Review

Next-generation sequencing technologies and their potential impact on CHO cell-based biomanufacturing

There is growing interest in the possibility of harnessing detailed, mechanistic understanding of the biology of CHO cells to enhance the use of these cells in the manufacturing of biologics. Among the important questions about CHO cells are issues related to productivity, product quality attributes and stability. The advent of next-generation DNA sequencing technologies provides an opportunity to characterize the genome of various host cells and to link genomic changes to phenotypes. In this review, we discuss some of the current and emerging technologies for genome sequencing, their initial applications to CHO bioprocessing, and provide context relative to other omic approaches.

The recombinant biopharmaceutical market 2012 global sales totaled more than US$125 billion, with mammalian cell line biopharmaceuticals responsible for $86 billion [1]. CHO cells are the host cell platform used to manufacture nearly 80% of the $86 billion of mammalian cell line-produced biopharmaceuticals [1]. The biomanufacturing and cell line development community has made great strides to increase the capacity of CHO cells to produce recombinant proteins by over 100-fold over the past 2–3 decades [2]. These improvements are a result of engineering improvements throughout the bioprocess; from cell line development, to process understanding and control, to purification [2–5]. In addition to enhancements in productivity, there have been improvements in the ability to control product quality attributes [5,6]. Nonetheless, it can be argued that gaps remain in the CHO community’s understanding of the detailed relationship between the cell culture conditions (raw materials, cell banking and expansion, growth in reactors) and stable, predictable outcomes in terms of both cell line productivity and product quality. Accordingly, there has been interest in the application of genomics, and related tools, to gain a better understanding of the biology that is happening inside CHO cells. The emergence of omic technologies over the past 20 years provides great opportunity for the field. However, the application of new and emerging technologies to the study of CHO cells for biomanufacturing invariably comes with a period of learning – to refine techniques, to understand benefits and drawbacks of various approaches, to develop new methods for data analysis – prior to substantial and transformative impact. Partnerships and collaboration among academic and industrial scientists will inevitably shorten the time needed to realize the benefits of the application of omics to biomanufacturing. In this review, we present an overview of some of the newest DNA sequencing technologies. Some of these technologies have already been applied to CHO cells whereas others have not. In cases where the method has been applied to CHO cells, we also discuss some of the relevant literature.
Traditional DNA sequencing

- Traditional sequencing technology
  The original ‘Sanger sequencing method’ was developed in 1977 [7] and some technological improvements have been made since then [8–10]. The basic approach as used today relies on a DNA amplification-based strategy with chain termination with dideoxynucleotides wherein individual molecules will terminate in a specific fluorescent molecule representing each of the nucleotide bases (A, T, G or C). An electrophoretic separation of the molecule and subsequent detection, results in the ability to read a DNA sequence by reading the sequence of fluorophores. Because of limits with electrophoresis and with the chemistry of the reaction, one can only ‘read’ a DNA sequence of a given length before the approach becomes unreliable. For modern versions of traditional sequencing, read lengths can reliably approach 900 base pairs (bp) of DNA. While 900 bp is much shorter than the typical gene or genome, it is long enough to provide the data needed to enable bioinformatic algorithms to assemble the data into gene sequences or into whole genomes (of $3 \times 10^9$ bp).

- Applications of traditional sequencing
  The most common uses of traditional sequencing involve routine analyses that are performed as part of many molecular biology experiments. In the context of genomics, an important contribution of traditional sequencing technology is the ability to sequence genes that are expressed within an organism to facilitate analyses of gene expression. The collection of cDNAs from CHO cells and the subsequent sequencing of expressed sequence tags (ESTs) can allow one to identify possible CHO genes even without a complete CHO genome. This approach was successfully applied to CHO cells and the resulting EST sequences were searched (using the BLAST algorithm) against mouse, rat and human genomes to demonstrate sequence alignment between CHO ESTs and other mammalian genomes [11]. In particular, it was observed that the strongest correlation existed between CHO and mouse sequences, resulting in the application of mouse microarrays to study CHO cells [12]. Such work also ultimately led to the accumulation of enough CHO-specific cDNA sequence information to permit the design of CHO-specific microarrays.

Microarrays were used to explore the relationship between CHO gene expression and phenotypes of interest including high-productivity [13], butyrate treatment effects [14,15] and the apoptosis pathways [16]. For example, microarray analysis of a CHO-K1 suspension culture during lag, exponential and stationary growth phases identified 1400 mRNAs as differentially regulated in stationary phase relative to the culture starting point [17]. Further clustering analysis revealed gene groups with similar expression patterns (e.g., homologous recombination, Jak-STAT signaling pathway, spliceosome). In another study, microarray analysis of an IgG-producing CHO cell line at low temperature and butyrate conditions identified more than 900 differentially expressed genes. Butyrate treatment was observed to increase protein production by inducing cell cycle arrest, which coincides with the down-regulation of many cell cycle control genes. The altered culture conditions resulted in an increased IgG production rate, likely caused by an elevated cellular secretory capacity [15].

Next-generation DNA sequencing

- Next-generation sequencing technologies
  While traditional sequencing provided the technology platform to sequence and assemble the complete human genome in 2001 [18–20], as well as the mouse [21] and rat [22] genomes, the cost of sequencing these genomes was in the billions and the time involved was over a decade [20]. Continued technology development of DNA sequencing methods have substantially reduced the time required to collect sequence data while also reducing the cost. Motivated in part by a public effort to develop technologies to sequence a human genome for $1000$ [20], next-generation sequencing (NGS) technologies have emerged as alternatives to traditional sequencing; they include various new approaches to collecting genome-scale sequence information faster and for a lower cost; however, they also place an increased emphasis on bioinformatics to process and organize the resulting data. Nonetheless, NGS methods are being applied to study biological systems in several different ways. When used to sequence genomic libraries, NGS methods are a powerful method to resequence an organism, sequence organisms whose genome can be compared with a well-defined reference genome, or to collect data to perform a de novo genome assembly. They are also used to perform quantitative measures of changes in gene expression (e.g., mRNA expression) among a number of samples using an approach commonly referred to as RNASeq. Moreover, when these techniques are applied to small RNAs, they are used to catalog and quantify differential expression in noncoding RNAs (ncRNAs) such as miRNAs.

NGS methods are commonly referred to by the name of the company associated with their development (even if that company was subsequently acquired
and the technology rebranded) as shown in Table 1. The two most often used approaches within the CHO community are ‘454’ (or Roche/454) and ‘Illumina’ (or Solexa). It is beyond the scope of this review to describe in detail the nuances of these approaches and the reader is referred to recent articles that describe these methods in more detail [23–27]. Nonetheless, we provide a basic overview of 454 and Illumina before describing applications in biomanufacturing.

454 is based upon pyrosequencing technology wherein the DNA sequence is determined by the emission of light that occurs when a complementary nucleotide is incorporated into a DNA molecule under certain conditions. By monitoring this process over hundreds of thousands of molecules in parallel, 454 takes advantage of multiplexing to accelerate the speed with which DNA sequence information is collected. The method has an average read length of 450–700 bp and can provide up to 1 million reads per run to yield approximately 450–700 Mb of data per run [23,101]. Among the advantages of this approach are that the technology provides high quality data, good genome coverage, and a lower cost (per bp of sequence information) than traditional sequencing. However, the method yields shorter read lengths than traditional sequencing which means that it is important to employ relevant bioinformatic algorithms to assemble DNA sequence information together. Other issues with this approach are that homopolymers are difficult to sequence and there is a decreased throughput of data collection compared with some other NGS methods.

Illumina technology is based upon ‘sequencing by synthesis’ and also takes advantage of parallel processing of DNA sequence strands to generate sequence information. Strands of DNA (from a genomic library) are attached to a surface and nucleotides are added sequentially. After each nucleotide addition, images are captured of the location of specific nucleotide addition based on fluorescence. A series of images can be analyzed to determine the DNA sequence of billions of strands of DNA in a single instrument run. Among the advantages of this approach are the very large amounts of data that can be collected (up to 300 or 600 Gb of data) [302], relatively low error rates and the very low cost of sequencing (on a per bp of sequence information basis). Among the limitations of the current form of the technology are the relatively smaller read lengths (below 250 bp) which makes sequencing through repeat regions, and assembly of information, even more challenging than other NGS methods [26]. A number of bioinformatic tools have been developed to address some of the challenges posed by the shorter NGS reads [28–32].

Applications of NGS

NGS approaches have been applied to biomanufacturing-related questions for a number of years. However, one could argue that the most important advance to the field has been the use of NGS to sequence CHO genomes [33–35] as well as the Chinese hamster [35] and its chromosomes [36]. Because NGS approaches are effective not only at sequencing genomes, but also at monitoring changes in expression profiles of mRNA and miRNA, their application to problems relevant to biomanufacturing began before the CHO and Chinese hamster genomes were made available. An up-to-date list of CHO-based NGS studies are provided in Table 2. For example, there were significant efforts directed at the use of 454 to generate as many genomic reads from CHO as possible to collect and expand knowledge about CHO ESTs for the development of tools to facilitate transcriptome analysis. A total of 400,000 reads from CHO with an average length of 212 bp [37] were aligned with more than 34,000 available ESTs to extend the sequence of 70% of the ESTs an average of 150 bp.

An IgG-producing CHO cell line grown under butyrate conditions was studied with Illumina to

Table 1. Sequencing method statistics.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Read length (bp)</th>
<th>Cost/million bases¹ (US$)</th>
<th>Throughput/run</th>
<th>Run time</th>
<th>Maximum reads/run</th>
<th>Advantage(s)</th>
<th>Drawback(s)</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>450–700</td>
<td>10</td>
<td>700 Mb</td>
<td>23 h</td>
<td>1 million</td>
<td>Moderate read length</td>
<td>High relative cost for NGS, homopolymer sequence error rate, low throughput for NGS</td>
<td>101</td>
</tr>
<tr>
<td>Illumina</td>
<td>36–250</td>
<td>0.10</td>
<td>47–600 Gb</td>
<td>1.5–11 days</td>
<td>1.5–6 billion</td>
<td>High throughput, low cost</td>
<td>Short read length</td>
<td>102</td>
</tr>
<tr>
<td>Traditional</td>
<td>500–900</td>
<td>2400</td>
<td>0.9–2.0 Gb</td>
<td>0.5–2 h</td>
<td>1, 16 or 96</td>
<td>High quality, long read length</td>
<td>High cost, low throughput</td>
<td>104</td>
</tr>
</tbody>
</table>

¹Sequencing technology statistics demonstrate the 100- to 1000-fold improvement of NGS compared with traditional Sanger sequencing. These technologies are commonly used for genomic and transcriptomic CHO studies.

NGS: Next-generation sequencing.
identify genes responsible for enhanced productivity [38]. More than 13,000 CHO genes were sequenced and annotated, using genomic information from similar organisms, and approximately 5000 novel CHO genes were identified and added to their CHO model. In this same study, the transcriptome was analyzed for gene clusters affected by butyrate treatment and it suggested that the down-regulated genes were related to cell cycle check point control, mitotic check point control, and the initiation and elongation phases of DNA replication processes [38]. While these observations demonstrated agreement with other butyrate transcriptomic analyses in literature [14,15,39], the additional knowledge gained may facilitate a better understanding of high productivity phenotypes for cell line development.

Illumina analysis of a secreted alkaline phosphatase-producing line yielded 3.57 million contigs and provided CHO-specific sequence information for 18–19,000 metabolic process, cellular signaling, and transport orthologs [32]. This approach identified nearly 5000 additional CHO genes without a reference CHO genome and demonstrated the possibility of using NGS to sequence an entire genome from a CHO cell line.

Transcriptomic analysis of an IgG-producing CHO line yielded 55 million sequencing reads that were mapped to an existing CHO EST-derived unigene set and several public sequence databases [40]. The transcript abundance varied up to six orders of magnitude, while the coverage across the transcript lengths varied to a far lesser extent [40]. While the sequencing was successful, methods for coefficient of variation reduction related to the use of NGS results for transcript measurements and gene expression were addressed, but not fully resolved.

Transcriptomes from multiple recombinant CHO cell lines under various cultivation conditions were investigated with 454 technology [41]. The findings reinforced the idea that there is a reasonable amount of CHO gene sequence similarity to mouse sequences. The gene transcript levels relevant to the central carbohydrate metabolism and glycosylation pathways were measured, which enabled construction of accurate model pathways. For each section of the N-glycosyl-

### Table 2. CHO-based next-generation sequencing publications.

<table>
<thead>
<tr>
<th>Omics</th>
<th>Cell line(s)</th>
<th>NGS technology used</th>
<th>Focus of the study</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>CHO: SEAP</td>
<td>Illumina</td>
<td>Genes</td>
<td>[33]</td>
</tr>
<tr>
<td>G/T</td>
<td>CHO-K1</td>
<td>Illumina</td>
<td>Genome, glycosylation and viral susceptibility pathway genes</td>
<td>[34]</td>
</tr>
<tr>
<td>G/T</td>
<td>Chinese hamster, CHO-K1, DG44, CHO-S</td>
<td>Illumina</td>
<td>Genomes, apoptosis pathway genes</td>
<td>[35]</td>
</tr>
<tr>
<td>G</td>
<td>Chinese hamster</td>
<td>Illumina</td>
<td>Genome (apoptosis pathway genes)</td>
<td>[36]</td>
</tr>
<tr>
<td>T</td>
<td>Parental CHO</td>
<td>454</td>
<td>Transcriptome (chromosomes)</td>
<td>[37]</td>
</tr>
<tr>
<td>T/G</td>
<td>CHO: IgG</td>
<td>Illumina</td>
<td>Transcriptome, butyrate-affected pathway genes</td>
<td>[38]</td>
</tr>
<tr>
<td>T</td>
<td>CHO: IgG</td>
<td>Illumina</td>
<td>Transcriptome</td>
<td>[39]</td>
</tr>
<tr>
<td>T</td>
<td>CHO-K1</td>
<td>454</td>
<td>Transcriptome, N-glycosylation pathway genes, and splice variants</td>
<td>[40]</td>
</tr>
<tr>
<td>T</td>
<td>CHO-K1, DG-44</td>
<td>Illumina</td>
<td>miRNA</td>
<td>[41]</td>
</tr>
<tr>
<td>T</td>
<td>CHO-K1, CHO-DUXB11</td>
<td>Illumina</td>
<td>miRNA transcriptome: novel and conserved</td>
<td>[42]</td>
</tr>
<tr>
<td>T</td>
<td>CHO-K1</td>
<td>Illumina</td>
<td>miRNA genomic loci and precursor miRNA</td>
<td>[43]</td>
</tr>
<tr>
<td>T</td>
<td>CHO-K1, CHO: SEAP, CHO: tPA</td>
<td>Illumina</td>
<td>miRNA</td>
<td>[44]</td>
</tr>
<tr>
<td>T</td>
<td>CHO-K1, CHO-DUXB-11</td>
<td>Illumina</td>
<td>piRNAs and piRNA clusters</td>
<td>[45]</td>
</tr>
<tr>
<td>P</td>
<td>CHO-K1: SEAP</td>
<td>N/A</td>
<td>Improved proteome identification based on genome sequences</td>
<td>[46]</td>
</tr>
<tr>
<td>P</td>
<td>CHO-K1</td>
<td>N/A</td>
<td>Codon frequency, gene ontology, post-translational modifications</td>
<td>[47]</td>
</tr>
</tbody>
</table>

Table of NGS-based CHO studies including type of study, cell lines, technology platform, and focus area. Proteomics studies are included if they rely on the NGS-based CHO genome.

*EST: Expressed sequence tags; G: Genomics; NGS: Next-generation sequencing; P: Proteomics; T: Transcriptomics.

*NGS technology was used on the data, but not in the publication.

**Key Term**

Transcriptomic analyses: The study of the RNAs expressed by a cell under a given set of conditions.
miRNA is one class of ncRNAs and has a unique and characteristic secondary structure [42]. miRNA precursor transcripts are nearly 70 nucleotides and the processed, mature sequences are typically between 20 and 24 nucleotides [43] in length, which makes them well-suited for analysis using the Illumina platform. miRNAs control the fate of gene expression via post-transcriptional repression of mRNA translation or destabilization [44] and as a result, they have great potential in cell characterization and engineering applications. For example, miRNA sequences that are expressed during cultivation may influence a range of cellular processes that control productivity, product quality attributes, and growth. Many miRNA sequences are conserved across species and others may be unique to a given species. One of the ongoing efforts in the life science community is to catalog and understand each of the known miRNAs. In the past few years, the number of known CHO miRNA sequences has significantly increased. Expressed miRNA can be extracted from CHO cell lines and sequenced. The resulting sequences can be compared against general databases of known miRNAs (e.g., miRBase) to confirm sequences already known or conserved across species. By linking the sequence to the experimental conditions used that resulted in the miRNA expression, one can begin to establish a link between miRNA and phenotype. The number of CHO miRNA sequences has gone from 260 [37] to 350 [43] to 387 [45] within the past few years alone. Today there are nearly 400 known CHO miRNA sequences, 350 of which are conserved and 235 of which have a specific function [45].

In addition to experimental approaches, bioinformatic algorithms can also be used to predict miRNA sequences from an established genome. With the sequencing of the CHO-K1 genome [34], the ability to computationally predict CHO miRNAs became possible and identified 415 miRNAs [46]. The locations of 365 structures were cataloged, 319 of which are expressed, mature CHO miRNAs that were verified and assigned to miRBase [46]. Relative genomic locations have also been used in the CHO-K1 genome in an attempt to discover additional miRNA sequences based upon identified miRNA scaffold organization [47]. Post CHO-K1 genome availability, bioinformatics tools identified 190 miRNAs as conserved CHO miRNAs between CHO-K1, mouse, rat and human genomes, of which more than 80% exhibited differential expression across two recombinant CHO cell lines [47]. Moving forward, NGS approaches offer an unprecedented ability to interrogate changes in both mRNA and miRNA expression that are related to phenotypes of interest [17].

Unlike miRNA, PIWI (a class of proteins) interacting RNAs (piRNAs) are a poorly understood class of small ncRNAs, which likely mediate RNA silencing and repress transposable elements, protecting the genome’s integrity [48,49]. piRNA function may affect the prolonged stability of genetically modified CHO cell lines, in addition to cellular processes and metabolic pathways. Computational analysis of small RNA sequencing data predicted 540 piRNA clusters, consisting of nearly 26,000 piRNA sequences [49]. piRNA sequence expression was measured across six CHO cell lines, including adherent, suspension adapted, and recombinant CHO-K1 and DUKXXB11 cell lines, using the published CHO-K1 genome as a reference [49]. This initial analysis of CHO piRNA indicated the potential of piRNAs as tools for cell line development and genetic engineering.

As mentioned earlier, perhaps the most important contribution of NGS to the CHO biomanufacturing community has been the establishment of the CHO-K1 and Chinese hamster genomes [34–36]. The draft CHO-K1 genomic sequence consisted of 2.45 Gb and was assembled into 24,383 genes [34]. This catalog of genes permitted genetic modification of CHO-specific target DNA sequences. The CHO-K1 genes were analyzed by comparative genomic analysis with the human, mouse and rat genomes, which confirmed that the mouse genome demonstrated the greatest similarity [34]. The draft genome enabled immediate analysis of genes relevant to CHO biomanufacturing such as those related to product quality attributes such as glycosylation. In that study [34], the number of CHO genes that were homologous to human glycosylation genes suggested that CHO cells have the potential to perform 99% of the glycosylation reactions that human cells perform. However, transcriptome analysis of CHO cells further suggested that only approximately half of the CHO glycosylation genes were actually expressed under any of a variety of conditions, suggesting the other half may be silenced. Such analyses based on a draft genome promise a more detailed, molecular understanding of the behavior of CHO cells leading to improvements in bioprocessing in the future. However, there are also a number of unaddressed challenges that emerged with the CHO genome, partly as a result of the use of NGS methods.

One important issue from the K1 genome [34] and the hamster genome [35] is that the assembled DNA
within a cell to investigate the intermediates of metabolism, small molecule metabolites, the Metabolomics: The study of set of conditions. within a cell at a specific time and complete complement of proteins Proteomics: Key Terms The study of the chromosomal variability among CHO cells, the identification and establishment of a definitive reference, the draft genomes for the community has recently established a framework to facilitate similar efforts among CHO-DG44, CHO-S, and three other CHO cell lines were sequenced, and the CHO-K1 cell line was resequenced after the CHO-K1 genome. Annotation of all cell lines and nucleotide-resolution analysis of the CHO cell line genotypic differences was completed. Comparative genomics identified copy number variations and 3.7 million single-nucleotide polymorphisms between the different cell lines, many of which affected genes relevant to bioprocessing pathways, such as apoptosis. In an attempt to determine the genomic structure, the sequences were aligned to published BACs and filtered; however, only 26% of the genomic sequence was reliably localized to specific hamster chromosomes.

Following the release of the CHO-K1 genome, an international academic and industrial collaboration developed ‘CHOgenome.org’ to facilitate accessibility of the genomic data and the development of genomic tools for the C. griseus and CHO cell communities. The current list of tools offered includes BLAST searches, individual gene searches, and visual representation of the CHO-K1 genome assemblies, along with a CHO proteome database. However, the sequencing of a number of CHO-related genomes creates challenges in terms of comparative genomics. That is, there is a need for tools to facilitate the analysis of multiple genomes by a given user. For example, it may be desirable to compare the genome of a host cell early in culture versus late in culture or to compare the genome of a proprietary host cell with that of CHO-K1 and of the Chinese hamster. Tools that facilitate analysis of events as large as chromosomal rearrangements and as small as single-nucleotide polymorphisms would provide users the opportunity to link genome information to observed phenotypes. However, such tools do not yet exist.

Emerging sequencing approaches
The pace of DNA sequencing technology development has continued and there is new a generation of technologies available. These approaches, which offer new and greater amounts of data for similar or less cost per run, are designed with single-molecule or electro-chemical platforms, different than the NGS platforms that use fluorescent signals and PCR amplification. Among the new methods are those developed by Pacific Biosciences (PacBio), Life Technologies (Ion Torrent) and Oxford Nanopore Technologies (nanopore). The PacBio approach involves single molecule sequencing based on an immobilized polymerase in a cell designed for single molecule, real time detection, which has potential applications in the bioprocessing field including studies of epigenetic regulation of heterologous gene regulation and genome-wide structural

sequences have not been aligned onto chromosomes. An important consideration in the CHO genome is the lack of chromosomal stability that has been observed. Moreover, there is significant genomic drift. Indeed these are properties that the biomanufacturing community has relied on in the application of CHO cells because of the ability to reasonably quickly adapt CHO cells to various growth conditions. However, once established, it would be advantageous to have host cells that have minimal chromosomal or genomic changes. To facilitate analysis of chromosomal or genomic rearrangements, it is important for the CHO community to have a physical mapping of a reference genome. One way to help build a physical map of the CHO genome is to use a bacterial artificial chromosome (BAC) library.

An initial CHO BAC-based map identified 20 different chromosomes and high aneuploidy was observed. This library was used to obtain a detailed physical chromosomal map of the CHO-DG44 cell line utilizing fluorescence in situ hybridization imaging of the randomly selected BAC clones. For eight of the 20 chromosomes identified, chromosomal rearrangements did not occur between CHO-DG44, CHO-K1, and Chinese hamster lung cells. The conservation without large rearrangement suggests their genetic importance and resultant stability; however, it was not possible to identify which genes were located on these eight chromosomes. The recent publication of sequences for individual Chinese hamster chromosomes provides a critical step forward for the community because it includes the sequences of each of the Chinese hamster chromosomes independent of the other chromosomes.

A second important issue from the K1 genome is the need for ongoing updates to the assembly and annotation. While the initial draft genome is assembled and annotated with the aid of humans, the majority of the work is performed by bioinformatic algorithms. The human genome and other genome communities have developed mechanisms to make updates and corrections to their genomes and the CHO genome community has recently established a framework to facilitate similar efforts. Given the diversity of cell lines and the known issues with significant genomic and chromosomal variability among CHO cells, the identification and establishment of a definitive reference genome is essential for the community.

To better understand the genomic diversity and help establish a reference, the draft genomes for the Chinese hamster (Cricetulus griseus), the CHO-DG44, CHO-S, and three other CHO cell lines were sequenced, and the CHO-K1 cell line was resequenced after the CHO-K1 genome. Annotation of all cell lines and nucleotide-resolution analysis of the CHO cell line genotypic differences was completed. Comparative genomics identified copy number variations and 3.7 million single-nucleotide polymorphisms between the different cell lines, many of which affected genes relevant to bioprocessing pathways, such as apoptosis. In an attempt to determine the genomic structure, the sequences were aligned to published BACs and filtered; however, only 26% of the genomic sequence was reliably localized to specific hamster chromosomes.
variation [61]. Ion Torrent is a semiconductor platform-based sequencing approach that relies on a pH probe for detection and takes advantage of the fact that hydrogen ions are released as nucleotides are incorporated into a growing DNA chain [57]. The nanopore sequencing approach also relies on an immobilized enzyme, staphylococcal α-hemolysin, as the nanopore through which a DNA molecule is sequenced by passing through the pore and across an electric potential field [56]. While none of the emerging and third-generation sequencing technologies have yet been applied to CHO studies relevant to the biomanufacturing community (at the time of this publication), these approaches will certainly see widespread application in the near future.

Other omics
While NGS has had the most dramatic and obvious impact on the CHO community in the past few years, there are a number of important and parallel omics approaches that are also necessary for a complete understanding of CHO biology in a way that enhances productivity and product quality attributes. There is a significant amount of literature on this topic [63–71]. The important issue to consider is that genome sequencing, and genomics alone, may not provide enough information about observed phenotypes. Certainly the genes and other features (e.g., miRNA) that are expressed significantly influence cell behaviors. The mRNAs lead to protein expression and proteins have diverse functions including structural roles, metabolism, and many other cell processes. Ultimately, a true understanding of the basis for a given phenotype may rely not only on the ability to capture the genome of the cell line at that moment in time, but also on transcript analysis, proteomics, metabolomics, and other measures, see Clarke for an example [72]. The CHO community has applied many of these other techniques to understand phenotypes, but very few studies to date have integrated data from NGS studies together with other omic methods – efforts which rely heavily on bioinformatics because of the large volume of data created by NGS.

In one metabolomics study, an in silico model was used with metabolomic analysis to understand CHO intracellular fed-batch culture mechanisms. The identified, growth limitation metabolites were associated with the glutathione, glycerophospholipid and energy pathways [73]. The in silico model was used to obtain a greater understanding of these affected pathway fluxes, the results of which were in good agreement with aging culture glycolysis and TCA cycle flux details, resulting in the identification of novel, growth-related mechanisms. The in silico model used was not originally developed from the annotated CHO genome, but rather from the mouse genome [74], refined and validated with mouse hybridoma cell observations [75], and expanded for CHO with annotated CHO cDNA, as the CHO genome and gene function identification was ongoing. An entirely CHO genome-based model would potentially enhance these results, support future metabolomics studies, and lead to new CHO culture improvements.

Proteomic analysis has also been applied to the study of biopharmaceutical production cell lines for many years but only recently has it been combined with NGS datasets. Proteomics studies typically involve the use of mass spectrometry to link changes in observed proteins to their underlying genes. The availability of a sequenced CHO genome, which was facilitated by NGS, has provided a means to improve the efficiency by which mass spectra are assigned to gene sequences [76]. For example, two CHO specific databases were used for CHO protein identification, including the CHO-K1 genome database. Identification using this database increased the number of identified proteins by 35%, which further increased to 47% with the addition of a second CHO-specific database [76]. In another recent analysis of the CHO proteome based on the CHO K1 genome, the proteome, secretome and glycoproteome contained 6164 grouped proteins [77], an eightfold increase in the number of CHO proteins identified. This increase was attributable to both an improved cell lysate fractionation method as well as the use of an organism-specific sequence database. More importantly, the availability of a detailed proteome dataset permits a better understanding of codon frequency in CHO, the degree of pathway enrichment, and of possible post-translational modifications. For example, codon frequency in CHO was observed to be distinct from human cells [77]. The degree of pathway enrichment was obtained from combined proteomic and transcriptomic (mRNA) data sets, highlighted by the enrichment of the protein processing and apoptosis pathways and depletion of the steroid hormone and glycosphingolipid metabolism pathways. The cataloged post-translational modifications included 504 N-acetylation proteins and 1292 N-glycosylated proteins.

Applications to industrial biomanufacturing
The many studies above offer a glimpse into possible applications of NGS to increase the understanding of CHO biology. A deeper understanding of CHO biology is an important prerequisite to achieve significant improvements in cell line development and process development by the industry. The applications of new
technologies have generally been initiated by the relevant academic community to build a foundation of relevant knowledge before applications are developed of industrial significance. As a result, we are unaware of any NGS-based CHO studies having yet revealed the identity of a gene, protein, or regulatory element that has transformed industrial practice. Indeed, the complex basis for phenotypes means that further study will be needed to achieve substantial improvements in CHO cell-based manufacturing. Nonetheless, it is only a matter of time before NGS-based discoveries provide clues about the capabilities of CHO cells through (for example) whole genome reconstructions, models of metabolism, or models of glycosylation and allow for reprogramming of cells with specific phenotypic characteristics. Efforts will be accelerated through collaboration and cooperation as discussed in the ‘Future perspective’ section.

Conclusion
The CHO biomanufacturing community is at the start of a new era. The availability of the first drafts of a number of relevant genomes and the low cost of sequencing various host cells provides a basic foundation for the community to better understand the molecular basis for issues related to productivity, product quality, and stability (of productivity, of product quality, and of viability). However, many challenges are also emerging. First, the community does not yet have a stable, well-defined and characterized reference genome. The recent sequencing of the Chinese hamster and its chromosomes provides the basis for this moving forward, but the need to correct annotations remains an ongoing challenge even in the human genome community. Second, there are relatively few tools available to compare genomes. For example, tools to easily compare the genome of a proprietary host cell versus CHO K1 versus the Chinese hamster, or to compare the genome of a host cell early in culture versus late in culture, do not exist. Third, having a genome is most useful when placed in the context of other omic data for a given cell (transcriptomics, proteomics, fluxomics) and there are not yet tools available to simply integrate information across these datasets. Despite these and other challenges, NGS has helped move the CHO biomanufacturing community forward towards a time when host cells can make any given product and can be reliably and predictably customized and designed to ensure high productivity of specific product quality attributes that are stably expressed by cells.

Future perspective
The CHO biomanufacturing community is at the beginning of the genomics era. The unprecedented ease with which one can collect DNA sequence information will enable a deeper understanding of the relationship between the genome and phenotypes. However, the pace with which data is generated is increasing and there are bottlenecks in the ability to analyze the data. As a result, there may be an increasing emphasis on bioinformaticians who can assist in the interpretation and understanding of these large datasets. Moreover, there is an urgent need for a well-defined reference genome that is stable and that the community can use as a foundation for genomics-based studies. Once established, individual teams can employ methods to study the genome, epigenome, transcriptome, proteome and metabolome as part of their efforts to understand cellular phenotypes. However, making biological inferences that will lead to targets for cellular engineering to modify cell productivity, product quality attributes, and the stability of cell lines in a predictable manner may take many years. Cooperation and collaboration among the academic and industrial scientific community will be essential for the CHO community to fully realize the potential of the genomics era for the production of biologics.

Executive summary

- Next-generation sequencing technologies may enable a deeper understanding of CHO cells in terms of issues related to productivity, product quality attributes and stability.
- Next-generation sequencing enabled the genome sequencing of CHO cell lines and the Chinese hamster – although a well-defined reference genome still needs to be established.
- Much more work needs to be done to fully realize the potential of genomics and may need to better integrate other omic approaches.
- Academic–industrial collaboration will be essential for transformative impact in the community.

Financial & competing interests disclosure
The authors thank the National Science Foundation, the National Institutes of Standards and Technologies, and the International CHO Genome Community for financial support. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.
NGS technologies & their potential impact on CHO cell-based biomanufacturing

References

Papers of special note have been highlighted as:

- of interest


- Provides some of the initial CHO genomic sequences and confirms CHO’s sequence similarity to mouse.


- Provides some of the initial CHO genomic sequences and confirms CHO’s sequence similarity to mouse.


- Defines the CHO-K1 genome as consisting of 24,383 genes and identifies human homologs that exist for many pharmaceutically relevant genes.
Describes the hamster genome consisting of 24,044 genes and many genomic variations that exist between CHO and hamster.

Reveals the sequences and genes associated with each Chinese hamster chromosome.

Shows the dynamic nature of the CHO genome and demonstrates the ability to evolve improved CHO cell variants.

Website with CHO genomic sequences, genes and visualization tools.


Luo J, Vijayasankaran N, Autsen J et al. Comparative metabolic analysis to understand lactate metabolism shift in


» Websites

101 454 Sequencing.  
http://454.com/products/gs-flx-system

102 Illumina.  
www.illumina.com/systems/hiseq_2000_1000.ilmn

103 CHOgenome.  
www.CHOgenome.org

104 Applied Biosystems – Sanger.  
www6.appliedbiosystems.com/products/abi3730xlspecs.cfm