Computational and Structural Biotechnology Journal 18 (2020) 3632–3648

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Predicting favorable landing pads for targeted integrations in Chinese hamster ovary cell lines by learning stability characteristics from random transgene integrations

Heena Dhiman [a,b,1], Marguerite Campbell [c,*], Michael Melcher [a], Kevin D. Smith [c], Nicole Borth [a,b,*]

[a] University of Natural Resources and Life Sciences, Vienna, Austria
[b] Austrian Centre of Industrial Biotechnology, Vienna, Austria
[c] Janssen Research & Development, PA, USA

ARTICLE INFO

Article history:
Received 24 July 2020
Received in revised form 4 November 2020
Accepted 4 November 2020
Available online 12 November 2020

Keywords:
Genomics
Transcriptomics
Integration sites
Stable transgene expression
Chinese hamster ovary cells

ABSTRACT

Chinese Hamster Ovary (CHO) cell lines are considered to be the preferred platform for the production of biotherapeutics, but issues related to expression instability remain unresolved. In this study, we investigated potential causes for an unstable phenotype by comparing cell lines that express stably to such that undergo loss in titer across 10 passages. Factors related to transgene integrity and copy number as well as the genomic profile around the integration sites were analyzed. Horizon Discovery CHO-K1 (HD-BIOP3) derived production cell lines selected for phenotypes with low, medium or high copy number, each with stable and unstable transgene expression, were sequenced to capture changes at genomic and transcriptomic levels. The exact sites of the random integration events in each cell line were also identified, followed by profiling of the genomic, transcriptomic and epigenetic patterns around them. Based on the information deduced from these random integration events, genomic loci that potentially favor reliable and stable transgene expression were reported for use as targeted transgene integration sites. By comparing stable vs unstable phenotypes across these parameters, we could establish that expression stability may be controlled at three levels: 1) Good choice of integration site, 2) Ensuring integrity of transgene and observing concatemerization pattern after integration, and 3) Checking for potential stress related cellular processes. Genome wide favorable and unfavorable genomic loci for targeted transgene integration can be browsed at https://www.borthlabchoresources.boku.ac.at/

1. Introduction

Biopharmaceuticals have become an integral part of modern medicine since the production of insulin using recombinant DNA technology in 1978 [1]. Since then various recombinant protein therapeutics have been industrially produced to treat conditions ranging from cancer to infertility. These approved products include monoclonal antibodies, hormones, clotting factors, enzymes, vaccines and nucleic acid-based products [2]. While a high product titer is desirable, product quality such as correct folding and post-translational modifications that are compatible with humans, is of key importance. This leads to very specific requirements that the production host needs to fulfill. The production of biologically active forms of tissue plasminogen activator (tPA) using CHO cells in 1987 marked the beginning of the CHO-based therapeutics era [3]. Due to their adaptability, the robustness of CHO cells also enables growth and protein production in a variety of culture conditions. The continuous increase in demand for biopharmaceuticals has been noticed over the past, with estimated total sales of $140 billion in 2013 [4] and $188 billion in 2017 [2]. Today, over 70% of recombinant biopharmaceuticals are produced in CHO cell factories [5]. Monoclonal antibodies, 84% of which are produced in CHO expression systems, have been reported to dominate the biopharmaceutical market [2,5].

While these cell lines are considered the preferred cell factory for the production of biotherapeutics, the instability of transgene expression during scale up and manufacture presents significant challenges to the biopharmaceutical industry in the context of reliable yields and regulatory approval for the product of interest.

* Corresponding authors at: University of Natural Resources and Life Sciences, Vienna, Austria (N. Borth).
E-mail address: nicole.borth@boku.ac.at (N. Borth).
[1] The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.

https://doi.org/10.1016/j.csbj.2020.11.008

Conventional strategies based on 'random transgene integration' have been considered to be one of the causes for instability and potential product heterogeneity [6,7]. This is primarily due to the impact of the genomic profile on the integration site. Over the years, many studies investigated factors associated with transgene integration and subsequent perturbation in molecular mechanisms that in turn may disturb protein production. It has been observed that genomic rearrangements at the transgene integration site can result in 'loss of transgene copies' over time [8]. Subsequent epigenetic modifications are further expected to be a cause of 'transcriptional silencing' of the promoters. Moreover, transgene proximity to telomeric or heterochromatin regions is anticipated to show a 'position effect' [9]. Susceptibility to apoptosis and a decreased efficiency of post-transcriptional processes have also been reported in CHO cell lines with unstable transgene expression [10]. A decrease in the frequency of unstable clones was observed with the use of apoptotic-resistant host cell lines.

To avoid unforeseen effects from the transgene integration sites, efficient CRISPR mediated targeted transgene integration is now discussed as a solution to this problem [7]. This necessitates precise annotation of genomic safe harbors that can be targeted for transgene integration. Knowledge of the genomic profile around random integration sites along with the effect on the product of interest can be used to learn from the cell. Lentiviral mediated random transgene integrations were performed in single copy by O'Brien *et al.* [11] and as multi-landing pad by Gaidukov *et al.* in 2018 [12]. Both studies observed that the integration sites for cell lines with stable expression are located within transcriptionally active regions. Most of these sites were found to be in the intergenic regions of the genome or intronic regions of expressed coding genes. While, so far, a handful of favorable sites have been reported across the genome, with recommendations on preferable criteria for transgene integration, the mechanisms underlying instability of transgene expression are still unclear.

Understanding the molecular mechanism of expression instability is extremely important to ensure that inefficient sites for transgene integration are avoided. Comparing genomic, transcriptomic and epigenetic data of cell lines with both stable and unstable transgene expression can help with getting an insight into the molecular drivers of phenotypic heterogeneity. With this objective, we focus here on characterizing the genomic profiles around the exact transgene integration sites identified by targeted locus amplification sequencing (TLA-Seq). Thirteen Horizon CHO cell lines were grouped on the basis of transgene copy number (low, medium or high copy number) and loss in titer of monoclonal antibody from passage 1 to passage 10. Comparison of expression profiles revealed upregulation of gene sets associated with apoptosis, cell signaling and extracellular matrix components in unfavorable (high copy number, unstable transgene expression) cell lines. Genes associated with glucose metabolism promoting Akt signaling were found to be enriched in all the expressing cell lines in contrast to the host. Based on the key factors observed while comparing stable and unstable cell lines, genome wide landing pads were reported both for the regions to target and for those to avoid during targeted transgene integration. These can be used to estimate the fate of integrated transgenes in the case of random transgene integration, or directly, for targeted integration to avoid failures due to position effects.

2. Materials and methods

2.1. Cell line development

Industrial production cell lines typically have higher specific productivities than those used in academic labs that undergo lower stress levels and thus potential differences in stability. Hence, we used production cell lines provided by the Janssen R&D Cell Line Development group (Springhouse, PA, USA) (Fig. 1). Horizon-Discovery CHO-K1 GS -/- cell lines were chosen since they are considered as one of the industry-standard selection systems today. This is because of the reduced timelines for identifying high expressing clones from glutamine synthetase negative cells [13].

Suspension Horizon Discovery CHO-K1 (HD-BIOP3) glutamine synthetase knockout cells were cultured in CDFortiCHO (Gibco/ThermoFisher) + 4 mM L-glutamine (Gibco/ThermoFisher) with 50 mL culture volume in 125 mL shake flasks at 130 rpm with a 25 mm orbit in a humidity controlled 37 °C incubator with 5% CO2 and passaged every 3–4 days at $3x10^5$ viable cells (vc)/mL. 15 μg of a Janssen-derived expression plasmid conferring a functional glutamine synthetase gene driven by a Simian Virus 40 Promotor as well as the monoclonal antibody heavy and light chain genes driven by two human cytomegalovirus (hCMV-MIE) promoters, were transfected by electroporation (Bio-Rad Gene Pulser) into $1x10^6$ cells in 1 mL Ex-Cell CD-CHO Fusion media (SAFC) by exponential decay at voltage of 300 V and capacitance of 950*u*F). Transfected cells were transferred to a T25 flask with 4 mL CDFortiCHO + 4 mM L-glutamine to recover for one day in a humidity controlled static incubator at 37 °C with 5% CO_2. 24 h post transfection, cells were centrifuged for media exchange for selection in 10 mL CD-CHO without glutamine (Gibco/ThermoFisher) in a T75 flask for 7–12 days. Selection media was glutamine-free and without supplemental L-methionine sulfoximine (MSX). Thereby, the transfection pool was selected for cells that have integrated both the functional glutamine synthetase gene and transgene and survived without amplification.

To select single cell colonies, transfectants were plated in 6-well plates as a cell suspension in custom glutamine-free Methocult medium containing 2.5% (w/v) methylcellulose in Dulbecco's Modified Eagle's Medium (DMEM) base medium (StemCell Technologies). The working solution also contained 30% (v/v) gamma-irradiated dialyzed fetal bovine serum (Hyclone), 1x GS Supplement (SAFC), 1.5 mg animal component-free protein G Alexa Fluor 488 conjugate (Invitrogen), and Dulbecco's Modified Eagle's Medium with F12 (Gibco/Invitrogen). The protein G binds to antibody secreted by the cells such that single colonies secreting the most antibody will show the highest levels of fluorescence.

After incubation for 10–14 days, a ClonePix FL colony picking instrument (Molecular Devices) was used to pick single colonies. The ClonePix FL system imaged each colony based on the parameters of fluorescence (high FITC level), colony size (0.05–2.0 mm^2), shape/roundness (Irregular 1 and 2 < 0.6), and proximity to other colonies (>1 mm). After ranking the colonies on all parameters, the most desirable colonies for expression and with a sufficient probability of monoclonality were transferred by micro-pins to 1–2 96-well plate containing 100 μl CD CHO selection media supplemented with phenol-red.

The 96-well plates were incubated for seven days with a medium feed of 100 μl CD CHO selection media 3–4 days post seeding. The 96-well plates were titered via Octet (ForteBio) using a Protein A dip and read sensors and IgG standard curve to measure recombinant protein. The cultures corresponding to the highest 96w titers were then scaled-up to 24-Deep Well plates to shake flasks in Janssen R&D proprietary medium without glutamine. For bolus fed-batch (BFB) shake flask productions, cells were seeded at 4×10^5 vc/mL in Janssen R&D proprietary medium without glutamine and fed with Janssen R&D proprietary feeds 4 times over the span of 12 days. After 14 days, fed-batch cultures were harvested and the supernatant was clarified by centrifugation at 1000 rpm for 5 min to remove the cell pellet and titered via Octet. For all clones expressing >1 g/L, 3 vials of research cell banks (2×10^6 cells in 1 mL CD CHO selection media with 10% dimethyl

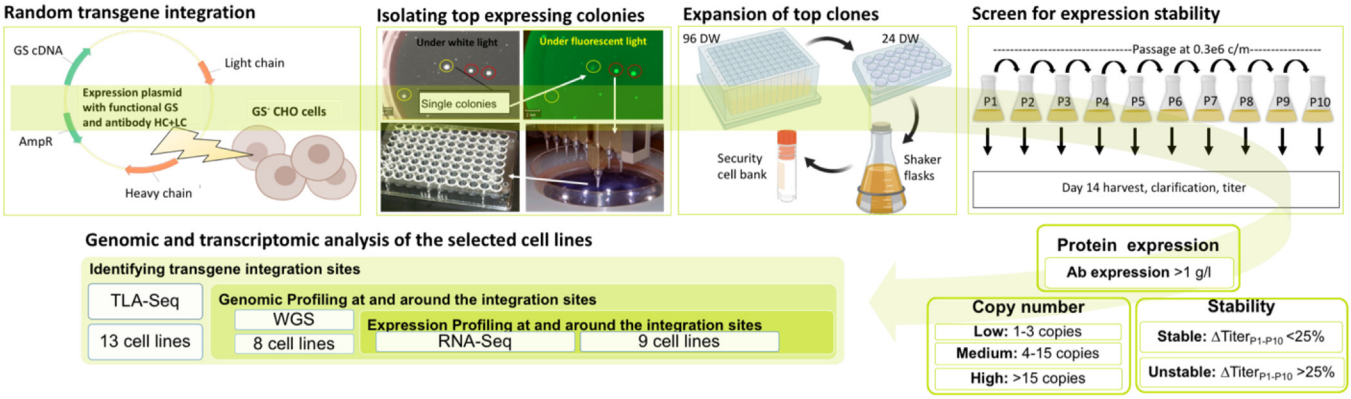

Fig. 1. Cell line development of antibody producing Horizon CHO-K1 GS -/- cell lines. After random integration, top expressing single colonies were isolated using ClonePix, followed by screening of antibody expression and expansion of the top clones. Samples for further analysis were selected based on the copy number (estimated using Digital Droplet PCR) and expression stability. Cell lines corresponding to low/medium/high copy number with stable/unstable transgene expression were sequenced for genome and transcriptome along with ATCC and Horizon host cell lines. Targeted locus amplification sequencing (TLA-Seq) by Cergentis allowed identification of transgene integration sites at early (P1) as well as late (P10) passages for further characterization.

sulfoxide (DMSO)) and 3 pellets of $1x10^8$ cells for genomic DNA preparation were generated and archived.

2.2. Gene copy number

Gene copy number was determined by Digital Droplet PCR (ddPCR) on a QX200 (Bio-Rad). Genomic DNA of clones expressing >1 g/L was isolated from a frozen pellet of $1x10^6$ cells generated at exponential growth phase using a DNeasy Blood & Tissue Kit (Qiagen). gDNA samples were fragmented by restriction digest with an enzyme that did not cut in the coding regions. Primer probe sets were designed with primers to specifically amplify heavy chain and light chain coding sequences of the transgene flanking a FAM labelled probe that binds the amplicon. Analysis included a Glucogon Receptor (GcgR) housekeeping gene primer and HEX labelled probe set for normalization. Genomic DNA of the CHO host was fragmented with the same restriction enzyme as samples as a negative control to ensure primer specificity of the transgene FAM labelled probe set and confirm the housekeeper Hex labelled probe set. A no template water negative control and a transgene expression plasmid positive control were also included. Clones were categorized by gene copy number: low (1–3 copies), medium (4–15 copies), and high (>15 copies). The copy number of selected cell lines was also determined for passage 10 cultures.

2.3. Cell line protein expression stability

Expression stability was characterized by determining the change in expression levels over 10 passages without selective pressure. For clones expressing >1 g/L, batch shake flask cultures were seeded at 3×10^5 vc/mL in Janssen R&D proprietary growth media and passaged every 3 to 4 days for 10 passages. After subculturing, the remaining culture at each passage was left as a batch culture and harvested on day 14 for titer determination by Octet. The titer change was calculated by:

$$\Delta\text{Titer} = \frac{Avg(P1,P2) - Avg(P9,P10)}{Avg(P1,P2)} \times 100$$

Clones with <25% titer change were characterized as stable cell lines and clones with >25% titer change as unstable cell lines. Security cell banks and cell pellets for genomic DNA preparation were generated at passage 10.

2.4. Screening and sample selection

The cell lines were categorized into six phenotypes corresponding to low, medium or high copy number with stable and unstable transgene expression, respectively (Table 1). For each category, a subclone was selected and sequenced for their genomic (whole genome sequencing – WGS) and transcriptomic (RNA-Seq) profiles (in triplicates). To know the precise location of the randomly integrated transgene, TLA-Seq was performed.

2.5. Whole genome sequencing

Genomic DNA of the cell lines selected for six phenotypes corresponding to low, medium or high copy number with stable and unstable transgene expression was isolated from a frozen pellet of $1x10^8$ cells generated at exponential growth phase using a Blood & Cell Culture Maxi Kit (Qiagen). gDNA samples were also prepared for three host cell lines (ATCC CHO-K1 GS +/−, Horizon Discovery HD-BIOP3 GS −/−, Horizon process evolved). The Horizon process evolved host was isolated after mock electroporation, recovery, single cell isolation and expansion.

gDNA samples were submitted to GeneWiz (South Plainfield, NJ) for library preparation and sequencing. gDNA was quantified by GeneWiz on a Qubit 2.0 Fluorometer (Life Technologies) and checked for integrity by agarose gel. Genomic library preparation was completed using an NEBNext® Ultra™ DNA Library Prep Kit for Illumina following manufacturer's recommendations. The library was validated using D1000 ScreenTape on an Agilent 4200 TapeStation and quantified on a Qubit 2.0 Fluorometer and real time PCR (Applied Biosystems). Samples were analyzed using Illumina HiSeq (2×150bp configuration, single index per lane, ~350 M raw paired-end reads per lane). Image analysis and base calling was completed with the HiSeq Control Software. Raw sequencing data was converted into FASTQ files and demultiplexed using Illumina's bcl2fastq software for data delivery.

2.6. RNA-seq

The cell lines selected for six phenotypes corresponding to low, medium or high copy number with stable and unstable transgene expression were cultured as biological triplicates. RNA was isolated in 1 mL RNAlater (Invitrogen/ThermoFisher) from a frozen pellet of $1x10^6$ cells generated at exponential growth phase (48 h post seeding). Extractions were completed using an RNAeasy Mini Kit (Qia-

Table 1
Horizon Discovery CHO-K1 derived cell lines under analysis. The table describes characteristics of the samples being used for WGS, RNA-Seq and TLA-Seq. The cell lines are distinct in terms of copy number and stability. Distinction in copy number: Low = 1–3 copies, Medium = 4–15 copies, High >= 15 copies. Distinction in stability based on drop in titer from passage 1 to passage 10: stable <25%, unstable >25%.

Sample Number	Copy Number	Stability	RNA-Seq	WGS	TLA
C1835A	ATCC	–	✓	–	–
C3234A	Horizon host	–	✓	✓	–
PE24	Pricess evolved horizon host	–	✓	✓	–
G9	Low	Stable	✓	✓	✓*
5G10	Low	Unstable	✓	✓	✓*
E3	Medium	Stable	✓	✓	✓*
1C11	Medium	Unstable	✓	✓	✓*
6A6	High	Stable	✓	✓	✓*
6H1	High	Unstable	✓	✓	✓*
E1	Low	Stable	–	–	✓
C5	Low	Stable	–	–	✓
F6	Low	Stable	–	–	✓
4B8	Low	Stable	–	–	✓
1F1	Low	Stable	–	–	✓
C2714B	Unknown	Stable	–	–	✓
2H7.13	Unknown	Unstable	–	–	✓

*TLA sequencing was performed for these samples both at passage 1 and passage 10.

gen) on the QIAcube (Qiagen) with DNAse treatment. RNA samples were also prepared from three host cell lines (ATCC, Horizon, Horizon process evolved) in biological triplicates.

RNA samples were submitted to GeneWiz (South Plainfield, NJ) for rRNA depletion, library preparation, and sequencing. RNA was quantified by GeneWiz on a Qubit 2.0 Fluorometer (Life Technologies) and checked for integrity on an Agilent 4200 TapeStation. rRNA depletion was completed using an Illumina Ribo-Zero rRNA removal kit. RNA library preparation was completed using an Illumina TruSeq Stranded Total RNA library kit following manufacturer's recommendations. The library was validated using DNA Analysis ScreenTape on an Agilent 2200 TapeStation and quantified on a Qubit 2.0 Fluorometer and real time PCR (Applied Biosystems). Samples were individually barcoded, pooled, and divided across lanes to avoid lane bias. Samples were analyzed using Illumina HiSeq (2x150bp configuration, single index per lane, ~350 M raw paired end reads per lane). Raw sequencing data was converted into FASTQ files and de-multiplexed using Illumina's bcl2fastq v. 2.17 software for data delivery.

2.7. Targeted locus amplification sequencing

The cell lines selected for six phenotypes corresponding to low, medium or high copy number with stable and unstable transgene expression were cultured to exponential growth phase to generate frozen pellet suspension samples of $1x10^6$ cells in 1 mL Dulbecco's phosphate-buffered saline (dPBS) (Gibco/ThermoFisher) + 10% DMSO (Sigma). Aged samples were also generated for the 6 phenotyped cell lines at passage 10 of the expression stability study. To obtain more information on the genomic profile around integration sites, seven additional Horizon CHO cell lines were selected with similar copy number and stability parameters for TLA sequencing.

Samples were submitted to Cergentis (Utrecht, The Netherlands) for sample preparation based on two transgene specific primer pairs. Cergentis completed NGS library preparation, sequencing, and data analysis to determine the transgene integration sites, estimated copy number, assessment of structural changes surrounding integration sites and genetic alterations in the transgene with their standard pipelines. Genomic regions around the integration sites were plotted with the Gviz package in R.

2.8. Analyzing genomic variability with WGS

Raw reads were trimmed for adapters and quality with Trimmomatic (version 0.36) setting minimum length to 25, trimming first 3 and last 3 bases and sliding window 4 bases with average quality of 20 [14]. After confirming the quality of processed reads with FastQC (version 0.11.5) [15], they were mapped to the latest Chinese hamster genome assembly (CriGri-PICR, RefSeq Accession: GCF 003668045.1 [16]) using BWA (version 0.7.17) [17] mem algorithm with default settings. Removal of polymerase chain reaction duplicates was done using Picard tools (version 2.3.0) [18]. Alignment coverage was then evaluated using Qualimap (version 2.2.1) [19]. This was followed by generating base recalibration files for Genomic Analysis Toolkit (GATK) haplotype caller (version 3.8) [20]. The uncalibrated bam files were used for the first round of variant calling with output mode set to emit all confident sites and standard call confidence of 30. High confidence single nucleotide polymorphism (SNP) calls were filtered based on 'variant confidence standardized by depth' (QD < 2.0), 'strand bias in support of REF vs ALT allele calls' (FS > 60), 'mapping quality of the SNP' (MQ < 40), 'Rank sum test for mapping qualities of REF vs ALT' (MQRankSum < −12.5), 'position of called SNP with respect to end of the read' (ReadPosRankSum < −8.0), 'sequencing bias of one DNA strand being favored over another' (SOR > 3.0). Similarly, high confidence insertion and deletion (INDEL) calls were filtered based on QD < 2.0, FS > 200.0, ReadPosRankSum < −20.0 and SOR > 10.0. These filtered SNP and INDEL calls were used as known SNPs to create the recalibration table. The final round of variant calling was done using the recalibrated data and the SNPs and INDELs were filtered again with the same filter expression.

The locations of deduced variant calls from all the samples were pooled and unique positions were intersected with 2 kb bins across the genome to record number of counts in each bin. High and low genome variability was calculated on the basis of median absolute deviation (MAD) of the recorded counts. The "M-value" was calculated as a score of variability by taking the ratio of the difference between each bin's frequency (f_i) and the median of the values across the binned genome (m) to the MAD value:

$$M_i = \frac{f_i - m}{MAD}$$

f_i = bin frequency (counts per bin),
m = median of the values (f_i) across the binned genome,
i = 1 to n;

where n = number of 2 kb bins across the genome

All the bins with M-value greater than 0.7 were used later as high variability regions while screening for landing pads. These scores were calculated for the samples individually as well and the corresponding bigwig files have been uploaded to the genome browser as well.

2.9. Deducing structural variations from WGS data

Recalibrated alignment maps that were used for calling small variants with GATK were used for calling out structural variations as well. Since specific algorithms are suitable for each type of structural variation to be identified, we used the tools that have been recommended to report variants with high precision and recall values in a benchmarking study [21]. Insertions, deletions, duplications and translocations were identified with Manta (version 1.6.0) using default settings and filtered for variants with PASS reports [22]. Inversions were identified with Delly (version 0.7.9) [23] using default settings and further filtered for paired end support quality (PE > 4) and median mapping quality of paired ends (MAPQ > 19). For deletions and translocations, variants identified both by Lumpy (version 0.2.13) [24] and Manta with overlapping coordinates within 100 bp distance were selected. After filtering all the "Imprecise" or "Low Quality" calls, the counts of each type were checked in all the cell lines under analysis. Coordinates for consensus calls reported by Manta were used for further analysis. StructuralVariantAnnotation package in R [25] was used for extracting genomic coordinates of all the deduced structural variations. Chromosomal context was further used to check inter or intra chromosomal translocations. Genes within 2 kb distance from the deduced structural variation were also identified for inferring potential phenotypic associations.

2.10. Exploring expression profiles with RNA-Seq

Raw reads for all the RNA-Seq samples were trimmed with Trimmomatic similar to WGS reads (version 0.36) and after quality check with FastQC (version 0.11.5) the reads were mapped to the latest Chinese hamster genome assembly (CriGri-PICR) [16] using Hisat2 (version 2.1.0) [26] with default settings. The mapped read counts were calculated with htseq-count (version 0.11.0) [27] corresponding to the 'gene' attribute of RefSeq version (GCF_003668045.1 release 103) of the gene annotation for CriGri-PICR. Expression peaks were deduced based on coverage reported by bamCoverage from deepTools (version 3.0.1) [28] which reports the number of reads per bin. These bins are made from short consecutive counting windows of 50 bp that are extended to reflect the actual fragment length. High and low expression peaks are deduced for individual cell lines based on the deviation from the median of the binned frequencies as described before for variant calls.

2.11. Differential expression analysis

A raw count matrix was prepared for the three stable and three unstable samples corresponding to low, medium and high copy numbers along with the three host cell lines - Horizon host, the process evolved horizon host and the ATCC host for each of the three replicates. Data normalization and further differential expression analysis was performed with the DESeq2 package (version 1.24.0) [29]. Variance stabilizing transformation (VST) of the raw counts was done to plot principal components and plotting the heat map for observing clustering of samples based on expression profiles.

Genes with at least 10 reads in more than three samples were classified as "expressed" while filtering out the rest of the lowly expressed genes. The design formulae were created for different comparisons to report up and down regulated genes in context to stability, copy number, favorability and host vs expressing cell lines. Shrunken log fold changes were calculated for genes in different comparisons with the lfcShrink function using 'apeglm' as shrinkage estimator. Genes with fold change greater than 1.5x and adjusted p-values less than 0.01 were selected as significantly differentially expressed.

The ranked lists of differentially expressed genes were created by using DESeq2 'stat' (Wald statistic) values. These were used to analyze gene set enrichments for up or down regulated genes in each comparison using GSEA software (version 3.0) [30]. The enrichments were observed with a background of gene sets curated from online pathway databases, biomedical literature and domain experts that represent biological processes (c2.cp.v6.2.symbols.gmt). Significantly enriched gene sets were reported with a false discovery rate (FDR) of 25 percent as threshold.

2.12. Identifying landing pads with plausible favorable and unfavorable transgene integration

Based on genomic, transcriptomic and epigenetic profiling around random integration sites deduced by TLA analysis, we hypothesized that transgene integration at sites 1) with large distance from high variability regions and repressed heterochromatin regions would increase the chances of "stable" expression, and 2) closer to high expression peaks and active chromatin states would increase the chances of "higher" expression.

Landing pads for favorable integrations: To find regions with favorable characteristics, a 20 kb window was browsed throughout the genome to record distances from positive parameters (high expression peak and active chromatin states - active transcription, promoter, enhancer) as well as negative parameters (high variability peak and repressed chromatin states) (Fig. 2). The bins in small scaffolds that didn't have any high expression peak or chromatin state annotation were filtered out. The bins overlapping structural variations, exonic regions of genes or high expression peaks were also removed. Based on the thresholds set for distribution of distances observed for random integration sites, the bins that miss favorable characteristics (with more than 250 kb distance from a high expression peak, less than 3 kb distance from a high variability region or having lesser distance from a repressive state than to the active chromatin states) were also excluded from further analysis. After filtering, the close-by bins (within 2 kb) were merged and distances were recorded again from the center of bigger merged regions. Parameters that would reduce the score to a low value were classified as "negative parameters" (np) and the ones that increase the score with low value were classified as "positive parameters" (pp). We scaled the values for each parameter (p i.e. np or pp) by maximum value (p_{max}) such that all values range from 0 to 1 ($p_q = p/p_{max}$). Score was calculated by adding the scaled values with preference for higher values of negative parameters ($np_{q.x} = np/np_{max.x}$; x → all negative parameters) and lower values for positive parameters ($pp_{q.y} = 1\text{-}pp/pp_{max.y}$; y → all positive parameters). Additional weight for longer regions was projected on the final score by multiplying the scaled size to the sum of scaled values for negative and positive parameters. The regions were finally sorted for higher score, assigned chromosome (2 > X > 7 > 3 > 5 > 1 > 4 > 8 > 6 > 10 > 9 > unplaced) and annotation (intergenic > intronic).

Landing pads for unfavorable integrations: Unfavorable landing pads were identified by enabling selection of small regions with unfavorable characteristics as well (unlike searching for larger regions with favorable characteristics in previous section). Smaller windows of 5 kb were browsed across the genome to capture different parameters and the filtered bins were merged within 10 kb

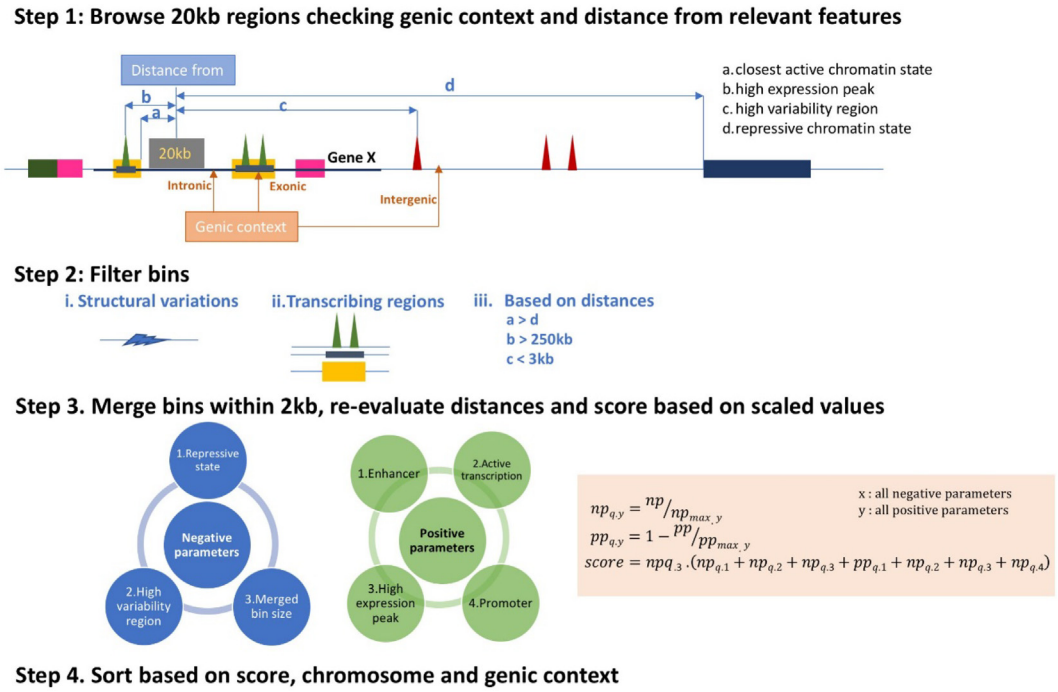

Fig. 2. Scanning the genome for favorable landing pads. The figure shows steps for deducing genome wide landing pads that are expected to deliver stable transgene expression if chosen as sites for targeted transgene integration.

genomic distance. Preliminary filtering was done to keep regions with a distance of less than 500 bp from high variability peaks and remove the ones with less than 700 kb distance from high expression peak or 500 kb from active transcription state (H3K36me3). Also, regions overlapping structural variations or having active states closer than the repressed states were excluded. After merging the filtered regions, the distances were re-evaluated from their respective centers and filtering was repeated again. Scores were calculated again without adding weight for longer size at the last step. Finally, the list was sorted for lower score and reversed order of chromosome assignment and genic context.

3. Results

This study characterizes consequences of random transgene integration in different cell lines with high specific productivity based on genomic, transcriptomic and epigenetic profiling of integration sites. Events with low, medium or high copy numbers and each resulting in both stable as well as unstable transgene expression were sampled resulting in six production clones of Horizon host CHO cell lines (Table 2). To identify factors associated with stable or unstable transgene expression, each of the cell lines was sequenced for genome (by WGS), transcriptome (by RNA-Seq) and targeted locus amplified around the transgene construct (by TLA-Seq), respectively. Genomic and transcriptomic variability in host cell lines was also analyzed with WGS and RNA-Seq data for the Horizon host, a process evolved Horizon host and CHO-K1 ATCC host cell lines. For investigating association of genomic rearrangements at integration level with expression stability, TLA-Seq data was analyzed at early as well as late passage for the producer cell lines. Additionally, another six stable and one unstable transgene expressing cell lines were TLA sequenced to investigate genomic profiles around the integration sites. Association with epigenetic factors were analyzed based on known chromatin states

Table 2
Copy number and titer information regarding cell lines under observation. Cell lines with transgene copy number <3 were classified as "Low copy", those with 4–15 copies as "Medium copy" and the ones with >15 copies as "High copy" cell lines. Change in titer from passage 1 to passage 10 of <25% was termed as "stable" and for >25% as "unstable".

Clone Name	Sample type	CD Forti CHO BFB Expression (mg/L)	P1 average copy number	P10 average copy number	P1 to P10 stability (% drop in titer)
G9	Low copy, Stable	2352	1.9	2.0	−17.4
5G10	Low copy, Unstable	1413	2.0	1.7	−33.9
1E3	Medium copy, Stable	1999	4.7	6.0	−16.3
1C11	Medium copy, Unstable	2118	9.9	9.8	−52.7
6A6	High copy, Stable	2090	53.2	66.5	−17.5
6H1	High copy, Unstable	2389	48.3	48.3	−38.2
E1	Medium copy, Stable	2117	4.4	4.3	−4.5
C5	Low copy, Stable	999	0.9	0.9	−14.5
F6	Low copy, Stable	1291	0.9	0.9	0.1
4B8	Low copy, Stable	1266	1.3	1.3	6.3
1F1	Low copy, Stable	1927	1.9	1.7	−11.8
C2714B	Stable	863	unknown	unknown	−22.9
2H7.13	Unstable	1010	unknown	unknown	−47.9

Table 3
Genomic profiling of integration sites. The table depicts 3 stable and 3 unstable cell lines with low, medium and high copy numbers respectively. Integration sites, rearrangements induced in the host genome, number and arrangement of TG-TG fusions are reported for the 6 cell lines at early (P1) and late passage (P10).

Cell line	Copy number	Stability	Integration sites (CriGri-PICR)		Sequence/structural variation (Host)		Sequence/structural variation (TG-TG fusions)		Number of each type of TG-TG fusion		Overlapping genomic region
			P1	P10	P1	P10	P1	P10	P1	P10	
G9	Low	Stable	NW_020822533.1:6601451–6601475	NW_020822533.1:6601451–6601475	_	_	3 TG-TG fusions	2 TG-TG fusions	H → H = 1; T → T = 1; T → H = 2 homologous	H → H = 1; T → T = 1	Just at the boundary of an exon of Fam219b
5G10	Low	Unstable	NW_020822425.1:6655503–6681964; NW_020822407.1:9302191–9324853	NW_020822425.1:6655503–6681964; NW_020822407.1:9302191–9324853; NW_020822603.1:1899019–1995809	Inversion; Deletion	Inversion; Deletion; Partial integration	_	1 (For primer pair 2 data on scaffold with partial integration)	_	H → H = 1	2nd integration site creates a deletion in host genome encoding Gabra4 gene
1E3	Medium	Stable	NW_020822529.1:13736142	NW_020822529.1:13736142	_	_	3 TG-TG fusions	1 TG-TG fusion	H → T = 1; T → T = 1; T → H = 1	T → H = 1	Intron of Robo1 gene
1C11	Medium	Unstable	NW_020822464.1:4510798	NW_020822464.1:4510798	_	_	15 TG-TG fusions, 1SNV	15 TG-TG fusions, 1SNV	H → H = 3; H → T = 4; T → T = 5; T → H = 4	H → H = 3; H → T = 4; T → T = 5; T → H = 4	Intergenic region downstream of a ncRNA, overlapping enhancer marks
6A6	High	Stable	NW_020822506.1:18313071	NW_020822506.1:18313071; NW_020822426.1:2132680	Partial integration with primer 1 data based on alignment coverage	_	17 TG-TG fusions	13 TG-TG fusions	H → H = 3; T → T = 3; T → H = 11	T → T = 3; T → H = 10	Intron of Nidogen-1; intron of ncRNA
6H1	High	Unstable	NW_020822506.1:18313071	NW_020822506.1:18313071; NW_020822426.1:2132680	Partial integration with primer 1 data based on alignment coverage	_	17 TG-TG fusions	17 TG-TG fusions	H → H = 3; T → T = 3; T → H = 11	T → T = 3; T → H = 10	Intron of Nidogen-1; intron of ncRNA

for a CHO-K1 cell line evaluated from six active and repressive histone modification marks [16,31].

3.1. Integrity of the transgenic loci as a source of phenotypic heterogeneity

Targeted locus amplification-based transgene sequencing data provides high coverage of the sequence around the integration site, including fusion reads that span host genome and transgene sequence. The alignment coverage and breakpoint analysis of these fusion reads can enable precise identification of the transgene integration site. The coverage peaks can also report the existence of possible rearrangements in the host genome as a result of transgene integration. Other than that, fusion reads representing two non-adjacent parts of the transgene can suggest fusion of two transgene copies or structural rearrangement of the transgene sequence.

These criteria were used to identify transgene integration sites and the surrounding sequence integrity. TLA sequencing was done for all 13 cell lines that were sampled from early passage culture. To capture sequence variations or structural changes that might have occurred during passaging of cells, the main six cell lines were also sampled at P10 (Table 1). The precise locations of transgene integration and rearrangement events were predicted by the well-established pipelines developed at Cergentis .

Table 3 illustrates the potential integration-related reasons for unstable transgene expression in cell lines with low (G9-Stable, 5G10-Unstable), medium (1E3-Stable, 1C11-Unstable) and high copy number (6A6-Stable, 6H1-Unstable). While the transgene integration in G9-Stable resulted in no structural variation in the host genome and only a few (<=3) transgene-transgene (TG-TG) fusions, one of the two sites observed in 5G10-Unstable resulted in an inversion and the other in deletion of part of the host genome sequence (Fig. 3, Suppl. Table 1). The genomic region marked with the deletion was found to overlap a portion of Gabra4 (Gamma-aminobutyric acid receptor subunit alpha-4), Gabrb1 (gamma-aminobutyric acid receptor subunit beta-1) and a 60S ribosomal protein L23a. Moreover, a partial integration was observed at a third site with primer 2 sequencing data in the sample taken from the late passage culture. A subsequent deletion and a TG-TG fusion were also observed at this site with partial integration.

Although no structural variations were observed at the integration site in the host genome for the medium copy cell lines, the unstably expressing cell line (1C11) was found to have 15 TG-TG fusions where its stable counterpart (1E3) had only three (Table 2, Suppl. Table 1). The integration sites for stable and unstable cell lines were in intronic and intergenic regions respectively (Suppl. Fig. 1). The transgene was found to be integrated at the same genomic locus for both the stable (6A6) as well as the unstable (6H1) cell lines with high copy number (Suppl. Fig. 2). Numerous TG-TG fusions were identified for both high copy cell lines at early as well as late passage (Table 3).

Hence, observations for the "low copy, unstable" cell line (5G10) clearly indicate that genomic rearrangements in the host and transgene sequence can be potentially associated with instability in protein expression of the integrated transgene. For medium and high copy cell lines, we observed that while a lower number of TG-TG fusions (as in G9, 5G10 and 1E3) by itself doesn't seem to have strong impact, a higher number (as in 1C11, 6A6 and 6H1) could be a potential cause of loss in titer with time in culture. The occurrence of a higher number of transgene fusions can have both positive [32] as well as negative correlation with gene expression [33,34]. The negative correlation is reasoned to be due to the transcriptional silencing of high copy number tandem integration events from random knock-in approach [35]. Insertions of multiple copies are likely to induce local heterochromatin-like properties, thereby increasing the chances of transcriptional repression around that region. Multi-copy integrations in head-to-head and tail-to-tail like arrangements also have a tendency to cause steric hindrance, thereby restricting replication and transcription, but also favoring rearrangements and deletions based on sequence repeats. In contrast, a positive correlation has been associated with head-to-tail multi-copy fusions of intact transgene sequence [35]. These findings can themselves be considered self-explanatory for the unstable phenotype and may be used for the early deselection of clones with increased likelihood of being unstable. However, to identify potential other causes of instability upon integration in some sites and not in others, whole genome and transcriptome analyses were also performed.

3.2. Contribution of genome-wide genomic rearrangements to expression stability

Structural variants are genomic rearrangements longer than 50 bp that include balanced forms like translocations (TRANS) and inversions (INVs) as well as imbalanced forms like deletions (DELs), duplications (DUPs) and insertions (INSs). Although structural variations are known to have high functional potential, there

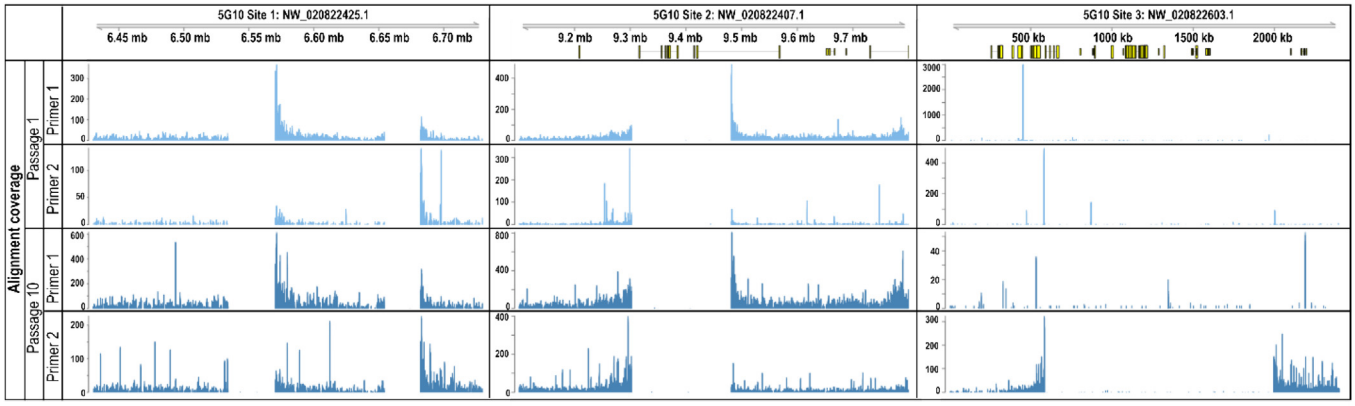

Fig. 3. Structural variations at the integration site are evident from TLA sequence coverage of "low copy, unstable" cell line (5G10). The figure represents coverage of reads from primer 1 and primer 2 of cell lines sampled at early (P1) and late (P10) passages. (A) Site 1: The genomic region between the two breakpoints was inverted during integration (B) Site2: The genomic region of around 170 kb has been deleted within the two sides of the integration site following the integration event. The deleted region overlaps gene annotations for Gabra4, Gabrb1 and a 60S ribosomal protein L23a (C) Site 3: A partial integration site was observed at site 3 where a part of the transgene including the primer 2 binding site got integrated. The integration resulted in deletion of approximately 1.4mb of genomic region within the two sides of the integration site.

are technical challenges in calling out such genomic rearrangements based on short read sequencing. The available tools to identify genomic loci with structural variations are based on using discordant alignment features, split alignment features or an assembly approach. Based on the variant type, different computational algorithms are required to accurately and sensitively detect each type. We identified structural variations in all the cell lines under analysis based on the recommendations from a benchmarking study by Kosugi *et al.* [21].

The presence of co-existing rearrangements identified in all the Horizon cell lines under analysis (250 INVs, 1261 INSs, 2403 DELs, 256 DUPs, 366 TRANS) depict the evolution of the host cell line from the Chinese hamster genome. Excluding these common regions, Fig. 4A shows similar frequencies of genomic rearrangements in each cell line. Minor differences in counts do not necessarily depict stress on the cell lines but could rather be due to the difference in alignment coverage for each cell line (Suppl. Table 2). Since the minimum value of mean coverage was around 40X, each sample per se was expected to provide reliable variant calls. No subsampling was performed to bring samples at the same coverage in order to avoid any unexpected bias or loss of information from the respective samples. Translocations have been associated with adaptation of CHO cell lines in various studies [36]. Fig. 4B depicts the load of intra-chromosomal translocations within each chromosome of different cell lines, which is proportional to the respective size of the chromosome. High frequencies for chromosome 9 and 10 indicate high genomic instability of these chromosomes. Lower frequency and smaller size of dots for chromosome 2, 7 and X indicate low genomic instability across different cell lines. Chromosomal preference for targeted transgene integration based on this information can help in avoiding expression instability. Inter-chromosomal translocations for the host cell

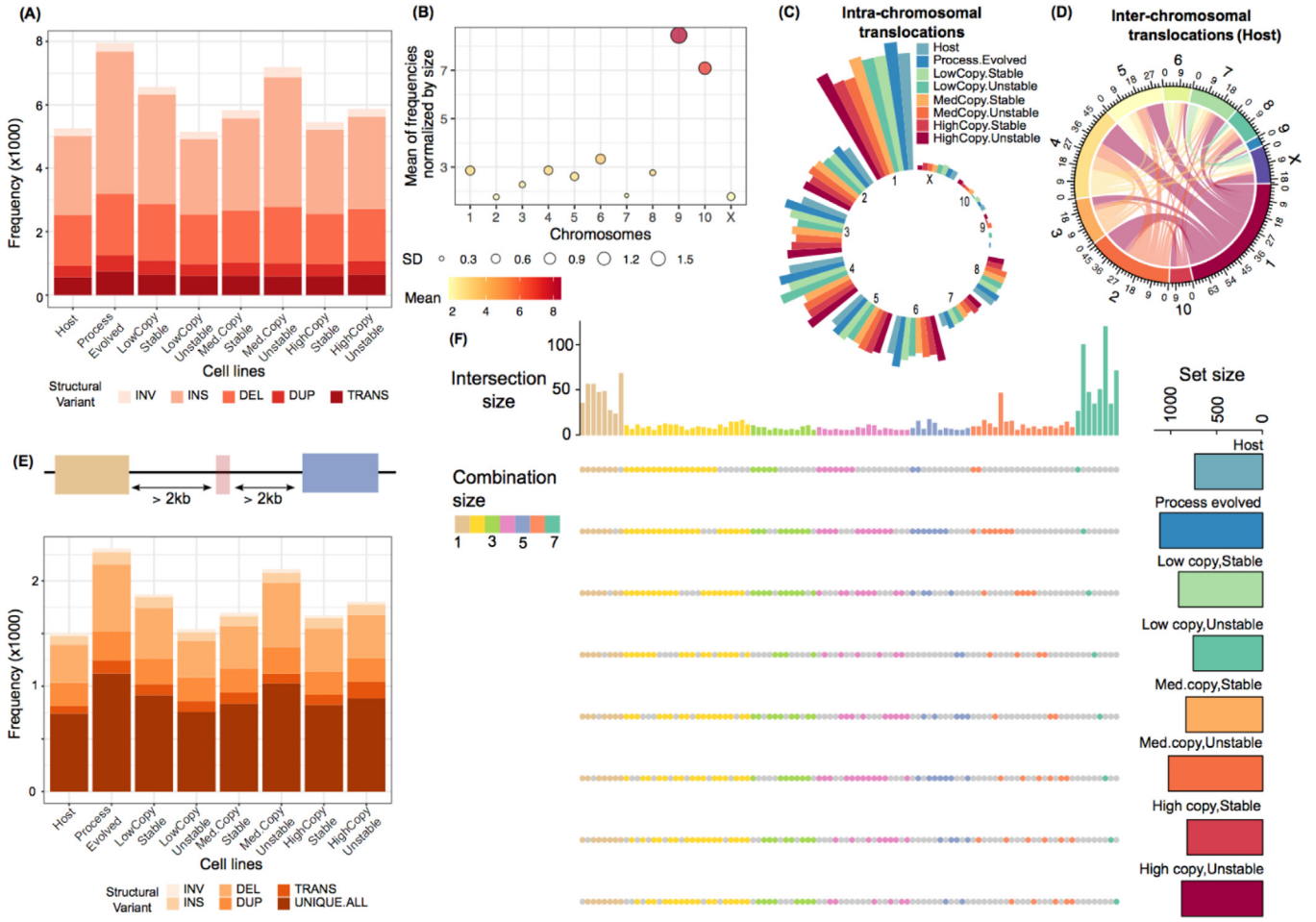

Fig. 4. Distribution of structural variations within host and expressing cell lines (A) Distribution of different types of structural variations in each cell line. Similar counts are observed across different cell lines with maximum counts for deletions followed by insertions, inversions, duplications and translocations. (B) Mean of total number of rearrangements normalized by chromosomal length in all the cell lines under consideration are plotted across different scaffolds. The size of the dots represents the standard deviation from the mean of different cell lines. The figure is an evidence for the high instability of chromosomes 9 and 10 in context of genomic rearrangements. Transgene integration in these chromosomes could be a source of unstable transgene expression. Chromosomes 2, 7 and X on the contrary seem to be highly stable with minimum genomic rearrangements across host and different expressing cell lines. (C) Number of intra-chromosomal translocations observed in each cell line. Chromosome assignment for each scaffold has been used and SVs on unassigned scaffolds are omitted. Counts with and without normalization for length are shown in the inner and outer track respectively. (D) Number of inter-chromosomal translocations (only on chromosome assigned scaffolds) in the host cell line are plotted as a chord diagram. The width of the lines from one chromosome (depicted in the outer ring of labels) to another represents the number of corresponding inter-chromosomal translocations as scaled in the inner ring that shows counts. A similar distribution of counts in different cell lines (see Suppl. Fig. 3 for the other cell lines) was observed with the maximum number of translocations observed between chromosome 1 and chromosomes 2 and 4. (E) Frequency of potentially affected genes by different types of structural variations as plotted in (A). As represented in the cartoon above, a gene was considered as potentially affected, if its distance from the neighboring SV is less than 2 kb (F) The "upset plot" displays frequency of cell line specific or co-existing genes (combination size >=5) that are potentially affected by the neighboring SVs (Refer Suppl. Fig. 4 for no filter on combination size). Intersection size on y-axis represents number of genes potentially effected in different combination of cell lines that are marked by dots below each frequency bar. Color code of frequency bar corresponds to the combination size. Set size on the right represents the total number of potentially effected genes in each cell line.

Table 4
Known integration sites in favorable cell lines within favorable landing pads. The overlap of integration sites reported for cell lines with known phenotypes (stable/unstable) with deduced (a) favorable and (b) unfavorable regions.

(a)			
G9	Stable	NW_020822533.1:6601451–6601475	NW_020822533.1:6560001–6620000
4B8	Stable	NW_020822530.1:888970–888992	NW_020822530.1:880001–900000
E1	Stable	NW_020822688.1:8025060–8077508	NW_020822688.1:8060001–8080000
mLP8(C5)*	Stable	NW_020822438.1:2314575–2314576	NW_020822438.1:2280001–2320000
mLP10(D9)*	Stable	NW_020822420.1:7297331–7297332	NW_020822420.1:7000001–7500000
mLP12(G8)*	Stable	NW_020822427.1:1973325–1973326	NW_020822427.1:1700001–2000000
mLP13(G11)*	Stable	NW_020822570.1:21228769–21228770	NW_020822570.1:21180001–21280000
mLP14(PL1.2)*	Stable	NW_020822531.1:4360473–4360474	NW_020822531.1:4340001–4420000
mLP15(PL1.3)*	Stable	NW_020822506.1:8862559–8862560	NW_020822506.1:8800001–8880000
mLP16(PL1.10)*	Stable	NW_020822533.1:220545–220546	NW_020822533.1:220001–260000
(b)			
5G10	Unstable	NW_020822425.1:6655503–6681964	NW_020822425.1:6650001–6690000
5G10	Unstable	NW_020822603.1:1899019–1995809	NW_020822603.1:1915001–2005000
2H7.13	Unstable	NW_020822638.1:1900568–1900569	NW_020822638.1:1860001–1905000
1E3	Stable	NW_020822529.1:13736142–13736143	NW_020822529.1:13510001–14095000
E1	Stable	NW_020822688.1:8025060–8077508	NW_020822688.1:8030001–8055000
C2714B	Stable	NW_020822638.1:1900568–1900569	NW_020822638.1:1860001–1905000
sLP1.2a*	Stable	NW_020822461.1:48834446–48834447	NW_020822461.1:48780001–48855000
mLP5(A8)*	Stable	NW_020822461.1:38271496–38271497	NW_020822461.1:38270001–38275000
mLP6(B5)*	Stable	NW_020822641.1:3962252–3962253	NW_020822641.1:3700001–4355000
mLP9(C10)*	Stable	NW_020822407.1:3134962–3134963	NW_020822407.1:3105001–3165000
mLP18(PL1.17)*	Stable	NW_020822407.1:21169568–21169569	NW_020822407.1:21045001–21275000

* Sites from the study by Gaidukov et al.

lines are also reported in the chord diagram as shown in Fig. 4C, where the width of each chord corresponds to the number of translocations identified within that chromosome pair.

The rearrangements co-existing in all the cell lines were found to be either overlapping or around (within 2 kb) the same 126 genes in total. Excluding these common genes, Fig. 4D represents the number of genes around each type of structural variation (SV) in all the cell lines. The number of co-existing or cell line specific genes are shown in Fig. 4E. Fig. 4F is an "upset plot" where the intersection size depicts the number of potentially affected genes from a rearrangement within 2 kb distance from it, in the cell lines labeled with connected dots below each bar. The x-axis of the intersection size bar plot is ordered with reducing number of overlapping cell lines. Frequency bars to the left of the plot indicate highly co-existing potentially affected genes, reducing with co-occurrence towards the right with cell-line specific ones. Suppl. Table 3.1 reports lists of potentially effected (within 2 kb of SV) differentially expressed genes exclusively found in either stable or unstable cell line with low, medium or high copy number separately. To observe a significant association of SVs with differential expression of neighboring genes, overlap in more than one cell line was considered. Suppl. Table 3.2 reports 15 potentially affected genes that were identified in two or more of the stable cell lines with no occurrence in their unstable counterparts, while 25 genes were identified only in the unstable cell lines. However, no significant differential expression or pathway enrichment was observed corresponding to these genes.

3.3. Influence of genomic characteristics around the integration sites on transgene expression

Potential factors responsible for expression instability were identified based on genomic, transcriptomic and epigenetic characterization of regions around the random transgene integration events. Chromatin states reporting promoter, enhancer, repressor and actively transcribing regions computed for the CriGri-PICR reference genome and analyzed for a suspension-adapted CHO-K1 ECACC cell line were utilized for this analysis [16,31]. Sequence variations including single nucleotide variants, small insertions and deletions as well as larger structural variations (>50 bp) comprising INDELs, translocations, inversions and duplications were identified based on the WGS data. Regions of low and high sequence variability were marked across the genome based on a score derived from the number of deduced sequence variations in a 2 kb bin-size across the genome. These scores were calculated in terms of measures of standard deviation from the median of frequencies observed in a particular sample. Similarly, regions of high and low expression levels were also calculated based on the RNA-seq reads mapped within 2 kb windows browsed throughout the genome (Table 4).

This analysis is based on 13 unique sites identified by TLA-Sequencing of the cell lines being discussed here (Suppl. Table 4). To increase the number of observations, an additional 20 sites with known stable transgene expression as reported by Gaidukov *et al.* were also included [12]. The level of genomic variability (M-value in 2 kb bin) at the site, distance from the closest high expression peak and chromatin state in the vicinity of the site were noted. The distribution of observed values was plotted for each factor to compare the levels in samples with stable vs unstable transgene expression. Cell lines with unstable transgene expression (Fig. 5A) showed a tendency towards higher genome variability, apart from some outliers. Lower distance from high expression peaks is more pronounced in the case of stable cell lines (Fig. 5B) and they are more likely to be in the vicinity of active chromatin states (Fig. 5C). Levels of chromatin states are plotted in Fig. 5C as positive scores on y-axis for active states (Enhancer = 1, Promoter = 2, Active transcription = 3), negative scores for repressive states (Repressed heterochromatin = -1, Polycomb repression = −2) and quiescent state scored as 0 (Quiescent = 0).

3.4. Akt signaling observed to be upregulated in mAb expressing cell lines in comparison to host

It has been reported that once inside the cell, integration of the transgene is highly dependent on how the cell reacts in relation to genes involved in pathways such as DNA repair, replication and recombination [35]. We were able to segregate the antibody expressing cell line samples from the host cell lines based on Prin-

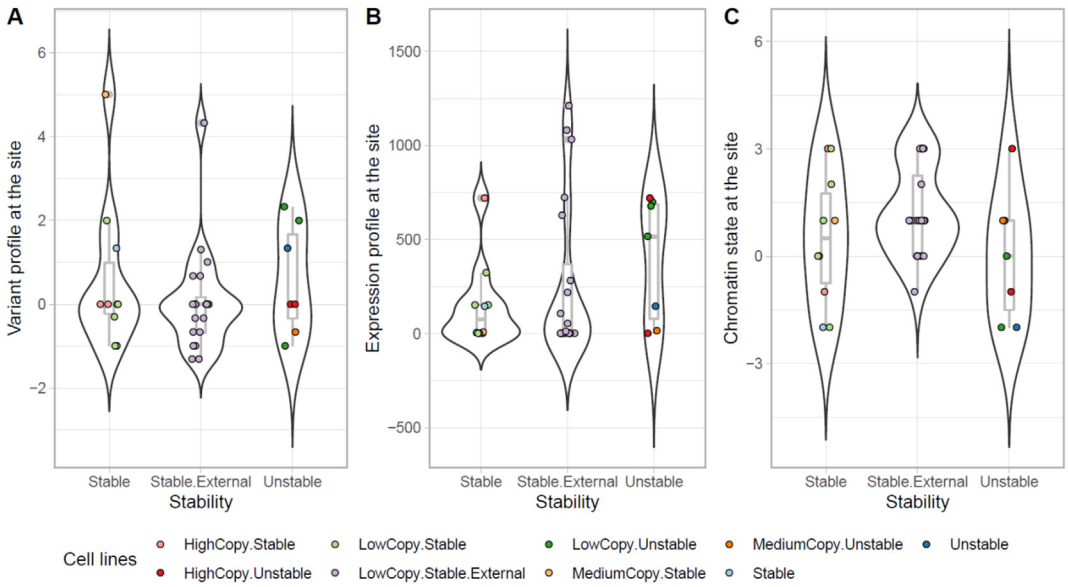

Fig. 5. Effect of genomic, transcriptomic and epigenetic profile at or around the integration sites. (A) Sequence variability in context to the M-value observed at the integration site is plotted for the stable and unstable cell lines separated on the x-axis. Similar distributions are plotted for (B) Distance of center of the integration site from the closest high expression peak (in kb) and (C) Chromatin state at the integration site. The dot plots inside the violin plots correspond to 13 unique integration sites deduced in the Horizon cell lines under observation and additional 20 sites from low copy number stable cell lines reported by Gaidukov et al. (12). Besides a few outliers, the boxplots within the figure show high variability, presence of repressive states and more distance from high expression peaks in case of unstable cell lines.

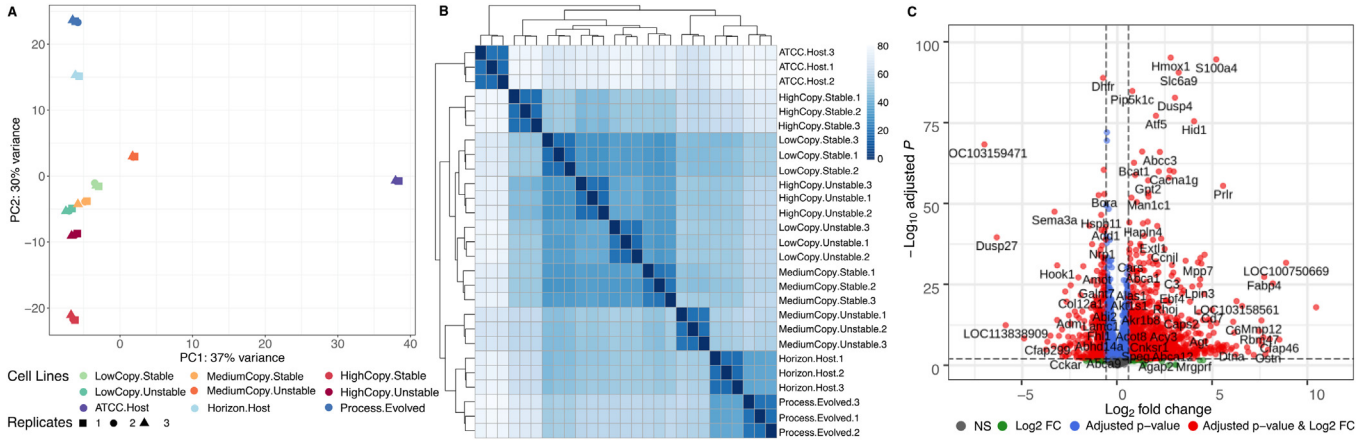

Fig. 6. Distinction of cell lines based on expression profiles. (A) Principal component analysis shows a clear distinction of the three host and six expressing cell lines based on the expression levels of top 500 genes with highest variations. This indicates the existence of differential regulation of gene expression in different phenotypes. (B) Heatmap with hierarchical clustering based on all the expressed genes also demonstrates clear separation of Horizon cell lines from ATCC cell line and the host from all expressing cell lines. (C) The volcano plot depicts more upregulated genes in the expressing cell lines in reference to the Horizon host genome than downregulated genes. Statistically significant differentially expressed genes that are above the threshold for fold change (absolute value of FC >=1.5) as well as adjusted p-value (FDR<=0.01) are shown in red. Those that pass the cut-off for only adjusted p-value are shown in blue, only FC in green and those that pass neither of the cut-offs are shown in grey. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cipal Component Analysis over top 500 protein coding genes with maximum variance. ATCC and Horizon hosts also had remarkable differences in their expression profiles, as shown in Fig. 6. This motivated us to check which genes are differentially regulated in the mAb expressing cell lines in reference to their host. In comparison to the Horizon host cell line, 806 genes were found to be upregulated and 485 genes down regulated in all the expressing cell lines. Interestingly, it was observed that apart from amino acid processing and synthesis which would be indicative of higher activity in translation processes, various gene sets associated with AKT and MAPK signaling pathways were also upregulated in the mAb expressing cell lines (Suppl. Table 5). While AKT signaling is known to correlate with increased glucose metabolism and to promote cell survival and growth, MAPK signaling pathways play an important role in cell proliferation, differentiation, development and apoptosis. This represents regulation of gene expression catering to the higher energy requirement of the expressing cells. The gene sets enriched for host cells on the contrary were mostly associated with processes like cell cycle, DNA replication, transcription and translation (Suppl. Table 6).

3.5. Role of post-translational factors in transgene expressing cell lines

Variability in samples for only expressing cell lines can be observed from principal components plotted in Fig. 7A based on the top 500 genes with high variability. The unfavorable cell lines with high copy number and unstable medium copy number separate from the remaining cell lines based on expression profiles.

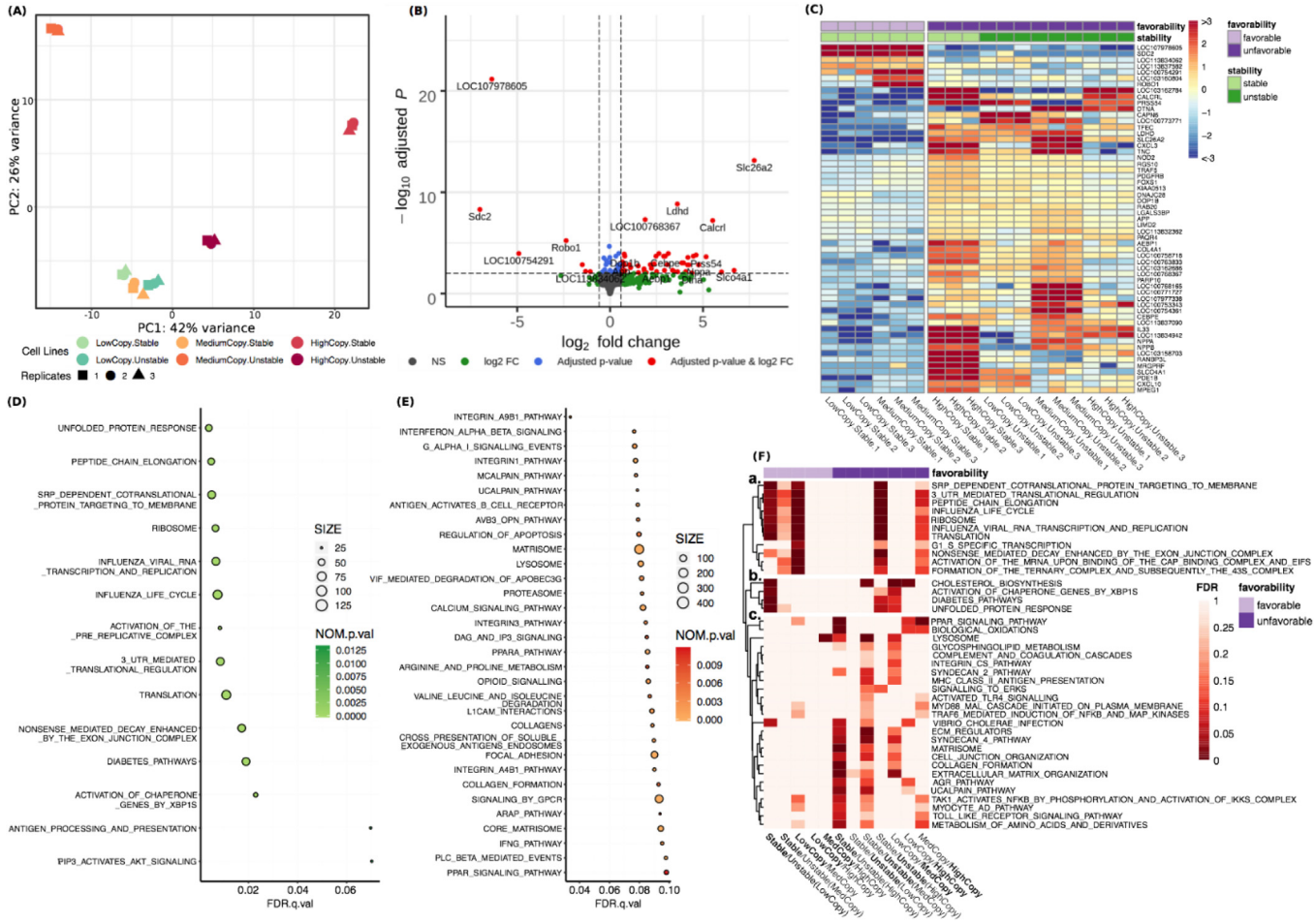

Fig. 7. Differences in expression profiles of cell lines with favorable and unfavorable phenotypes in reference to the host cell line. (A) Principal component analysis of the 6 transgene expressing cell lines separates the high copy number cell lines and medium copy unstable cell line from the rest. (B) The volcano plot depicts more upregulated genes in the unfavorable cell lines than the favorable ones. Statistically significant differentially expressed genes that are above the threshold for fold change (absolute value of FC >=1.5) as well as adjusted p-value (FDR<=0.01) are shown in red. Those that pass the cut-off of only adjusted p-value are shown in red, only FC in green and those that pass neither of the cut-offs are shown in grey. (C) The heatmap shows comparative levels of transcript expression (z-scores) in different cell lines (in triplicates). Contrast in expression levels of favorable and unfavorable cell lines is evident from the plot. Significantly enriched gene sets for differentially expressed genes in (D) favorable and (E) unfavorable cell lines are shown as dot plots with the radius of dots being proportional to number of DE genes in the gene set and color intensity proportional to p-value. Based on the significantly differentially expressed genes the enrichment of pathways in different phenotypes is depictive of the cellular state that can be utilized for estimating the factors involved in unstable transgene expression. (F) Significantly enriched gene sets in all the unfavorable groups corresponding to expression contrast based on stability or copy number. FDR values of gene sets upregulated in more than two favorable contrast groups are shown in the heatmap to represent similar transcriptional regulation while comparing similar phenotypes in distinct group of cell lines. X-axis represents different comparisons that are taken into consideration with the group in bold corresponding to the plotted enrichment. The gene-sets that are not enriched certain groups have been designated an FDR value of 1.00. A similar plot for gene set enrichment in favorable groups is reported as Suppl. Fig. 8. Part (c) of the plot represents enrichment of gene sets like extracellular matrix (ECM) regulators and various signaling pathways mostly in the group of unfavorable samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Hence, we further analyzed differential expression between favorable (low/medium copy number, stable: G9, 1E3) and unfavorable (high copy number and/or unstable: 5G10, 1C11, 6A6, 6H1) phenotype groups. We observed seven genes to be significantly upregulated in the favorable group (Fig. 7B, 7C and Suppl. Table 7). Functional association of trends in gene expression was checked with gene set enrichment analysis within these two groups. Considering the genotoxic stress with transgene integration into the cells, survival and proliferation necessitates intrinsic adaptive mechanisms and stress response pathways. Interestingly, such pathways are found to be enriched in the favorable phenotypes (Fig. 7D, Suppl. Table 8). These include CDK regulation of DNA replication by mini-chromosome maintenance (MCM) complex, 3′ UTR mediated translational regulation, signal recognition particle (SRP) dependent peptide chain elongation, unfolded protein response (UPR) and finally N-linked glycosylation.

To survive genotoxic stress, cells slow down cell cycle progression by inhibiting cyclin dependent kinases (CDKs) [37]. In response to replication stalling, ataxia telangiectasia and Rad3-related (ATR) kinase is activated that also regulates DNA damage repair [38]. Other than CDKs, mini-chromosome maintenance complex (MCM), a DNA helicase, is also important for genomic DNA replication [39]. Enrichment of gene sets - "Activation of ATR in response to replication stress", "MCM pathway" and "G2M checkpoints" clearly indicate cellular response for survival in response to genomic stress.

For transcriptional regulation, we found enrichment of gene set for nonsense-mediated mRNA decay (NMD) which is a universally conserved RNA quality control process [40]. Interaction of RNA binding proteins at 3′ untranslated regions (UTRs) and initiation factors at the 5′ end further influence mRNA translation. Peptide chain elongation and folding seems to be further controlled by

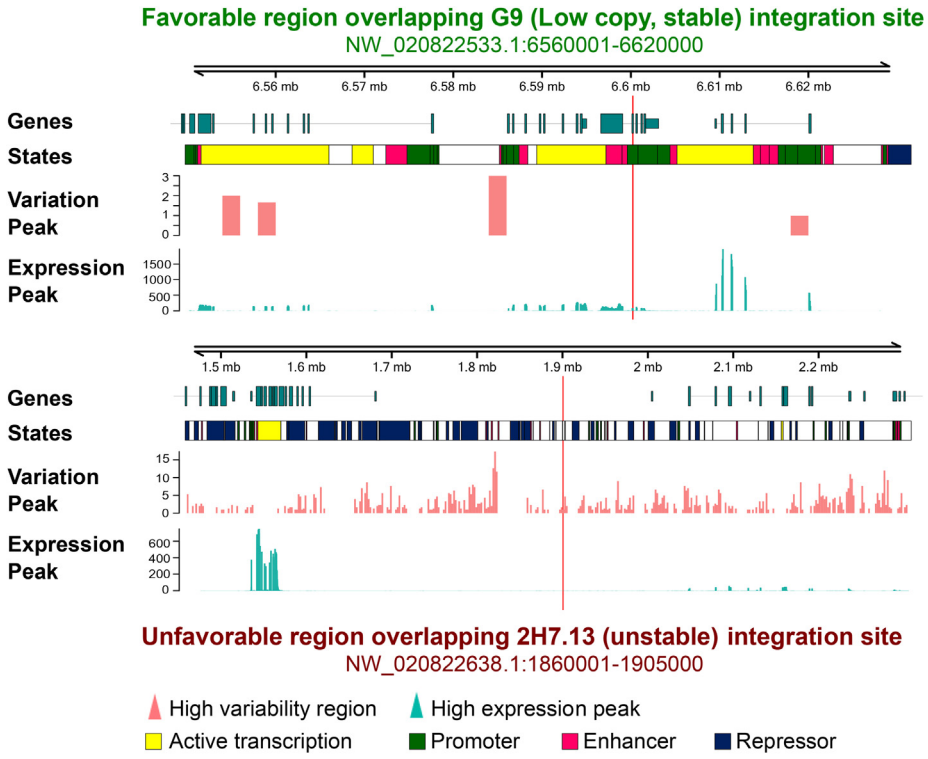

Fig. 8. Genomic profiles representing favorable and unfavorable landing pads in the genome. The figure shows genomic profile within a favorable and an unfavorable region overlapping known integration sites of stable (G9) and an unstable (2H7.13) transgene expressing cell lines respectively. The favorable landing pads were identified by browsing 20 kb regions that are close to high expression peaks and active chromatin states while being distant from high variability genomic regions and repressive chromatin states. The unfavorable landing pads were identified by zooming in with smaller window size of 5 kb bin to locate regions that are close to high variability regions and repressed chromatin states while being distant from high expression peaks and active chromatin states.

the signal recognition particle (SRP) that enables transfer of nascent peptide chain across ER membrane to the lumen [41]. Unfolded protein response (UPR), a dynamic intracellular signaling pathway [42], takes over the next level of control to manage protein folding capacity of the ER. This stress response pathway is also enriched in the favorable cell lines which is plausibly contributing towards cellular homeostasis and thereby stability of transgene expression. Finally, the enrichment of N-linked glycosylation in favorable cells adds another level that promotes the cells in favorable phenotype to stable transgene expression since addition of N-glycans to a protein is known to enhance its stability and solubility in the ER.

Fifty genes were found to be upregulated in the unfavorable group (Suppl. Table 7). These genes were significantly enriched in gene sets associated with extra-cellular matrix (ECM) regulators, interferon signaling, cancer and apoptosis regulation (Fig. 7E, Suppl. Table 9). Interestingly, it has been shown before that unstable cell lines are more prone to apoptosis [11]. Accumulation of cell debris and free DNA from cell lysis can lead to formation of large clumps [43], which connects to differentially expressed ECM regulators. Other than providing mechanical and structural support, ECM molecules like integrins, collagens and proteoglycans play a major role in cell signaling processes, developmental process, proliferation, differentiation and survival [44,45]. Since ECM proteins possess binding sites for growth factors as well as cell adhesion, defects in ECM assembly can induce differential regulation of signaling pathways and changes in cell architecture and motility that eventually triggers transcriptional regulation also [45]. As interferons are actively involved in anti-proliferative and apoptotic activities [46], the results indicate that in the unfavorable phenotypes cells are more prone to die due to some unknown stress signals. Also, the upregulation of phospholipase C mediated cascade indicates kinase activity that suggests the occurrence of differential regulation of proteins in the unfavorable cell lines.

Pairwise comparison for stable vs unstable (G9/5G10; 1E3/1C11; 6A6/6H1) and lower vs higher copy number ([G9,5G10] / [1E3,1C11]; [G9,5G10] / [6A6,6H1]; [1E3,1C11] / [6A6,6H1]) was also performed. Stable and lower copy number phenotype in each comparison was considered as favorable and the other as unfavorable (Suppl. Table 8). Fig. 7F shows gene sets upregulated in two or more unfavorable groups. Cluster "a" and "b" represent upregulation of transcription and translation related gene sets in unfavorable groups that were common to those in a few favorable groups as well. However, various ECM and signaling related gene sets in cluster "c" were mostly upregulated for more than two unfavorable groups specifically. We believe that these phenomena and associated genes can be utilized for the early prediction of instability of subclones during cell line development (Suppl. Figs. 5–7). Similar enrichment for favorable groups is shown in Suppl. Fig. 8.

3.6. Landing pads for targeted transgene integration identified across the genome

Transgene integration sites have been reported to frequently happen in regions of naturally occurring chromosomal breaks and within transcriptionally active regions [35]. Stable and high transgene expression has been observed mostly in introns or intergenic regions and in the vicinity of actively transcribed genes [7,11]. It has also been suggested that transgenic sites within regions with high sequence or structural variability and heterochromatin regions adjacent to endogenous silencers or insulators could result in unstable transgene expression [8,35,47]. The instability could also be due to physical steric hindrance that elicits

recombination and DNA break repair and thereby causes deletions that can scramble local endogenous sequences, resulting in loss of protein production [35].

All these reports were confirmed while identifying factors responsible for unfavorable phenotype (high copy number, unstable transgene expression) in the antibody expressing cell lines used in this study (Fig. 5). Taking our results and literature reports into consideration, genome-wide landing pads for transgene integration were predicted with the hypothesis that having an integration site in low variability regions would result in stable transgene expression, where a preference to sites flanked by promoters, enhancers and high expression peaks would increase the chances of higher transgene expression (Fig. 8). Based on the 3rd quantile values of boxplots in Fig. 5, thresholds for the contributing factors were decided and the landing pads for detecting favorable and unfavorable transgene integrations were determined.

We could extract 7,166 regions with favorable features that are expected to support stable and high transgene expression ranging from 20 kb to 960 kb in length (Suppl. Table 11). The known loci of random integration sites of antibody expressing cell lines were overlapped with the deduced list of genome wide favorable regions. Amongst the 28 known integration sites that resulted in stable transgene expression, 10 overlapped with our list. Considering the outliers observed in Fig. 5, our list cautiously missed the remaining sites reported with stable expression. None of the seven known integration sites from the unstable transgene expressing cell lines showed any overlap. This supports the predicted outcome of our favorable regions (Table 4a).

Similar overlap was checked for 16,048 predicted unfavorable regions that range from 5 kb to 2.455mb in size (Suppl. Table 12). Amongst the seven integration sites that report unstable transgene integration in four cell lines, three sites were found to overlap with our list of unfavorable regions (Table 4b). These three correspond to the cell line with low copy number (5G10) and another for which the copy number is unknown (2H7.13). Considering that a higher copy number together with more transgene-transgene fusions could also have resulted in expression instability for the other cell lines (Medium copy, Stable – 1E3; High copy, Stable – 6A6; High copy, Unstable – 6H1), we expect their phenotype is probably independent of the surrounding genomic profile and more related to the integration event. Even though eight integration sites reported for cell lines with stable transgene expression were also found to overlap with the predicted unfavorable regions, all these sites are either within high variability regions or very distant from high expression peaks and some of them are also surrounded by repressed chromatin states. Moreover, genomic rearrangements were induced in the host genome by transgene integration in 5G10 and E1 cell lines (Suppl. Table 1). More than seven TG-TG fusions were also observed for E1, 2H7.13 and C2714B cell lines.

3.7. Genome browser

All the information deduced from the genomic and transcriptomic profiles along with previously computed chromatin states were uploaded to a genome browser accessible at https://www.borthlabchoresources.boku.ac.at/. As shown in Fig. 9, information tracks can be selected from the left to be overlaid with other profiles for getting a comprehensive view of a particular genomic region. The screenshot displays the genomic profile within a 960 kb region (marked as site 1 in "Favorable regions" track) having active chromatin states, very low sequence variability, high expressing regions and no genomic rearrangement. This site was ranked on top of the favorable regions for targeted transgene integration. Based on the results shown here, we hypothesize that transgene(s) integrated 1) in a low genomic variability region will avoid expression instability, and 2) in close proximity to an expressing region with active chromatin states and away from

Fig. 9. Genome browser for observing genomic, transcriptomic and epigenetic profiles across the genome. Menu on the left side presents all the available categories of information that has been loaded on the browser for access. Selecting each category can further show different sub-categories or a check list of tracks that can be visualized. The screenshot displays favorable landing pad, ranked first in the list, spanning 960 kb with very low genomic variability, high expression and active chromatin states.

repressed chromatin states will allow for higher transgene expression. The genomic profile around the integration sites for randomly integrated transgenes can also be visualized by selecting the corresponding regions listed in Suppl. Table 4.

4. Discussion and conclusion

Considering the extensive use of Chinese hamster ovary cell lines in the biopharmaceutical industry for the production of mAb like and complex therapeutics, this study is focused to address the challenge of unreliable transgene expression stability associated with these cell lines. In this study, we identified factors associated with expression instability. To this end we performed random transgene integrations, selected cell lines which show stable and unstable protein expression and analyzed them via genomic and transcriptomic profiling. The potential genomic factors associated with stable and unstable phenotypes were then browsed across the genome to catalogue genomic regions where transgene integration can result in stable or unstable transgene expression. Such sites can also be used for targeted transgene integration to avoid any potential factors resulting in expression instability.

To ensure capturing relevant industrial phenotypes of CHO cells, which are often distinct from the cell lines cultured in academic labs with different stress levels, the Horizon CHO-GS −/− cell lines developed at Janssen R&D Cell Line Development group were analyzed. The analysis was based on studying various potential causes associated with expression instability of the product – 1) transgene copy number, 2) genomic rearrangements within the integrated transgene or vector, or induced in the host genome, 3) position effects around the transgene integration site, 4) genomic variability or rearrangements around the integration site, 5) upregulation of certain genes or cellular processes. To capture all this information, 13 cell lines with low (<3), medium [4–15] or high copy number (>15) with both stable (Δ Titer < 25%) and unstable (Δ Titer > 25%) transgene expression were selected. Six of these cell lines (set 1) with both copy number and stability phenotype were sampled at early (P1) as well as late (P10) passage. TLA analysis around the integration site was then performed to observe any changes at the transgene integration level. Factors linked with transgene integration were also assessed by TLA analysis for seven additional cell lines (set 2) sampled at early passage. Early passage samples from set 1 were also analyzed for comparing whole genome (WGS based) and transcriptome (RNA-Seq) changes amongst each other and the host (Horizon-host, process evolved horizon, ATCC-host) cell lines. Previously reported chromatin states were also utilized to assess potential epigenetic factors influencing the expression of transgenes integrated nearby.

TLA analysis of the low copy cell line with unstable transgene expression (5G10) reported two integration sites, one of which induced an inversion and the other a deletion in the host genome. Moreover, partial integration at a third site was observed at passage 10 (Fig. 3). Products from multiple integration sites resulting in rearrangements of the host genome and incomplete transcripts expressed at late passage with the risk of interfering with the host transcriptome can without doubt be associated with unstable transgene expression for this cell line. Investigating any position effect of the integration site on the transgene, we could notice high variability bins and a large distance from high expression peaks for two sites in the low copy unstable cell line (Fig. 5, Suppl. Table 4). The third site was found to lie within a repressed heterochromatin region. In contrast, the integration site for the stable counterpart of this low copy cell line shows no genomic rearrangement and falls within a low variability and actively transcribing region. Apart from this, differential expression of genes associated with collagens, N-cadherin pathway and extracellular matrix organization along with Peroxisome Proliferator-Activated Receptors (PPAR) and Calcium signaling pathways were also observed within the unstable cell line in reference to the stable cell line (Suppl. Table 10). These processes mediate a cross talk between intracellular cytoskeleton to the environment in the ECM, thereby playing a role in cell survival, development, differentiation and proliferation [45]. Hence, they are additional evidence that indicate adaptation in cellular processes to deal with stress which finally results in an unstable phenotype.

Although these factors revealed clear evidences distinguishing stable and unstable phenotypes in low copy cell lines, synergistic effects from multiple factors are expected to result in unstable transgene expression for cell lines with high transgene copy number. Multi-copy integrations are mostly associated with transgene silencing because of head-to-head or tail-to-tail TG-TG fusions that result in physical steric hindrance or undesirable transcriptional products that lead to transcriptional silencing of otherwise expressed regions of the host genome [35]. This phenomenon is evidently noticeable in medium copy cell lines where only three TG-TG fusions were observed for the cell line with stable transgene expression, but 16 fusions in the unstable cell line. Amongst those, four were head-to-tail, four tail-to-head, five tail-to-tail (two identical because of homology within the transgene) and three head-to-head fusions. However, the stable transgene expression with an integration site in a high variability region for the 1E3 cell line (medium copy, stable) is surprising (Suppl. Table 4). This indicates that although genomic variability at the integration site is a major factor in deciding the fate of transgene expression (5G10, 2H7.13), it might not be the sole responsible criteria to result in loss of titer (1C11, 6H1). Interestingly when investigating transcriptional differences between stable and unstable phenotypes for medium copy cell lines, we observed enrichment of gene sets associated with ECM regulators and cell receptor signaling pathways similar to the observation for low copy cell lines. Additionally, differentially expressed genes were also enriched for gene sets associated with apoptosis and DNA damage response (Suppl. Table 10).

While the potential reasons for loss in titer were more evident for low copy integrations in G9 and 5G10 cell lines with loss in clarity for the genomic variability criteria in medium copy cell lines, the uncertainty increases for high copy integrations (6A6, 6H1). While multiple copy integrations themselves are one of the major causes for transgene silencing and hence loss in titer over passaging, the other genomic factors (such as induced SVs) seem to be less relevant. Both cell lines with stable and unstable phenotype with high copy number were reported to have identical transgene integration sites (Suppl. Table 1). This might indicate the genome's susceptibility to allow such huge load to be integrated only in a few selected sites which were targeted in these cases. Both cell lines were reported to have two integration sites amongst which one overlaps the intron of an actively transcribing gene – Nidogen1, and the other the intron of a silent non-coding RNA. The latter being surrounded by Polycomb repressed marks and repressed heterochromatin states could be the potential cause of loss in titer after 10 passages for the unstable cell line. However, the same profile around the integration site resulted in a stable phenotype (Suppl. Table 4). Upregulation of gene sets associated with unstable phenotypes in lower copy cell lines like ECM regulators and cell signaling were also observed in the stable phenotype of high copy cell lines. Thus, it was surprising to observe the same gene sets to have opposite trends of enrichment between the lower and high copy number cell line subsets (Suppl. Table 10). One could speculate that the rules that hold for low to medium copy number integrations are of less importance in the case of high copy number, where the effect of copy number may be predominant. It is

also possible that the 10 passages used in this study to assess stability were too few to exhibit an unstable phenotype.

Taken together, these results suggest that expression stability may be controlled at 3 levels: 1) the choice of an integration – site with low genomic variability and high transcriptional activity, 2) the organization of the transgenic locus, with low transgene fusions and no genomic rearrangement upon integration, and 3) the absence of differential expression of genes that indicate stress related cellular processes. Low copy targeted integration of the transgene into regions with low genomic variability, high expression profile and open chromatin structure with good accessibility to transcription factors and regulatory enzymes increases the chances for cell lines to show consistent productivity and stability [35,48]. At the same time, the precise arrangement of the transgene and vector within the integration site is an important factor that contributes to stable or unstable outlook irrespective of the suitability of the integration site. It thus should be analyzed and taken into consideration for the choice of a subclone already at early stages of cell line development.

Raw data availability

Whole genome sequencing, RNA sequencing and TLA sequencing data can be accessed at European Nucleotide Archive (ENA) using the project code PRJEB39258. Scripts for analyzing all the data are available at https://github.com/hd4git/StabilityAnalysis.

Author statement

Heena Dhiman performed all bioinformatics analyses and wrote the manuscript.

Marguerite Campbell generated the cells lines and performed all lab experiments.

Michael Melcher supervised the statistical analyses.

Kevin Smith and Nicole Borth devised and supervised the study.

All authors read and corrected the manuscript.

Funding

This study was supported by the Austrian Center of Industrial Biotechnology, a COMET K2 competence center of the Austrian Research Promotion Agency FFG. HD received support from the "eCHO Systems" ITN PhD Program funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 642663.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank David Ruckerbauer, Dr. Nicholas Marx and Neža Novak from University of Natural Resources and Life Sciences for scientific discussions and their insightful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.11.008.

References

[1] First Successful Laboratory Production of Human Insulin Announced (1978) Genentech Press Releases.
[2] Walsh G. Biopharmaceutical benchmarks 2018. Nat Biotechnol 2018;36:1136–45.
[3] Wurm FM. Production of recombinant protein therapeutics in cultivated mammalian cells. Nat Biotechnol 2004;22:1393–8.
[4] Walsh G. Biopharmaceutical benchmarks 2014. Nat Biotechnol 2014;32:992–1000.
[5] Jayapal K, Wlaschin K, Hu W, Yap M. Recombinant protein therapeutics from CHO Cells - 20 years and counting. Chem Eng Prog 2007;103:40–7.
[6] Hamaker NK, Lee KH. Site-specific integration ushers in a new era of precise CHO cell line engineering. Curr Opin Chem Eng 2018;22:152–60.
[7] Lee JS, Kildegaard HF, Lewis NE, Lee GM. Mitigating Clonal Variation in Recombinant Mammalian Cell Lines. Trends Biotechnol 2019. https://doi.org/10.1016/j.tibtech.2019.02.007.
[8] Bandyopadhyay AA, O'Brien SA, Zhao L, Fu H-Y, Vishwanathan N, Hu W-S. Recurring genomic structural variation leads to clonal instability and loss of productivity. Biotechnol. Bioeng.
[9] Wilson C, Bellen HJ, Gehring WJ. Position effects on eukaryotic gene expression. Annu Rev Cell Biol 1990;6:679–714.
[10] Dorai H, Corisdeo S, Ellis D, Kinney C, Chomo M, Hawley-Nelson P, et al. Early prediction of instability of Chinese hamster ovary cell lines expressing recombinant antibodies and antibody-fusion proteins. Biotechnol Bioeng 2012;109:1016–30.
[11] O'Brien SA, Lee K, Fu H-Y, Lee Z, Le TS, Stach CS, et al. Single copy transgene integration in a transcriptionally active site for recombinant protein synthesis. Biotechnol J 2018;13:1800226.
[12] Gaidukov L, Wroblewska L, Teague B, Nelson T, Zhang X, Liu Y, et al. A multi-landing pad DNA integration platform for mammalian cell engineering. Nucleic Acids Res 2018;46:4072–86.
[13] Fan L, Kadura I, Krebs LE, Hatfield CC, Shaw MM, Frye CC. Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. Biotechnol Bioeng 2012;109:1007–15.
[14] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.
[15] Andrews S. FastQC: A quality control tool for high throughput sequence data.
[16] Rupp O, MacDonald ML, Li S, Dhiman H, Polson S, Griep S, et al. A reference genome of the Chinese hamster based on a hybrid assembly strategy. Biotechnol Bioeng 2018. https://doi.org/10.1002/bit.26722.
[17] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
[18] Broad Institute (2018) Picard tools.
[19] Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinforma Oxf Engl 2016;32:292–4.
[20] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
[21] Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol 2019;20:117.
[22] Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinforma Oxf Engl 2016;32:1220–2.
[23] Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012;28:i333–9.
[24] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 2014;15:R84.
[25] Cameron D, Dong R. Structural Variant Annotation: Variant annotations for structural variants; 2019.
[26] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357–60.
[27] Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–9.
[28] Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 2014;42:W187–91.
[29] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
[30] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005;102:15545–50.
[31] Feichtinger J, Hernández I, Fischer C, Hanscho M, Auer N, Hackl M, et al. Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. Biotechnol Bioeng 2016;113:2241–53.
[32] Kong Q, Wu M, Huan Y, Zhang L, Liu H, Bou G, et al. Transgene expression is associated with copy number and cytomegalovirus promoter methylation in transgenic pigs. PLoS ONE 2009;4:e6679.
[33] Strathdee D, Ibbotson H, Grant SGN. Expression of transgenes targeted to the Gt(ROSA)26Sor locus is orientation dependent. PLoS ONE 2006;1:e4.

[34] Tang W, Newton RJ, Weidner DA. Genetic transformation and gene silencing mediated by multiple copies of a transgene in eastern white pine. J Exp Bot 2006;58:545–54.
[35] Kohli A, Melendi PG, Abranches R, Capell T, Stoger E, Christou P. The quest to understand the basis and mechanisms that control expression of introduced transgenes in crop plants. Plant Signal Behav 2006;1:185–95.
[36] Vcelar S, Jadhav V, Melcher M, Auer N, Hrdina A, Sagmeister R, Heffner K, Puklowski A, Betenbaugh M, Wenger T, et al. Karyotype variation of CHO host cell lines over time in culture characterized by chromosome counting and chromosome painting. Biotechnol. Bioeng., 10.1002/bit.26453.
[37] Lim S, Kaldis P. Cdks, cyclins and CKIs: roles beyond cell cycle regulation. Development 2013;140:3079–93.
[38] Maréchal A, Zou L. DNA damage sensing by the ATM and ATR kinases. Perspect. Biol.: Cold Spring Harb; 2013. p. 5.
[39] Forsburg SL. Eukaryotic MCM proteins: beyond replication initiation. Microbiol Mol Biol Rev MMBR 2004;68:109–31.
[40] Zhao Y, Ye X, Shehata M, Dunker W, Xie Z, Karijolich J. The RNA quality control pathway nonsense-mediated mRNA decay targets cellular and viral RNAs to restrict KSHV. Nat Commun 2020;11:3345.
[41] Fewell SW, Brodsky JL. Entry into the endoplasmic reticulum: protein translocation, folding and quality control. In: Trafficking inside cells. New York, New York, NY: Springer; 2009. p. 119–42.
[42] Hetz C, Papa FR. The unfolded protein response and cell fate control. Mol Cell 2018;69:169–81.
[43] Merck Common Cell Culture Problems: Cell Clumping. Common Cell Cult. Probl. Cell Clumping.
[44] Kumar A, Baycin-Hizal D, Wolozny D, Pedersen LE, Lewis NE, Heffner K, et al. Elucidation of the CHO Super-Ome (CHO-SO) by Proteoinformatics. J Proteome Res 2015;14:4687–703.
[45] Kim S-H, Turnbull J, Guimond S. Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. J Endocrinol 2011;209:139–51.
[46] Kotredes KP, Gamero AM. Interferons as inducers of apoptosis in malignant cells. J Interferon Cytokine Res 2013;33:162–70.
[47] Stam M. Review article: the silence of genes in transgenic plants. Ann Bot 1997;79:3–12.
[48] Papapetrou EP, Schambach A. Gene insertion into genomic safe harbors for human gene therapy. Mol Ther 2016;24:678–84.