Analysis of Aneuploidy Spectrum From Whole-Genome Sequencing Provides Rapid Assessment of Clonal Variation Within Established Cancer Cell Lines

Ahmed Ibrahim Samir Khalil¹, Anupam Chattopadhyay¹ and Amartya Sanyal²

¹School of Computer Science and Engineering, Nanyang Technological University, Singapore. Singapore. ²School of Biological Sciences, Nanyang Technological University, Singapore, Singapore.

ABSTRACT

BACKGROUND: The revolution in next-generation sequencing (NGS) technology has allowed easy access and sharing of high-throughput sequencing datasets of cancer cell lines and their integrative analyses. However, long-term passaging and culture conditions introduce high levels of genomic and phenotypic diversity in established cell lines resulting in strain differences. Thus, clonal variation in cultured cell lines with respect to the reference standard is a major barrier in systems biology data analyses. Therefore, there is a pressing need for a fast and entry-level assessment of clonal variations within cell lines using their high-throughput sequencing data.

RESULTS: We developed a Python-based software, AStra, for de novo estimation of the genome-wide segmental aneuploidy to measure and visually interpret strain-level similarities or differences of cancer cell lines from whole-genome sequencing (WGS). We demonstrated that aneuploidy spectrum can capture the genetic variations in 27 strains of MCF7 breast cancer cell line collected from different laboratories. Performance evaluation of AStra using several cancer sequencing datasets revealed that cancer cell lines exhibit distinct aneuploidy spectra which reflect their previously-reported karyotypic observations. Similarly, AStra successfully identified large-scale DNA copy number variations (CNVs) artificially introduced in simulated WGS datasets.

CONCLUSIONS: AStra provides an analytical and visualization platform for rapid and easy comparison between different strains or between cell lines based on their aneuploidy spectra solely using the raw BAM files representing mapped reads. We recommend AStra for rapid first-pass quality assessment of cancer cell lines before integrating scientific datasets that employ deep sequencing. AStra is an open-source software and is available at https://github.com/AISKhalil/AStra.

KEYWORDS: Aneuploidy spectrum, cancer cell lines, clonal variation, copy number state, whole-genome sequencing, computational tool

Introduction

Cell lines are the cornerstone of cancer research and drug screening studies. The availability of a variety of NGS-based molecular datasets generated in different laboratories using cancer cell lines has paved the way for their integrative analysis and interpretation.¹ However, even established cell lines undergo continuous genetic alterations over the cumulative number of passaging and culture conditions to display significant clonal variations.²⁻⁶ Recently, Ben–David et al.³ have provided compelling evidence of heterogeneity in cancer cell lines cultured in different laboratories in terms of their mutational and chromosomal aberrations, gene expression, and drug response. Consequently, clonal variations in cancer cell lines due to genetic drift and culture conditions can adversely impact experimental outcomes. Moreover, the integration of genomic datasets derived from different clones/strains will yield unreliable conclusions. Therefore, there is a strong need for a rapid first-pass assessment method that can help scientists to measure and compare the variations in a particular cell line obtained from different sources without the involvement of any elaborate computational algorithms or any detailed and complex experimental investigations. The main aim is to perform a simple quality control check of cell line stocks using WGS reads to measure the extent of their variability and help assist in simplifying decision-making process while combining or integrating cancer datasets or their downstream results such as gene expression and drug response.

Given adequate bioinformatics support, WGS-derived genetic profiles reproduce the results of experimental approaches at higher resolution. Therefore, in theory, strain differences in cancer cell lines can be established by identifying single-nucleotide variants and/or CNVs through computational analyses of their WGS data.⁷⁻¹¹ Over the years, several read depth (RD)-based tools⁹,¹₂,¹₃ were available to detect the ploidy level and CNVs from WGS data which can be utilized for analysis of strain difference. However, some tools are built on licensed software and some necessitate advanced computational skills and knowledge-based tunings, such as a matched/reference control
or ploidy information. Additionally, many tools require high-coverage data or auxiliary information (mappability, GC-content, gap regions, etc.) or a specific format of input data. Therefore, there is a space for developing a computational tool that offers a simple data visualization platform for measuring clonal variability of cancer cell lines from raw WGS reads without going into too much technical detail.

Hence, we developed a Python-based software, AStrA (Aneuploidy Spectrum [detection] through read depth analysis) for rapid analysis of clonal variations in cancer cell lines by extracting their copy-number information directly from input WGS BAM (Binary Alignment Map) files without any requirement of a matched sample or auxiliary information. AStrA first computes the segmental aneuploidy profile that captures the elementary knowledge of the copy number state (interval) of large genomic segments. For that, AStrA estimates the copy number (CN) reference (CN = 2), even from low-pass sequencing data, by employing a multimodal distribution that streamlines the majority of genomic segments in the range of 2N to 4N based on the knowledge base of karyotypic information of majority of cancer cell lines.\(^{14}\) The aneuploidy information of genomic segments is then projected as a frequency distribution to obtain the aneuploidy spectrum that can be visually interpreted to distinguish different cancer cell lines as well as clonal variants.

In a proof-of-concept study, using exclusively WGS datasets of 27 strains of MCF7 breast cancer cell line obtained from different sources, AStrA reproduced strain-specific differences in aneuploidy profiles as reported earlier utilizing diploid reference samples and previously-reported ploidy information.\(^{3}\) Using simulated genomes with variable degrees of aneuploidy, we further established that AStrA can accurately trace the artificially-introduced changes in the aneuploidy spectrum. Additionally, using 19 cancer cell line datasets with variable levels of structural variations and data coverage, we showed that AStrA can effectively capture aneuploidy spectra of cancer cell lines at high speed (in minutes). Therefore, AStrA-based aneuploidy spectrum analysis provides a rapid and reliable first-pass quality control method to assess the impact of genetic drift or strain differences in cancer cell lines.

**Methods**

**AStra framework**

AStra utilizes the RD frequency distribution and the RD segments (binned at 100 kb) to identify the most-fitted segmental aneuploidy profile. In the absence of karyotype information, AStrA robustly estimates the segmental aneuploidy profile of a cell line based on 2 assumptions. First, CN reference (CN = 2) is the RD value that best allocates genomic segments into integer CN states.\(^{15}\) Second, the majority of genomic segments should have CN states ranging from 2N to 4N.\(^{14}\)

AStra first scans the RD frequency distribution to infer the candidate CN reference. For that, we employed 6 models (m1-m6) of unimodal/multimodal distributions for fitting the RD signal. Unimodal models are normal distributions with mean at 2N, 3N and 4N CN states, whereas multimodal models are generated by combining these unimodal models (Figure 1a). We utilize each model (m) separately to find the model-specific CN reference (CN\(_m\)) in 2 steps. First, we compute initial CN reference (CN\(_{mi}\)) that achieves the maximum overlap between each model and the RD frequency distribution of the input cell line. Second, we use the RD segments, computed using Pruned Exact Linear Time (PELT) method,\(^{16}\) to find the model-associated CN reference (CN\(_{mj}\)) by scanning the CN reference interval (1.9N to 2.1N) around the CN\(_{mi}\) (Figure 1b). This CN\(_{mj}\) best assigns the majority of genomic segments to integer/near-integer CN states and achieves the minimum centralization error (CE), the weighted summation of differences between the estimated copy numbers of CN-designated RD segments and their CN states. The final CN reference is then chosen out of these 6 candidate references (CN\(_{mj} \ldots m_6\)) that yields the lowest CE (Figure 1b). In the end, we obtain a collection of CN-designated RD segments within their corresponding CN state. Each CN state (1N, 2N, 3N, . . .) is a representative value for each CN interval (from 0.5N–1.5N, 1.5N–2.5N, 2.5N–3.5N, and so on respectively). Next, we compute the aneuploidy spectrum that displays contributions of different genomic segments with different CN states. Finally, we extract the CN-associated features such as cellular ploidy and whole-genome ploidy level.

**RD frequency distribution.** Cancer cells exhibit a multimodal distribution of RD signal due to widespread genetic and chromosomal aberrations. Given that most cancer cell lines are either diploid, triploid or tetraploid,\(^{14}\) the majority of genomic segments will therefore have copy numbers ranging between 2N to 4N. Hence, we build 6 RD frequency distributions as the weighted summation of Gaussian/normal distributions centered at 2N, 3N and 4N CN states to compute the candidate CN reference (Figure 1a):

\[
f(x) = \sum_i c_i \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-i)^2}{2\sigma^2}} \right), x \geq 0
\]

such as

\[
\int_0^\infty f(x) dx = 1
\]

\[
\int_0^\infty \sum_i c_i \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-i)^2}{2\sigma^2}} \right) dx = 1
\]

\[
\sum_i c_i \int_0^\infty \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-i)^2}{2\sigma^2}} \right) dx = 1
\]
Khalil et al

\[ \sum_{i} c_i = 1 \]

\[ c = \frac{1}{\sum_{i} 1} \]

where \( i \) is the common CN state (2, 3, 4) and \( c \) is a constant for the normalization of the probability distribution function. The standard deviation (\( \sigma \)) is chosen as 0.5/3 to make 99% of each Gaussian distribution within a single CN interval of unit width. This ensures the complete isolation between CN states by setting \( f'(x) = 0 \) at the boundaries of CN intervals (1.5, 2.5, 3.5, and 4.5).

Estimation of candidate CN reference. For each model (m), the initial CN reference (\( CN_{m_i} \)) is defined as the RD value that best allocates the RD segments around integer CN states. Final CN reference is selected with lowest centralization error.

\[ \text{Figure 1. AStra framework: (a) RD frequency distribution, extracted from WGS reads, is scanned against 6 prospective models (m1-m6) to identify the initial CN reference candidates (CN}_{m_{i,6}}). (b) RD segments are utilized for fine-tuning the initial CN reference candidates by searching their narrow intervals (1.9CN-2.1CN). For each interval, the CN reference candidate (CN}_{m_{i,6}} are the RD value that best allocates the RD segments around integer CN states. Final CN reference is selected with lowest centralization error.} \]
achieves the maximum matching between the RD frequency distribution \( r(x) \) and the model \( (m) \) frequency distribution \( f(x) \) (Figure 1a). The RD interval \([s, c]\) is divided into \( n \) equally-spaced RD values. At each RD value \( k \), \( f_i(x) \) is generated assuming this is the CN reference (2N) and a rank \( R_k \) is computed as:

\[
R_k = \sum_{j=s}^{c} r(j) \ast f_j(k), \quad k \in [s, c]
\]

Finally, RD value \( k \) with the maximum rank \( R_k \) is chosen as the initial CN reference (\( CN_{mi} \)).

**Centralization error.** Given a candidate CN reference, we first merge the neighboring genomic segments under a specific CN interval to divide the genome into contiguous RD segments with distinct integer CN states. Then, we compute the centralization error (CE) to measure the degree of localization of these segments around integer CN states (Supplemental Figure 1a):

\[
CN_i = RD_i \ast \frac{2}{\text{CNR}}, \quad S_j = RNE(CN_i),
\]

\[
CE_j = \sum_i |S_i - CN_i| \ast W_i, \quad S_i = j
\]

\[
CE = \sum_{j=1}^{n} CE_j,
\]

where \( i \) is the RD segment index, \( j \) is the CN state, \( RD_i \) is the median RD per segment, \( CN_i \) is the copy number of the segment, \( CNR \) is the candidate CN reference, \( S_j \) is the CN state of the segment (round-to-nearest value of \( CN_i \)), \( W_i \) is the width of the segment \( i \), and \( CE_j \) is the centralization error of segments of CN state \( j \).

**Features of the aneuploidy profile.** Many attributes can be extracted from the aneuploidy profile. First, the centralization score (CS) is computed as the percentage of RD segments that are close to their integer CN state (Supplemental Figure 1b):

\[
CS = \frac{\sum_i W_i \ast |S_i - CN_i| \leq 0.25}{\sum_i W_i}
\]

where \( i \) is the RD segment index, \( CN_i \) is the copy number of the segment, \( S_i \) is the CN state of the segment, and \( W_i \) is the width of the segment \( i \). Second, we compute the cellular (tumor) ploidy as the average of the observed copy numbers across the entire genome.\(^{17,18}\) Third, we define the whole-genome ploidy level of cancer cell lines as diploid, triploid or tetraploid based on the CN state that harbors the maximum percentage of genomic segments.

**Results**

**Astra framework provides a pragmatic solution for effective comparison of different strains of cancer cell lines**

RD signal provides a coverage-dependent numerical count of reads. However, for effective comparison among strains/cell lines of different coverages, the RD signal must be scaled to the standardized CN state. For that, Astra computes aneuploidy profile which is the normalized version of RD signal with the CN state information of every genomic bin (locus). Additionally, aneuploidy spectrum is the normalized RD signal frequency distribution that summarizes the percentage contribution of genomic bins with different CN states. For aneuploidy profiling, our fundamental assumption is that most genomic segments have near-integer copy number values.\(^{15,19}\) Correct estimation of CN reference (\( CN = 2 \)) is a prerequisite for accurate computation of CN states. We envisage that any tool will estimate CN reference using either of the 2 guiding principles. The first one is termed as RD scanning (RDS) method which is inspired by ABSOLUTE\(^{19}\) and the second guiding principle is the multimodal distribution scanning (MMDS) method inspired by nQuire\(^{20}\) (Supplemental Figure 2). In the RDS method, the RD signal range of the input sample is divided into \( m \) equally spaced RD values. Each RD value is considered as a candidate CN reference and the centralization error (CE) is computed. The CN reference is selected which yields the least CE. In the case of MMDS, a single multimodal distribution, comprising summation of normal distributions centralized at 2N to 10N, is used to fit the input RD signal distribution. CN reference is computed as the RD value that achieves the maximum overlapping between the 2 distributions. However, we believe that these methods may not find the accurate CN reference because they bestowed equal weightage to different CN states. Hence, a combination of 1N and 2N states is equally probable as the combination of 2N and 4N states. In other words, if we have 2 genomic loci with RD values of 100 and 200 reads/bin, CN of these segments can be inferred equally likely to be 1N and 2N, or 2N and 4N, or 3N and 6N, and so on, respectively. To overcome this problem, we have taken advantage of the common knowledge that almost all established and widely-used cancer cell lines have cellular ploidy ranging from diploid to tetraploid based on American Type Culture Collection (ATCC), CCLE (Cancer Cell Line Encyclopedia)\(^{21}\) and COSMIC (Catalogue Of Somatic Mutations In Cancer)\(^{22}\) database information. Therefore, Astra presents a pragmatic solution by narrowing down the RD values to specifically target combinations that favor allocation of the majority of genomic segments to 2N, 3N and 4N CN states. This underscores the reason for choosing 6 prospective models (m1-m6) for estimating CN reference (Figure 1a).

For evaluating Astra approach, we first computed the genome-wide RD signals of 27 strains of MCF7 cell line reported earlier\(^2\) by counting the number of WGS reads mapped to the genomic locus of specified bin size of 100 kb. The
primary advantage of large binning is that the RD signal can be readily estimated from low-coverage (<1x) WGS data. Visual inspection revealed that genome-wide RD signals of 27 strains vary significantly suggesting pervasive chromosomal alterations in MCF7 cell line cultured in different labs (Figure 2). Then, we compared AStra-derived aneuploidy profiles of 27 MCF7 strains (Supplemental Table 1) with their CNV profiles reported earlier.3 In the original study, Ben-David et al. utilized known

Figure 2. Aneuploidy signature reveals the genetic variations among the MCF7 strains. The genome-wide aneuploidy profiles (left) and aneuploidy spectra (right) for MCF7 strain C, G, H, P, and S are shown. Genomic loci in aneuploidy profiles are colored based on their copy number states (CN state 1 and 2: black, CN state 3 and 4: blue, CN state 4 and above: red). The aneuploidy spectrum shows the normalized RD frequency distribution where the dotted black lines denote the CN states whereas the red line denotes the median RD signal.
The genome-wide ploidy level of MCF7 (triploid to tetraploid) and a panel of normal samples as a diploid reference to compute the relative CNV profiles of these strains. Remarkably, AStrA successfully identified the correct aneuploidy profiles of all 27 strains as interpreted in the original study without the requirement of any reference control or ploidy information. Visually, both aneuploidy profiles and aneuploidy spectra demonstrated remarkable differences between MCF7 strains (Figure 2, Supplemental Figure 3). For example, chromosome (chr) 2 showed variable copy numbers (3N or 4N) among strains C, G, H, P, and S while chr 4 exhibited copy number of 3N for all strains (Figure 2). Similarly, we noticed considerable variations in aneuploidy spectra of 27 strains in terms of the number and amplitude of CN peaks. Thus, aneuploidy spectrum provides a simple histogram-based graphical signature of aneuploidy profile of cancer cell lines that can be effectively used for visual comparison of different strains of a cell line.

Notably, the cytogenetic analysis of metaphase chromosome spreads showed that the ploidy level of MCF7 strains is either hypertriploid or hypotetraploid with a modal chromosome number of 82 (range 66-87) as reported by ATCC. CN states are detected as distinct peaks in aneuploidy spectrum. We found that majority of genomic segments of the 27 MCF7 strains have CNs around 3N (e.g., C, G and H) or 4N (e.g., P and S) (Figure 2, Supplemental Figure 3). Therefore, given only whole-genome ploidy level of a cell line, aneuploidy spectrum can validate the correctness of AStrA result. For example, if karyotyping suggests the modal chromosome number of a cell line to be triploid/near-triploid, then the majority of genomic segments should be allocated around 3N CN state in aneuploidy spectrum.

It is important to note that these 27 MCF7 strains have different gene expression profiles and display differential sensitivity to cancer drug treatments. These phenotypic differences can be possibly attributed to the variations in their aneuploidy spectra. Therefore, we recommend to employ aneuploidy spectrum analysis for rapid assessment of strain-level variations.

**AStra successfully identifies the characteristic aneuploidy spectra of cancer cell lines**

We next evaluated the ability of AStrA to detect aneuploidy spectrum in a robust manner using simulated data as well as using publicly available WGS datasets (Supplemental Table 2). For the simulated data, we used an in-house-derived method (see Extended Methods under Supplemental Information) to manipulate the WGS reads of HG00119 (1000 Genomes Project sample of a diploid male) preserving the inherent systematic biases of the sequencing data. Using this approach, we generated 21 "artificial chromosomes" (chr 2 to chr 22) by introducing large-scale (>1 Mb) CN gain or loss regions randomly in HG00119 genome. We further created 21 "neo-genomes" comprising 23 chromosomes by merging the original HG00119 chromosomes and the "artificial chromosome(s)" in different combinations. In the first neo-genome (A), for example, we replaced the "original" chr 2 of HG00119 with "artificial" chr 2 keeping other chromosomes intact. Similarly, the second neo-genome (B) contains artificial chr 2 and 3, while the rest are from HG00119. We then progressively added more artificial chromosomes to create additional neo-genomes (C to U). We intentionally excluded 2 chromosomes (chr 1 and chr X) from any manipulation to evaluate the robustness of AStrA's CN reference estimation using these chromosomes as positive controls. These neo-genomes represent copy number aberrations of different complexity that can be used for evaluating AStrA's performance. We repeated this simulation 4 times with different combinations of artificially-introduced CN gain or loss regions to create 84 neo-genomes. Our evaluation showed that AStrA could identify the segmental aneuploidy of all neo-genomes (A to U) and compute their CN references (CN = 2) accurately (Supplemental Figure 4) as evidenced by the correct estimation of CN states of "unmodified" chr 1 and chr X as 2N and 1N respectively.

Next, we applied AStrA on 22 public WGS datasets that include 19 cancer cell lines and three 1000 Genomes Project samples of varying genome-wide ploidy levels that were experimentally verified (Supplemental Table 2). As demonstrated for MCF7 strains, AStrA successfully assigned the genomic segments to correct CN states in all cancer cell lines. These CN states are detected as peaks of input RD signal distribution (Figure 3a; frequency distribution histograms). Second, for most cell lines, the majority of genomic segments are assigned around the CN state corresponding to their reported genome-wide ploidy levels (Figure 3a; Supplemental Figure 5–7 and Table 2). However, aneuploidy spectra derived from WGS data of few cancer cell lines, such as MDA-MB-468, T47D, PC3, K562, etc., do not conform to the reported ploidy. We cannot verify if these 4 cancer WGS datasets were generated using cell line stocks that have undergone extensive clonal variation and therefore not reflecting the reported karyotype information. Nevertheless, AStrA could compute aneuploidy spectra of these cancer cell lines with a wide range of aneuploidy levels, e.g., from no/negligible aneuploidy (Supplemental Figure 5) to low aneuploidy (Supplemental Figure 6) to high-level aneuploidy (Supplemental Figure 7).

It is noteworthy to mention that WGS datasets for the majority of cancer cell lines were collected from low-coverage "input" DNA control data of ChIP-seq experiments. This supports the idea that aneuploidy profiles can be easily extracted from diverse genome-wide NGS datasets that are publicly available without the additional cost of performing targeted WGS. As far as computation time is concerned, AStrA computes the aneuploidy profiles and aneuploidy spectra of low-coverage (<3x) data in less than 3 minutes and high-coverage (~28x) in about 15 minutes (Supplemental Table 2). Taking together, AStrA provides a visual framework for rapidly studying aneuploidy signatures of cancer cell lines that should be taken into consideration before integrative analyses of cancer datasets generated by different scientific laboratories across the globe.
Figure 3. Centralization errors as a function of copy number reference (CN Ref) are computed for various cancer cell lines (a) and different MCF7 strains (b). (Top panel) The centralization errors based on CN reference (RD value corresponding to CN=2) computed by AStra, RDS, and MMDS methods are denoted by the red circle, black square, and blue square respectively. (Bottom panel) The corresponding aneuploidy spectra of cancer cell lines and MCF7 strains are shown where the dotted black lines denote the CN states whereas the red line denotes the median RD signal.
To illustrate the computational advantage of AStra approach over RDS and MMDS methods, we applied them on simulated datasets (neo-genomes A to U) as well as cancer datasets. We plotted the CEs for all CN candidates by scanning the entire range of values of the RD signal (Figure 3; Supplemental Figures 8 and 9). In general, for simulated datasets, we observed that CN references computed by AStra, RDS, and MMDS methods are identical (Supplemental Figure 8). In contrast, for 19 cancer datasets and MCF7 strains, CN reference varies considerably across these 3 methods (Figure 3; Supplemental Figure 9). As shown previously, CN reference and CN states computed by AStra match accurately for all cancer datasets. However, RDS method failed to identify the correct CN reference and consequently correct aneuploidy spectra of 12 MCF7 strains (A, F, K, R, S, T, U, V, W, X, Y, and Z) and 8 cancer cell lines (697, CAL-51, K562, MDA-MB-231, MOLT-4, SK-N-SH, SUM159, and T47D) (Supplemental Tables 1 and 2). Similarly, the MDMDS method was unable to find the correct CN reference of 8 MCF7 strains (A, F, S, T, U, W, Y, and Z) and 6 cancer cell lines (697, CAL-51, K562, 22Rv1, MOLT-4, SUM159, and T47D) (Supplemental Tables 1 and 2). Interestingly, we found that CE is a non-convex function and has many local minima. Therefore, the CN reference corresponding to the global minimum of the CE may not be always correct (Figure 3; Supplemental Figure 9).

The successful allocation of genomic segments into distinct CN states depends mainly on the degree of separation of RD signal corresponding to these segments. In other words, if the RD signal peaks corresponding to genomic segments are well separated (e.g., simulated data in Supplemental Figure 4), all the 3 methods can accurately estimate the CN states of these segments. Therefore, the discordance between 3 methods, in the case of real cancer datasets, may be attributed to the centralization score (CS) that approximately measures the degree of separation between different CN states (peaks) in the aneuploidy spectrum.

We analyzed the CSs of 27 MCF7 strains and 19 cancer cell lines. Although all 27 strains have similar coverage (<0.5x), their CSs vary remarkably (Supplemental Figure 10a). Interestingly, we noticed that strains whose CN references were wrongly identified by both RDS and MMDS methods generally have low CSs. We observed the same problem in the case of T47D, K562, 22Rv1, and CAL-51 cancer cell lines as well (Supplemental Figure 10b). Interestingly, the variation in CS across strains and cancer cell lines is independent of their whole-genome ploidy level and coverage (Supplemental Figure 10). Overall, the CS reflects an interesting feature of the WGS data that can be interpreted from aneuploidy spectrum. We opined that CS holds the clue for assessment of the quality of the sample where low CS may be attributed to sample-driven bias such as sample heterogeneity or cross-contamination.

Discussion

The meteoric rise of NGS-based techniques has provided easy access to WGS reads from various experiments. Even in the absence of targeted WGS of cancer cell lines, the genome sequencing reads are readily available from alternate sources such as ChIP-seq or Hi-C experiments. For example, the input DNA control of a ChIP-seq experiment provides the WGS data albeit at a low coverage. Similarly, reads from Hi-C experiments can be effectively used for computing RD signal. Therefore, the free access of NGS reads from hundreds of cancer cell lines has paved the way for easy sharing as well as integrative analyses of data. However, caution should be employed before integrating different datasets generated using the same cell line in different labs in the light of widespread concerns about genetic drift in cultured cell lines over time. Therefore, there is an urgent need for a “visual indicator” for rapid assessment of concordance/discordance of scientific datasets generated using cancer cell lines.

Several RD-based computational tools have been developed to compute ploidy level and CNVs. These genomic and genetic alterations represent unique features that can be used as a “digital signature” for the identification and verification of cancer cell lines. Many RD-based CNV detection tools are around for comprehensive discovery and genotyping of copy-number events by employing various statistical and modeling parameters, with each based on different software and requiring a specific format of input files and/or additional information (such as reference sample, auxiliary files for bias correction, cellular ploidy, whole-genome ploidy level, gain/loss percentage, etc.)

AStra is a simple and easy-to-use tool where the user needs to only input the BAM file. The output files comprising aneuploidy profile, aneuploidy spectrum, and CN-associated features (centralization score, cellular ploidy, genome-wide ploidy level, etc.) are generated within minutes. Although similar tools are available to estimate the ploidy level and assess the discordance between strains/samples, that are functionally analogous to AStra, they have specific limitations. For example, ConPADE computes ploidy level by assessing the distribution of allele frequency at biallelic single-nucleotide polymorphisms (SNPs), which requires very high coverage SNP data. Recently, nQuad has been developed to estimate the whole-genome ploidy level by fitting the RD frequency distribution to pre-determined distributions.

AStra offers a robust and accurate solution for comparing the relatedness of strains of cell lines. This approach provides an atlas of aneuploid segments in a matter of minutes. Therefore, there is an urgent need for a “visual indicator” for rapid assessment of concordance/discordance of scientific datasets generated using cancer cell lines.
under each ploidy assumption. However, as demonstrated, cancer cell lines harbor variable combinations of ploidy levels that cannot be modeled using these fixed distributions. On the other hand, ABSOLUTE and Cell Strainer require segmented RD/CNV profile as input, and cannot be applied on BAM file directly implying their dependence on additional tools to compute, normalize, and segment the RD signal. Additionally, Cell Strainer computes the copy number discordance of a test sample relative to the reference CCLE (Cancer Cell Line Encyclopedia) cell line data generated using Affymetrix SNP6.0 arrays. However, it is noteworthy to mention that the ideal reference sample is subjective and debatable. On the contrary, AStRa can compute the absolute CN state of genomic loci independent of any reference sample.

On a different note, cell line authentication has attracted increased attention in recent years in the scientific community to ensure reproducibility of results that is currently verified using the standard STR (short tandem repeat) profiling. We believe authentication is a collective process to verify the identity of a cell line. The multi-level numerical, structural, and sequence alterations of chromosomes are difficult to capture using a single technique. Moreover, even the authenticated cell lines may evolve with time in continuous culture due to genetic drift. In this situation, in silico estimation of genetic alterations using genome sequencing reads can provide an alternative approach for the assessment of various aspects of the quality of cell lines. Therefore, we envisage that our NGS-based solution, AStRa, can be a precursor for developing new tools for digital authentication of cell lines that can complement laboratory-based verification techniques.

Acknowledgements
We like to thank Costerwell Kyriem for helpful discussions. We acknowledge Sanjal and Chattopadhyay lab members for their valuable comments.

Author Contributions
AS, AC, and AISK conceived the project. AISK developed the AStRa software, with inputs from AC and AS, and performed all the analyses. AISK, AC, and AS analyzed the data and prepared the manuscript. All authors read and approved the final manuscript.

ORCID iD
Amartya Sanjal
https://orcid.org/0000-0002-2109-4478

Supplemental material
Supplemental material for this article is available online.

REFERENCES
1. Park ST, Kim J. Trends in next-generation sequencing and a new era for whole genome sequencing. Int J Neurourol J. 2016;20:576–583.
2. Kleivi K, Teixeira MR, Elnaas M, et al. Genome signatures of colon carcinoma cell lines. Cancer Genet Cytogenet. 2004;155:119–131.
3. Ben-David U, Siranosian B, Ha G, et al. Genetic and transcriptional evolution alters cancer cell line drug response. Nature. 2018;560:325–330.
4. Hynds RE, Vladimirou E, Janes SM. The Secret Lives of Cancer Cell Lines. The Company of Biologists Ltd; 2018.
5. Spans L, Arak ZK, Van Nieuwerburgh F, et al. Variations in the exome of the LNCaP prostate cancer cell line. Prostate. 2012;72:1317–1322.
6. Thompson SL, Compton DA. Chromosomes and cancer cells. Chromosoma Res. 2011;19:433–444.
7. Otto R, Sers C, Leser U. Robust in-silico identification of cancer cell lines based on next generation sequencing. Oncotarget. 2017;8:34310–34320.
8. Petljak M, Alexandrov LB, Brammeld JS, et al. Characterising mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. Cell. 2019;176:1282–1294.e20.
9. Duan J, Zhang J-G, Deng H-W, Wang Y-P. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. PLoS One. 2013;8:e61928.
10. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the cancer cell line encyclopedia. Nature. 2019;569:503–508.
11. Almeida JJ, Cole KD, Plant AL. Standards for cell line authentication and beyond. PLoS Biol. 2016;14(1):e1002476.
12. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinformatics. 2013;14:81.
13. Alkodi A, Louhimoi R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Brief Bioinform. 2015;16:242–254.
14. Nicholson JM, Cimini D. Cancer karyotypes: survival of the fittest. Front Oncol. 2013;3:148.
15. Khalil A1S, Khryiem C, Chattopadhyay A, Sanjal A. Hierarchical discovery of large-scale and focal copy number alterations in low-coverage cancer genomes. BMC Bioinformatics. 2020;21:147.
16. Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc. 2012;107:1590–1598.
17. Luo Z, Fan X, Su Y, Huang YS. Accuracy: accurate tumor purity and ploidy inference from tumor-normal WGS data by joint modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants. Bioinformatics. 2018;34:2004–2011.
18. Yu Z, Liu Y, Shen Y, Wang M, Li A. CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. Bioinformatics. 2014;30:2576–2583.
19. Carter SL, Pin M, Cano LM, Kamoun S, Burbano HA. nQuire: a statistical framework for ploidy estimation using next generation sequencing. BMC Bioinformatics. 2018;19:122.
20. Barretina J, Caponigro G, Stramski N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–607.
21. Forbes SA, Beare D, Boutroselakis H, et al. COSMIC: somatic cancers at high-resolution. Nucleic Acids Res. 2017;45:D777–D783.
22. Rondon-Lagos M, Verdun Di Cantogno L, Marchio C, et al. Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis. Mol Cytogenet. 2014;7:8.
23. Li R, Hsuan JT, Zhang Y, et al. Computational tools for copy number variation in Hi-C data. Nucleic Acids Res. 2012;40:886–896.
24. Khalil A1S, Muzaki SRB, Chattopadhyay A, Sanjal A. Identification and utilization of copy number information for correcting Hi-C contact map of cancer cell lines. BMC Bioinformatics. 2020;21:506.
25. Vidal E, Díly F, Quiles J, et al. OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. Nucleic Acids Res. 2018;46:e49.
26. Servaert N, Vasequazius N, Heard E, Barilior E, Vert JP. Effective normalization for copy number variation in Hi-C data. BMC Bioinformatics. 2018;19:313.
27. Chakraborty A, Ay F. Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics. 2018;34:338.
28. Khalil A1S, Muzaki S, Chattopadhyay A, Sanjal A. Identification and utilization of copy number information for correcting Hi-C contact map of cancer cell lines. BMC Bioinformatics. 2020;21:506.
29. Boeza J, Nataf T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number alteration and allelic content using next-generation sequencing data. Bioinformatics. 2012;28:433–435.
30. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS One. 2011;6:e16327.
31. Margarido GR, Heckerman D. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. BMC Bioinformatics. 2013;14(1):e100429.
32. Augusto Correa Dos Santos R, Goldman GH, Riano-Pachón DM. ploidyNGS: visually exploring ploidy with next generation sequencing data. Bioinformatics. 2017;33:2575–2576.