Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Toward product attribute control: developments from genome sequencing

Jong Youn Baik and Kelvin H Lee

Chinese hamster ovary (CHO) cells are important hosts for the production of therapeutic proteins. Recent genome sequencing studies provide an initial baseline of information useful for understanding cell line performance in terms of product quality attributes. However, the lack of a well-established reference genome together with concerns about genome stability have not yet permitted the community to define the detailed relationship between the genome and cell line performance. Emerging efforts to define a new reference genome, together with new data on genome stability, herald an era where cell line's with defined genomes can be combined with defined process parameters to yield product quality attribute control.

Addresses
Department of Chemical and Biomolecular Engineering and Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711, United States

Corresponding author: Lee, Kelvin H (KHL@udel.edu)

Current Opinion in Biotechnology 2014, **30**:40–44

This review comes from a themed issue on **Pharmaceutical biotechnology**

Edited by **Beth Junker** and **Jamey D Young**

http://dx.doi.org/10.1016/j.copbio.2014.05.001

Introduction

Chinese hamster ovary (CHO) cells have emerged as the key platform host for the production of therapeutic proteins. As a result, there is community-wide interest in ways to improve and control CHO cell characteristics as a host for heterologous protein expression. In particular, there has been effort invested on enhancing productivity via cell line development, genetic engineering, metabolic engineering, and bioprocess engineering [1,2] and there is now a growing interest in understanding how to engineer CHO cells and their growth to help define product quality attributes. Well-defined product attribute control will ultimately arise from a detailed understanding of the underlying cell biology which can provide a relationship between genotype (sequence, annotation, and gene expression levels) and quality attributes of the product (e.g. glycosylation). However, unlike other important genomes, the CHO genome has only recently been sequenced [3]. This development has been enabled by new technologies which have significantly reduced the cost required to sequence mammalian genomes. For example, it had cost approximately $300 million to sequence the human genome in 2001 [4].

The advent of so-called second-generation sequencing (SGS) technology enabled relatively low-cost, high-throughput sequencing [4,5]. While there are different SGS technologies, most are characterized by massively parallel sequencing (millions of concurrent reactions) and generation of DNA libraries without vector-based cloning steps (for a detailed review, see [6]). The rates of data generation, compilation, and processing have advanced rapidly with developments in sequencer technologies, computing power, and assembly algorithms [5] which heralds the possibility of linking genome sequence to host cell characteristics including the range of possible product quality attributes. Looking forward, emerging technologies will provide insights not only on the relationship between genome and host cell, but also on epigenetic changes and their impact on cell behavior.

In this review, we summarize recent developments in CHO (and Chinese hamster) genomics studies and the emerging understanding of distinct characteristics of the CHO (and Chinese hamster) genome. We will also discuss future applications of CHO genomics toward bioprocess engineering but limit our discussion to the genome.

Genome sequencing of CHO cells

Traditionally (before SGS), CHO DNA sequencing efforts were limited to a small set of genes (less than 700 sequences found in the GenBank database in 2004) individually isolated and sequenced by the traditional Sanger method [7]. In 2005, Wlaschin et al. sequenced expressed sequence tags (ESTs) from a cDNA library of the CHO DXB11 cell line to create a CHO specific cDNA microarray [7]. The sequencing of 4219 ESTs yielded 2602 unique sequence assemblies, of which 76% were successfully annotated by sequence comparison with other organisms including human, mouse, and rat. In 2007, Partridge et al. reported a full mitochondrial CHO DNA sequence (16.3 kb) which encoded for a subset of the proteins involved in the electron transport chain of mitochondria [8]. SGS technology has changed the landscape for CHO genomics by providing the means to generate sequence information reasonably quickly.

Hammond et al. sequenced the genome of a secreted alkaline phosphatase (SEAP) expressing cell line at low sequencing coverage (~1X) using Illumina technology [9]. The data were presented without assembly or annotation, but provided proof of principle that such sequencing was possible. Short read mapping against the mouse and rat genomes showed that the genome of that cell line had higher homology with the mouse genome than the rat genome [9]. Subsequently the CHO K1 genome was published in 2011 [3]. They reported a genome of 2.45 Gb that harbored 24,383 genes, similar to those of mouse and rat [3]. In an indicator of the rapid pace of the field, two different teams have sequenced the Chinese hamster and one team sequenced six additional CHO cell lines since that first CHO K1 publication [3,10•,11••]. One of the key outcomes of these initial efforts (coupled with important insights gleaned from karyotyping over the years) is that the CHO K1 genome may not be the best reference genome for the community because different CHO cell lines contain different chromosome numbers and have significant sequence variation [12,13••].

Because a CHO-derived cell line may not serve as an effective reference genome, there has been an emphasis on the use of the Chinese hamster genome as the reference genome for the community and there is currently an active effort to annotate Chinese hamster genes and chromosomes. Lewis et al. sequenced and assembled the genome of Chinese hamster with ~128X genome coverage and also reported low coverage (<9X) sequences from five other host cell lines (DG44, CHO-S, CHO-K1 ECACC, and two derivative cell lines of CHO-K1 ECACC) and one production cell line [11••]. The Chinese hamster genome was estimated to be 2.7;Gb and harbor 24,044 genes (Table 1). Short sequence reads from the cell lines (and from CHO K1 [3]) were mapped against a *de novo*-assembled Chinese hamster genome and identified ~3.7 million single nucleotide variations (SNVs) and ~550,000 short insertions/deletions [11••]. Sequence variation analysis also indicated a significant amount of frameshifts and gene duplications and deletions. Brinkrolf et al. physically isolated each chromosome of the Chinese hamster by flow cytometry and then sequenced each chromosome separately to create chromosome specific scaffold sequences [10•]. The Chinese hamster has 22 chromosomes ($n = 10 + 1$), and its average chromosome length is much longer than that of other organisms of similar genome size. For example, the length of Chinese hamster chromosome 1 was estimated to be 563 Mb, whereas human chromosome 1 is 249 Mb [10•,14]. A key outcome of these efforts is the knowledge of all possible glycosylation related enzymes encoded by each of the genomes which helps define the range of certain product quality attributes. However, there is not yet a detailed way to organize this information nor an understanding of the control of expression of relevant genes adequate to relate genome to cell line characteristics.

Table 1

Summary of the CHO-K1 and Chinese hamster genome.

	CHO-K1	Chinese hamster	
		by Lewis et al.	by Brinkrolf et al.
Number of contigs	265,786	458,620	319,162
Number of scaffolds	109,151	287,210	28,764
Total contig length (Mb)	2367	2332	2089
Total scaffold length (Mb)	2447	2393	2333
N50 contig length (kb)	38.3	26.8	11.9
N50 scaffold length (kb)	1116	1545	1245
Percentage of Ns	3.27	2.46	10.45
Sequence coverage	~95X	~128X	~70X
Chromosome information	No	No	Yes

N50 contig (scaffold) size is the length of the smallest contig (scaffold) contributing 50% of the entire genome.

Despite having two Chinese hamster reference genomes, there is now an effort underway by the community to synergize these two distinct datasets and to apply third generation sequencing technology to create a new reference assembly and annotation for the Chinese hamster. All of the relevant sequences that are available in the public domain are hosted at the CHO community website (www.CHOgenome.org, [15•]) with associated data and tools, including homolog information, a genome browser, and BLAST service.

Chromosomal rearrangements and genomic instability

One feature of CHO cells is the significant number of chromosomal rearrangements among different cell lines — rearrangements that can occur over time or during cell line development within the same cell line [12,13••,16] and contribute to issues with cell line (in)stability. While sequence variations, such as SNVs and small insertions/deletions, may affect gene expression (by changing the sequences of mRNAs and proteins), larger structural variations can impact gene expression by varying gene copy numbers (duplication, deletion), creating chimeric genes (translocation), or changing the physical location of genes in the nucleus [17,18,19]. Therefore, knowledge and functional analysis of the sequence and structural variations for any newly developed CHO cell line is important to predict and/or control for desirable phenotypes such as high productivity and high product quality (Figure 1).

Chromosomal level sequence characterization of genomes can be gleaned by either of two approaches: 'finishing' *de novo* assembled scaffolds or 'resequencing' short sequence reads. Finishing is a process by which the order and orientation information is assigned to the scaffolds in a genome and the gaps between scaffolds are filled [20]. Finishing a genome sequence is cost-intensive and labor-intensive because it involves targeted sequencing at contig/scaffold ends and manual alignments of extended