Published in final edited form as:
Biotechnol Bioeng. 2018 August ; 115(8): 2087–2100. doi:10.1002/bit.26722.

A reference genome of the Chinese hamster based on a hybrid assembly strategy

Oliver Rupp[1,13], **Madolyn L. MacDonald**[2,3,13], **Shangzhong Li**[4,5,13], **Heena Dhiman**[6,13], **Shawn Polson**[2,3], **Sven Griep**[1], **Kelley Heffner**[7], **Inmaculada Hernandez**[8], **Karina Brinkrolf**[9], **Vaibhav Jadhav**[6], **Mojtaba Samoudi**[5,10], **Haiping Hao**[11], **Brewster Kingham**[3], **Alexander Goesmann**[1], **Michael J. Betenbaugh**[7,14], **Nathan E. Lewis**[4,5,10,14], **Nicole Borth**[6,8,14], and **Kelvin H. Lee**[3,12,14]

[1]Bioinformatics and Systems Biology, Justus-Liebig-University Giessen, Giessen, Germany

[2]Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, USA

[3]Delaware Biotechnology Institute, Newark, Delaware, USA

[4]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA

[5]Novo Nordisk Foundation Center for Biosustainability, University of California, San Diego, La Jolla, CA, USA

[6]Austrian Center of Industrial Biotechnology, Vienna, Austria

[7]Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland, USA

[8]Department of Biotechnology, University of Natural Resources and Life Sciences Vienna, Austria

[9]Department of Biorescources, Fraunhofer Institute for Molecular Biology and Applied Ecology, Giessen, Germany

[10]Department of Pediatrics, University of California, San Diego, La Jolla, CA USA

[11]Johns Hopkins University Deep Sequencing and Microarray Core, Johns Hopkins University, Baltimore, Maryland, USA

[12]Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware, USA

Correspondence to: Kelvin H. Lee (KHL@udel.edu), Nicole Borth (nicole.borth@boku.ac.at), Nathan E. Lewis (nlewisres@ucsd.edu), Michael J. Betenbaugh (beten@jhu.edu).

[13]Co-first authors

[14]Co-corresponding authors

5 Author Contributions

OR, SG and KB conducted genome assembly, and aided in writing. MLM functionally annotated the genome, and aided in writing. HD and IH performed the mitochondrial gene and chromatin state analysis. SL conducted gap analysis, prepared CHO pgsA-745 DNA for sequencing, and aided in writing. KH isolated hamster tissue and prepared DNA for sequencing. MJB conceived of the project and oversaw the hamster DNA preparation. SP provided valuable guidance and aided in the sequencing and analysis. HH, BK and MS aided in the sequencing and analysis. VJ evaluated approaches to separate scaffolds for chromosomes 9 and 10. AG oversaw genome assembly efforts. NEL conceived of the project, wrote the manuscript, and guided the gap analysis. NB conceived of and coordinated the project, oversaw the chromatin state analysis and wrote the manuscript. KL conceived of and coordinated the project, and oversaw the genome annotation.

Abstract

Accurate and complete genome sequences are essential in biotechnology to facilitate genome-based cell engineering efforts. The current genome assemblies for *Cricetulus griseus*, the Chinese hamster, are fragmented and replete with gap sequences and misassemblies, consistent with most short-read based assemblies. Here, we completely resequenced *C. griseus* using Single Molecule Real Time (SMRT) sequencing and merged this with Illumina-based assemblies. This generated a more contiguous and complete genome assembly than either technology alone, reducing the number of scaffolds by >28-fold, with 90% of the sequence in the 122 longest scaffolds. Most genes are now found in single scaffolds, including up- and downstream regulatory elements, enabling improved study of noncoding regions. With >95% of the gap sequence filled, important CHO cell mutations have been detected in draft assembly gaps. This new assembly will be an invaluable resource for continued basic and pharmaceutical research.

Keywords

Chinese hamster; genome; assembly; biopharmaceuticals

1 Introduction

For decades, Chinese hamster ovary (CHO) cells have been the primary recombinant protein production host across the biopharmaceutical industry [Walsh, 2014]. Characteristics such as glycosylation, fast growth, and ease of genetic manipulation help explain their prevalence. The history of CHO cells dates back to the 1950s when ovarian connective tissue was harvested from the Chinese hamster and derivative cells spontaneously became immortal [Tjio and Puck, 1958]. Since then, CHO has diverged into different adherent and suspension cell lines, such as CHO-K1, CHO-S, and CHO DG44 [Lewis et al., 2013]. Their protein production capacity has been greatly enhanced through decades of refinements in bioprocessing strategies, media optimization, and engineering of transgenes and expression vectors. However, little engineering was done on the host cell itself which remained poorly characterized for decades. Increasing demands on quantities of difficult-to-express-proteins, protein quality, and time-to-market now require new strategies that involve cell engineering.

To facilitate CHO cell research and development, the community relies on published genomes for the CHO-K1 cell line and the parent Chinese hamster, sequenced using short-read Illumina technologies [Xu et al., 2011], [Brinkrolf et al., 2013], [Lewis et al., 2013], [Yusufi et al., 2017]. These resources have enhanced the use of transcriptomics, proteomics, genetic engineering, and other technologies [Kildegaard et al., 2013], [Lee et al., 2015a], [Richelle and Lewis, 2017] to understand and engineer desired traits in cells. However, to improve accuracy in such endeavors, there is a need for genomic resources with far more contiguous sequence and less pervasive gaps. The acquisition of such contiguous sequences is possible now with third generation sequencing technologies, such as Single Molecule Real Time (SMRT) sequencing technology [Eid et al., 2009], which provide mean read lengths that are more than an order of magnitude larger than earlier sequencing technologies. The reads can span repetitive elements, resulting in longer contigs and minimal gaps within scaffolds [Bickhart et al., 2017], [Jiao et al., 2017], [Gordon et al., 2016]. This enables the routine assembly of mammalian genomes approaching the current quality of the human genome.

To obtain a higher quality reference assembly of the Chinese hamster, we have resequenced Chinese hamster liver tissue using long-read SMRT technology at 45x coverage. Assemblies generated with Illumina or SMRT sequencing data were merged with the existing publicly available assemblies. Assembly merging yielded four candidate assemblies, which were evaluated for completeness and quality using 80 assembly metrics. Merging the platform-specific assemblies results in a more contiguous, accurate, and complete genome assembly than using either technology alone. The final assembly presented is the most complete Chinese hamster genome to-date, with the number of scaffolds reduced to fewer than 3–6% the number in earlier works, and the mean contig length 16–29x longer. The new genome has substantial improvement in gene completeness and extent of flanking noncoding DNA, thereby allowing for identification of promoters and enhancers. Finally, 95% of the sequence gaps were filled, exposing hundreds of cell line-specific mutations in coding regions of the genome for several CHO cell lines. For example, an important SNP in the glycosyltransferase, Xylt2, which impacts glycosylation and which was hidden in gaps in previous assemblies, can now be detected. Thus, this resource will serve as an important reference genome for researchers across the biotechnology industry and scientific community.

2 Results

2.1 Platform-specific assemblies of the Chinese hamster genome

2.1.1 Pooled Illumina assembly—In two independent previous attempts, the Chinese hamster genome was generated using Illumina sequencing from DNA isolated from liver tissue acquired from the same hamster colony as used for deriving CHO cells in 1957 [Brinkrolf et al., 2013], [Lewis et al., 2013]. The current RefSeq assembly originated from whole-genome libraries with varying insert sizes [Lewis et al., 2013]. A second assembly (CSA) using chromosome sorted sequencing libraries is also publicly available [Brinkrolf et al., 2013]. The different libraries combined yielded about two billion read pairs with read lengths from 99 to 150 bp, in total 442.22 Gb (see Supplementary material for details). K-mer based genome size estimations of different libraries and k-mers ranged between 2.55 Gb and 2.75 Gb.

We *de novo* assembled the pooled Illumina reads from both previous assemblies using ALLPATHS-LG. This Illumina assembly contained 2.39 Gb of scaffolds with 2.66% gaps. The scaffold N50 number (minimal number of scaffolds needed to cover 50% of the assembled genome) was 128 with an N50 length (length of the smallest N50 scaffold) of 5.95 Mb (Table 1), which was much greater than the previously published assemblies.

2.1.2 PacBio SMRT sequencing assembly—Pacific Biosciences SMRT sequencing yielded 107.45 Gb total sequence from 13.49 million subreads, corresponding to ~45x coverage of the 2.4 Gb genome (after filtering and adapter trimming). Pooled and corrected Illumina reads were used to correct sequencing errors of the SMRT reads. Specifically, overlapping paired-end reads were merged and error-corrected as part of the ALLPATHS-

LG [Gnerre et al., 2011] assembly process. This created about 836 million single reads with a mean size of 171 bp and 143.75 Gb total. These were reused in the SMRT error-correction which was done in two steps using proovread [Hackl et al., 2014] and LoRDEC [Salmela and Rivals, 2014], leading to a reduction of the indel-ratio (number of indels divided by the number of matches in the alignments against the Illumina contigs) from 0.18 to 0.04. SMRT reads were assembled using HGAP [Chin et al., 2013], resulting in the assembly hereafter referred as the PacBio SMRT assembly. After removal of duplicate contigs (see Supplementary Table S1), the assembly resulted in 2.3 Gb of non-redundant sequence with an N50 scaffold number of 223 and an N50 size of 2.9 Mb (Table 1).

2.2 A highly contiguous metassembly is obtained by merging draft assemblies

Recent studies have highlighted the improvements of SMRT-only assemblies compared to Illumina-only assemblies [Bickhart et al., 2017], [Gordon et al., 2016], [Jiao et al., 2017], [Zhang et al., 2016], [Shi et al., 2016]. Here we found that both the pooled Illumina assembly (with mixed read length) and the PacBio SMRT-only assembly resulted in substantially improved assembly statistics compared to the two published hamster genome assemblies (Table 1), with an order of magnitude fewer scaffolds and 2–4x larger N50 values. However, the longer PacBio SMRT reads and the larger Illumina insert libraries should provide unique strengths that could be captured through assembly merging. Therefore, we aligned the scaffolds and contigs from four independent assemblies: the Illumina-based chromosome sorted assembly (CSA) [Brinkrolf et al., 2013], the RefSeq assembly [Lewis et al., 2013], the pooled Illumina assembly developed here, and our *de novo* uncurated PacBio SMRT assembly. The Metassembler tool [Wences and Schatz, 2015] uses the first assembly provided as the base and subsequently merges additional assemblies. The tool was applied to the four assemblies using four different orders of merging, resulting in four different metassemblies as shown in Table 2.

All metassemblies showed considerable improvement over all initial draft assemblies (Table 3), with far fewer N50 scaffolds (only 32–34 compared with 223 for the PacBio SMRT and 128–501 for the Illumina-based assemblies), and a significant decrease in gap sequence compared to the Illumina-only assemblies. Improvements in many metrics in all the intermediate merging stages show that all four initial draft assemblies contribute to the improvement of the final assemblies (Supplementary Figure S5). However, the metassemblies starting with the PacBio SMRT assembly outperformed the ones starting with the Illumina assembly in almost all metrics.

To validate assembly accuracy, chromosome-separated sequencing libraries [Brinkrolf et al., 2013] were aligned to the scaffolds. Misassemblies can be easily identified by decreased read coverage from one chromosome and a rise in coverage from another (Supplementary Figure S1). Manual inspection of all scaffolds larger than 1 Mb showed only one scaffold with a clear misassembly in the PacBio SMRT-starting (PICR and PIRC) metassemblies and 11 in the metassemblies starting with Illumina scaffolds (IPCR and IPRC), while the current RefSeq assembly has >24 (Supplementary Figure S2). Inspection of the chromosome coverage at the error-region (Supplementary Figure S3) showed a 30 kb region that contained low and mixed coverage, along with scaffolding gaps. This region was manually cut, and two new scaffolds were created. Ultimately, 96.6% of the sequence could be unambiguously assigned to a specific chromosome (Supplementary Table S2).

2.3 The best assembly is identified using 80 assembly metrics

To quantify and compare the quality of our eight assemblies (including the four initial assemblies and the four metassemblies), we computed 80 different metrics (see Supplementary Material), split into six classes covering different aspects of an assembly (Figure 1.a; Supplementary Figures S4 and S5; Supplementary Table S3) and ranked the assemblies for each class individually. The PICR metassembly had the best overall rank in four of the six classes followed by PIRC with two best overall ranks. Based on this evaluation, PICR was chosen for further analyses.

The PICR metassembly has substantially longer contigs (contiguous sequences with 'N'-regions smaller than 100 bp) than the previous RefSeq assembly and even assemblies of some model organisms such as the rat (*Rattus norvegicus*, assembly Rnor_6.0). In addition, PICR is approaching the continuity seen in the murine reference assembly (*Mus musculus*, assembly GRCm38.p5) (Figure 1.b and Supplementary material).

2.4 Polishing the final assembly

2.4.1 Chromosomes are assigned using reads from flow-sorted DNA—To assign each scaffold to a chromosome, we aligned all chromosome-separated reads to the PICR metassembly. 307 scaffolds were uniquely assigned to a chromosome, accounting for 94% of the genome (or 2.23 Gb). Unassigned scaffolds and scaffolds assigned to the unseparated hamster chromosome 9 and 10 library were instead mapped to the mouse genome. Scaffolds that could be aligned uniquely were assigned to a hamster chromosome based on published hamster chromosome localization [Yang et al., 2000], [Wlaschin and Hu, 2007]. Fifteen scaffolds (18.79 Mb) could be assigned to chromosome 9 and 2 scaffolds (32.58 Mb) to chromosome 10. A detailed list of assigned scaffold numbers and sizes is shown in Supplementary Table S2. The final PICR assembly and the associated raw PacBio SMRT sequencing read data is available under NCBI BioProject PRJNA389969. The existing Illumina assemblies are available under NCBI BioProjects PRJNA167053 (RefSeq) and PRJNA189319 (CSA). Illumina sequencing data for BioProject PRJNA167053 is available from the Sequence Read Archive under SRP020466

2.4.2 Repeat masking, gene prediction and annotation—We annotated the PICR and IPCR metassemblies using the Maker annotation tool [Holt and Yandell, 2011] (Table 4 and Supplementary Table S4). Due to the similarity of the PICR and PIRC assemblies, we decided to compare the annotation of PICR and IPCR. This comparison demonstrated the impact of using assemblies built from different sequencing methods as the primary assembly in Metaassembler. Repeat-masker [Smit et al., 2015] masked approximately 5.5 million repeats in PICR and 5.7 million in IPCR (Supplementary Table S5).

The Maker annotation yielded ~1,300 more genes and transcripts in PICR than in IPCR. Functional annotations were assigned for 23,153 transcripts/proteins in PICR, but only 21,839 transcripts/proteins in IPCR. The annotations of PICR and IPCR demonstrate that beginning assembly merging with the PacBio SMRT assembly, rather than the Illumina assembly, led to the identification and functional annotation of more genes.

The predicted proteins from PICR were searched using BLAST (e-value ≤ 0.001) against the proteins from IPCR and vice versa to compare the annotation of the two assemblies. 24,578 proteins in PICR have a BLAST hit in IPCR and 22,970 of these proteins have a functional annotation assigned from the top BLAST hit against the Swiss-Prot database, compared to 23,420 proteins in IPCR with a BLAST hit in PICR.

Analysis of the 236 proteins in IPCR, but not PICR, showed that most were not functionally annotated, or were duplicates or isoforms of genes in PICR. Some proteins unique to the IPCR assembly include the protease carboxypeptidase Q (Cpq), the histone H3 threonine kinase Haspin (Gsg2), the antioxidant Sulfiredoxin-1 (Srxn1), and the possible ortholog of DNA-directed RNA polymerase III subunit RPC9 (CRCP). Analysis of the 367 proteins in PICR, but not IPCR, showed that about half were not functionally annotated. Proteins of interest unique to the PICR metassembly include posphatidylglycerophosphate (pgp or pgs1) which is involved in phospholipid biosynthesis in mammalian cells [Kawasaki et al., 1999], and two DNA repair related proteins, breast cancer type 1 susceptibility protein (Brca1) and non-homologous end-joining factor 1 (NHEJ1). In addition, Bcl-2-like protein 10 (Bcl2l10), a signaling molecule involved in apoptosis, and stress-associated endoplasmic reticulum protein 1 (Serp1) are both in PICR but not IPCR. MicroRNAs targeting these two proteins in CHO cells have been developed [Jadhav et al., 2013].

2.4.3 The PICR metassembly has more contiguous genes and non-coding regulatory elements—

In the previous genome assemblies, many genes were fragmented or separated from their functional genomic elements (e.g., promoters, enhancers, or regions of active or repressed transcription). Thus, efforts to define the chromatin states of genes and their regulatory units were error-prone [Feichtinger et al., 2016]. We therefore recalculated the chromatin states for the PICR assembly using the ChiPSeq-derived histone mark reads obtained by Feichtinger et al., (2016). In comparison to the previously deduced chromatin states, the emission profile of the new chromatin states matched better with those obtained for the well assembled human epigenome [Kundaje et al., 2015] (Figure 2.a).

To test whether continuity of genes and their regulatory regions is improved in the PICR metassembly, we extracted a shortlist of 1,538 mitochondria-associated genes, localized to 1,654 sites in the mouse genome. We mapped the sequences between the mouse transcription start site (TSS) to transcription end site (TES) against the PICR metassembly, the RefSeq assembly, and the CSA [Lewis et al., 2013], [Brinkrolf et al., 2013]. Genes were called present if both the TSS and TES were found on the same scaffold. Due to the high variance in UTRs across species, few genes were identified (Figure 2.b), demonstrating the importance of a species-specific genome. We subsequently searched for both the start and end of the coding sequences on the same scaffold (Figure 2.c). Of the complete genes found in PICR (1,011), 85% were annotated and localized to 900 unique locations. The corresponding sequences in PICR were elongated to include UTRs, 5 kb upstream, and 1.5 kb downstream, to capture potential regulatory regions such as promoters or repressive elements. These elongated sequences were mapped against the previously published Chinese hamster genomes [Lewis et al., 2013], [Brinkrolf et al., 2013] and again checked for presence on a single scaffold (Figure 2.d).

Several genes had their elongated sequence in properly assembled in earlier assemblies, despite having the coding sequence on a single scaffold in each of the three assemblies (Supplementary Table S6). Examples for three genes, Rab4b, a member of the Ras family of oncogenes, the mitochondrial ribosome protein MRPL27, and TIMM50, a translocase responsible for targeting proteins into the mitochondria, are shown. In all cases, the scaffold in the CSA assembly contained histone marks for active transcription or a genic enhancer, but lacked flanking enhancers and promoter regions. In the new assembly, these are now correctly annotated (Figure 2.e). The correct assembly of coding and non-coding regions is of increasing importance to better understand their regulatory function and enable engineering applications. A browser with all PICR scaffolds, the preliminary annotation and the chromatin states throughout a batch culture is available at http://cgr-referencegenome.boku.ac.at/jb/.

2.5 Pervasive gaps are filled by SMRT sequencing

The RefSeq assembly [Lewis et al., 2013] contains 166,152 gaps with total length of 58.8 Mb, representing 2.5% of the entire genome. The PICR metassembly has eliminated most gaps, with only 3,238 remaining (Figure 3.a). These gaps account for 2.9 Mb, or 0.1%, of the genome. By aligning the RefSeq assembly to PICR using MUMmer3.0 [Kurtz et al., 2004], we identified missing sequence for 125,812 (76%) of the RefSeq gaps (Figure 3.b and Methods). Sequence for a subset of RefSeq gaps was not identified in the PICR metassembly. Of this subset, 90% could not be unambiguously identified because the flanking fragments did not both align to the new assembly, likely due in part to misassemblies in the RefSeq genome (Figure 3).

The elimination of most gaps in the PICR metassembly enables more accurate and complete genome editing and genomic analyses, since 2,252 genes in the PICR metassembly had their RefSeq assembly gaps filled. We called variants from whole-genome resequencing data for 13 representative resequenced CHO cell lines [Lewis et al., 2013], [Feichtinger et al., 2016] to identify genes that have newly discovered mutations in the RefSeq coding gaps. Each sample has ~300 mutations in coding gaps, 90% of which are SNPs (Supplementary Table S7). Across 13 cell lines, 885 novel variants in coding gaps were found in 134 genes (Figure 3.c).

Gene classes with the highest gap filling success included genes associated with protein binding, RNA binding, and transcription (Supplementary Figure S6), including genes containing zinc finger motifs and ribosomal genes. Previously, such genes were replete with gaps due to their conserved domains shared across many other genes in the genome. We further explored which genes had coding mutations in their filled gaps. The top GO terms for these 225 genes are also enriched in DNA binding and transcription (Supplementary Figure S7). In summary, the gaps in the previous assembly could potentially confound genomic studies in CHO, especially those involving mutations associated with DNA or RNA binding, including transcription factors.

2.5.1 An important mutation in Xylt2 is found within a filled sequence gap—

Beyond their importance in biopharmaceutical production, CHO cells were fundamental to cell biology and biochemistry research for many decades. For example, genetic screens of many CHO cell lines were used to identify glycosyltransferases [Stanley, 2014], [Patnaik and Stanley, 2006], [Maeda et al., 2006], [Zhang et al., 2006] and genetic mapping efforts were deployed to identify causal mutations. The pgsA745 cell line [van Wijk et al., 2017] has been used for decades in the glycobiology field due to its deficiencies in glycosaminoglycan synthesis [Esko et al., 1985], due to a truncation of the xylosyltransferase 2 (Xylt2) protein [Cuellar et al., 2007]. However, upon variant calling from whole-genome resequencing data for the pgsA745 cell line [van Wijk et al., 2017] using the RefSeq assembly, we failed to identify the causal mutation, while a G->T SNP encoding a premature stop codon was found in exon 1 of Xylt2 when using the PICR genome assembly (Figure 3.d). This mutation was previously missed since the RefSeq assembly has a gap of 447 bp that spanned the first exon on scaffold NW_003613846.1. However, this gap was filled in PICR enabling the identification of the mutation. Thus, filling of gap sequence adds a valuable improvement to genomic studies, including the identification of causal variants in CHO cell lines.

3 Discussion

For 60 years, CHO cells have been invaluable for biomedical research and fundamental to the study of several biological processes, such as glycosylation [Goh et al., 2014] and DNA repair [Thompson et al., 1987]. In addition, for >30 years, they have been the host cell of choice for production of most biotherapeutics. While the aforementioned research was carried out without genomic resources, new opportunities are arising with published CHO genome sequences [Xu et al., 2011], [Brinkrolf et al., 2013], [Lewis et al., 2013], [Yusufi et al., 2017]. However, the draft nature of these genome sequences pose challenges for many applications. Here we present a major step forward in further facilitating the adoption of cutting-edge technologies for cell line development and engineering.

The primary outcome here is a substantially improved reference genome sequence for the Chinese hamster. Specifically, the N50 of the PICR metassembly is 13x the length of the RefSeq assembly N50, and we reduced the number of scaffolds to 1/29 the number in RefSeq. Furthermore, we demonstrated that the initial PICR assembly only had one detected misassembly, while the RefSeq assembly had at least 24 >1 Mb scaffolds with cross-chromosome misassemblies (Supplementary Figure S2). Finally, we eliminated more than 95% of the gap sequence in the current RefSeq assembly, and provide a more complete and contiguous view of the genomic sequence of the Chinese hamster.

Various aspects of the genome assembly were improved by merging the different datasets and data types. First, merging the Illumina reads from two different genome sequencing efforts resulted in a higher quality genome than the starting assemblies. Second, further improvements of the assembly attributes were obtained by merging the single-platform assemblies. Previously, assembly merging with Metassembler was found to modestly improve the starting assemblies [Bradnam et al., 2013]. Here, we obtained large gains in the N50, with the PICR metassembly being ~4x more contiguous than the starting assemblies.

Medium and longer scaffolds were successfully merged, thus, reducing the number of N50 and N90 scaffolds. However, by including Illumina-based assemblies, many short scaffolds remained, as seen in the lower median scaffold length in the PICR metassembly compared with the curated PacBio SMRT assembly. The merged assembly thus benefited both from the longer reads from the PacBio SMRT contigs, and the longer scaffolds from the large insert size libraries used for the Illumina assemblies. It is anticipated that the use of optical mapping and chromatin interaction mapping [Bickhart et al., 2017] would further extend the scaffolds and span large repeat regions, resulting in more complete chromosomal maps for the Chinese hamster.

Despite the absence of genomic resources, CHO-based bioprocessing advanced substantially for ~30 years. Massive improvements in protein titer were predominantly achieved through media and process optimization. Systematic optimization of CHO cell lines itself has lagged behind *E. coli* and *Pichia pastoris* and has only recovered traction with the comparatively late release of draft genomes. The availability of genomic data now enables improved control over product quality and more predictable culture phenotypes. For example, more contiguous and complete sequences will facilitate the identification of sites for targeted integration of transgenes, enabling more reproducible productivity across clones [Lee et al., 2015b] and reducing the burden of stability testing. In addition, the elimination of gap sequence regions enables the improved identification of genomic variants and design of genome editing tools. Furthermore, by sequencing through repetitive elements, endogenous retroviral elements can be deleted. This could substantially reduce the retroviral particles secreted in mammalian cell culture [Wheatley, 1974], [Anderson et al., 1991], increase biopharmaceutical safety, and decrease the burden of adventitious agent testing and purification. Comparable efforts have successfully cleaned up similar elements in the porcine genome [Yang et al., 2015].

The full benefit of this more contiguous genome is apparent as novel genome editing tools are applied to control cell phenotypes. These include efforts to delete larger tracts of sequence, including genes, promoters and other regulatory elements using paired gRNAs that remove the entire sequence rather than only introducing frameshifts [Schmieder et al., 2017]. Thus, genes can be removed or promoters can be replaced with synthetic or inducible elements. Furthermore, with more complete regulatory element sequences, one could use CRISPRa/i to regulate gene expression levels. Finally, tools can be deployed that modify the methylation of endogenous promoters to activate or silence gene expression [Morita et al., 2016], [Vojta et al., 2016]. Overall, these strategies enhance our control over cell phenotype. As demonstrated, these precision engineering tools are highly dependent on the availability of a contiguous and well-assembled genome, as presented here, to the entire scientific and industrial community.

6 Materials and Methods

6.1 Sequencing

6.1.1 Illumina sequencing—Short read data from Chinese hamster liver tissue were generated using Illumina’s sequencing technology in two previously published studies. These included chromosome separated paired-end libraries and mate-pair short read data

[Brinkrolf et al., 2013], and whole genome libraries with different insert sizes [Lewis et al., 2013]. The size and coverage of sequencing libraries are in Supplementary Table S8.

6.1.2 Pacific Biosciences SMRT sequencing

Preparation of Chinese hamster tissue: Five female Chinese hamsters (strain 17A/gy) were raised under certified conditions. At 10 weeks of age, the individuals were euthanized by CO2 asphyxiation, and verified by puncture wound to the abdomen. Livers were removed and cut into multiple pieces, flash frozen in liquid nitrogen, and stored at −80C until further processing.

High molecular weight (HMW) genomic DNA extraction: HMW genomic DNA extraction and purification from randomized liver samples was performed using MagAttract HMW DNA Kit (Qiagen Inc., Venlo, Netherlands) as per manufacturer’s instructions. HMW DNA was confirmed using a Fragment Analyzer (Advanced Analytical Technologies Inc., Ankeny, IA).

Single molecule real-time sequencing (SMRT) library preparation from genomic DNA samples: HMW DNA (10 μg aliquots) were converted to SMRTbell templates using the Pacific Biosciences RS DNA Template Preparation Kit 1.0 (Pacific Biosciences, Menlo Park, CA) as per manufacturer’s instructions. In summary, samples were end-repaired and ligated to blunt adapters. Exonuclease treatment was performed to remove unligated adapters and damaged DNA fragments. Samples were purified using 0.6x AMPureXP beads (Beckman Coulter Inc., Brea, CA). The purified SMRTbell libraries were eluted in 10 μl elution buffer. Eluted SMRTbell libraries were size-selected on the BluePippin (Sage Science Inc., Beverly, MA) to eliminate library fragments below 5 kb. Final library quantification and sizing was carried out on a Fragment Analyzer (Advanced Analytical Technologies Inc., Ankeny, IA) using 1 μl of library. SMRTbell templates were aliquoted, shipped, and prepared for sequencing at the University of Delaware Sequencing & Genotyping Center and the Johns Hopkins University Deep Sequencing and Microarray Core.

SMRT sequencing on the Pacific Biosciences RSII: The amount of primer and polymerase required for the binding reaction was determined using the SMRTbell concentration and library insert size. Primers were annealed and polymerase was bound to SMRTbell templates using the DNA/Polymerase Binding Kit P5 and P6 (Pacific Biosciences, Menlo Park, CA). Sequencing was performed using DNA sequencing reagent C3 and C4 (Pacific Biosciences, Menlo Park, CA), with Pacific Biosciences RSII sequencers and SMRT Cell V3 (Pacific Biosciences, Menlo Park, CA) at the University of Delaware Sequencing & Genotyping Center (DBI) and the Johns Hopkins University Deep Sequencing and Microarray Core (JHU). RSII loading efficiency was optimized for each individual library utilizing a standardized titration protocol. Over the course of the project, data capture time for the sequencing runs was initially set at 4 hours. This was extended to 6 hours following software upgrades.

SMRT data metrics: The two sequencing centers ran a total of 202 SMRT cells (92 DBI, 110 JHU). 65 SMRT cells were run using P5/C3 chemistry, while 137 SMRT cells were run using P6/C4 chemistry. After filtering and adapter trimming, a total yield of 107.45 Gb was generated from 13.49 million sequence reads, or approximately 45x coverage of the 2.4 Gb genome. Mean read length calculated from all generated reads was 11.55 kb. N50 read length calculated from all generated reads was 15.9 kb.

6.1.3 SMRT read error-correction—Prior to assembly, SMRT reads were error-corrected (SMRT reads have 15% errors pre-correction). As insufficient SMRT coverage was obtained for self-correction of SMRT reads, we used Illumina paired-end reads [Brinkrolf et al., 2013], [Lewis et al., 2013] for SMRT read error correction. The reads were preprocessed with the ALLPATHS-LG error-correction module for fragment libraries [Gnerre et al., 2011]. The reads from the same pair are joined, possible gaps are filled and the read is error-corrected, resulting in a longer, single, error-free read. Two different tools for error correction were tested with different parameters, proovread [Hackl et al., 2014] and LoRDEC [Salmela and Rivals, 2014]. The tools were tested separately and in combination. Best results were achieved when, in the first step, proovread was run on the initial reads with a single iteration on the complete Illumina reads. All Illumina reads were mapped to all SMRT reads (allowing for multi-mappings) using the modified version of BWA in the proovread tool. Then, the bam2cns algorithm in proovread was applied to correct the reads based on majority decision of the Illumina mappings. In the second step, the proovread-results were further processed with LoRDEC. Using the corrected reads, LoRDEC created a de Bruijn graph from the Illumina reads, mapped the nodes (k-mers of size 85) to the SMRT reads and corrected the unmapped regions following a path in the de Bruijn graph. See Supplementary text and Figures S10–S11 for more details.

6.2 Genome size estimation

Genome size was estimated by the k-mer frequency of the Illumina read data, using (1) all Illumina whole-genome paired-end libraries with an insert-size of 500, (2) the libraries with an insert size of 800, and (3) a combination of sets one and two. Jellyfish [Marçais and Kingsford, 2011] was used to count the frequencies for k-mers of 17, 25 and 31. The GCE tool [Liu et al., 2013] was used to estimate the genome size.

6.3 Genome assembly

The final genome assembly was conducted in two stages. In the first stage, four different assemblies were built with different tools and library combinations, using the raw Illumina or the error-corrected SMRT reads. In the second stage, the four primary assemblies were iteratively merged in four different orders using the Metassembler tool (see Supplementary Figure S8) [Wences and Schatz, 2015]. Various quality metrics (Figure 1) were used to assess the quality of the eight assemblies (four primary assemblies and four metassemblies). These metrics were further used to rank the assemblies and select the assembly with the best overall rank. Finally, the PICR was used as the reference assembly following polishing by correcting the single detected misassembly and minor gap filling from the PIRC assembly (see Supplementary material).

6.3.1 Primary assemblies

Assembly 1: Illumina-based chromosome sorted assembly (CSA): The ten chromosome sorted libraries were assembled separately, including the whole-genome mate-pair library to each assembly, with the ALLPATHS-LG tool [Gnerre et al., 2011]. The resulting scaffolds were filtered for possible contaminations of other chromosomes. The final assembly was previously published [Brinkrolf et al., 2013], and is available at the NCBI assembly archive (accession: GCA_000448345.1).

Assembly 2: Whole-genome Illumina assembly (RefSeq): The RefSeq reference genome of the Chinese hamster is based on the SOAPdenovo2 [Luo et al., 2012] assembly [Lewis et al., 2013]. The different paired-end and mate-pair Illumina libraries were assembled using SOAPdenovo2 [Luo et al., 2012]. The assembly is accessible at the NCBI assembly archive (accession: GCA_000419365.1).

Assembly 3: Whole-genome and chromosome sorted assembly (Illumina): Sequence data originating from the published chromosome-sorted Illumina libraries and whole-genome Illumina libraries [Brinkrolf et al., 2013], [Lewis et al., 2013] were combined and assembled with the ALLPATHS-LG tool (version 51927) [Gnerre et al., 2011].

Assembly 4: Pacific Biosciences SMRT assembly (PacBio SMRT): The ALLPATHS-LG tool was used to merge and error-correct overlapping paired-end Illumina reads, and these reads were further extracted and converted to fasta format to aid in the SMRT error-correction process. The error-corrected SMRT reads were assembled following the HGAP-3 pipeline [Chin et al., 2013] without the error-correction step. For better control over the workflow, we used the customizable makefile-based smrtmake workflow [smr, 2016].

6.3.2 Merged assemblies—The four primary assemblies were iteratively merged with the Metassembler [Wences and Schatz, 2015] tool. For each metassembly, one assembly is selected as the primary assembly. The scaffolds of a second assembly are subsequently mapped to the primary scaffolds using NUCmer [Kurtz et al., 2004]. A CE-statistic, based on the distance of mate-pair reads, is computed for both assemblies. Primary scaffolds are joined and gaps are closed with the sequence of the second assembly. If the CE-statistics of the primary scaffolds indicate potential errors, the sequence in this area is replaced by the sequence in the second assembly. The resulting scaffolds are then used as primary scaffolds for the next iteration. Changes to the default parameters were applied for the merging step (asseMerge). The minimal range for finding links between scaffolds was raised to 50,000 and the minimal coverage of the secondary scaffold was lowered to 1. The minimal gap size for closure was lowered to 1 (asseMerge -e 50000 -L 1 -t 1). The order in which the assemblies are merged influences the result of the final metassembly, and four different orders were tested (see table 2).

6.4 Chromosome assignment

Scaffolds were assigned to chromosomes using chromosome-sorted library coverage, computed for 1 kb regions. Specifically, for each 1 kb region of each scaffold, the coverage of each chromosome library was computed. If at least 90% of the 1 kb region of a scaffold showed a normalized coverage between 0.5 and 2 of the same chromosome, the scaffold was assigned to this chromosome. Scaffolds assigned to the pooled chromosome 9 and 10 library and all unassigned scaffolds were mapped to the mouse genome using NUCmer [Kurtz et al., 2004]. Yang *et al.* [Yang et al., 2000] and Wlaschin *et al.* [Wlaschin and Hu, 2007] described the localization of the hamster chromosomes on the mouse chromosomes. This information was used to assign the mapped scaffolds to a chromosome by manually comparing the mapped position with the localization from Yang and Wlaschin.

6.5 Gene prediction and annotation

6.5.1 Gene annotation—Annotation of the PICR and IPCR metassemblies was completed using Maker v2.31.8. Chinese hamster ESTs (40 million reads) from SRA (SRR823966) were assembled using Trinity v2.0.6 [Grabherr et al., 2011]. The resulting transcripts were aligned to the previously published hamster transcriptome assembly [Lewis et al., 2013], which had used Trinity v. r2011-08-20. NUCmer [Kurtz et al., 2004] was used for the alignment with default parameters. 91,027 transcripts were found in both transcriptomes and used as evidence for gene prediction within Maker. In addition, all proteins from the 2014 RefSeq annotation (GCF000419365.1) of the hamster genome were used as evidence. A comparison to mouse, rat and the RefSeq hamster annotations are provided in Supplementary Table S9 and Figures S12, S13, S14, S15, S16 and S17.

Repeat masking was done within the Maker pipeline. To identify repeat regions, we used Repeat-Masker version open-4.0.6 [Smit et al., 2015]. Dfam v2.0 (2015-09-23), a database of eukaryotic transposable element and other repetitive DNA sequence alignments, and the RepeatMasker database (release 2015-08-07, derived from RepBase v20.08). Once repeat masking was completed, BLAST v2.2.28 [Camacho et al., 2009] and Exonerate v2.2.0 [Slater and Birney, 2005] were run within Maker for evidence based alignments and SNAP v2006-07-28 [Korf, 2004] and Augustus v3.2.2 [Keller et al., 2011] for *ab initio* gene prediction.

The resulting annotation only included genes with more than one type of evidence supporting the prediction, i.e. both an *ab initio* prediction and an evidence-based alignment. Functional annotation of Maker's output was done as described in "Support Protocol 3: Assigning putative gene function" of "Genome Annotation and Curation Using MAKER and MAKER-P"[Campbell et al., 2014]. BLAST was used (e-value <0.001) for each predicted gene against the Swiss-Prot release-2016-02 database, where the best hit was used as the putative function of that gene.

Further comparisons were made based on the NCBI annotation pipeline, as detailed in the supplementary text and Supplementary Tables S10–S11.

6.6 Gap analysis

6.6.1 Identification of filled-gap sequence—We aligned the Chinese hamster RefSeq genome sequence to the PICR genome sequence using NUCmer [Kurtz et al., 2004] to identify gap sequence (see Supplementary Figure S9). Briefly, NUCmer clusters a set of maximally exact matches as an anchor, and then extends alignments between the clustered matches. Gaps are represented using letters N in the genome, and since they differ between the RefSeq and PICR metassembly, the MUMmer alignments stop at gaps larger than 89 bp. This means that if two fragments that flank both ends of a gap are found on the same PICR scaffold in the same orientation, the sequence between the two matches correspond to the sequence of the gap. Since sequence errors may occur near gap regions, we consider matches flanking a gap if the distance between fragment and gap is less than 10 bp. When the gap is shorter than 90 bp, MUMmer clusters the gap together with the two matches on both ends and only reports the merged long fragment as mapping. In this case, we first used the show-aligns method in MUMmer to output the alignment details between the RefSeq hamster and PICR, and then we extracted the corresponding gap sequence by parsing the alignments. The gap analysis was performed using PICR and RefSeq hamster assembly, except the gap in Xylt2 gene which was visualized using the RefSeq CHO-K1 genome assembly.

6.6.2 Identification of genes with gaps and mutations—We called variants in whole-genome resequencing data from various CHO cell lines [Lewis et al., 2013], [Feichtinger et al., 2016], [van Wijk et al., 2017]. GATK v3.5 [McKenna et al., 2010], [DePristo et al., 2011], [Auwera et al., 2013] was used with the GATK manual-recommended parameters. We also called variants using the reads from the RefSeq assembly project [Lewis et al., 2013] to identify and filter false positive variants. Pybedtools [Quinlan and Hall, 2010], [Dale et al., 2011] identified genes with gaps in their coding regions. GO term analysis was performed using DAVID [Huang et al., 2009b], [Huang et al., 2009a].

First, to identify classes of genes with gaps in the RefSeq assembly, we mapped all hamster genes to their human homologs. The functional enrichment analysis for all the 2,252 genes with coding gap regions was performed using the human genes with hamster homologs as the background. Second, to identify classes of genes with a higher frequency of mutations in gaps, we looked for over-representation of the 132 genes with variants in coding gaps, while using the 2,252 gap-filled genes as background. GO terms with a p-value smaller than 0.01 were visualized using REViGO [Supek et al., 2011]. Code for the gap analysis can be acquired here https://github.com/LewisLabUCSD/assembly_gaps.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge George Yarganian for providing hamster tissue. Valerie Schneider and Françoise Thibaud-Nissen from NCBI aided in running a ‘light’ version of the NCBI annotation pipeline on the assemblies. This work was supported with generous funding from Biogen, Genentech, Eli Lilly and Company, Dublin City University, University of Queensland, and University of Tokushima. Grant support was provided by the Novo Nordisk Foundation to the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517 and NNF16CC0021858), NIGMS (R35 GM119850), a FISP fellowship from UC San Diego to SL, and NSF (NSF1144726, NSF1412365, NSF1539359, NSF1736123) to the University of Delaware. HD is supported by the EU Horizon 2020 MSCA ITN grant No 642663. OR, SG, IH, VJ, KB, AG and NB received support from the Austrian Center of Industrial Biotechnology acib, a COMET K2 competence center of the Austrian Research Promotion Agency. Bioinformatics support by the BMBF-funded project “Bielefeld-Gießen Center for Microbial Bioinformatics—BiGi (Grant Number 031A533)” within the German Network for Bioinformatics Infrastructure (de.NBI) is gratefully acknowledged. SP and the computational infrastructure provided by the University of

Delaware Center for Bioinformatics and Computational Biology Core Facility is supported through Delaware INBRE, NIGMS (P20 GM103446).

References

1. smrtmake: Hackable smrtpipe workflows using makefiles instead of smrtpipe.py (commit 29a9c75). original-date: 2014-06-13T22:32:12Z.

2. Anderson KP, Low MA, Lie YS, Keller GA, Dinowitz M. Endogenous origin of defective retroviruslike particles from a recombinant Chinese hamster ovary cell line. Virology. 1991; 181(1): 305–311. [PubMed: 1704658]

3. Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. 2013; 11(1110):11.10.1–11.10.33.

4. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nature Genetics. 2017; 49(4):643–650. [PubMed: 28263316]

5. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience. 2013; 2:10. [PubMed: 23870653]

6. Brinkrolf K, Rupp O, Laux H, Kollin F, Ernst W, Linke B, Kofler R, Romand S, Hesse F, Budach WE, et al. Chinese hamster genome sequenced from sorted chromosomes. Nature Biotechnology. 2013; 31(8):694–695.

7. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST plus: architecture and applications. BMC Bioinformatics. 2009; 10(421):1. [PubMed: 19118496]

8. Campbell MS, Holt C, Moore B, Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinform. 2014:48.

9. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 2013; 10(6):563–569. [PubMed: 23644548]

10. Cuellar K, Chuong H, Hubbell SM, Hinsdale ME. Biosynthesis of chondroitin and heparan sulfate in chinese hamster ovary cells depends on xylosyltransferase ii. Journal of Biological Chemistry. 2007; 282(8):5195–5200. [PubMed: 17189266]

11. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible python library for manipulating genomic datasets and annotations. Bioinformatics. 2011; 27(24):3423–3424. [PubMed: 21949271]

12. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. Nature Genetics. 2011; 43(5):491–498. [PubMed: 21478889]

13. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009; 323(5910):133–138. [PubMed: 19023044]

14. Esko JD, Stewart TE, Taylor WH. Animal cell mutants defective in glycosaminoglycan biosynthesis. Proceedings of the National Academy of Sciences. 1985; 82(10):3197–3201.

15. Feichtinger J, Hernández I, Fischer C, Hanscho M, Auer N, Hackl M, Jadhav V, Baumann M, Krempl PM, Schmidl C, et al. Comprehensive genome and epigenome characterization of cho cells in response to evolutionary pressures and over time. Biotechnology and Bioengineering. 2016; 113(10):2241–2253. [PubMed: 27072894]

16. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(4):1513–1518. [PubMed: 21187386]

17. Goh JS, Liu Y, Chan KF, Wan C, Teo G, Zhang P, Zhang Y, Song Z. Producing recombinant therapeutic glycoproteins with enhanced sialylation using CHO-gmt4 glycosylation mutant cells. Bioengineered. 2014; 5(4):269–273. [PubMed: 24911584]

18. Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. Long-read sequence assembly of the gorilla genome. Science. 2016; 352(6281):aae0344. [PubMed: 27034376]

19. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011; 29(7):644–52.

20. Hackl T, Hedrich R, Schultz J, Förster F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics. 2014; 30(21):3004– 3011. [PubMed: 25015988]

21. Holt C, Yandell M. MAKER2: an annotation pipeline and genome database management tool for second-generation genome projects. BMC Bioinformatics. 2011; 12(1):491. [PubMed: 22192575]

22. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009a; 37(1):1–13. [PubMed: 19033363]

23. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature Protocols. 2009b; 4(1):44– 57. [PubMed: 19131956]

24. Jadhav V, Hackl M, Druz A, Shridhar S, Chung CY, Heffner KM, Kreil P, Betenbaugh M, Shiloach J, Barron N, et al. CHO microRNA engineering is growing up: Recent successes and future challenges. Biotechnology Advances. 2013; 31(8):1501–1513. [PubMed: 23916872]

25. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017; 546(7659):524–527. [PubMed: 28605751]

26. Kawasaki K, Kuge O, Chang SC, Heacock PN, Rho M, Suzuki K, Nishijima M, Dowhan W. Isolation of a Chinese hamster ovary (CHO) cDNA encoding phosphatidylglycerophosphate (PGP) synthase, expression of which corrects the mitochondrial abnormalities of a PGP synthase-defective mutant of CHO-K1 cells. Journal of Biological Chemistry. 1999; 274(3):1828–1834. [PubMed: 9880566]

27. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011; 27(6):757–763. [PubMed: 21216780]

28. Kildegaard HF, Baycin-Hizal D, Lewis NE, Betenbaugh MJ. The emerging CHO systems biology era: harnessing the ’omics revolution for biotechnology. Current Opinion in Biotechnology. 2013; 24(6):1102–1107. [PubMed: 23523260]

29. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004; 5:59. [PubMed: 15144565]

30. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518(7539):317–30. [PubMed: 25693563]

31. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biology. 2004; 5(2):R12. [PubMed: 14759262]

32. Lee JS, Grav LM, Lewis NE, Faustrup Kildegaard H. CRISPR/Cas9-mediated genome engineering of CHO cell factories: Application and perspectives. Biotechnology Journal. 2015a; 10(7):979–994. [PubMed: 26058577]

33. Lee JS, Kallehauge TB, Pedersen LE, Kildegaard HF. Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. Scientific Reports. 2015b; 5:srep08572.

34. Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O’Brien E, Bordbar A, Roth AM, Rosenbloom J, Bian C, et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. Nature Biotechnology. 2013; 31(8):759–765.

35. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W, et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. 2013 arXiv: 1308.2012 [q-bio],. arXiv: 1308.2012.

36. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012; 1(1):18. [PubMed: 23587118]

37. Maeda Y, Ashida H, Kinoshita T. Cho glycosylation mutants: Gpi anchor. Methods in enzymology. 2006; 416:182–205. [PubMed: 17113867]

38. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011; 27(6):764–770. [PubMed: 21217122]

39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome research. 2010; 20(9):1297–1303. [PubMed: 20644199]

40. Morita S, Noguchi H, Horii T, Nakabayashi K, Kimura M, Okamura K, Sakai A, Nakashima H, Hata K, Nakashima K, et al. Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. Nature Biotechnology. 2016; 34(10):1060–1065.

41. Patnaik SK, Stanley P. Lectin-resistant cho glycosylation mutants. Methods in enzymology. 2006; 416:159–182. [PubMed: 17113866]

42. Quinlan AR, Hall IM. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. [PubMed: 20110278]

43. Richelle A, Lewis NE. Improvements in protein production in mammalian cells from targeted metabolic engineering. Current Opinion in Systems Biology. 2017; 6:1–6. [PubMed: 29104947]

44. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014:btu538.

45. Schmieder V, Bydlinski N, Strasser R, Baumann M, Kildegaard H, Jadhav V, Borth N. Enhanced genome editing tools for multi-gene deletion knock-out approaches using paired CRISPR sgRNAs in CHO cells. Biotechnology Journal. 2017

46. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nature Communications. 2016:7.

47. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005; 6(1):31. [PubMed: 15713233]

48. Smit A, , Hubley R, , Green P. RepeatMasker Open-4.02015

49. Stanley P. Chinese hamster ovary mutants for glycosylation engineering of biopharmaceuticals. Pharmaceutical Bioprocessing. 2014; 2(5):359–361.

50. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. PloS one. 2011; 6(7):e21800. [PubMed: 21789182]

51. Thompson LH, Salazar EP, Brookman KW, Collins CC, Stewart SA, Busch DB, Weber CA. Recent progress with the DNA repair mutants of Chinese hamster ovary cells. Journal of Cell Science. Supplement. 1987; 6:97–110. [PubMed: 3477565]

52. Tjio JH, Puck TT. Genetics of somatic mammalian cells. The Journal of Experimental Medicine. 1958; 108(2):259–268. [PubMed: 13563760]

53. van Wijk XM, Döhrmann S, Hallström BM, Li S, Voldborg BG, Meng BX, McKee KK, van Kuppevelt TH, Yurchenco PD, Palsson BO, et al. Whole-genome sequencing of invasion-resistant cells identifies laminin α2 as a host factor for bacterial invasion. mBio. 2017; 8(1):e02128–16. [PubMed: 28074024]

54. Vojta A, Dobrinić P, Tadić V, Bočkor L, Korać P, Julg B, Klasić M, Zoldoš V. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Research. 2016; 44(12):5615–5628. [PubMed: 26969735]

55. Walsh G. Biopharmaceutical benchmarks 2014. Nature Biotechnology. 2014; 32(10):992–1000.

56. Wences AH, Schatz MC. Metassembler: merging and optimizing de novo genome assemblies. Genome Biology. 2015; 16:207. [PubMed: 26403281]

57. Wheatley DN. Pericentriolar virus-like particles in Chinese hamster ovary cells. The Journal of General Virology. 1974; 24(2):395–399. [PubMed: 4853583]

58. Wlaschin KF, Hu WS. A scaffold for the Chinese hamster genome. Biotechnology and Bioengineering. 2007; 98(2):429–439. [PubMed: 17390381]

59. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S, et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nature Biotechnology. 2011; 29(8):735–741.

60. Yang F, O'Brien PC, Ferguson-Smith MA. Comparative chromosome map of the laboratory mouse and Chinese hamster defined by reciprocal chromosome painting. Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology. 2000; 8(3):219–227.

61. Yang L, Güell M, Niu D, George H, Lesha E, Grishin D, Aach J, Shrock E, Xu W, Poci J, et al. Genome-wide inactivation of porcine endogenous retroviruses (PERVs). Science. 2015; 350(6264):1101–1104. [PubMed: 26456528]

62. Yusufi FNK, Lakshmanan M, Ho YS, Loo BLW, Ariyaratne P, Yang Y, Ng SK, Tan TRM, Yeo HC, Lim HL, et al. Mammalian systems biotechnology reveals global cellular adaptations in a recombinant cho cell line. Cell Systems. 2017; 4(5):530–542. [PubMed: 28544881]

63. Zhang J, Chen L-L, Sun S, Kudrna D, Copetti D, Li W, Mu T, Jiao WB, Xing F, Lee S, et al. Building two indica rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. Scientific Data. 2016:3.

64. Zhang L, Lawrence R, Frazier BA, Esko JD. CHO glycosylation mutants: proteoglycans. Methods in Enzymology. 2006; 416:205–221. [PubMed: 17113868]

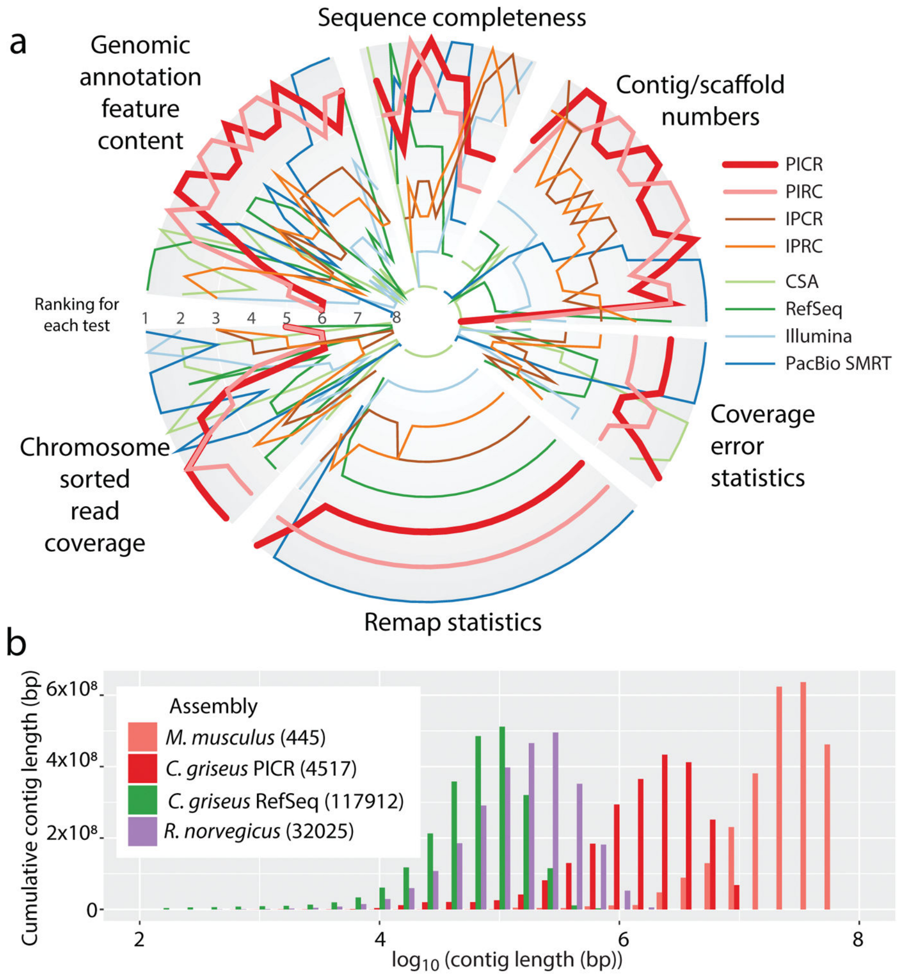

Figure 1.

The PICR assembly ranked against other mammalian assemblies.(a) The PICR assembly was compared to other candidate assemblies of *C. griseus* based on 80 different assembly metrics. This shows for each test how the assemblies compare. The best assembly for each test is plotted on the outer rim, while the worst is near the center. Eighty tests were defined (see Supplementary Table S3) in six different categories. On average, the PICR assembly was the most highly ranked with the PIRC assembly closely following. (b) Weighted histogram of the contig lengths for the PICR assembly (red) compared to the Ensemble mouse (salmon), rat (purple), and the prior Chinese hamster RefSeq assemblies (green).

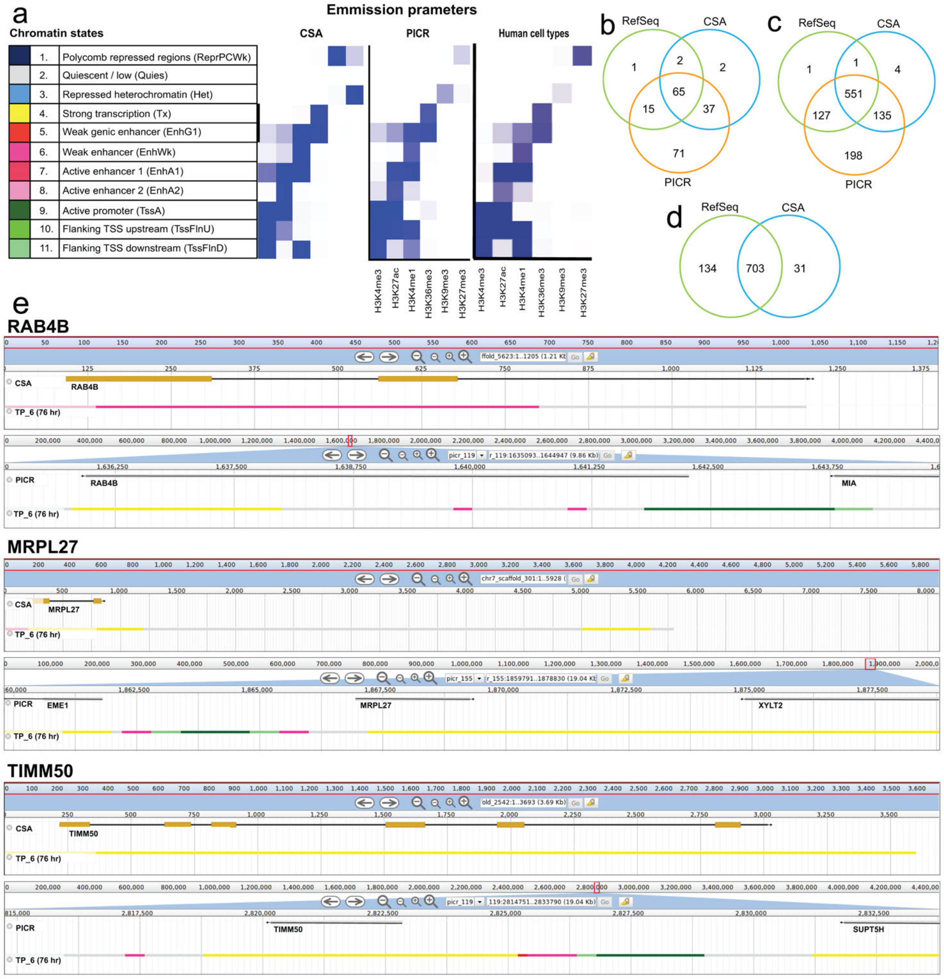

Figure 2.

Importance of correct assembly of genes and non-coding regions. (a) Chromatin states defined by histone marks: Left: histone marks for CSA assembly [Brinkrolf et al., 2013], [Feichtinger et al., 2016]; center: histone marks for PICR assembly; right: histone marks from the Human Epigenome Project [Kundaje et al., 2015] (b) 1,538 genes associated with mitochondria were blasted from TSS to TES against the CSA and RefSeq assemblies. The number of hits completely found on a single scaffold is displayed for each assembly. (c) Mouse coding sequences were blasted against Chinese hamster assemblies from translation start to end. (d) The 1,011 complete genes found in PICR were extended 5 kb upstream and 1.5 kb downstream to include promoters and other regulatory non-coding regions and blasted against existing assemblies. (e) Chromatin states around three genes as found in the previously published CSA-based chromatin state model [Feichtinger et al., 2016] (top for each gene) and the PICR assembly (bottom for each gene), showing promoter and regulatory elements in addition to active transcription.

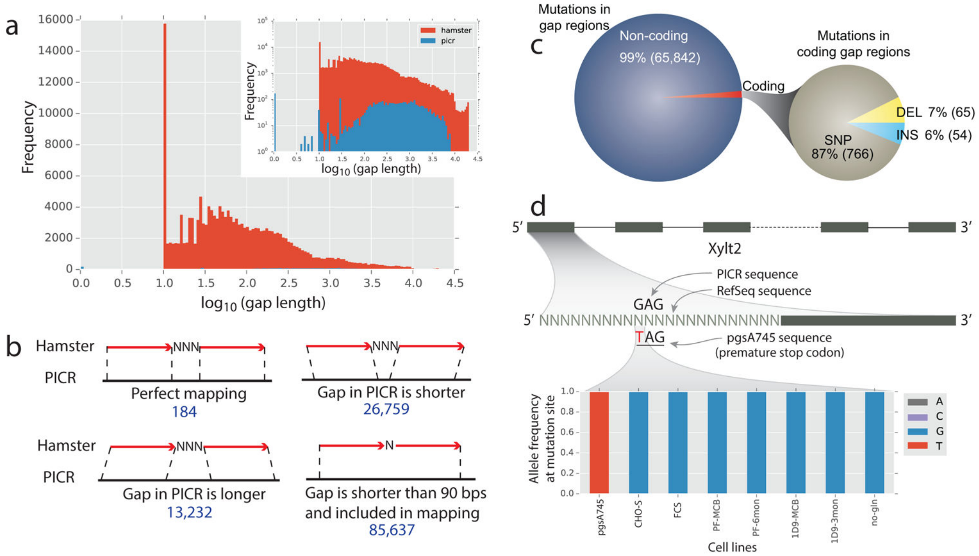

Figure 3.

Important variants are located in sequence gaps in previous assemblies. (a) >95% of sequence gaps were filled in the PICR metassembly (inset shows the log frequency of gaps to highlight the low frequency of PICR gaps not visible in the normal histogram). (b) The missing sequence in gaps in the RefSeq assembly was identified by aligning RefSeq sequence flanking the gaps to the PICR sequence.

(c) Across 13 cell lines, we found 65,842 SNP and indel mutations in the RefSeq gap regions, and 1.3% of these were found in coding regions. (d) A legacy CHO cell line, pgsA745, identified Xylt2 as the glycosyltransferase responsible for the first step in glycosaminoglycan biosynthesis, as this cell line is deficient in glycosaminoglycan biosynthesis. Because of a gap in the RefSeq assembly, only in the new PICR metassembly can the causal variant be identified. A G->T mutation introduces an early stop codon in exon 1, resulting in a loss in Xylt2 activity. The genotype is shown for a variety of CHO cell lines [Lewis et al., 2013], [Feichtinger et al., 2016], [van Wijk et al., 2017], with only pgsA745 showing the early stop codon.

Table 1

Assembly metrics of the Illumina scaffolds and PacBio SMRT curated assembly compared to the previously published assemblies.

	RefSeq (Lewis 2013)	CSA (Brinkrolf 2013)	Pooled Illumina scaffolds	Curated PacBio SMRT contigs
Scaffolds [#]	52,710	28,749	17,373	1,659
Length [Gb]	2.36	2.33	2.39	2.31
Min length [bp]	201	830	898	100,560
Max length [Mb]	8.32	14.66	25.84	16.08
Mean length [kb]	44.78	81.14	137.45	1,394.69
Median length [bp]	363	1,927	2,063	693,156
N50 length [kb]	1,558.30	1,236.52	5,951.71	2,906.73
N50 [#]	450	501	128	223
N90 length [kb]	395.29	180.69	1,003.29	623.9
N90 [#]	1,558	2,251	468	884
Total N gaps [#]	166,152	290,660	110,314	0
Total N [%]	2.49	10.45	2.66	0

Table 2

Four different orders were used to merge the four initial assemblies with the Metassembler tool, where PICR starts with the PacBio SMRT assembly, after which the Illumina assembly is merged into it, followed by the CSA assembly and the RefSeq assembly.

Base assembly	Added in step 1	Step 2	Step 3	Name
PacBio SMRT	**I**llumina	**C**SA	**R**efSeq	**PICR**
PacBio SMRT	**I**llumina	**R**efSeq	**C**SA	**PIRC**
Illumina	**P**acBio SMRT	**C**SA	**R**efSeq	**IPCR**
Illumina	**P**acBio SMRT	**R**efSeq	**C**SA	**IPRC**

Table 3

Assembly metrics of the four merged assemblies.

	PICR	PIRC	IPCR	IPRC
Scaffolds [#]	1,829	1,825	2,317	2,304
Length [Gb]	2.37	2.37	2.36	2.36
Min length [bp]	568	568	915	915
Max length [Mb]	80.58	80.58	66.35	66.35
Mean length [kb]	1,295.21	1,298.43	1,019.33	1,024.64
Median length [bp]	37,019	38,181	13,201	14,241
N50 length [kb]	20,188.72	19,582.71	21,744.88	21,262.79
N50 [#]	32	33	33	34
N90 length [kb]	4,400.57	4,422.38	3,545.61	3,650.27
N90 [#]	121	122	122	122
Total N gaps [#]	3,237	3,250	72,528	72,536
Total Ns [%]	0.12	0.12	1.13	1.13

Table 4

Gene and transcript information from the Maker annotation of the PICR and IPCR genome assemblies.

Assembly		
All genes	**PICR**	**IPCR**
Gene count	24,686	23,410
Transcript count	24,948	23,656
Transcripts per gene	1.01	1.01
Avg. length transcript	17,615.04	18,089.17
Total length transcript	439,460,104	427,917,413
Avg. coding length	1,324.93	1,316.11
Total coding length	33,054,355	31,133,905
Avg. exons per transcript	7.49	7.54
Total exons	186,939	178,277
Complete transcripts		
Transcript count	18,476	17,557
Total exons	138,358	131,262
Incomplete transcripts		
Transcript count	6,472	6,099
Total exons	48,581	47,015