Computational and Structural Biotechnology Journal 27 (2025) 1511–1517

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review article

Next-generation sequencing: A powerful multi-purpose tool in cell line development for biologics production

Luigi Grassi [a,*], Claire Harris [a], Jie Zhu [b], Diane Hatton [a], Sarah Dunn [a]

[a] Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK
[b] Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, USA

ARTICLE INFO

Keywords:
NGS
RNA-seq
DNA-seq
Transcriptomics
Multiomics
GOI
SNVs

ABSTRACT

Within the biopharmaceutical industry, the cell line development (CLD) process generates recombinant mammalian cell lines for the expression of therapeutic proteins. Analytical methods for the extensive characterisation of the protein product are well established; however, over recent years, next-generation sequencing (NGS) technologies have rapidly become an integral part of the CLD workflow. NGS can be used for different applications to characterise the genome, epigenome and transcriptome of cell lines. The resulting extensive datasets, especially when integrated with systems biology models, can give comprehensive insights that can be applied to optimize cell lines, media, and fermentation processes. NGS also provides comprehensive methods to monitor genetic variability during CLD. High coverage NGS experiments can indeed be used to ensure the integrity of plasmids, identify integration sites, and verify monoclonality of the cell lines. This review summarises the role of NGS in advancing biopharmaceutical production to ensure safety and efficacy of therapeutic proteins.

A cell line development (CLD) campaign is the process used to establish a recombinant mammalian cell line suitable for the manufacture of a therapeutic protein. It starts with the transfection of a host cell line with a plasmid containing one or more genes of interest (GOIs), encoding the therapeutic protein sequences and a selectable marker [1]. The plasmid DNA is incorporated into the host cell chromosomes through either random, or site-specific integration. Cells with one or more integrated copies of the plasmid are selected by growth in a medium containing a selective agent for the marker gene. The resulting transfectants should also express the linked GOIs encoding the therapeutic protein polypeptides. An increased concentration of the selective agents for marker genes, such as dihydrofolate reductase (DHFR) and glutamine synthetase (GS), can be used as a gene amplification strategy to increase the copy number of marker genes, thereby increasing the copy number and expression of the linked GOIs and generating more productive cell lines [2–4]. Alternatively, transposon-based strategies can be used to generate highly productive cell lines without the need for time consuming amplification steps [5,6]. To isolate high yielding clonally derived cell lines that meet regulatory requirements, transfectant cells undergo a round of cell cloning where cell lines are isolated from a single progenitor cell and then screened to assess productivity. Clonal cell lines expressing high titers of recombinant protein are chosen for progressive expansion and further evaluation of characteristics, such as the production stability of the clone (i.e. maintenance of high levels of product expression over successive generations) and the quality of the recombinant protein, before cell banking.

The majority of approved biotherapeutic proteins are expressed in mammalian cells, with Chinese hamster ovary (CHO) cells being the predominant host cell line [7]. The first CHO cell line was isolated in 1957 [8], and several different lineages, including CHO-K1, CHO-S, DUXB11, and DG44, were subsequently derived [9,10].

CHO cell lines are an attractive host as they can express high levels of recombinant proteins with human-like folding and post-translational modification patterns and can be scaled up in suspension for manufacturing in large-scale bioreactors [11,12]. However, the high productivity and versatility of CHO cells comes with genome plasticity and an error-prone DNA replication system [13] that is typical of immortalised cell lines. This results in chromosomal heterogeneity of the cell line populations, with cells having different complements of abnormal chromosome rearrangements and numbers [14].

The random integration of the expression plasmid at different CHO genomic loci can result in unwanted structural rearrangements affecting both the plasmid and host genome. Moreover, due to the inherent genetic instability of the CHO cell host, one or multiple copies of the GOI

[*] Corresponding author.
E-mail address: luigi.grassi@astrazeneca.com (L. Grassi).

https://doi.org/10.1016/j.csbj.2025.04.006
Received 9 December 2024; Received in revised form 1 April 2025; Accepted 2 April 2025
Available online 3 April 2025

can be lost [15] or can accumulate variants in a subpopulation of a clonal cell line [16], affecting purity, titer, and quality of the final product, with subsequent effects on manufacturing and drug supply [17]. Mutations can also spontaneously occur in the host genes involved in protein post-translational modifications, such as those encoding glycosylation enzymes, and they can hence impact the product quality [18].

Next-generation sequencing (NGS) is a technology based on the high-throughput sequencing of DNA and RNA fragments and can be used to monitor different cell line characteristics during the steps of a CLD campaign (Fig. 1). Parallel sequencing of thousands of fragments is superior in coverage, sensitivity, quality, and data output [19–21] compared with traditional Sanger sequencing analyses. Several NGS chemistries are used to characterize fragments of different lengths [22–24]. In this review, we use 'short-read technologies' to refer to Illumina or Ion Torrent platforms and 'long-read technologies' to refer to Oxford Nanopore Technologies (ONT) or Pacific Bioscience (PacBio) platforms.

NGS is a very powerful tool that can be deployed to monitor cell line quality during the CLD process and can also provide increased molecular understanding of CHO cell lines and cell culture processes, helping to build foundations for rational engineering strategies to continue to improve the performance of CHO cell systems. In this review, we summarise the applications of NGS in four main sections: 1) characterization of the host or clone genome; 2) RNA-seq analysis to characterize the host, clone, and culture process transcriptomes; 3) RNA-seq applications to detect variants in the GOI; 4) host, clone, and culture process epigenomic characterization. NGS can also be used to detect adventitious agents that can potentially infect CHO cells. Testing of the manufacturing cell banks of the production cell lines for these adventitious agents is a mandatory regulatory requirement to safeguard patients. NGS is now replacing animal-based viral safety testing methods, as encouraged in the recent revision to the ICH Q5A viral safety guidance [25]. As this topic has already been covered by several publications [26–31], it will not be discussed further in this review.

1. Characterization of the host or clone genome

The genome refers to the complete set of DNA of an organism. It includes protein-coding and non-protein-coding genes, intergenic regions, and specifies all the instructions necessary for an individual to develop, function, and reproduce. High-quality genome references of Chinese hamster [32,33] CHO-K1 [34] and CHOZN [35] are fundamental for transcriptome and genome analysis of CHO cell lines. These resources build upon data, insights, and methodologies developed in previous studies [36,37] and make the detection of single nucleotide variants (SNVs) in cell lines both feasible and straightforward [38,39], facilitating the consequent selection of cell lines with desired genetic traits [40]. Whole genome characterization is a very data-rich assay that can discover important features of a given host or clone, but it is not typically performed during a CLD campaign. In fact, the primary focus is on the sequence verification of the plasmid prior to transfection, the detection of the plasmid integration sites in the CHO genome after cell cloning, and the assessment of the integrity of the GOIs and their associated regulatory regions, along with clonality verification in the subsequent steps of the CLD campaign. The verification of the sequences of the GOIs on the plasmid used for transfection is an important requirement and can be generated using capillary-based Sanger sequencing with sets of primers specifically designed for each plasmid. The advent of long-read-based methods has made it possible to speed up and reduce the cost of this process [41–43]. The integration of the plasmid into the host genome, which occurs following transfection and selection, can result in plasmid sequence concatemers and truncated and inverted repeats. For this reason, it is very important to monitor plasmid integrity and sequence identity. Cartwright *et al.* used PacBio single-molecule real-time technology to sequence amplimers of the plasmid before transfection and in stably transfected CHO cells at early and late generations. The study confirmed the genetic plasticity of the immortalized cell lines, reporting several low-frequency variants not identified in the plasmid before transfection, in one-third of the assessed samples [44], highlighting the potential for point mutations to impact recombinant gene sequences.

Paired-end short reads derived from RNA-seq or DNA-seq experiments of a recombinant cell line can also be specifically analysed to identify integration sites [45]. Enrichment strategies such as targeted locus amplification (TLA) or Cas-9 enrichment can be used with NGS to characterize the integrated plasmids and their surrounding regions, delivering a greater breadth of information than an amplicon-seq approach and being at the same time quicker and lower-cost than high-coverage genome sequencing. TLA is a cross-linking-based technique that generates complex DNA libraries, covering more than 100 kilobases surrounding a specific sequence [46,47]. In combination with NGS, this technique enables the complete sequencing and haplotyping of targeted regions of interest and is suitable for integration site

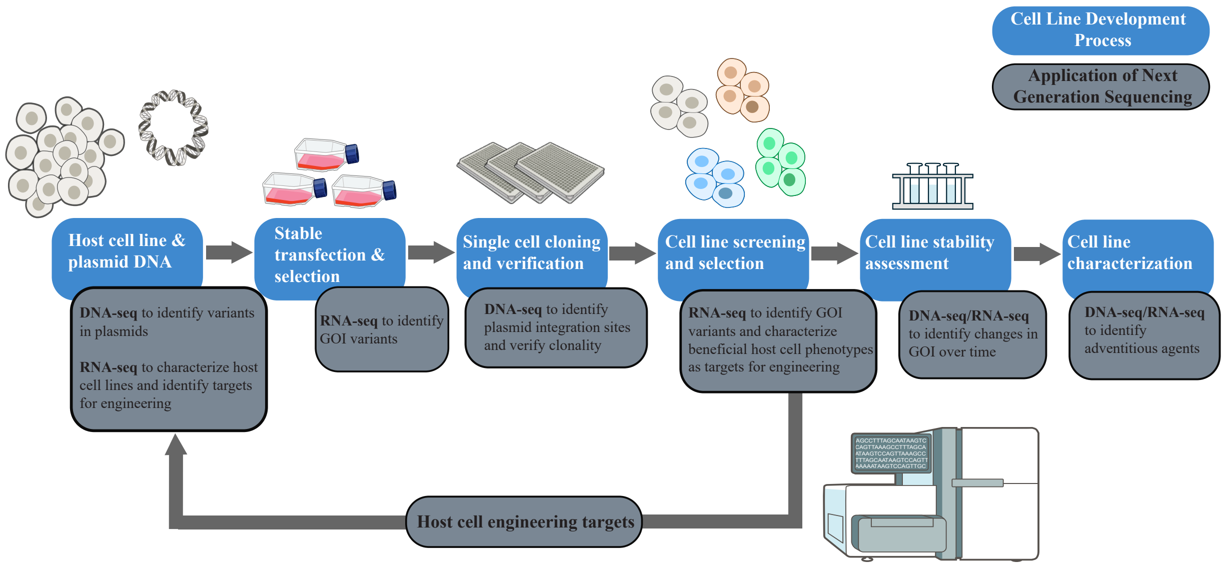
Cell Line Development Process
Application of Next Generation Sequencing
Host cell line & plasmid DNA
Stable transfection & selection
Single cell cloning and verification
Cell line screening and selection
Cell line stability assessment
Cell line characterization
DNA-seq to identify variants in plasmids
RNA-seq to characterize host cell lines and identify targets for engineering
RNA-seq to identify GOI variants
DNA-seq to identify plasmid integration sites and verify clonality
RNA-seq to identify GOI variants and characterize beneficial host cell phenotypes as targets for engineering
DNA-seq/RNA-seq to identify changes in GOI over time
DNA-seq/RNA-seq to identify adventitious agents
Host cell engineering targets

Fig. 1. This overview of cell line development (CLD) procedures highlights where next-generation sequencing (NGS) techniques can be applied throughout the process.

identification [48,49]. Cas9 is a component of the bacterial clustered regularly interspaced short palindromic repeats (CRISPR) system and can be used for targeted enrichment of specific genomic regions followed by long-read sequencing [50–52]. Owing to the length of the fragments, this approach results in more detailed information on the copy number, orientation, and structure of transgenes integrated into the same locus, which is otherwise difficult to infer with short-read sequencing. Where non-site-specific integration methods are used for cell line development, the integration sites, derived by the TLA/NGS or by the long-read sequencing experiments, are 'molecular fingerprints' univocally associated with an individual clonal cell line. Hence, integration site sequences can be used to verify the clonal origin of a cell line by using specifically designed qPCR primers to assess the integration sites of a defined number of subclones derived from the clone [53]. Clonality assurance can also be inferred by the statistical analysis of single-nucleotide variants derived by whole-genome sequencing characterizations. Khun *et al.* demonstrated that the most specific SNVs fixed in a clonally-derived cell line can be used to confirm clonal derivation of a cell line with high confidence and to quantify the clonal fractions present in non-clonal samples [54].

As with other mammalian cells, some genetic information related to mitochondrial function, including oxidative phosphorylation, is encoded by the mitochondrial genome in CHO cells. An analysis by Kelly *et al.* [55] revealed widespread mitochondrial DNA heteroplasmy in a panel of 22 CHO cell lines, in which 197 variants were identified with an allele frequency between 1 % and 99 %. Interestingly, loss-of-function mutations have been found with different frequencies among clones, suggesting a phenotypic effect of the mtDNA variation and providing a justification for the metabolic variability frequently observed in extended fed-batch cell culture [56].

The knowledge of the CHO genome and its annotation [57] provides the foundations for important CHO technology initiatives, such as the development of a genome-reduced CHO host [58] or the whole-genome CRISPR-based functional assays [35], aimed at developing the next-generation of CHO cell factories. Similarly, multiple genomic studies have been geared towards identifying expression hotspots suitable for targeted integration [59–61].

2. RNA-seq analysis to characterize the host, clone, and culture process transcriptomes

The transcriptome refers to the complete set of RNA transcripts produced by an organism at a given time and under specific conditions. It is composed of different types of RNAs (protein-coding and non-protein-coding) and reflects the expression of the genes at a given moment, providing insights into cellular functions and responses to environmental changes. Transcriptome analyses can be used to investigate the differential gene expression profiles that correlate with the phenotypic differences observed between hosts, as described in the comparison of CHO-S, CHO-K1, and DG44 CHO cell lines [62]. Könitzer *et al.* found that expression levels of sialyl transferases and enzymes synthesizing sialic acid precursors are correlated with differences in antibody glycosylation between CHO cell lines [63]. In a comparison of clones expressing monoclonal antibodies (mAbs), Sha *et al.* [64] found a good correlation between product titer and the transcript level of the GOI. They also identified overexpression of genes involved in secretion and protein transportation in the high-producer clones. Similar results have been confirmed by a recent study on different clonal cell lines expressing mAbs and bispecific antibodies during fed-batch production [65]. In another study [66], cell lines with high specific productivity (qP) for a mAb showed upregulation of genes involved in protein intracellular transport and surprisingly also upregulation of genes associated with apoptosis and cell death. The same study also identified significant phenotypic differences between subclones derived from the same parental cell, explained by the genome or epigenome plasticity of the CHO cell lines [67]. Another transcriptome analysis [68] demonstrated that the transcript level of the GOI was the major qP determinant, whilst the expression of many genes related to cell growth and housekeeping functions was negatively correlated with it. Transcriptomic analyses have also been useful to understand the mechanisms of action of some media components and to better characterize feeding strategies. Kretzmer *et al.* demonstrated that the addition of bromodeoxyuridine to the media influences the electron transport chain [69], whilst Schulze *et al.* demonstrated that the supplementation of butyric acid regulates the transcription of genes involved in cell proliferation and histone modification in an intensified CHO cell fed-batch process [70]. Genes involved in cell cycle and primary metabolism have also been found upregulated in feed-spiked cultures with a higher productivity [71]. RNA-seq was also able to identify furin as the main enzyme responsible for the undesirable cleavage of an IgG4 Fc-fusion protein [72].

MicroRNAs (miRNAs) and long non-coding RNAs (lncRNAs) are two major families of the non-protein-coding transcripts and are important regulators of gene expression.

MiRNAs are endogenous short RNAs (21–25 nucleotides) that promote cleavage or translational repression of their targets; they have important post-transcriptional regulatory roles in animals and plants [73] and have also been identified in CHO cell lines producing recombinant proteins. Their short length and ability to simultaneously target multiple genes make them ideal candidates for gene manipulation in biotherapeutic protein-expressing cell lines [74–76]. Several studies uncovering different mechanisms used by microRNAs to influence the production of recombinant proteins in CHO cell lines have been summarised in a review by Liu *et al.* [77]. LncRNAs are non-protein-coding transcripts greater than 500 nucleotides in length, mostly generated by RNA polymerase II, and have varied functions [78]. They have been shown to have a key role in gene regulation [79] and a recent CRISPR-Cas13 screen in five human cell lines demonstrated essentiality for some of them, independently of their nearest protein-coding genes [80]. RNA-seq has been used to identify lncRNAs in CHO cell lines [81,82], and some studies correlated their expression with cell growth and productivity [83,84].

Single-cell RNA-seq (scRNA-seq) is a high-resolution measurement of gene expression, made at the level of individual cells, capable of identifying rare cell types and unravelling the cellular dynamics that shape tissue heterogeneity. Whilst bulk RNA-seq provides a robust and cost-effective method to assess the overall gene expression landscape of a sample, scRNA-seq maximises the resolution of the assay, exploring cellular heterogeneity and discovering novel cell types, but it comes with a higher cost and a more complex data analysis and interpretation. Tzani *et al.* [85] used single cell transcriptomics to investigate the production instability of a CHO cell line expressing a mAb. They identified transcriptional heterogeneity in the cell population, with a reduced expression of the transgene and of the genes associated with protein production and assembly. Another scRNA study demonstrated transcriptome homogeneity in suspension CHO-K1 and adherent HEK293FT cells under standard culture conditions, with most of the observed differences primarily driven by the cell cycle [86].

Molecular mechanisms determining an optimal producer cell line are often complex and, generally, not simply related to a single gene or gene product [87,88]. Transcriptome analyses can be used to better describe system-wide properties of a given host, clone, or culture process, and this information is valuable for the optimization of cell lines, media, and fermentation processes. Nevertheless, the high number of potential target genes and the lack of information on the mechanistic models governing biotechnological processes can be a limitation for transcriptomic analyses and, in general, of all omics experiments, challenging the observational results to become actionable [89]. Multi-omics approaches integrated with system biology models represent a valid strategy to provide a holistic view of the biological state of a sample. They can indeed capture the complexity of the processes more completely than any single omics approach and have been successfully applied to better understand diseases [90–92], to personalise treatments and optimize therapies in precision medicine [93–95], and also to better characterise recombinant CHO cell lines [96–99].

3. RNA-seq applications to detect variants in the GOI

Protein primary structure is a critical quality attribute for biotherapeutics [100,101]. A sequence variant of the GOI can indeed generate unwanted impurities, with consequences for the efficacy, stability, and safety of the final product. Different types of variants exist: single-nucleotide variants (SNVs) are the smallest genomic variants and consist of differences in a single nucleotide, whilst insertion/deletion variants (indels) are caused by extra or missing nucleotides and typically affect fewer than 50 nucleotides [102]. Structural variants (SV) are DNA rearrangements involving larger regions, which can include inversions and balanced translocations or large insertions or deletions [103]. The application of NGS analyses to characterize cell lines can be equally informative and, at the same time, faster than the traditional mass spectrometry (MS)-based techniques [104] used to detect sequence variants in the product itself. Furthermore, NGS data analyses can inform on the underlying causes of sequence variants.

Differences in the library preparation, sequencing chemistry, and bioinformatic analysis used can result in varying levels of sensitivity and specificity of the analyses. Monitoring strategies based on amplicon sequencing of the GOI have been proven successful in CLD campaigns with confirmatory evidence from high-throughput protein analytical assays for variants with frequencies higher than 5 % [105]. Zhang *et al.* [106] demonstrated an increase in the occurrence of low-frequency (0.1 %-0.5 %) SNVs proportional to the population doubling levels (PDLs), confirming the genomic instability of the characterized cell lines. Interestingly, they also established a high-confidence lower limit of detection of 0.1 % in RNA-seq experiments.

Unique molecular identifiers (UMIs) are short arbitrary oligonucleotide sequences attached to the library fragments before the PCR amplification step and can be used to increase the sensitivity and specificity of the assay. These unique tags, specific for each fragment, can be used to distinguish PCR duplicates of a single fragment from different fragments derived from a given transcript in RNA sequencing experiments [107]. Collapsing the amplimers and generating a high-quality consensus of PCR amplimer is also a valid strategy for reducing the rate of false positives with both short- and long-read sequencing technologies [108–112]. Untargeted RNA-seq experiments have, by definition, a less biased design compared with amplicon-seq experiments and hence can reveal the presence of chimeric transcripts or non-designed splicing products. The de novo reconstruction of untargeted RNA-seq short reads revealed mis-splicing and intron retention events in the heavy chain gene of a mAb-producing CHO clone [113]. These transcripts produced truncated heavy chain products and their assembly with light chain products mimicked the appearance of fragments identified by routine purity assays. In another mAb-producing CHO cell line, a similar approach combining MS analysis and genome sequencing was able to identify an abnormal heavy chain fusion transcript as the underlying cause of an aberrant heavy chain peptide with an extra mass of 11 kDa [114].

NGS analyses have been successfully employed for a wide range of applications and have demonstrated great sensitivity; however, these analyses are not exempt from challenges. The sequencing coverage in the experiment design and technical variables introduced during sample preparation, library construction, sequencing and bioinformatic analysis are factors determining false-positive and false-negative results [115]. CHO cell lines are easy to culture and generate samples, and this offers the possibility to perform experiments in replicates and/or in parallel to proteomics experiments. These can be integrated to provide a proteogenomic analysis of the results [116].

4. Host, clone, and culture process epigenomic characterization

The epigenome refers to the reversible modifications, on the DNA and/or its binding proteins (like histones), affecting gene expression without altering the DNA sequence [117]. These events can influence the stability of the GOI expression during long-term culture [118,119] and have hence been the subject of many important studies. A seminal study pointed out differences among the genomes of six related CHO cell lines and demonstrated that culture conditions, media, and specific phenotypes are associated with distinct states of DNA methylation, maintained in the transition between exponential and stationary growth phases [120]. A subsequent study [121] demonstrated that DNA methylation in the promoter region is an ON/OFF switch for gene expression and it is followed by other layers of regulation involving histones and differentially expressed lncRNAs. Marx *et al.* [122] demonstrated that DNA methylation of promoters, either recombinant or endogenous, leads to the loss of active histone marks in favour of distinct repressive heterochromatin marks. These results are in agreement with the findings that targeting the DNA methyltransferase is a successful epigenetic strategy to enhance the expression level of the GOI and to increase the stability of expression in stable [123] and transient [124] systems. Dhiman *et al.* compared the genome and the transcriptome of stable and unstable CHO cell lines, with low, medium, or high copy number of the GOI [125]. They demonstrated that the integration site is a major determinant for the expression stability and developed a genome catalogue of loci that are favourable and unfavourable for targeted transgene integration.

NGS is the basis of multiple assays to characterize recombinant cell lines expressing therapeutic proteins and provides a comprehensive, high-throughput approach to several genetic analyses. Applications, largely employed in different biomedical fields, have the potential to uncover important aspects of the CHO biology. One example of this is the use of chromosome conformation capture techniques, which are employed to determine the physical interaction of genomic regions from the same or different chromosomes with each other and with nuclear structures [126]. In particular promoter capture Hi-C has been used to identify important transcriptional mechanisms mediated by long-range interactions between regulatory elements and gene promoters [127,128]. This approach, combined with RNA-seq, can disclose the promoter-enhancer interactions, determining actively transcribed regions, also named transcription factories [129], and can be employed to describe the dynamic interactions of the GOI promoters during CHO cell culture processes. This analysis could also take advantage of the transcription starting sites atlas in CHO-K1 cell lines and Chinese hamster tissues published in a recent study [130].

Here, this review has summarised several studies using genome, epigenome and transcriptome analyses to uncover important features of recombinant CHO cell lines, disclosing mechanisms determining cell productivity and quality of the final product. These NGS assays aim to identify bottlenecks and opportunities to optimise the efficiency, and the productivity of biotechnological processes used to produce biotherapeutics. Typically, these assays are applied to pilot projects to support increased mechanistic understanding and platform development, due to cost and the uncertainty associated with execution time, interpretation and the actionability of the results. On the other hand, other assays like variant analysis of the GOI and clonality assays based on TLA analysis followed by specific insertion site qPCR can be routinely executed during CLD. This is possible thanks to multiple factors, including contained cost, defined execution timelines and standardised protocols for sample preparation, experimental set up and bioinformatics analyses. Overall NGS analyses constitute a comprehensive toolset to ensure quality, clonality, and stability of cell lines expressing biotherapeutics and can be used to enhance cell line development and cell culture process optimization which in turn have the potential to improve development timelines and reduce production costs for biologics.

CRediT authorship contribution statement

Hatton Diane: Writing – review & editing, Writing – original draft, Conceptualization. **Dunn Sarah:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Grassi Luigi:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Harris Claire:** Writing – review & editing, Visualization, Conceptualization. **Zhu Jie:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships that may be considered as potential competing interests: This work was supported by Biopharmaceutical Development, AstraZeneca. Authors L. G., C. H., J. Z., D. H. and S. D. are employees of AstraZeneca and have stock and/or stock interests or options in AstraZeneca.

Acknowledgements

Figures used in Fig. 1 have been downloaded from NIAID NIH BIOART [131] and licensed as 'Public Domain'.

References

[1] Yang W, Zhang J, Xiao Y, Li W, Wang T. Screening strategies for high-yield chinese hamster ovary cell clones. Front Bioeng Biotechnol 2022;10:858478.

[2] Mortensen R, Chesnut JD, Hoeffler JP, Kingston RE. Selection of transfected mammalian cells. Curr Protoc Mol Biol Chapter 9 2003:5.

[3] Kingston RE. Stable transfer of genes into mammalian cells. Curr Protoc Immunol Chapter 10 2001.

[4] Mortensen RM, Kingston RE. Selection of transfected mammalian cells. Curr Protoc Mol Biol Chapter 9, Unit 9.5 2009.

[5] Rajendran S, et al. Accelerating and de-risking CMC development with transposon-derived manufacturing cell lines. Biotechnol Bioeng 2021;118:2301–11.

[6] Wei M, Mi C-L, Jing C-Q, Wang T-Y. Progress of transposon vector system for production of recombinant therapeutic proteins in mammalian cells. Front Bioeng Biotechnol 2022;10:879222.

[7] Walsh G, Walsh E. Biopharmaceutical benchmarks 2022. Nat Biotechnol 2022;40:1722–60.

[8] Puck TT. The genetics of somatic mammalian cells. Adv Biol Med Phys 1957;5:75–101.

[9] Wurm FM. CHO Quasispecies—Implications for Manufacturing Processes. Processes 2013;1:296–311.

[10] Wurm MJ, Wurm FM. Naming CHO cells for bio-manufacturing: Genome plasticity and variant phenotypes of cell populations in bioreactors question the relevance of old names. Biotechnol J 2021;16:e2100165.

[11] Kim JY, Kim Y-G, Lee GM. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl Microbiol Biotechnol 2012;93:917–30.

[12] Wiberg FC, et al. Production of target-specific recombinant human polyclonal antibodies in mammalian cells. Biotechnol Bioeng 2006;94:396–405.

[13] Spahn PN, et al. Restoration of DNA repair mitigates genome instability and increases productivity of Chinese hamster ovary cells. Biotechnol Bioeng 2022;119:963–82.

[14] Baik JY, Lee KH. Growth rate changes in CHO host cells are associated with karyotypic heterogeneity. Biotechnol J 2018;13:e1700230.

[15] Bandyopadhyay AA, et al. Recurring genomic structural variation leads to clonal instability and loss of productivity. Biotechnol Bioeng 2019;116:41–53.

[16] Zhang S, et al. Identifying low-level sequence variants via next generation sequencing to aid stable CHO cell line screening. Biotechnol Prog 2015;31:1077–85.

[17] Dahodwala H, Lee KH. The fickle CHO: a review of the causes, implications, and potential alleviation of the CHO cell line instability problem. Curr Opin Biotechnol 2019;60:128–37.

[18] Sha S, Agarabi C, Brorson K, Lee D-Y, Yoon S. N-Glycosylation design and control of therapeutic monoclonal antibodies. Trends Biotechnol 2016;34:835–46.

[19] Zhong Y, Xu F, Wu J, Schubert J, Li MM. Application of next generation sequencing in laboratory medicine. Ann Lab Med 2021;41:25–43.

[20] Alekseyev YO, et al. A next-generation sequencing primer-how does it work and what can it do?. 2374289518766521 Acad Pathol 2018;5. 2374289518766521.

[21] Satam H, et al. Next-generation sequencing technology: current trends and advancements. Biology 2023;12:997.

[22] Marx V. Method of the year: long-read sequencing. Nat Methods 2023;20:6–11.

[23] Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: an overview. Hum Immunol 2021;82:801–11.

[24] Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Curr Protoc Mol Biol 2018;122:e59.

[25] ICH Q5A viral safety guidance. ⟨https://www.ema.europa.eu/en/documents/scientific-guideline/ich-q5ar2-guideline-viral-safety-evaluation-biotechnology-products-derived-cell-lines-human-or-animal-origin-step-5_en.pdf⟩.

[26] Bova RA, et al. Validation of a next generation sequencing method for adventitious virus detection: Demonstration of sensitivity in multiple cell lines. Biol J Int Assoc Biol Stand 2024;86:101771.

[27] Khan AS, et al. Report of the third conference on next-generation sequencing for adventitious virus detection in biologics for humans and animals. Biol J Int Assoc Biol Stand 2023;83:101696.

[28] Khan AS, et al. A multicenter study to evaluate the performance of high-throughput sequencing for virus detection. mSphere 2017;2:e00307-17.

[29] Ng SH, et al. Current perspectives on high-throughput sequencing (HTS) for adventitious virus detection: upstream sample processing and library preparation. Viruses 2018;10:566.

[30] Hirai T, et al. Evaluation of next-generation sequencing performance for in vitro detection of viruses in biological products. Biol J Int Assoc Biol Stand 2024;85:101739.

[31] Wilson CA, Simonyan V. FDA's activities supporting regulatory application of 'next gen' sequencing technologies. PDA J Pharm Sci Technol 2014;68:626–30.

[32] Hilliard W, MacDonald ML, Lee KH. Chromosome-scale scaffolds for the Chinese hamster reference genome assembly to facilitate the study of the CHO epigenome. Biotechnol Bioeng 2020;117:2331–9.

[33] Lewis NE, et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. Nat Biotechnol 2013;31:759–65.

[34] Xu X, et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotechnol 2011;29:735–41.

[35] Kretzmer C, et al. De novo assembly and annotation of the CHOZN® GS-/- genome supports high-throughput genome-scale screening. Biotechnol Bioeng 2022;119:3632–46.

[36] Brinkrolf K, et al. Chinese hamster genome sequenced from sorted chromosomes. Nat Biotechnol 2013;31:694–5.

[37] Rupp O, et al. A reference genome of the Chinese hamster based on a hybrid assembly strategy. Biotechnol Bioeng 2018;115:2087–100.

[38] Hammond S, Swanberg JC, Kaplarevic M, Lee KH. Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology. BMC Genom 2011;12:67.

[39] Kaas CS, Kristensen C, Betenbaugh MJ, Andersen MR. Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy. BMC Genom 2015;16:160.

[40] Holst-Jensen A, et al. Application of whole genome shotgun sequencing for detection and characterization of genetically modified organisms and derived products. Anal Bioanal Chem 2016;408:4595–614.

[41] Li W, et al. Arrayed in vivo barcoding for multiplexed sequence verification of plasmid DNA and demultiplexing of pooled libraries. Nucleic Acids Res 2024;52:e47.

[42] Brown SD, Dreolini L, Wilson JF, Balasundaram M, Holt RA. Complete sequence verification of plasmid DNA using the Oxford Nanopore Technologies' MinION device. BMC Bioinforma 2023;24:116.

[43] Mumm C, et al. Multiplexed long-read plasmid validation and analysis using OnRamp. Genome Res 2023;33:741–9.

[44] Cartwright JF, Anderson K, Longworth J, Lobb P, James DC. Highly sensitive detection of mutations in CHO cell recombinant DNA using multi-parallel single molecule real-time DNA sequencing. Biotechnol Bioeng 2018;115:1485–98.

[45] Grassi L, Harris C, Zhu J, Hardman C, Hatton D. DetectIS: a pipeline to rapidly detect exogenous DNA integration sites using DNA or RNA paired-end sequencing data. Bioinforma Oxf Engl 2021;37:4230–2.

[46] de Vree PJP, et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat Biotechnol 2014;32:1019–25.

[47] Hottentot QP, van Min M, Splinter E, White SJ. Targeted locus amplification and next-generation sequencing. Methods Mol Biol Clifton NJ 2017;1492:185–96.

[48] Stadermann A, et al. Structural analysis of random transgene integration in CHO manufacturing cell lines by targeted sequencing. Biotechnol Bioeng 2022;119:868–80.

[49] O'Brien SA, Ojha J, Wu P, Hu W-S. Multiplexed clonality verification of cell lines for protein biologic production. Biotechnol Prog 2020;36:e2978.

[50] Slesarev A, et al. CRISPR/CAS9 targeted CAPTURE of mammalian genomic regions for characterization by NGS. Sci Rep 2019;9:3587.

[51] Leitner K, Motheramgari K, Borth N, Marx N. Nanopore Cas9-targeted sequencing enables accurate and simultaneous identification of transgene integration sites, their structure and epigenetic status in recombinant Chinese hamster ovary cells. Biotechnol Bioeng 2023;120:2403–18.

[52] Clappier C, et al. Deciphering integration loci of CHO manufacturing cell lines using long read nanopore sequencing. N Biotechnol 2023;75:31–9.

[53] Aebischer-Gumy C, Moretti P, Little TA, Bertschinger M. Analytical assessment of clonal derivation of eukaryotic/CHO cell populations. J Biotechnol 2018;286:17–26.

[54] Kuhn A, Le Fourn V, Fisch I, Mermod N. Genome-wide analysis of single nucleotide variants allows for robust and accurate assessment of clonal derivation in cell lines used to produce biologics. Biotechnol Bioeng 2020;117:3628–38.

[55] Kelly PS, et al. Ultra-deep next generation mitochondrial genome sequencing reveals widespread heteroplasmy in Chinese hamster ovary cells. Metab Eng 2017;41:11–22.

[56] Gilbert A, McElearney K, Kshirsagar R, Sinacore MS, Ryll T. Investigation of metabolic variability observed in extended fed batch cell culture. Biotechnol Prog 2013;29:1519–27.

[57] Hefzi H, et al. A Consensus genome-scale reconstruction of chinese hamster ovary cell metabolism. Cell Syst 2016;3:434–443.e8.

[58] Jerabek T, et al. In pursuit of a minimal CHO genome: establishment of large-scale genome deletions. N Biotechnol 2024;79:100–10.

[59] Hilliard W, Lee KH. A compendium of stable hotspots in the CHO genome. Biotechnol Bioeng 2023;120:2133–43.

[60] Hilliard W, Lee KH. Systematic identification of safe harbor regions in the CHO genome through a comprehensive epigenome analysis. Biotechnol Bioeng 2021; 118:659–75.

[61] Woo HJ, et al. Context-dependent genomic locus effects on antibody production in recombinant Chinese hamster ovary cells generated through random integration. Comput Struct Biotechnol J 2024;23:1654–65.

[62] Singh A, Kildegaard HF, Andersen MR. An online compendium of CHO RNA-seq data allows identification of CHO cell line-specific transcriptomic signatures. Biotechnol J 2018;13:e1800070.

[63] Könitzer JD, et al. A global RNA-seq-driven analysis of CHO host and production cell lines reveals distinct differential expression patterns of genes contributing to recombinant antibody glycosylation. Biotechnol J 2015;10:1412–23.

[64] Sha S, Bhatia H, Yoon S. An RNA-seq based transcriptomic investigation into the productivity and growth variants with Chinese hamster ovary cells. J Biotechnol 2018;271:37–46.

[65] Bai Y, et al. Identification of cellular signatures associated with chinese hamster ovary cell adaptation for secretion of antibodies. Comput Struct Biotechnol J 2025;27:17–31.

[66] Orellana CA, et al. RNA-seq highlights high clonal variation in monoclonal antibody producing CHO cells. Biotechnol J 2018;13:e1700231.

[67] Weinguny M, et al. Subcloning induces changes in the DNA-methylation pattern of outgrowing Chinese hamster ovary cell colonies. Biotechnol J 2021;16: e2000350.

[68] Fomina-Yadlin D, et al. Transcriptome analysis of a CHO cell line expressing a recombinant therapeutic protein treated with inducers of protein expression. J Biotechnol 2015;212:106–15.

[69] Kretzmer C, et al. Chemical and genetic modulation of complex I of the electron transport chain enhances the biotherapeutic protein production capacity of CHO cells. Cells 2023;12:2661.

[70] Schulze M, et al. Transcriptomic analysis reveals mode of action of butyric acid supplementation in an intensified CHO cell fed-batch process. Biotechnol Bioeng 2022;119:2359–73.

[71] Reinhart D, Damjanovic L, Castan A, Ernst W, Kunert R. Differential gene expression of a feed-spiked super-producing CHO cell line. J Biotechnol 2018; 285:23–37.

[72] Clarke C, et al. Transcriptomic analysis of IgG4 Fc-fusion protein degradation in a panel of clonally-derived CHO cell lines using RNASeq. Biotechnol Bioeng 2019; 116:1556–62.

[73] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281–97.

[74] Müller D, Katinger H, Grillari J. MicroRNAs as targets for engineering of CHO cell factories. Trends Biotechnol 2008;26:359–65.

[75] Hackl M, et al. Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: Identification, annotation and profiling of microRNAs as targets for cellular engineering. J Biotechnol 2011;153:62–75.

[76] Jadhav V, et al. CHO microRNA engineering is growing up: recent successes and future challenges. Biotechnol Adv 2013;31:1501–13.

[77] Liu H-N, Dong W-H, Lin Y, Zhang Z-H, Wang T-Y. The effect of microRNA on the production of recombinant protein in CHO cells and its mechanism. Front Bioeng Biotechnol 2022;10:832065.

[78] Mattick JS, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 2023;24:430–47.

[79] Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 2021;22:96–118.

[80] Liang W-W, et al. Transcriptome-scale RNA-targeting CRISPR screens reveal essential lncRNAs in human cells. Cell 2024;187:7637–7654.e29.

[81] Motheramgari K, et al. Expanding the Chinese hamster ovary cell long noncoding RNA transcriptome using RNASeq. Biotechnol Bioeng 2020;117:3224–31.

[82] Vito D, Smales CM. The long non-coding RNA transcriptome landscape in CHO cells under batch and fed-batch conditions. Biotechnol J 2018;13:e1800122.

[83] Vito D, et al. Defining lncRNAs correlated with CHO cell growth and IgG productivity by RNA-Seq. iScience 2020;23:100785.

[84] Novak N, et al. LncRNA analysis of mAb producing CHO clones reveals marker and engineering potential. Metab Eng 2023;78:26–40.

[85] Tzani I, et al. Tracing production instability in a clonally derived CHO cell line using single-cell transcriptomics. Biotechnol Bioeng 2021;118:2016–30.

[86] Borsi G, et al. Single-cell RNA sequencing reveals homogeneous transcriptome patterns and low variance in a suspension CHO-K1 and an adherent HEK293FT cell line in culture conditions. J Biotechnol 2023;364:13–22.

[87] Griffin TJ, Seth G, Xie H, Bandhakavi S, Hu W-S. Advancing mammalian cell culture engineering using genome-scale technologies. Trends Biotechnol 2007;25: 401–8.

[88] Monger C, et al. Towards next generation CHO cell biology: bioinformatics methods for RNA-Seq-based expression profiling. Biotechnol J 2015;10:950–66.

[89] Masson HO, Karottki, la C KJ, Tat J, Hefzi H, Lewis NE. From observational to actionable: rethinking omics in biologics production. Trends Biotechnol 2023;41: 1127–38.

[90] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017; 18:83.

[91] Wang X, Fridley BL. Multi-omics data deconvolution and integration: new methods, insights, and translational implications. Methods Mol Biol Clifton NJ 2023;2629:1–9.

[92] Chen C, et al. Applications of multi-omics analysis in human diseases. MedComm 2023;4:e315.

[93] Bernal-Casas D, Serrano-Marín J, Sánchez-Navés J, Oller JM, Franco R. Advancing personalized medicine by analytical means: selection of three. Metab That Allows Discrim Glaucoma, Diabetes, Controls *Metab* 2024;14:149.

[94] Acharya D, Mukhopadhyay A. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. Brief Funct Genom 2024;23:549–60.

[95] Lunke S, et al. Integrated multi-omics for rapid rare disease diagnosis on a national scale. Nat Med 2023;29:1681–91.

[96] Yusufi FNK, et al. Mammalian systems biotechnology reveals global cellular adaptations in a recombinant CHO cell line. Cell Syst 2017;4:530–542.e6.

[97] Lin D, et al. CHOmics: a web-based tool for multi-omics data analysis and interactive visualization in CHO cell lines. PLoS Comput Biol 2020;16:e1008498.

[98] Lee AP, et al. Multi-omics profiling of a CHO cell culture system unravels the effect of culture pH on cell growth, antibody titer, and product quality. Biotechnol Bioeng 2021;118:4305–16.

[99] Gopalakrishnan S, et al. Multi-omic characterization of antibody-producing CHO cell lines elucidates metabolic reprogramming and nutrient uptake bottlenecks. Metab Eng 2024;85:94–104.

[100] Alt N, et al. Determination of critical quality attributes for monoclonal antibodies using quality by design principles. Biol J Int Assoc Biol Stand 2016;44:291–305.

[101] Gutierrez L, Cauchon NS, Christian TR, Giffin MJ, Abernathy MJ. The confluence of innovation in therapeutics and regulation: recent CMC considerations. J Pharm Sci 2020;109:3524–34.

[102] genome.gov. ⟨https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation⟩.

[103] SV -ncbi. ncbi ⟨https://www.ncbi.nlm.nih.gov/dbvar/content/overview/#:~:text=I.-,Introduction,copy%20number%20variants%20⟩(CNVs).

[104] Lin TJ, et al. Evolution of a comprehensive, orthogonal approach to sequence variant analysis for biotherapeutics. mAbs 2019;11:1–12.

[105] Wright C, et al. Genetic mutation analysis at early stages of cell line development using next generation sequencing. Biotechnol Prog 2016;32:813–7.

[106] Zhang S, et al. Mutation detection in an antibody-producing chinese hamster ovary cell line by targeted RNA sequencing. BioMed Res Int 2016;2016:8356435.

[107] Islam S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 2014;11:163–6.

[108] Sun J, et al. Correcting PCR amplification errors in unique molecular identifiers to generate accurate numbers of sequencing molecules. Nat Methods 2024;21: 401–5.

[109] Lin X, et al. High accuracy meets high throughput for near full-length 16S ribosomal RNA amplicon sequencing on the Nanopore platform. PNAS Nexus 2024;3:pgae411.

[110] Girardot C, Scholtalbers J, Sauer S, Su S-Y, Furlong EEM. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. BMC Bioinforma 2016;17:419.

[111] Clement K, Farouni R, Bauer DE, Pinello L. AmpUMI: design and analysis of unique molecular identifiers for deep amplicon sequencing. Bioinforma Oxf Engl 2018;34.

[112] Kou R, et al. Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations. PloS One 2016;11: e0146638.

[113] Delmar JA, et al. Monoclonal antibody sequence variants disguised as fragments: identification, characterization, and their removal by purification process optimization. J Pharm Sci 2022;111:3009–16.

[114] Harris C, et al. Identification and characterization of an IgG sequence variant with an 11 kDa heavy chain C-terminal extension using a combination of mass spectrometry and high-throughput sequencing analysis. mAbs 2019;11:1452–63.

[115] Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet 2017;18:473–84.

[116] Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11:1114–25.

[117] Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell 2007;128: 669–81.

[118] Kim M, O'Callaghan PM, Droms KA, James DC. A mechanistic understanding of production instability in CHO cell lines expressing recombinant monoclonal antibodies. Biotechnol Bioeng 2011;108:2434–46.

[119] Veith N, Ziehr H, MacLeod RAF, Reamon-Buettner SM. Mechanisms underlying epigenetic and transcriptional heterogeneity in Chinese hamster ovary (CHO) cell lines. BMC Biotechnol 2016;16:6.

[120] Feichtinger J, et al. Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. Biotechnol Bioeng 2016;113:2241–53.

[121] Hernandez I, et al. Epigenetic regulation of gene expression in Chinese Hamster Ovary cells in response to the changing environment of a batch culture. Biotechnol Bioeng 2019;116:677–92.

[122] Marx N, et al. Enhanced targeted DNA methylation of the CMV and endogenous promoters with dCas9-DNMT3A3L entails distinct subsequent histone modification changes in CHO cells. Metab Eng 2021;66:268–82.

[123] Jia Y-L, et al. CRISPR/Cas9-mediated gene knockout for DNA methyltransferase Dnmt3a in CHO cells displays enhanced transgenic expression and long-term stability. J Cell Mol Med 2018;22:4106–16.

[124] Wang X-Y, et al. Enhancing expression level and stability of transgene mediated by episomal vector via buffering DNA methyltransferase in transfected CHO cells. J Cell Biochem 2019;120:15661–70.

[125] Dhiman H, Campbell M, Melcher M, Smith KD, Borth N. Predicting favorable landing pads for targeted integrations in Chinese hamster ovary cell lines by learning stability characteristics from random transgene integrations. Comput Struct Biotechnol J 2020;18:3632–48.

[126] Hakim O, Misteli T. SnapShot: chromosome confirmation capture. 1068.e1–2 Cell 2012;148. 1068.e1–2.

[127] Song M, et al. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nat Genet 2019;51:1252–62.

[128] Javierre BM, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 2016;167:1369–1384.e19.

[129] Rieder D, Trajanoski Z, McNally JG. Transcription factories. Front Genet 2012;3:221.

[130] Shamie I, et al. A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells. NAR Genom Bioinforma 2021;3.

[131] bioart. ⟨https://bioart.niaid.nih.gov⟩.

Glossary

Alignment: The process of comparing a sequence of DNA or RNA with a reference genome or transcriptome. It is required by bioinformatic workflows used for different analyses likewise variant calling, gene expression quantification, RNA-seq, CHIP-seq, ATAC-seq.

ATAC-sequencing: Assay for transposase-accessible chromatin sequencing that combines active transposase treatment of the genome with NGS. It is used to unravel, genome-wide chromatin accessible regions.

ChIP-sequencing: Chromatin immune precipitation sequencing (ChIP-Seq) that combines chromatin immunoprecipitation with NGS to identify genome-wide DNA binding sites for proteins (transcription factors, histones or other proteins interacting with the DNA).

Coverage: Also known as sequencing depth, reports the number of times a nucleotide is sequenced, averaged on the size of the sequenced target (genome, transcriptome etc.). Higher coverage increases the accuracy of sequencing results.

Chromosome conformation capture: Techniques used to determine physical interactions of genomic regions. Alternative methods have been developed with different scopes: 3 C detects interaction between known sites; 4 C detects unknown interactors of a known bait sequence; 5 C can identify all regions interacting within a given genome domain; and Hi-C is a genome-wide assay.

Library Preparation: Preparation of DNA or RNA samples used for the sequencing, which might include the fragmentation of the material and the addition of adapters to the ends of the fragments for sequencing.

Long-read sequencing: Technology able to decipher the sequence of long DNA or RNA fragments (typically at least 10,000 nucleotides) in a single continuous string.

RNA-sequencing: RNA-sequencing (RNA-seq) is the characterization of the RNA transcripts present in a cell or tissue at a given time. It provides insights into gene expression profiles and functional characteristics of transcripts, likewise splicing events or presence of variants.

Short-read sequencing: Technology used to decipher the sequence of DNA or RNA fragments in continuous strings of short length (hundreds of nucleotides). Single-end technology employs a single read to cover only one end of the sequenced fragment, whilst the paired-end provides more detailed information with two reads covering both ends of the fragment.

Variant: A difference between the sequenced DNA or RNA and the reference genome or transcriptome. Variants consisting in substitutions of single nucleotides are defined single nucleotide variants (SNVs), when they affect the coding sequence of a gene they can be further classified as synonymous or non-synonymous or nonsense. Duplications, insertions, inversions, and translocations describe other types of variants. Structural variants (SVs) are alterations encompassing at least 50 base pairs.