Pharmaceutical BIOPROCESSING

Next-generation bioprocess: an industry perspective of how the 'omics era will affect future biotherapeutic development

"Next-generation sequencing techniques are too powerful to ignore, and offer advantages over current methods of contaminant detection and cell bank characterization."

Keywords: CHO bioprocess • CHO genomics • next-generation sequencing • 'omics techniques

Chapman Wright
Author for correspondence:
Biogen Idec, 15 Cambridge Center, Cambridge, MA 02142, USA
Chapman.Wright@biogenidec.com

Scott Estes
Biogen Idec, 15 Cambridge Center, Cambridge, MA 02142, USA

The past decade has seen significant instrumentation and methodological advances enabling much broader profiling of metabolites, proteins and nucleic acids. This in turn has been leveraged to enhance our sophistication and understanding of Chinese hamster ovary (CHO) cell metabolism in a bioprocessing environment. Of the three aforementioned fields, genomics has established itself as the preeminent technology platform for the foreseeable future largely due to a combination of throughput, comprehensive coverage and a relatively simple workflow. This has come to pass as a result of the tremendous advancements in next-generation sequencing (NGS) technology. These breakthroughs have lowered the barriers of cost and time associated with whole genome (DNA-Seq) and transcriptome (RNA-Seq) sequencing projects at an unprecedented rate, giving way to the 'sequencing revolution' [1].

The era of CHO genomics was ushered in by Xu and co-workers who applied NGS technology to create the first publically available CHO-K1 draft genome in 2011 [2]. CHO-K1 is but one of a handful of CHO host cell lines utilized by the bioprocessing industry. Indeed considering the extended time in culture and variety of adaptation strategies applied by the numerous labs working with CHO, every CHO host should be considered a unique cell line, irrespective of a shared common lineage [3]. Therefore, in subsequent publications by Lewis *et al.* [4] and Brinkrolf *et al.* [5] the CHO-K1 draft genome was expanded upon by sequencing the Chinese hamster genome. Having a Chinese hamster reference genome to facilitate the assembly of additional CHO genomes will benefit the community as a whole. However, the work is far from done as the quality of the genome is curtailed by gaps in sequencing coverage and incomplete gene annotations. A community of scientists working with CHOgenome.org is currently in the process of updating and correcting the CHO/hamster draft genomes, with particular interest in the sequencing gaps and annotations. This work will be critical to unlock the full benefits of having a high-quality genome to work with. Acknowledging the work that remains, the question becomes: how will NGS impact future bioprocess and what is the potential role of other 'omics platforms?

Current limitations of proteomics & metabolomics

Proteomics and metabolomics are technologies currently used to generate molecular profiles of CHO bioprocesses and have the advantage of closely reflecting the biological phenotype of a given system. However, the present reality is that only NGS is well suited for routine applications where speed is of

FUTURE SCIENCE part of fsg

the essence and comprehensive coverage of the analyte, genes/transcripts in the case of NGS, is desired. The technological advancements in workflow simplicity, coverage and throughput that have occurred in NGS far exceed those achieved with proteomics and metabolomics. For instance, one major limitation of proteomics is the restricted dynamic range of current platforms, which are only able to identify a subset of proteins from exceedingly complex spectra of masses. Coverage of the proteome can be enhanced by first fractionating lysates prior to analysis but this will add time and complexity; a factor which limits the utility of the approach. This contrasts to NGS which can sequence an entire genome or transcriptome from a single sample in a relatively short period of time. An additional barrier to ease of application is the lack of a well-annotated CHO proteome database. The publication of CHO genome databases [2,4,5] as well as the human proteome is a great first step [6]. This could elicit a similar period of advancement that followed the publication of the human genome [7,8].

> "Given these limitations as they exist today, the application of metabolomics and proteomics to bioprocessing will be limited in scope relative to the applications of next-generation sequencing."

Metabolomics and flux analysis have also been powerful techniques to decipher the metabolic profile of recombinant CHO cells and how it plays a key role in cell growth and productivity. The barrier to routine profiling of the entire metabolome remains the fact that the chemical diversity of metabolites present in cells makes their comprehensive detection extremely challenging. Thus, if global profiling of metabolites is desired, different extraction methods and separation instrumentation (e.g., LC–MS and GC–MS) will need to be applied. An additional barrier to routine application is that although care must be taken when extracting nucleic acids, particularly RNA for NGS analysis, this concern is further exacerbated for metabolomics. The instability of many metabolites coupled with the need to sustain representative partitioning between intracellular and extracellular compartments makes sample prep far more challenging and critical to ensure meaningful data are extracted.

Given these limitations as they exist today, the application of metabolomics and proteomics to bioprocessing will be limited in scope relative to the applications of NGS. However, not all bioprocess insights can come from NGS data. Key bioprocess information will be found in the proteome and metabolome as we consider all three 'omics areas to be crucial for developing broad biological understanding of CHO bioprocess. The proteome contains rich data that can provide more insight into bioprocess. Confirmation that transcripts of interest have been translated and detection of misincorporated amino acids and the biologically active post-translational modifications are key pieces of information. The ability to detect a broader spectrum of metabolites as they accumulate or are depleted during the bioprocess adds considerable value when implementing media or bioprocess optimization strategies.

Foundation building

The term 'quasispecies' has been applied to the diversity of untransfected CHO hosts available today [3]. Hence it is critical to regard each user's CHO host as unique. The dynamics are compounded when considering the variability in extracellular environment introduced by different media and process conditions. There are many examples in the literature, too many to cite here, of engineering targets and culture conditions that have made a positive impact on bioprocess. Although some of these findings may be relevant to the majority of CHO cell lines, it is important to note that these findings may not be universal, even when the same host lineage is employed. Clone, media and culture conditions will all influence the output. Thus, the most relevant 'omics data will be found in samples that derive from the user's cells and process. We encourage each user to collect 'omics data at relevant points during their cell culture process. This represents one of those applications where investment in all three 'omics technologies can add value by establishing a comprehensive map of how their host interfaces with the process platform. This practice, which we refer to as 'foundation building', will give the user the best opportunity to make meaningful discoveries in 'omics data that can be used to inform decisions around media and process for future cell lines.

To truly profit from 'omics investigations, it is best to form collaborations with computational biologists. Once the goals of the project are articulated, computational biologists will be able to aid with experimental design from a replicate and time-point perspective and, most importantly, distilling the vast amount of data down to a manageable summary of metabolic pathways of potential interest. The amount of data generated from an 'omics experiment can be overwhelming, and having a scientist that is trained to work with these large data sets and the software necessary to manipulate them is essential.

Omics for cell line development

Bioprocess begins with the cell line. The core objectives of cell line development groups are to consistently deliver high productivity clones with the necessary product quality. It is an appealing concept to be able

to screen early in the cell line development process and identify clones that display a predefined 'omics profile that is predictive of productivity and product quality. Given the advances in speed and throughput, RNA-Seq has the potential to make this vision a reality. Through multiplexing, each well of a 96-well plate can essentially be 'barcoded' with unique nucleotide sequences, allowing for distinct RNA-Seq data sets to be collected on hundreds of clones. Although feasible, the limitation of time is still a consideration. Results may require days to weeks to be in hand, depending on the instrumentation available, as well as the number of genes and/or clones being characterized. With run time of about a day, top-end instruments such as the HiSeq 2500 platform (Illumina, CA, USA) has output ranges in the hundreds of giga base pairs (Gbp) and the Ion Proton™ will boast similar numbers with the new Ion PII™ chip (Life Technologies, NY, USA). This would be enough output to analyze around 100 clones monitoring a few hundred genes at sufficient depth for analysis. This level of throughput makes the technology practical as an early clone-screening tool. The critical question is which genes should the user monitor? Clarke *et al.* using a proprietary microarray have illustrated that a set of ~300 genes could predict downstream productivity [9]. In contrast, Kang *et al.* have identified a much shorter list of transcripts and proteins correlative of titer using a combined DNA microarray and LC–MS shotgun proteomics approach [10]. The user will need to determine whether the throughput and breadth of coverage offered by NGS or the specificity offered by multiple 'omics techniques would be more beneficial in their platform. Both methods will require input from 'foundation building' but could significantly increase the quality of clones moving forward in the development process.

One of the most straightforward and useful applications of NGS that capitalizes on its rapid, high-throughput capabilities is to screen hundreds of clones using targeted sequence confirmation of the biotherapeutic to detect undesirable sequence variants. In this case, a series of primers specific to the transgene(s) are designed to interrogate only the gene(s) of interest. Clones with detectable sequence variants can be discarded early in the process to concentrate efforts on clones with correct sequences. This is a capability that prior to NGS was limited to a much smaller number of clones analyzed by time-consuming mass spectrometry-based peptide mapping. It is important to point out the limitation of this method which is its inability to detect post-transcriptional amino acid misincorporations. Hence the two methods become complimentary; NGS can be applied as an early high-throughput tool to weed out clones with genetic mutations rendering them unsuitable and then peptide mapping can be applied to a small subset of top clones to ensure the clones and process are not resulting in an unacceptable level of substitutions.

Early screening of product-quality attributes is an interesting area, especially with efforts to develop biosimilars expanding. Establishing a cell line and process that produces a product with acceptable comparability to that of the innovator molecule is a significant challenge. Can screening for particular attributes be predictive of the ultimate comparability to the innovator molecule? This could be tricky given the influence raw materials and process conditions can have on glycan structure but one speculative path forward would be to monitor the transcript levels of key glycan remodeling enzymes (e.g., mannosidases and salidases) in clones early in the screening process. In this manner, the user could select clones predisposed to producing the desired glycan. While this process will require much effort on the front end to convince the user of its merit, the advancement of NGS technology makes this a possibility. Thus, if the user decides that there is utility in screening clones for genes predictive of productivity, it makes sense to include a glycan marker screen to create a single experiment that could be predictive for both productivity and product quality.

Monitoring bioprocess

The goals of a successful bioprocess are to have consistent and reliable cell culture process in which the biotherapeutic is produced with the same titer and product quality each run. Historically, CHO cell metabolic responses to traditional fed-batch conditions were not well understood due to the lack of foundation work and a narrow breadth of analytes captured in a typical run. Here, 'omics techniques can provide a powerful tool to greatly expand the number of molecules monitored, providing rich content on cellular metabolism to facilitate process and media development. Researchers have made progress in detecting changes in the CHO metabolome [11,12] and proteome [13,14] over the course of cell culture processes and identified hallmarks of high productivity. These works highlight the importance of fluxes through the tricarboxylic acid (TCA) cycle and the redox state of the cell and provide a starting point for molecules and pathways that should be considered for inclusion in a targeted profiling strategy. Ideally, this approach would also be supported by the 'foundation building' work described above.

We also envision NGS and perhaps targeted proteomics and metabolomics providing valuable insights during tech transfer and scale-down/-up model development. Despite our growing bioprocessing sophistication, when it comes time to transfer a bioprocess to a new facility or develop scale-down models, there are a limited number of readouts that are currently used

as metrics of comparability (e.g., titer and resulting product quality). The application of these 'omics platforms can provide a much richer data set to confirm comparability or alternatively provide insights as to the biological underpinnings that result in differences at scale. The challenge of this approach is establishing the difference between inherent biological variability of the system versus a bona fide lack of comparability.

Automation & robotics

Robotics will be a driving force to establish 'omics as a routine development or characterization technology in the bioprocess lab. Some engineers may be reading this commentary with trepidation, and although it is true NGS experiments will need the support of computational and molecular biologists, the goal is to have NGS experiments automated for bioprocess applications. Currently, there are robots on the market that are able to process cell samples, perform the library prep steps and, subsequently, load the sample libraries onto the sequencing instrument. While this is advantageous to free scientists' time, this is also important for consistency. Robotics will eliminate the 'user-to-user' variability that can amplify the noise in NGS experiments. Software/pipeline development is another area that will greatly affect the adoption of 'omics in the bioprocess lab. The recent report of the software program 'Sailfish' is representative of what is to come [15]. This program is able to estimate transcript levels without mapping reads, greatly increasing the speed of RNA-Seq analysis. These sorts of advancements in consistency, speed and hands-free processing is what will drive 'omics techniques into bioprocess laboratories.

Regulatory perspective

One of the most interesting areas in the 'omics field is the use of NGS for regulatory purposes. The speed and power of NGS technology cannot be ignored in this area. As an identity test, it is difficult to imagine any analytical readout being more convincing than that of the cell line genome. The ability to theoretically screen for any microbial or viral contaminant in a Good Manufacturing Practices (GMP) cell bank in a matter of days is an attractive alternative to current direct cultivation protocols that can take weeks to a month for a final readout. However, the integration of NGS techniques into regulatory workflows will not be trivial. There needs to be consensus on controls, sample prep protocols, NGS pipelines and genome databases for mapping reads to ensure the process is standardized. Industry leaders and regulatory agencies have recently begun the dialog on how best to implement NGS technology into bioprocess and quality-control departments. How and when regulatory agencies will require NGS-sourced contaminant data for filings is an open question; however, it seems inevitable that the industry is headed in this direction.

NGS technology can also be used to assess the genetic stability of production clones. Integration-site analysis prior to and post-bioprocess could provide nucleotide resolution of transgene stability, a level of precision not achievable with the current approach of Southern blot analysis to address this question. With the current reliance on short reads to perform NGS, there exists the challenge of creating an accurate contig of the integration sites in cell lines where multiple copies of the transgene integrated at a single site. Recent technology advances, such as Single Molecule, Real-Time (SMRT®; Pacific Biosciences, CA, USA) sequencing platform [16], generate much longer sequence reads and could potentially solve this dilemma. In general, all signs point to a future role for NGS in regulatory filings to support the safety and integrity of GMP cell banks.

Conclusion & future perspective

The 'omics era has undoubtedly reached biotechnology and biotherapeutic development. In the short term, RNA-Seq and transcriptome profiling of CHO cells will be the focus of the bioprocessing community in an effort to understand how we can achieve optimal cell culture performance. There will be 'omics discoveries that are applicable to all CHO hosts, but it is the 'internal' discoveries, observed in the user's relevant conditions, that we believe could be just as impactful. Computational biologist will be an integral part of this development process. Their training in big data-set manipulation and experimental design will be vital to uncovering actionable aspects from 'omics investigations. The role in which NGS will play in regulatory filings is of particular interest in the coming years. NGS techniques are too powerful to ignore, and offer advantages over current methods of contaminant detection and cell bank characterization. Whether all or any of the 'omics techniques discussed become routine in bioprocess development remains to be seen. However, with the pace in which our community is adopting 'omics it is clear that these techniques will be central to the bioprocess of the future.

Financial & competing interests disclosure

The authors have no 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. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1 Koboldt DC, Steinberg KM, Larson DE *et al.* The next-generation sequencing revolution and its impact on genomics. *Cell* 155(1), 27–38 (2013).

2 Xu X, Nagarajan H, Lewis NE *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* 29(8), 735–741 (2011).

3 Wurm FM. CHO quasispecies – implications for manufacturing processes. *Processes* 1, 296–311 (2013).

4 Lewis NE, Liu X, Li Y *et al.* Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. *Nat. Biotechnol.* 31(8), 759–765 (2013).

5 Brinkrolf K, Rupp O, Laux H *et al.* Chinese hamster genome sequenced from sorted chromosomes. *Nat. Biotechnol.* 31(8), 694–695 (2013).

6 Wilheim M, Schlegl J, Hahne H *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587 (2014).

7 Venter JC, Adams MD, Myers EW *et al.* The sequence of the human genome. *Science* 291(5507), 1304–1351 (2001).

8 International Human Genome Sequencing Consortium; Lander ES, Linton LM. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).

9 Clarke C, Doolan P, Barron N *et al.* Predicting cell-specific productivity from CHO gene expression. *J. Biotechnol.* 151(2), 159–165 (2011).

10 Kang S, Ren D, Xiao G *et al.* Cell line profiling to improve monoclonal antibody production. *Biotechnol. Bioeng.* 111(4), 748–760 (2014).

11 Dean J, Reddy P. Metabolic analysis of antibody producing CHO cells in fed-batch production. *Biotechnol. Bioeng.* 110(6), 1735–1747 (2013).

12 Chong W, Thng S, Hiu A *et al.* LC-MS based metabolic characterization of high monoclonal antibody-producing Chinese hamster ovary cells. *Biotechnol. Bioeng.* 109(12), 3103–3111 (2012).

13 Carlage T, Kshirsagar R, Zang L *et al.* Analysis of dynamic changes in the proteome of a Bcl-XL overexpressing Chinese hamster ovary cell culture during exponential and stationary phases. *Biotechnol. Prog.* 28(3), 814–823 (2012).

14 Pascoe D, Arnott D, Papoutsakis E *et al.* Proteome analysis of antibody-producing CHO cell lines with different metabolic profiles. *Biotechnol. Bioeng.* 98(2), 391–410 (2007).

15 Patro R, Mount SM and Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32(5), 462–464 (2014).

16 Utturkar SM, Klingeman DM, Land ML *et al.* Evaluation and validation of *de novo* and hybrid assembly techniques to derive high quality genome sequences. *Bioinformatics* 30(19), 2709–2716 (2014).