked genome assembly. We used AUGUSTUS (version 2.03)[56], GlimmerHMM (version 3.02) and Genscan (version 1.0) for *de novo* gene annotation. For homology-based prediction, we mapped the protein sequences from the CHO-K1 cell line using BLAT, with an E-value cutoff of 10^{-2}, followed by Genewise[57] (version 2.2.0) for gene annotation. Genes with less than 70% identity and 80% coverage in the BLAT alignment were filtered. Transcriptome-aided annotation was done by mapping all RNA-seq reads back to the reference genome using Tophat[58] (version 1.3.3), implemented with bowtie[59] (version 0.12.5). The transcripts were assembled using Cufflinks[60] (version 1.2.1). Taken together with the assembled transcripts from Cufflinks, we identified the genomic regions covered by the transcriptome. *De novo* genes with less than 50% coverage in the transcriptome data were filtered. Finally, the nonredundant gene sets were merged with the homology-based method genes and *de novo* genes, while filtering transposable element genes identified in the functional annotation. Gene functions were assigned according to the best match of the alignments using blastp (E-value ≤ 10^{-5}) against the Swiss-Prot and UniProt databases (release 15.10). The motifs and domains of genes were determined by InterProScan[61] (version 4.5) against protein databases. Gene Ontology IDs for each gene were obtained from the corresponding InterPro entry. All genes were aligned against KEGG (release 48.2) proteins, and the pathway in which the gene might be involved was derived from the matching genes in KEGG. If the best hit of a gene was "function unknown," "putative," etc., the second best hit was used to assign function until there were no more hits meeting the alignment criteria (then this gene would be annotated as functionally unknown). Repeat features, transposible elements and endogenous retroviral genes were also identified and annotated (**Supplementary Notes** and **Supplementary Figs. 4** and **5**).

Genome comparison. The assembled Chinese hamster and CHO-K1 scaffolds (>1 kb) were masked by RepeatMasker to remove repeat elements. The repeat-masked mouse genome[62] was downloaded from ENSEMBL (release 60). The repeat-masked hamster and the CHO-K1 assemblies were aligned to the mouse genome as previously described[63]. The LASTZ pair-wise whole genome alignment software (http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html) was used with the parameters: $K = 4{,}500$ $l = 3{,}000$ $Y = 15{,}000$ $E = 150$ $H = 0$ O = 600 $T = 2$. The Chain/Net package[64] was subsequently used to process the alignment. With the hamster chromosomal assignments (**Supplementary Fig. 6**) for many scaffolds, comparisons on chromosomal localization were made between the mouse and hamster (**Fig. 2a**, **Supplementary Notes** and **Supplementary Fig. 7**). Structural variations between the hamster and CHO-K1 genomes were found using a procedure previously applied to compare two human genomes[65]. Large masked scaffolds (larger than 1 megabase in length) were processed with LASTZ using the aforementioned parameter set. These alignments between the hamster and CHO-K1 were corrected for inaccurately predicted gaps in the assembly and other alignment errors. Using the corrected alignments, the best match for each location on the CHO-K1 scaffolds was chosen by the option "axtBest." This deploys a dynamic programming algorithm using the same substitution matrix as used during the alignment. The hits that contributed most to the colinearity between the large scaffolds of the Chinese hamster and CHO-K1 were selected, and discrepancies between the aligned sections were called as insertions and deletions, exhibiting a wide range of lengths (**Supplementary Fig. 8**).

Detection of sequence variation among cell lines. We sequenced six different CHO cell lines to assess the extent of genomic divergence from the hamster genome. The cell lines were grown on their respective media (**Supplementary Table 13**), after which their DNA was harvested and sequenced to greater

doi:10.1038/nbt.2624

than the minimum recommended depth of 9× for each cell line, to assure that enough coverage was obtained to resolve heterozygous SNPs. Sequencing data can be obtained from the NCBI short read archive (see **Supplementary Table 25** for accession numbers).

Missing genes in the six resequenced cell lines and the previously sequenced CHO-K1 ATCC genome[5] were detected as follows. Sequencing reads from the seven cell lines and hamster were mapped to hamster assembly with BWA (version 0.5.9). Read depth of genes was calculated using 'depth' tool of SAMtools (version 0.1.18). A gene was declared to be deleted if it conformed to the following criteria. First, when mapping the hamster reads to the assembled hamster genome scaffolds, the read depth of the gene had to be greater than half of the mean read depth across all hamster genes. Second, the read depth of the gene for a given cell line had to be less than 0.1. SOAPaligner (version 2.21) was also used for a repeat trial. The resulting read depth distribution was consistent with that derived from BWA.

To detect SNPs, indels and CNVs, the raw reads from each cell line were mapped to the hamster genome assembly to determine sequence variations. To aid the process of variant detection, the hamster scaffolds were concatenated in a random fashion to obtain 12 pseudo chromosomes. SOAP was used to align the sequencing reads from each cell line to the reference hamster assembly. The alignments were subsequently split into pseudochromosomes and sorted according to the mapped position. SOAPsnp was used to identify SNPs in each cell line. To further refine the predicted SNPs, we adopted an alternative approach using BWA to align the reads to the hamster assembly. The 'mpileup' tool of SAMtools was applied to get the information of each genomic position in the different samples and BCFtools in the same package was used for variant calling. The two SNP data sets were subsequently combined to make the final SNP data set. For each library, we filtered SNPs with depth less than half of the mean depth. We also filtered SNPs that were located within 5 bp of another SNP. In total, we identified 3,715,639 SNPs. SNPs were used to reconstruct the phylogeny of the CHO cell lines. The Jukes-Cantor pairwise distance was computed between all strains and the phylogenetic tree was built using the unweighted pair group method average. The alignments were further processed using SOAPindel (http://soap.genomics.org.cn/soapindel.html) to identify indels and analyzed using CNVnator to detect CNVs[66].

Nonsynonymous SNPs, frame-shifting indels, and gene-containing CNVs were identified and analyzed. The hypergeometric test was used to identify gene classes that were over- or under-represented in mutations in all Gene Ontology classes and KEGG pathways, based on our genome annotation (**Supplementary Tables 17**–**20**). More detailed analysis on apoptotic pathways was based on KEGG ortholog assignments (**Supplementary Table 21**). Additional analysis on glycosylation and viral succeptability genes (**Supplementary Notes** and **Supplementary Fig. 9**) were based on homology to gene lists published previously5 (**Supplementary Notes** and **Supplementary Tables 26** and **27**).

50. Peng, J., Wang, H., Haley, S.D., Peairs, F.B. & Lapitan, N.L.V. Molecular mapping of the Russian wheat aphid resistance gene in wheat. *Crop Sci.* **47**, 2418–2429 (2007).
51. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* **31**, 135–141 (2013).
52. Schwartz, D.C. *et al.* Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–114 (1993).
53. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
54. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
55. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
56. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
57. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
58. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
59. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
60. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
61. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
62. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
63. Jex, A.R. *et al. Ascaris suum* draft genome. *Nature* **479**, 529–533 (2011).
64. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100**, 11484–11489 (2003).
65. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. *Nat. Biotechnol.* **29**, 723–730 (2011).
66. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).