Digital Karyotyping for Rapid Authentication of Cell Lines

Ahmed Ibrahim Samir Khalil
Nanyang Technological University
Anupam Chattopadhyay
Nanyang Technological University
Amartya Sanyal (asanyal@ntu.edu.sg)
Nanyang Technological University https://orcid.org/0000-0002-2109-4478

Research article

Keywords:

Posted Date: December 16th, 2019

DOI: https://doi.org/10.21203/rs.2.18908/v1

Version of Record: A version of this preprint was published at Cancer Informatics on October 16th, 2021.
See the published version at https://doi.org/10.1177/11769351211049236.

Digital Karyotyping for Rapid Authentication of Cell Lines

Ahmed Ibrahim Samir Khalil[1], Anupam Chattopadhyay[1,*], Amartya Sanyal[2,*]

[1]School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798.
[2]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551.

Email:
Ahmed Ibrahim Samir Khalil: ahmed.ibrahim@ntu.edu.sg
Anupam Chattopadhyay: anupam@ntu.edu.sg
Amartya Sanyal: asanyal@ntu.edu.sg

*** corresponding authors**
Amartya Sanyal, Ph.D.
School of Biological Sciences
Nanyang Technological University
60 Nanyang Drive, SBS-05n-22,
Singapore 637551
Tel: (+65) 6513 8270

Anupam Chattopadhyay, Ph.D.
School of Computer Science and Engineering
Nanyang Technological University
50 Nanyang Avenue N4-02c-105
Singapore 639798
Tel: (+65) 6790 6092

Abstract

Background

The widespread concern about genetic drift and cross-contamination of cell lines calls for a pressing need for their authentication. The current genetic techniques for authentication are time-consuming and require specific documentary standard and laboratory protocols. Given the fact that whole-genome sequencing (WGS) data are readily available, read depth (RD)-based computational analyses has allowed the estimation of genetic profiles of cell lines.

Results

We propose WGS-derived aneuploidy profiling as a prototype of digital karyotyping for authentication of cancer cell lines. Here, we describe a Python-based software AStra for *de novo* estimation of the genome-wide aneuploidy profile, the copy number of every genomic loci, from raw WGS reads. We demonstrated that aneuploidy profile offers a unique signature that can distinguish the clonal variants (strains) of a cell line. We evaluated our approach using simulated data and variety of cancer cell lines. We further showed that cell lines exhibit distinct aneuploidy patterns which corroborate well with the experimental observations.

Conclusions

AStra is a simple, user-friendly, and free tool that provides the elementary information about the chromosomal aneuploidy for cell line authentication. AStra provides an analytical and visualization platform for rapid and easy comparison between different cell lines/strains. We recommend AStra for rapid first-pass quality assessment of scientific data that employ cancer cell lines. AStra is an open source software and is available at https://github.com/AISKhalil/AStra.

1. Introduction

Cancer cell lines are the cornerstone of cancer research and drug screening. However, even established cancer cell lines undergo continuous genetic alterations over passaging time and culture conditions to display significant clonal variations (strain differences) [1-5]. Additionally, in laboratory conditions cancer cell lines face the risk of contamination with other cell lines (intraspecies contamination). Scientific data collected using cross-contaminated or genetically-drifted cell lines can lead to irreproducible results. Therefore, verification of the identity and purity of the cell lines should be a standard practice for reproducibility of research and integrative analysis of datasets produced by different labs using same cell line [6].

Comprehensive cell line authentication is a collective process using different genetic and genomic techniques. Genetic characterization of cell lines has been conventionally performed using different experimental techniques that include karyotyping, short tandem repeat (STR) and single nucleotide polymorphism (SNP) profiling [6]. Current gold standard authentication services based on STR profiling are provided by trained professionals using controlled experimental protocols which are time consuming. Additionally, authentication is not a one-time service but needs to be carried out in regular intervals with currently-handled sample. Despite their availability and advantage, none of these authentication methods can capture the complete repertoire of genetic profiles to address all the quality assessment metrics. Moreover, many techniques require prior knowledge of reference standards to select targeted probes for authentication.

The advancement of next generation sequencing (NGS) technology, augmented by development of numerous computational methods, has paved the way for comprehensive characterization of human cell lines [7]. Recent explosion in whole-genome sequencing (WGS) data of cancer cell lines have provided powerful methods to identify different strains. Given adequate bioinformatics support, WGS can address almost all quality assessment metrics for cell line authentication, such as intra- and interspecies contamination, chromosomal alterations, genetic drift, etc. in a genome-wide manner [6]. The availability of WGS data have been successfully used for mutational and variant analyses and CNV detection of cancer cell lines [8-11]. These WGS-derived genetic profiles can reproduce the results of experimental approaches at higher resolution. Therefore, the verification of cancer cell lines can be established from computational analyses of the WGS data.

Here, we introduce WGS-derived aneuploidy profiling of cancer cells as a method for digital karyotyping. Aneuploidy profile provides the copy number information of every loci of the genome. Additionally, the aneuploidy spectrum is the percentage representation of genomic segments with particular copy number (CN) state. We have developed AStra (**A**neuploidy **S**pectrum (detection) **t**hrough **r**ead depth **a**nalysis) software for aneuploidy profiling which does not require any control/reference sample or any prior information about the cell line. AStra evaluation of 27 strains of MCF7 breast cancer cell line revealed strain-specific differences in aneuploidy signatures. Similarly, analysis of additional 19 cancer cell lines revealed that aneuploidy signature is distinct for every cell line and represents a characteristic feature that can be easily captured even from low-coverage (<1x) data. Therefore, digital karyotyping using aneuploidy profiling is a rapid *in silico* method to authenticate the cell lines from NGS-based experimental data. Comparison of aneuploidy profiles from NGS data produced by different labs will provide a quick and reliable quality control check for genetic drift or strain differences.

2. Methods

2.1 AStra framework

AStra utilizes the RD frequency distribution and the RD segments as the input data for identifying the most-fitted aneuploidy profile. In the absence of karyotype information, AStra estimates the CN reference (CN=2) of a cell line based on two assumptions. First, CN reference is the RD value that best allocates genomic segments into integer CN states. Second, majority of genomic segments should have copy number states ranging from 2N to 4N based on the karyotype information of most cancer cell lines.

AStra first scans the RD frequency distribution to approximately identify the candidate CN reference. To accomplish this, we employed six models (m1-m6) of unimodal/multimodal distributions for fitting the RD signal. Unimodal models are normal distributions with mean at 2N, 3N and 4N, whereas multimodal models are generated by combining these unimodal models (Fig. 1a). We utilize each model (m) separately to find the model-specific CN reference (CN_m) in two steps. First, we compute intial 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 [12], to find the model-associated CN reference (CN_m) by scanning the CN reference interval (1.9N to 2.1N) around the CN_{mi} (Fig. 1b). This CN_m best assigns integer/near-integer copy numbers to the majority of genomic segments and achieves the minimal centralization error (CE), the weighted summation of differences between the estimated copy number of CN-designated RD segments and their CN states. The final CN reference is chosen out of these six candidate references ($CN_{1\text{-}6}$), that yields the lowest CE (Fig. 1b). At the end, we obtain a collecton of CN-designated RD segments within their corresponding CN interval. Next, we compute the aneuploidy spectrum that displays contribution of different ploidy state into the complete aneuploidy profile.

2.1.1 RD frequency distribution

Cancer cells exhibit multimodal distribution of RD signal. Considering majority of genomic segments from cancer cell lines have copy numbers ranging from 2N to 4N, we build six RD frequency distributions as weighted summation of Gaussian/normal distributions centralized at 2N, 3N, and 4N to compute the candidate CN reference (Fig. 1a):

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

such as

$$\int_0^{\infty} f(x).\,dx = 1$$

$$\int_0^{\infty} \sum_i c * \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right).\,dx = 1$$

$$\sum_i c * \int_0^{\infty} \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right).\,dx = 1$$

$$\sum_i c = 1$$

$$c = \frac{1}{\sum_i 1},$$

where i is the common CN state (2, 3 and 4) and c is a constant for normalization of the probability distribution function. The standard deviation (σ) is chosen as 0.5/3 to make 99% of each Gaussian distribution is 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).

2.1.2 Estimation of candidate CN reference

For each model (m), initial CN reference (CN_{mi}) is defined as the RD value that achieves the maximum matching between the RD frequency distribution $r(x)$ and the model (m) frequency distribution $f(x)$ (Fig. 1a). The RD interval $[s, e]$ is divided into n equally spaced RD values. At each RD value k, $f_k(x)$ is generated assuming this is the copy number reference (2N) and a rank R_k is computed as:

$$R_k = \sum_{j=s}^{e} r(j) * f_k(j)\,, \quad k \in [s, e]$$

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

2.1.3 Centralization error

Given a candidate CN reference, we first merge the subsequent genomic segments with same CN state to divide the genome into contiguous RD segments of distinct CN states. Then, we compute the centralization error (CE) to measure the degree of localization of these segments around CN states (Supplementary Fig. 1a):

$$CN_i = RD_i * \frac{2}{CNR},$$

$$S_i = RNE(CN_i),$$

$$CE_j = \sum_i |S_i - CN_i| * W_i \quad , \quad S_i == j$$

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

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

2.1.4 Features of the aneuploidy profile

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

$$CS = \frac{\sum_i W_i \,,\ |S_i - CN_i| \leq 0.25}{\sum_i W_i}$$

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 whole-genome ploidy number (N) as average copy number across the entire RD segments:

$$N = \frac{\sum_i W_i * CN_i}{\sum_i W_i}$$

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. Third, we infer the genome-wide ploidy level (‘diploid’, ‘triploid’, ’tetraploid’) as the CN state harboring maximum percentage of genomic segments.

3. Results

3.1 RD signal captures the strain-specific karyotype of MCF7 cell line

Aneuploidy is paradoxically associated with both antiproliferative cellular response as well as uncontrolled cellular growth and cancer [13-16]. Aneuploidy is the hallmark of human cancers reported in 90% of solid tumors and 75% of blood cancers [17, 18]. Chromosome missegregation and instability have been implicated for manifestation of this complex genetic makeup of cancer cells and their evolution [19-21]. Consequently, the cancer cell lines show widespread aneuploidy at whole-chromosome and segment levels [5]. Interestingly, cancer cell lines spontaneously acquire new chromosomal alterations at a high rate even under controlled culture conditions resulting in their phenotypic evolution and strain difference [2, 22]. Hence, same cell line cultured in different labs can exhibit karyogram and mutational changes [1, 2, 4, 5].

Currently, NGS-based experimentation has improved upon the conventional genetic techniques for unbiased, reliable and high-resolution results. The explosion of NGS-based data has allowed easy access and sharing of data among researchers and their integrative analyses. However, care must be taken while integrating datasets from different labs generated using same cell line. Current genomics efforts have allowed a simple and rapid method for 'digital' karyotyping of cancer cell lines. Additionally, one can interrogate karyotype evolution of cancer cells by comparing samples from different labs. Therefore, we employed WGS data of 27 strains of a commonly used breast cancer cell line, MCF7 [2] to interrogate strain-specific karyotype differences.

We computed the genome-wide RD signal of 27 strains by counting the number of NGS reads mapped to a genomic locus of specified size (bin). Therefore, RD signal measures the *relative* frequency of genomic regions present in a cell. Visual analysis suggests that genome-wide RD signal of strains varies significantly showing pervasive chromosomal alterations in MCF7 (Fig. 2). This suggests that WGS data serves as an easy and valuable resource for visually interpret strain differences. The primary advantage of this approach is that the RD signal can be readily estimated even from low-depth sequencing (<1x) data.

3.2 RD signal can be decoded into biologically-relevant aneuploidy information

RD signal provides a coverage-dependent numerical count of reads. However, for effective comparison among strains/cell lines of different coverage, the RD signal should be scaled to standardized copy number state. To solve this, we propose to compute aneuploidy profile and aneuploidy spectrum from raw RD signal. Aneuploidy profile 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. Therefore, we developed AStra for rapid aneuploidy profiling-based digital karyotyping of cancer cells from NGS data without the aid of control/reference sample or prior knowledge of karyotype of input sample.

For evaluation, we compared AStra aneuploidy profiles of the 27 MCF7 strains (Supplementary Table 1) with their CNV profiles reported earlier using a panel of normal samples as reference [2]. AStra successfully identified the correct aneuploidy profiles of all strains as interpreted in the original study [2] without including any additional information or reference control. Visually, both the aneuploidy profile and aneuploidy spectrum demonstrated remarkable differences between MCF7 strains (Fig. 2, Supplementary Fig. 2). For example, chromosome (chr) 2 shows variable copy number (3 or 4) among strains C, G, H, P and S while chr 4 shows copy number of 3 for all strains (Fig. 2). Similarly, we noticed considerable variations in aneuploidy spectrum 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.

Notably, the cytogenetic analysis of MCF7 metaphase chromosome spreads showed that the genome-wide ploidy level is hypertriploid to hypotetraploid [23]. Integer 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) (Fig. 2, Supplementary Fig. 2). Therefore, given only genome-wide ploidy level of a cell line, aneuploidy spectrum can provide a rough validation of the correctness of AStra result. For example, if a cell line is known to be triploid/near-triploid, the majority of genomic segments should be allocated around 3N copy number.

It should be emphasized that these 27 MCF7 strains have different gene expression profiles and displayed differential sensitivity to cancer drug treatments [2]. Their phenotypic differences can be attributed to the variations in aneuploidy profiles [22]. Therefore, AStra offers a good alternative for cell line authentication using aneuploidy profiling.

3.3 AStra successfully identifies the aneuploidy signatures of cancer cell lines

We next evaluated the ability of AStra to detect aneuploidy profiles in a robust manner using simulated data as well as using publicly available WGS datasets (Supplementary Table 2). For the simulated data, we used in-house-derived method [24] [see Extended Method under Supplementary Information] to manipulate the WGS reads of HG00119 (1000 Genomes Project sample of diploid male) preserving the inherent systematic biases of the WGS data. Using this approach, we generated 21 'artificial chromosomes' (chr 2 to chr 22) by introducing copy number gain or loss regions randomly in HG00119 genome. We further created 21 'neo-genomes' comprising 23 chromosomes by mixing the original HG00119 chromosomes and the 'artificial chromosome(s)' using different combinations. In the first neo-genome (A), for example, we incorporated only artificial chr2 keeping the rest chromosomes of HG00119. Similarly, second neo-genome (B) contains artificial chr 2 and 3, while rest are from HG00119. We then progressively added more artificial chromosomes to create additional neo-genomes (C to U). We intentionally exclude two chromosomes (chr 1 and chr X) from any manipulation to evaluate the robustness of AStra's copy number estimation using these chromosomes as CN state controls. These neo-genomes represent aneuploidy profiles of different complexity that can be used for evaluating AStra's performance. We repeated this simulation 4 times with different combinations of induced structural variations to create 84 neo-genomes. Our evaluation showed that AStra could track the aneuploidy changes of neo-genome A to neo-genome U with correct CN reference (CN=2) estimation (Supplementary Fig. 3). This is evidenced by the correct estimation of the copy number of unaltered chr 1 and chr X as 2N and 1N respectively in all neo-genomes.

Next, we applied AStra on 22 public WGS datasets that include three diploid 1000 Genomes Project samples and 19 established cancer cell lines of varying genome-wide ploidy levels as reported by American Type Culture Collection (ATCC) (Supplementary Table 2). As demonstrated for MCF7 strains, AStra successfully identified the accurate aneuploidy spectrum of all cancer cell lines fulfilling two conditions. First, CN states are detected as peaks of input RD signal distribution (Fig. 3a; frequency distribution histograms). Second, majority of genomic segments have CN around the reported genome-wide ploidy level of these cells (Fig. 3a; Supplementary Fig. 4; Supplementary Table 2). It is noteworthy to mention that WGS data for majority of the cancer cell lines were collected from 'input' DNA control data of ChIP-seq experiments. This supports the idea that aneuploidy profile can be easily extracted from any genome-wide NGS datasets without additional cost of performing

targeted WGS. Taking together, AStra provides a glimpse of the cellular karyotype for verification or authentication.

As far as computation time is concerned, AStra computes the aneuploidy profiles and spectrum sequencing data of low-coverage (<3x) in less than 3 minutes and high-coverage (~28x) in about 15 minutes (Supplementary Table 2).

3.4 AStra framework provides a pragmatic solution for computing CN reference

Cancer cells harbor different degrees of hyperploidy with genomic segments belonging to different CN states. Nevertheless, the fundamental assumption for aneuploidy profiling is that most genomic segments should have integer copy number states [25]. Correct estimation of CN reference (CN=2), a prerequisite for accurate computation of CN states, can be achieved based on two guiding principles: 1) RD scanning (RDS) and 2) multimodal distribution scanning (MMDS) methods (Supplementary Fig. 5). In 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 centralization error (CE) is computed. The CN reference is selected which yields least CE. In case of MMDS, a single multimodal distribution, comprising summation of normal distributions centralized at 2N, 3N, …, 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 two distributions. However, we believe that these methods may not find the accurate CN reference because they bestowed equal weightage to the different CN states. Hence, a combination of 1N and 2N states is equally probable as 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 diploid to tetraploid karyotypes based on American Type Culture Collection (ATCC), CCLE [26] and COSMIC [27] database information. Therefore, AStra provides a pragmatic solution by narrowing down the RD values to specifically target combinations which favor majority of genomic segments with 2N, 3N and 4N CN states. That is the reason for choosing 6 prospective models (m1-m6) for AStra framework (Fig. 1a).

To illustrate the advantage of AStra aproach over RDS and MMDS methods, we applied them on simulated datasets (neo-genomes A to U) as well as cancer datasets. We plotted the CE for all CN candidates by scanning the entire range of values of the RD signal (Fig. 3, Supplementary Fig. 6 and 7). In general, we observed that CN reference (RD value corresponding to CN = 2) computed by AStra, RDS and MMDS methods are identical for simulated datasets (Supplementary Fig. 6). In contrast, CN reference varies considerably across methods in many cases for cancer datasets and MCF7 strains (Fig. 3; Supplementary Fig. 7). As shown previously, CN reference and CN states computed by AStra matches 100% for all cancer datasets. However, RDS method failed to identify the correct CN reference and subsequently correct aneuploidy spectrum 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) (Supplementary Table 1 and 2). Similarly, MMDS failed 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) (Supplementary Table 1 and 2). Interestingly, we found that CE is a non-convex function and has many local minima (Fig. 3; Supplementary Fig. 7). Moreover, the CN reference corresponding to global minimum of the CE may not be always correct.

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 corresponding to genomic segments are well separated (e.g. simulated data in Supplementary Fig. 3), all the three methods can accurately estimate the CNs of these segments. Therefore, the discordance between three methods, in case of real cancer datasets, may be attributed to the centralization score (CS) that approximately measures the degree of separation between different copy number states (peaks) of the cancer data. This score provides an empirical measure of the localization of genomic segments around CN states.

We analyzed the CS of the 27 MCF7 strains and 19 cancer cell lines. Although all 27 strains have similar coverage (~0.5x), their CS vary remarkably (Supplementary Fig. 8a). Interestingly, we noticed that strains whose CN references were wrongly identified by both RDS and MMDS methods generally have lower CS. We observed similar trend in case of cancer cell lines as well (*viz.* T47D, K562, 22Rv1 and CAL-51) (Supplementary Fig. 8b). It is important to note that the variation in CS across strains and cancer cell lines is independent of their whole-genome ploidy number and coverage (Supplementary Fig. 8). Overall, the CS reflects an interesting feature of the WGS data that can be interpreted from digital

158 karyotyping. We believe that CS may hold clue for assessment of the quality of the sample
159 where low CS may indicate sample- and/or sequencing-driven biases such as sample
160 heterogeneity, inter-tumor contamination, overdispersion, etc.

161 **3.5 Aneuploidy spectrum for calibration of single-sample CNV detection tools**

162 Several RD-based computational tools have been developed for CNV analyses [28-30]. Out
163 of these, single-sample CNV detection tools, which do not rely on matched/control reference,
164 require additional information. For example, FREEC [31] uses the whole-genome ploidy
165 number to define the CN states while ReadDepth [32] uses gain/loss percentage to adjust the
166 underlying Poisson/negative binomial distribution. Similarly, CNVnator [33] assumes that
167 99% of the genome is CNV-free. However, these information change across cell lines
168 (Supplementary Table 1 and 2). AStra provides the whole-genome ploidy number and the
169 percentage of genomic regions per CN state, which can be used as inputs for CNV detection
170 tools.

4. Discussion

Authentication of a cell line is a collective process to verify the cell line's identity and check that they are contamination-free. However, despite the availability of different genetic techniques for authentication, the numerical and structural alterations of chromosomes at multitude of length-scales are difficult to detect using a single technique. For example, verification by karyotyping cannot reveal alterations at the short tandem repeat (STR) and SNP levels and similarly STR/SNP profiling cannot determine chromosomal-level variations. Even the authenticated cell lines in continuous culture may evolve over time because they undergo genetic drift with continuous passaging. Therefore, this dynamic genetic alterations should be tracked to capture their genomic identity and strain differences.

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 control of any ChIP-seq experiment provides the WGS data, albeit at low coverage. Similarly, Hi-C reads can be effectively used for computing RD signal [34-36]. Therefore, the free access of NGS raw reads from hundreds of cancer cell lines has paved the way for easy sharing as well as their integrative analysis. However, before such analysis caution should be employed that the data are of good quality. Our results demonstrated that cell lines have characteristic aneuploidy profiles. Therefore, aneuploidy signature can serve as an ideal prototype of digital karyotyping of cell line. Additionally, aneuploidy pattern can be rapidly and accurately captured from low-depth sequencing datasets. This makes aneuploidy profiling a simple, quick and excellent alternative to SNP or CNV profiling which demand relatively higher coverage NGS data.

In conclusion, we believe that cell lines need to be authenticated by a holistic approach combining different genetic methods to assess various aspects of the quality of cell line. Though everybody acknowledges the importance of reporting cell line authentication data for all cell line-based studies, there is a general lack of enthusiasm in the scientific community in this regard. We have provided an *in silico* solution for cancer cell line authentication through digital karyotyping. We developed AStra as a standalone Python-based software to compute aneuploidy profile from individual sample. We have demonstrated that the digital karyotyping is a rapid and effective way to visually spot the differences among different cancer cell lines as well as different strain/variants of a cell line. Thus, AStra is a go-to tool

that can capture the dynamics of chromosomal alterations that represent timestamp ‘barcode’ of cancer cell lines. We are hopeful that AStra screening will be a routine test for quality assessment of cell line-based NGS data.

Declarations

Acknowledgment

We like to thank Costerwell Khyriem for helpful discussion. We acknowledge Sanyal and Chattopadhyay lab members for their valuable comments.

Funding

This work was supported by Nanyang Technological University's Nanyang Assistant Professorship grant and Singapore Ministry of Education Academic Research Fund Tier 1 grant to AS, as well as, Nanyang Technological University Start-up grant to AC.

Author's contributions

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

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Kleivi, K., et al., *Genome signatures of colon carcinoma cell lines.* Cancer Genet Cytogenet, 2004. **155**(2): p. 119-31.
2. Ben-David, U., et al., *Genetic and transcriptional evolution alters cancer cell line drug response.* Nature, 2018. **560**(7718): p. 325-330.
3. Hynds, R.E., E. Vladimirou, and S.M. Janes, *The secret lives of cancer cell lines*. 2018, The Company of Biologists Ltd.
4. Spans, L., et al., *Variations in the exome of the LNCaP prostate cancer cell line.* Prostate, 2012. **72**(12): p. 1317-27.
5. Thompson, S.L. and D.A. Compton, *Chromosomes and cancer cells.* Chromosome Res, 2011. **19**(3): p. 433-44.
6. Almeida, J.L., K.D. Cole, and A.L. Plant, *Standards for Cell Line Authentication and Beyond.* PLoS Biol, 2016. **14**(6): p. e1002476.
7. Park, S.T. and J. Kim, *Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing.* Int Neurourol J, 2016. **20**(Suppl 2): p. S76-83.
8. Otto, R., C. Sers, and U. Leser, *Robust in-silico identification of cancer cell lines based on next generation sequencing.* Oncotarget, 2017. **8**(21): p. 34310-34320.
9. Petljak, M., et al., *Characterizing mutational signatures in human Cancer cell lines reveals episodic APOBEC mutagenesis.* Cell, 2019. **176**(6): p. 1282-1294. e20.
10. Duan, J., et al., *Comparative studies of copy number variation detection methods for next-generation sequencing technologies.* PloS one, 2013. **8**(3): p. e59128.
11. Ghandi, M., et al., *Next-generation characterization of the Cancer Cell Line Encyclopedia.* Nature, 2019. **569**(7757): p. 503.
12. Killick, R., P. Fearnhead, and I.A. Eckley, *Optimal detection of changepoints with a linear computational cost.* Journal of the American Statistical Association, 2012. **107**(500): p. 1590-1598.
13. Weaver, B.A. and D.W. Cleveland, *The aneuploidy paradox in cell growth and tumorigenesis.* Cancer Cell, 2008. **14**(6): p. 431-3.
14. Weaver, B.A., et al., *Aneuploidy acts both oncogenically and as a tumor suppressor.* Cancer Cell, 2007. **11**(1): p. 25-36.
15. Sotillo, R., et al., *Mad2 overexpression promotes aneuploidy and tumorigenesis in mice.* Cancer Cell, 2007. **11**(1): p. 9-23.
16. Williams, B.R., et al., *Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells.* Science, 2008. **322**(5902): p. 703-9.
17. Weaver, B.A. and D.W. Cleveland, *Does aneuploidy cause cancer?* Curr Opin Cell Biol, 2006. **18**(6): p. 658-67.
18. Albertson, D.G., et al., *Chromosome aberrations in solid tumors.* Nat Genet, 2003. **34**(4): p. 369-76.
19. Vargas-Rondon, N., V.E. Villegas, and M. Rondon-Lagos, *The Role of Chromosomal Instability in Cancer and Therapeutic Responses.* Cancers (Basel), 2017. **10**(1).
20. Thompson, S.L. and D.A. Compton, *Examining the link between chromosomal instability and aneuploidy in human cells.* J Cell Biol, 2008. **180**(4): p. 665-72.
21. Gordon, D.J., B. Resio, and D. Pellman, *Causes and consequences of aneuploidy in cancer.* Nat Rev Genet, 2012. **13**(3): p. 189-203.
22. Li, R., et al., *Chromosomal alterations cause the high rates and wide ranges of drug resistance in cancer cells.* Cancer genetics and cytogenetics, 2005. **163**(1): p. 44-56.
23. Rondon-Lagos, M., et al., *Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis.* Mol Cytogenet, 2014. **7**(1): p. 8.

24. Samir Khalil, A.I., et al., *Hierarchical Discovery of Large-scale and Focal Copy Number Alterations in Low-coverage Cancer Genomes.* bioRxiv, 2019: p. 639294.
25. Carter, S.L., et al., *Absolute quantification of somatic DNA alterations in human cancer.* Nat Biotechnol, 2012. **30**(5): p. 413-21.
26. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
27. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.
28. Alkodsi, A., R. Louhimo, and S. Hautaniemi, *Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data.* Brief Bioinform, 2015. **16**(2): p. 242-54.
29. Duan, J., et al., *Comparative studies of copy number variation detection methods for next-generation sequencing technologies.* PLoS One, 2013. **8**(3): p. e59128.
30. Zhao, M., et al., *Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives.* BMC Bioinformatics, 2013. **14 Suppl 11**: p. S1.
31. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. **28**(3): p. 423-5.
32. Miller, C.A., et al., *ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads.* PLoS One, 2011. **6**(1): p. e16327.
33. Abyzov, A., et al., *CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.* Genome Res, 2011. **21**(6): p. 974-84.
34. Vidal, E., et al., *OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes.* Nucleic Acids Res, 2018. **46**(8): p. e49.
35. Servant, N., et al., *Effective normalization for copy number variation in Hi-C data.* BMC Bioinformatics, 2018. **19**(1): p. 313.
36. Chakraborty, A. and F. Ay, *Identification of copy number variations and translocations in cancer cells from Hi-C data.* Bioinformatics, 2017.

Figures

Fig. 1

AStra framework. (a) RD frequency distribution, extracted from WGS reads, is scanned against six prospective models (m1 to m6) to identify the initial CN reference candidates ($CN_{1i\text{-}6i}$). **(b)** RD segments are utilized for fine-tuning the initial CN reference candidates by searching their narrow intervals (1.9CN to 2.1CN). For each interval, the CN reference candidate ($CN_{1\text{-}6}$) are the RD value that best allocates the RD segments around integer CN states. Final CN reference is selected with minimum CE.

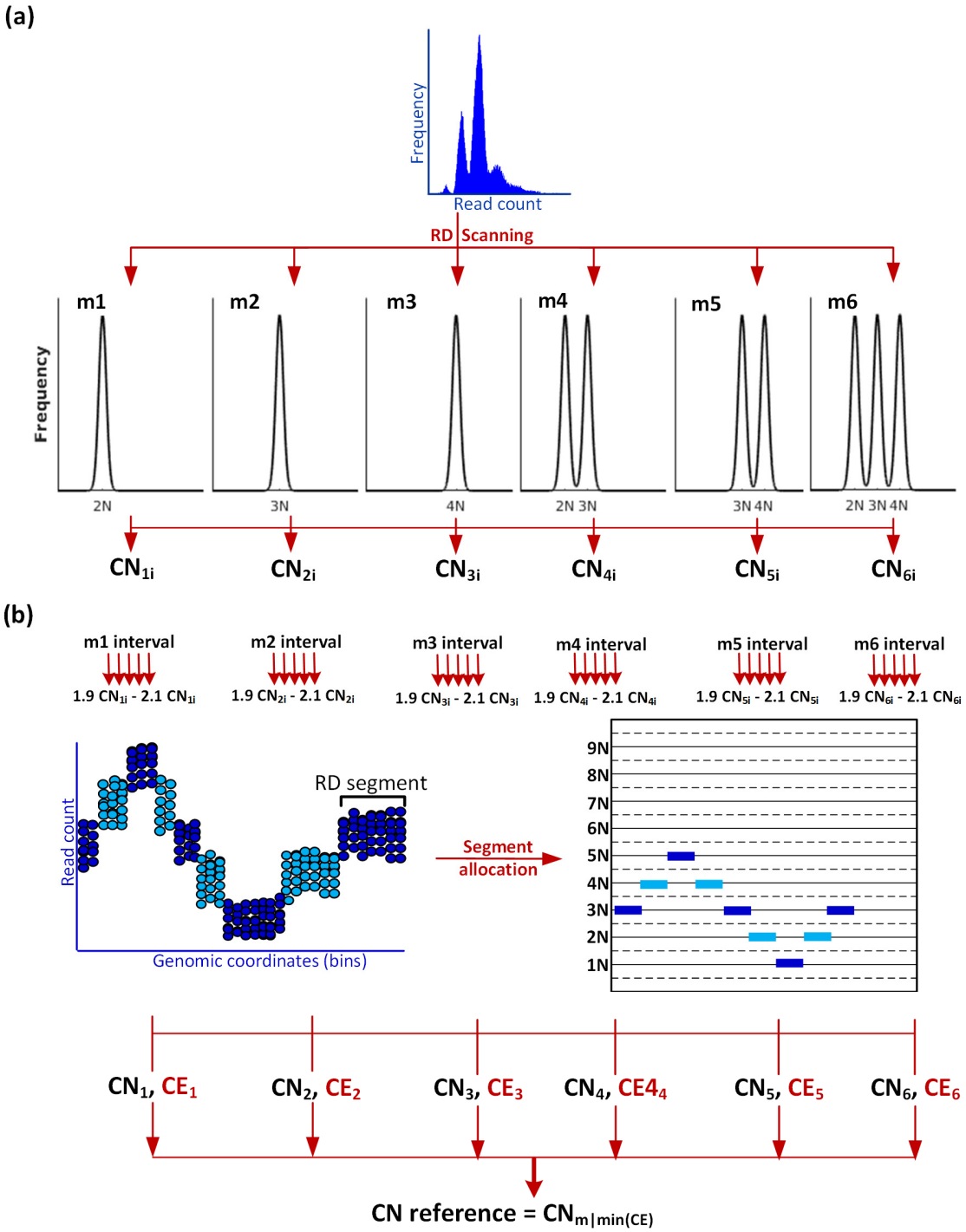

(a)
Frequency
Read count
RD Scanning
m1
m2
m3
m4
m5
m6
Frequency
2N
3N
4N
2N 3N
3N 4N
2N 3N 4N
CN1i
CN2i
CN3i
CN4i
CN5i
CN6i
(b)
m1 interval
m2 interval
m3 interval
m4 interval
m5 interval
m6 interval
1.9 CN1i - 2.1 CN1i
1.9 CN2i - 2.1 CN2i
1.9 CN3i - 2.1 CN3i
1.9 CN4i - 2.1 CN4i
1.9 CN5i - 2.1 CN5i
1.9 CN6i - 2.1 CN6i
RD segment
Read count
Genomic coordinates (bins)
Segment allocation
9N
8N
7N
6N
5N
4N
3N
2N
1N
CN1, CE1
CN2, CE2
CN3, CE3
CN4, CE44
CN5, CE5
CN6, CE6
CN reference = CNm|min(CE)

Fig. 2

Aneuploidy signature revealed the genetic variations among the MCF7 strains. The genome-wide aneuploidy profile (left) and aneuploidy spectrum (right) for strain C, G, H, P and S are shown. Genomic loci are colored based on their copy numbers (CN ≤ 2: black, 3 ≤ CN ≤ 4: blue, CN > 4: 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.

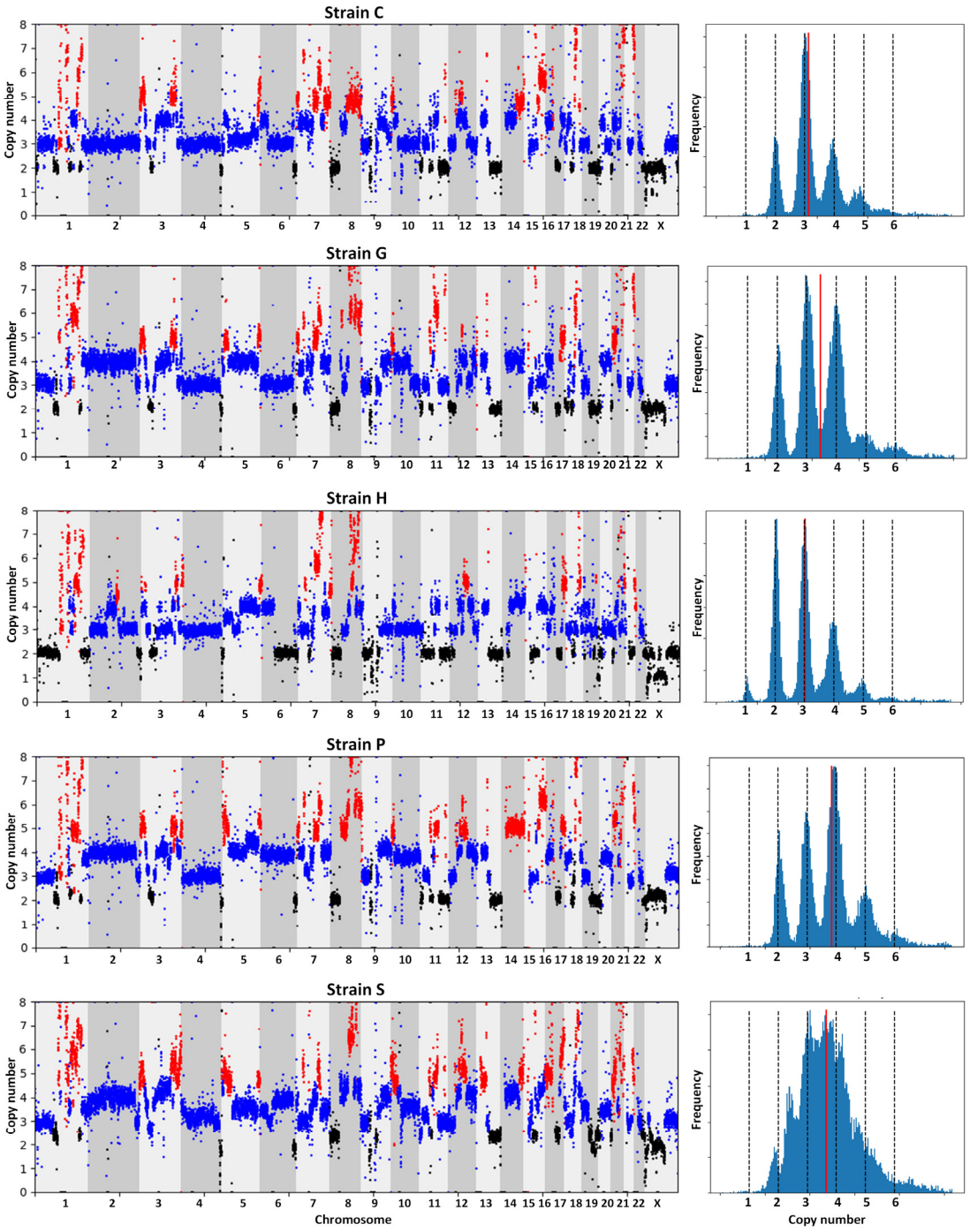
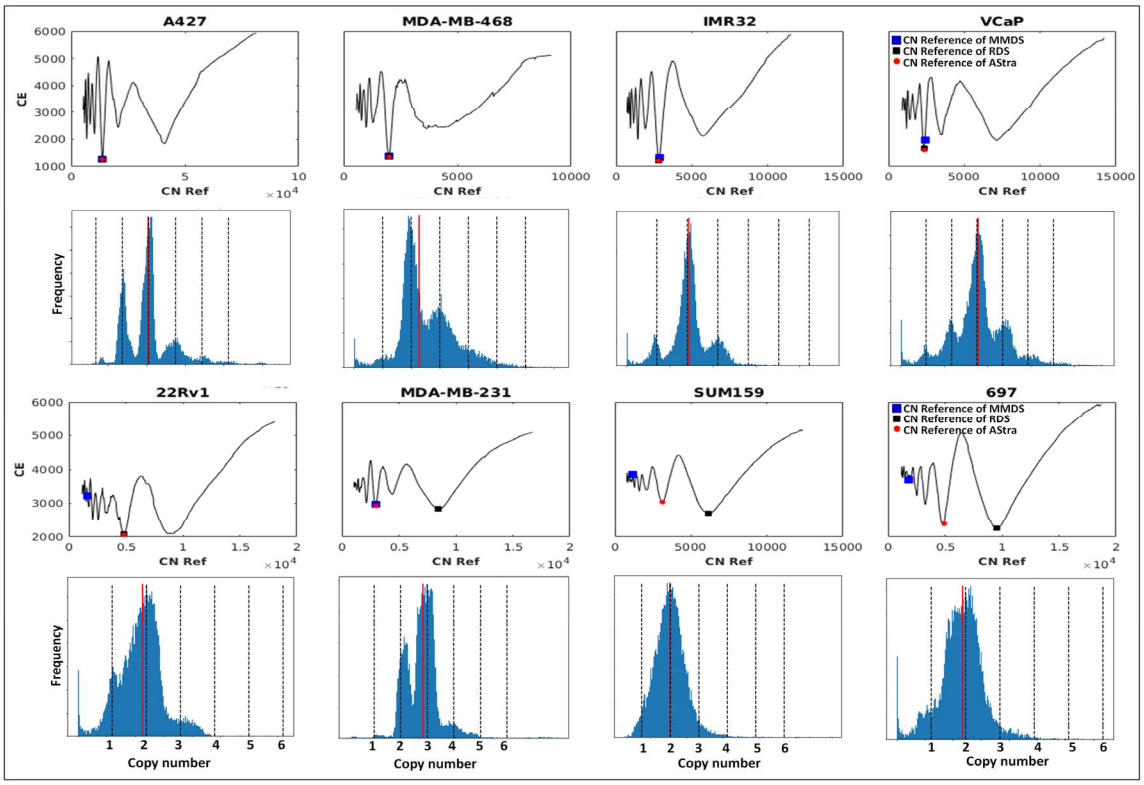
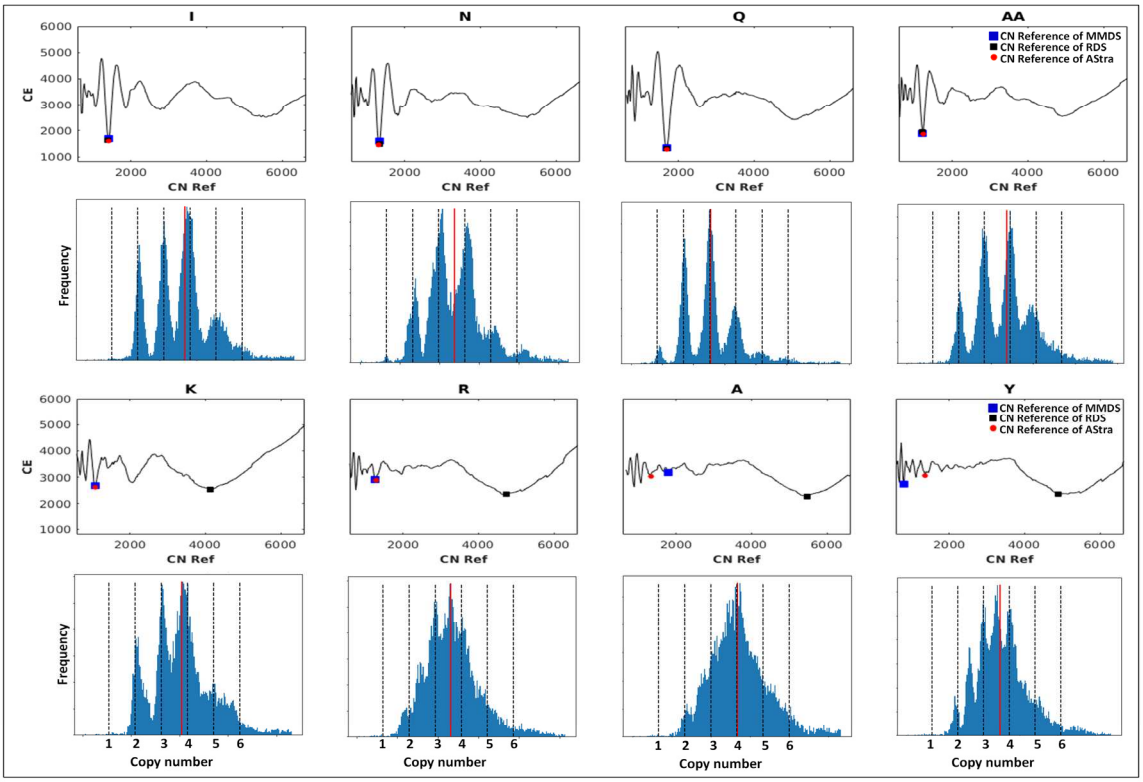

Strain C
Strain G
Strain H
Strain P
Strain S
Copy number
Frequency
Chromosome
Copy number
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

Fig. 3

Centralization error as a function of copy number reference in cancer cells lines (a) and MCF7 strains (b). (Top) The centralization error 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) The corresponding aneuploidy spectrum of cancer cell lines/MCF7 strains are shown where the dotted black lines denote the CN states whereas the red line denotes the median RD signal.



(a) Cancer cell lines



(b) MCF7 strains

Supplementary Information:

Digital Karyotyping for Rapid Authentication of Cell Lines

Ahmed Ibrahim Samir Khalil[1], Anupam Chattopadhyay[1,*], Amartya Sanyal[2,*]

[1]School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798.
[2]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551.

[*]Corresponding e-mail: anupam@ntu.edu.sg; asanyal@ntu.edu.sg

EXTENDED METHOD

1. Generation of simulated aneuploidy profile

For the simulated data, we used in-house-developed method [1] to artificially introduce large-scale copy number gain/loss regions by manipulating the WGS reads of HG00119 (1000 Genomes Project sample of diploid male). In this study, we simulated M random number of large copy number variations (LCVs) only and excluded the focal alterations. The pipeline of simulated aneuploidy profile generation contains two steps – 1) random selection of candidate location, and 2) artificial read spike-in. We divide the chromosome into M contiguous large segments randomly and then we proceed for spike-in of artificial reads. In order to modify a selected region (R) with original copy number C1 to a new copy number C2, we add or remove $\left[X * \left(\frac{C2}{C1} - 1\right)\right]$ reads, where X is the initial number of reads of that region. For amplified region, artificial reads were then spiked into the R region by randomly shifting the coordinates of the original reads by 10-500 bp. The original reads and the spiked-in artificial reads were merged into a new BAM file and used as input for CNV evaluation.

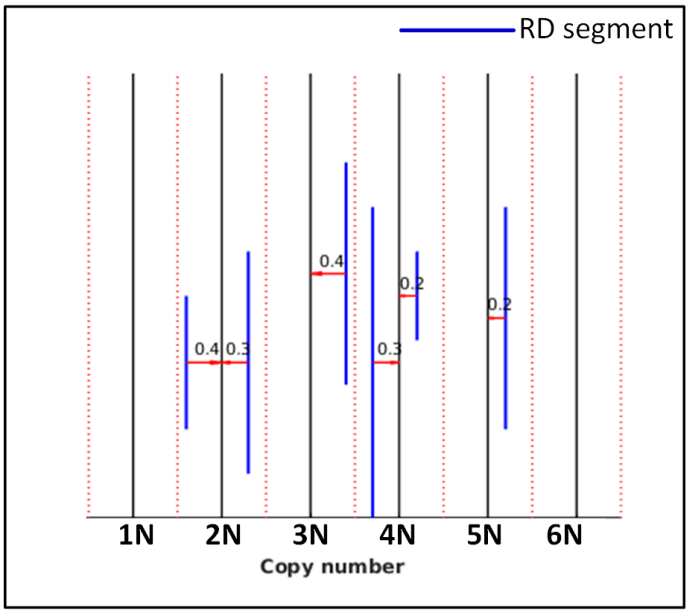
References

1. Samir Khalil, A.I., et al., *Hierarchical Discovery of Large-scale and Focal Copy Number Alterations in Low-coverage Cancer Genomes.* bioRxiv, 2019: p. 639294.

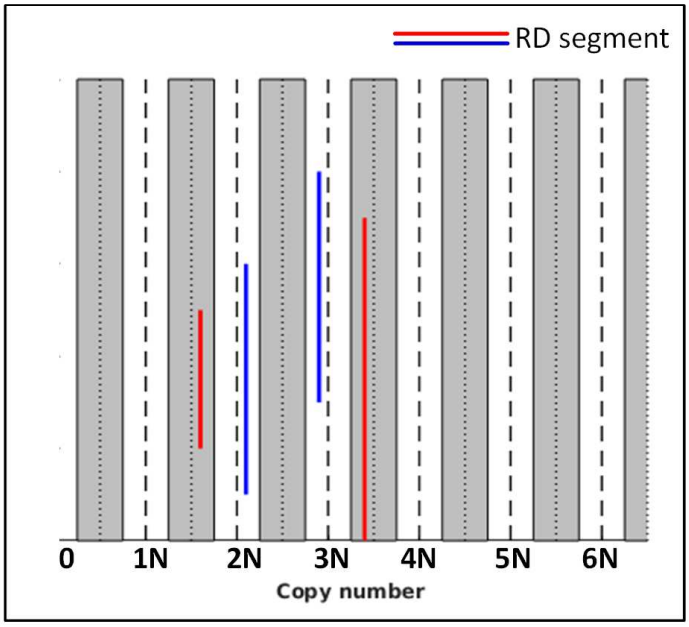
SUPPLEMENTARY FIGURES

Figure S1

Visual illustration to compute centralization error (CE) and centralization score (CS). For CE (a), the blue lines represent RD segments and the dotted red lines show the boundary of CN states (0.5N and 1.5N for CN = 1, and so on). The red arrow is the error weight of each segment (the difference between the CN of that segment and the nearest integer CN state). For CS (b), the blue lines represent RD segments with estimated CN within 0.25N from the nearest integer CN state whereas the red lines represent RD segments with estimated CN >0.25N from the nearest CN state.



(a) Centralization error



(b) Centralization score

Figure S2

Aneuploidy signatures of MCF7 strains. The genome-wide aneuploidy profile for all MCF7 strains (except C, G, H, P and S) are shown. Genomic loci are colored based on their copy numbers ($CN \leq 2$: black, $3 \leq CN \leq 4$: blue, $CN > 4$: red).

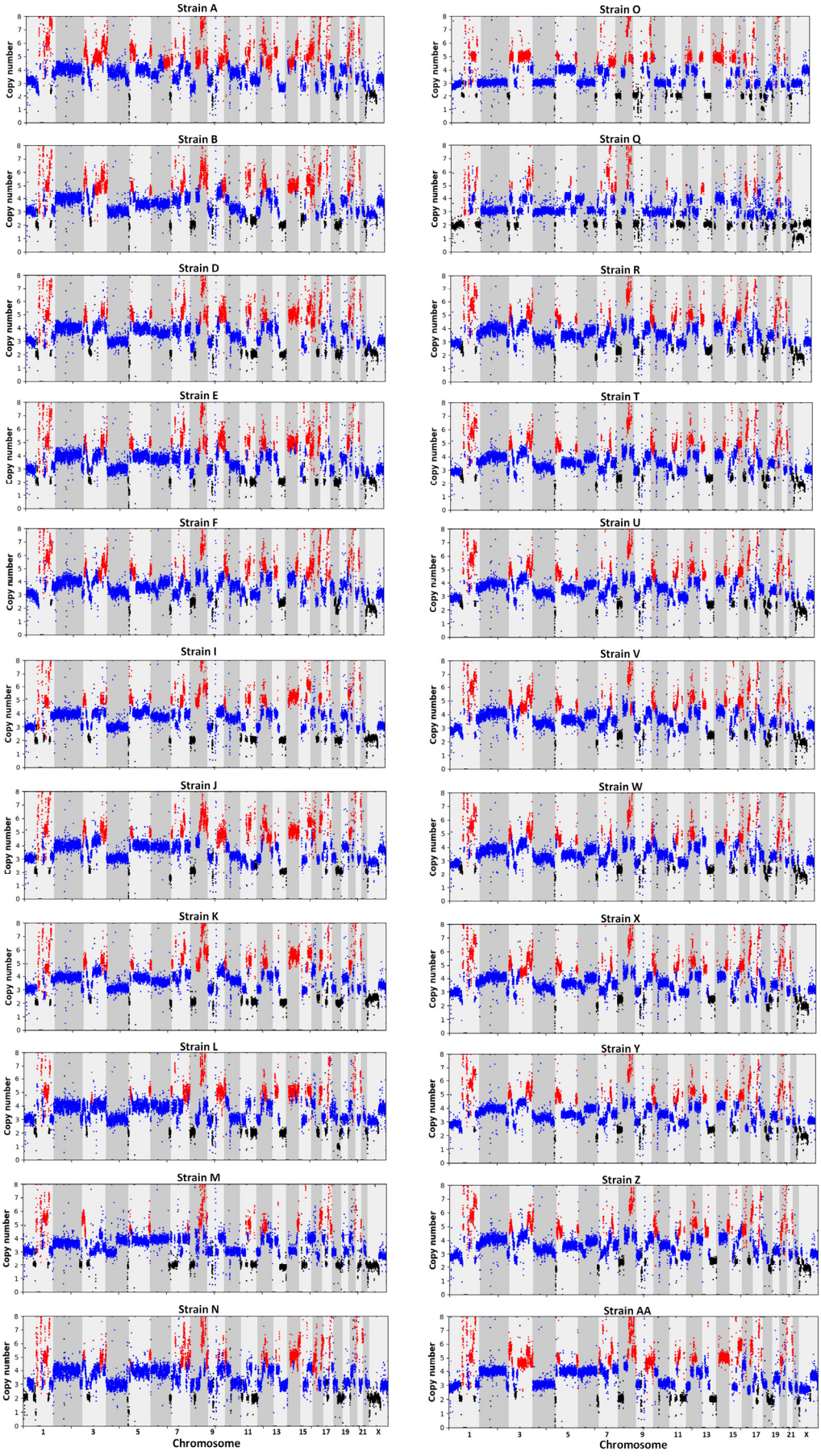

Strain A
Strain O
Strain B
Strain Q
Strain D
Strain R
Strain E
Strain T
Strain F
Strain U
Strain I
Strain V
Strain J
Strain W
Strain K
Strain X
Strain L
Strain Y
Strain M
Strain Z
Strain N
Strain AA
Copy number
0 1 2 3 4 5 6 7 8
Chromosome
1 3 5 7 9 11 13 15 17 19 21 X

Figure S3

Aneuploidy signatures of neo-genomes generated using simulated data. The genome-wide aneuploidy profile (left) and aneuploidy spectrum (right) for original HG00119 (diploid male) sample and the simulated neo-genomes A, B, G and U are shown. Genomic loci are colored based on their copy numbers (CN $\leq$ 1: black, CN = 2: blue, CN $\geq$ 3: 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 copy numbers of chr 1 (CN =2) and chr X (CN=1), which were excluded from simulation, are estimated correctly from all neo-genomes.

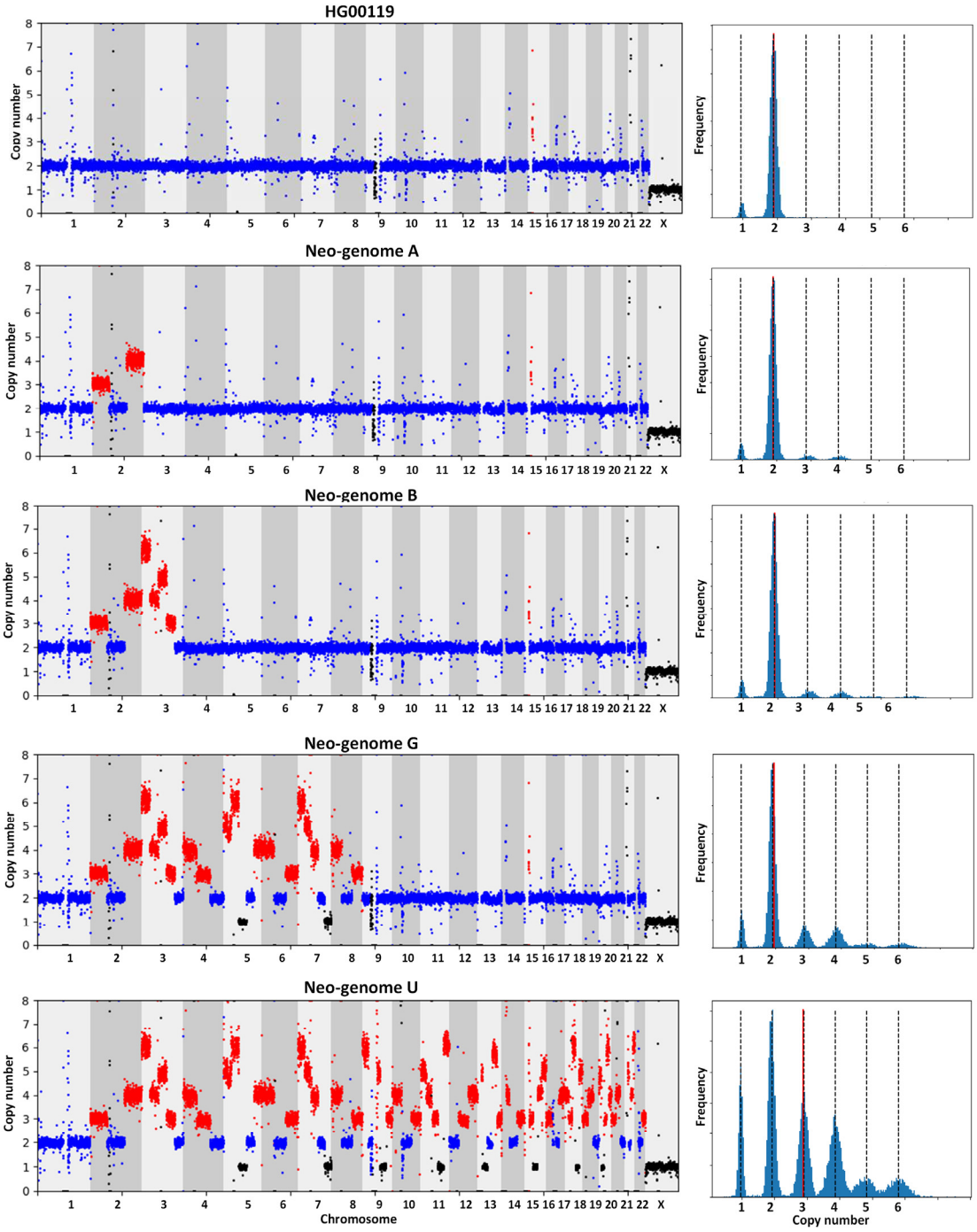

HG00119
Copy number
Frequency
Neo-genome A
Copy number
Frequency
Neo-genome B
Copy number
Frequency
Neo-genome G
Copy number
Frequency
Neo-genome U
Copy number
Frequency
Chromosome
Copy number

Figure S4

Aneuploidy signatures of cancer cell lines. The genome-wide aneuploidy profile of a variety of cancer cell lines are shown. Genomic loci are colored based on their copy numbers (CN ≤ 1: black, CN = 2: blue, CN ≥ 3: red).

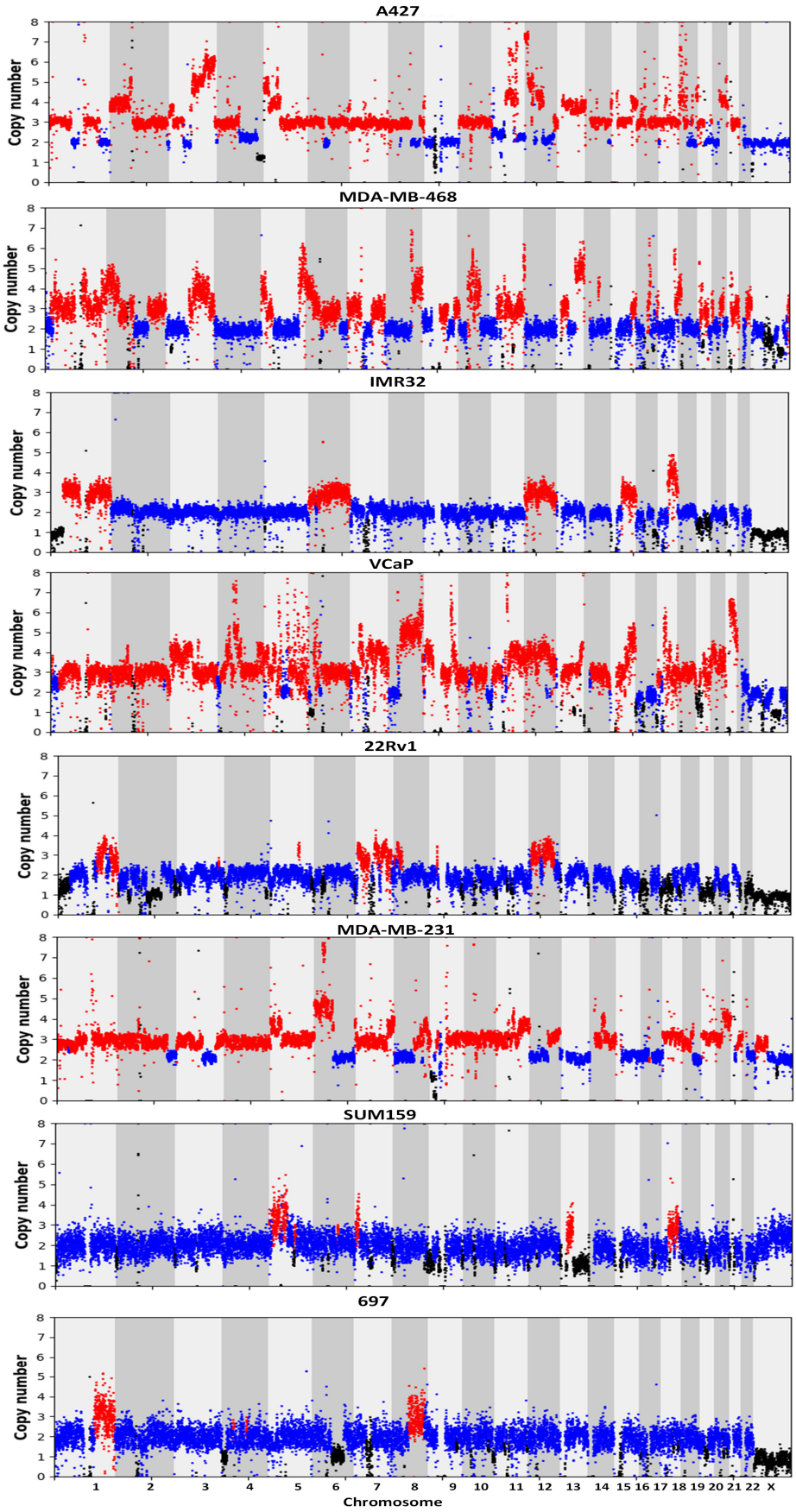

A427
MDA-MB-468
IMR32
VCaP
22Rv1
MDA-MB-231
SUM159
697
Copy number
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

Figure S5

Flowchart of RDS and MMDS methods. (a) In RDS method, RD segments are utilized for computing the CN reference by searching the entire range of RD values as candidate CN reference. The CN reference is selected with lowest CE. (b) In MMDS method, CN reference is computed as the RD value that achievs the maximum overlap between the input RD frequency distribution and the multimodal distribution.

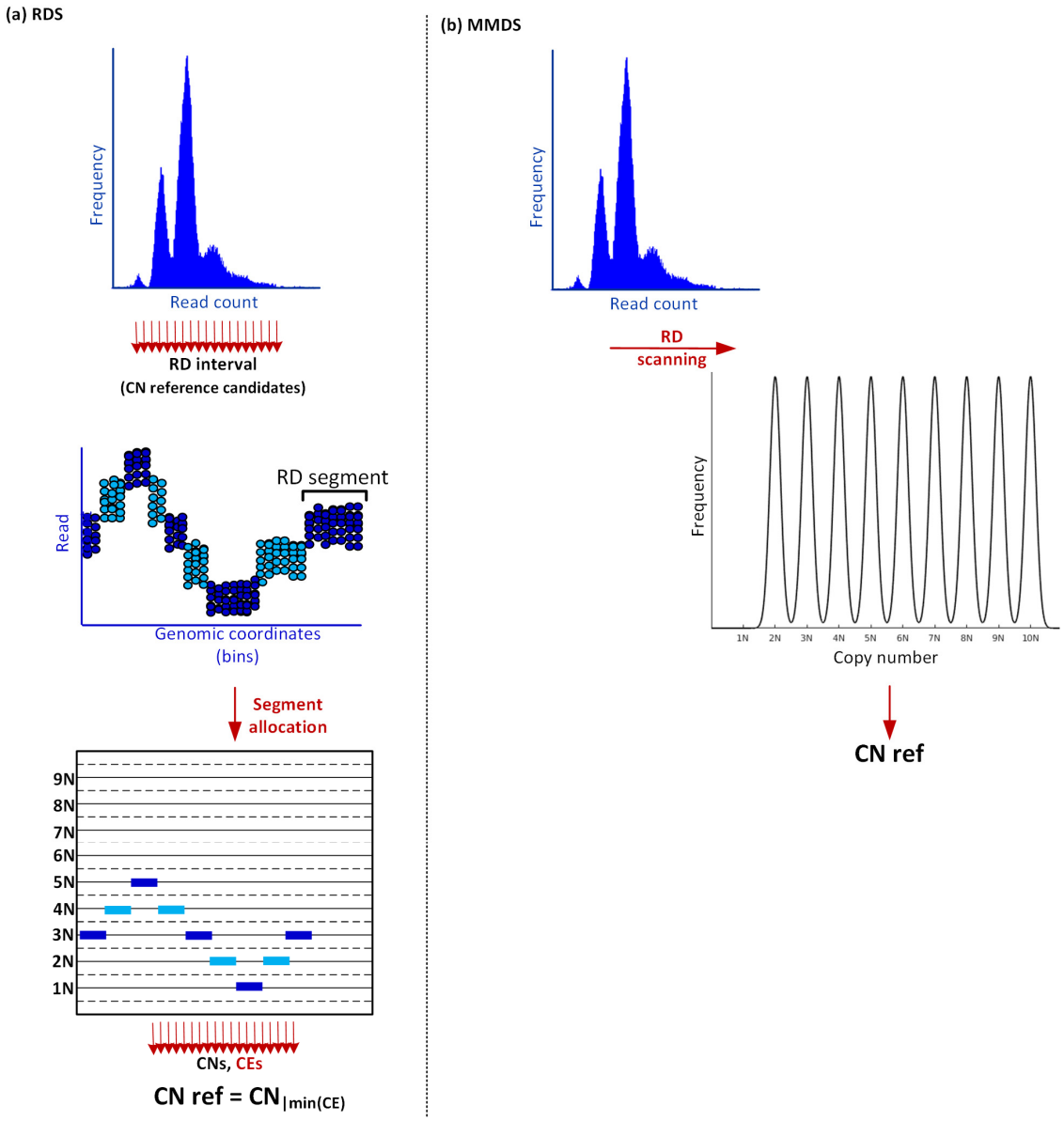

(a) RDS
Frequency
Read count
RD interval
(CN reference candidates)
Read
RD segment
Genomic coordinates
(bins)
Segment
allocation
9N
8N
7N
6N
5N
4N
3N
2N
1N
CNs, CEs
CN ref = CN|min(CE)
(b) MMDS
Frequency
Read count
RD
scanning
Frequency
1N
2N
3N
4N
5N
6N
7N
8N
9N
10N
Copy number
CN ref

Figure S6

Centralization error as a function of copy number reference in simulated neo-genomes (A to U). The centralization error (CE) 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.

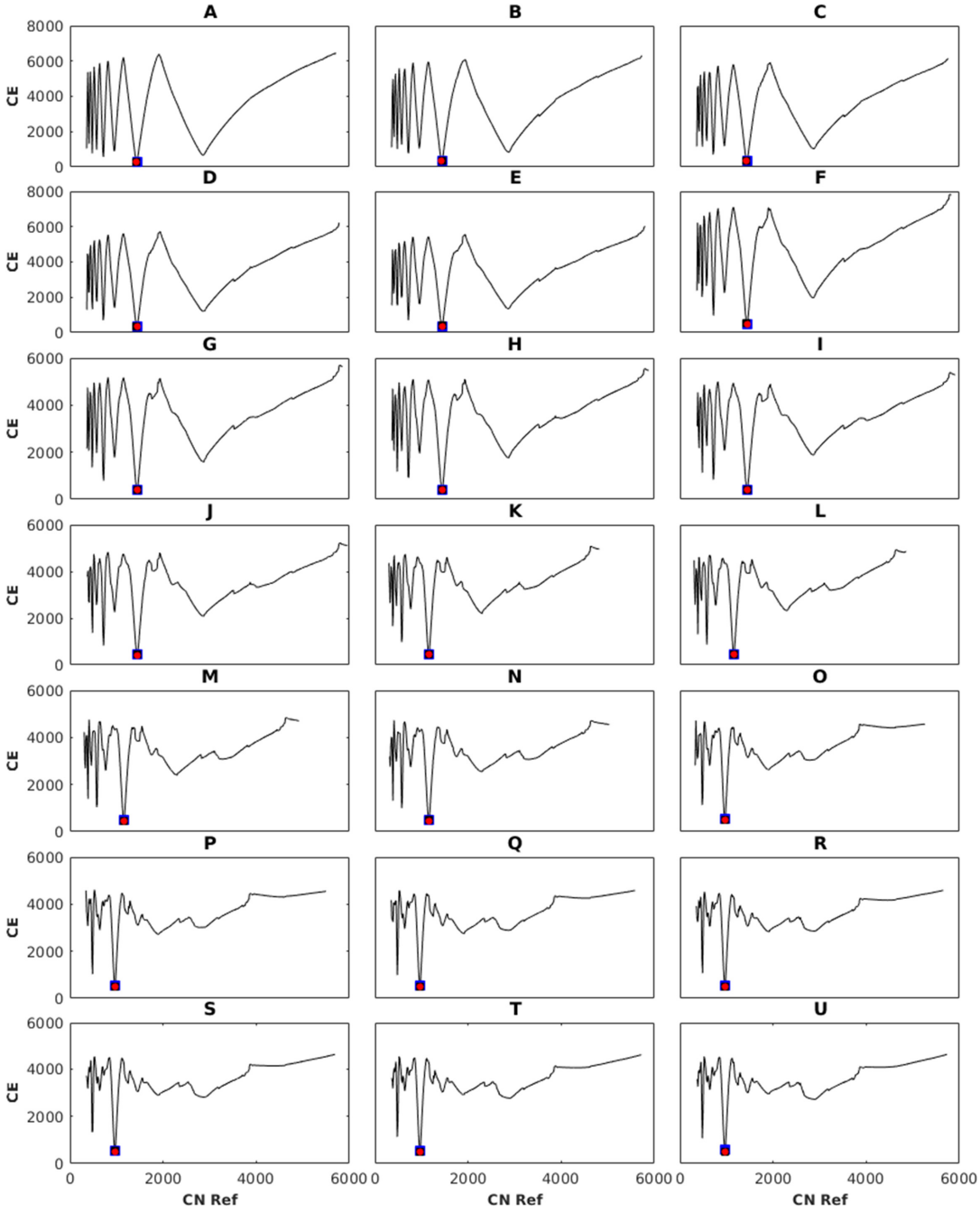
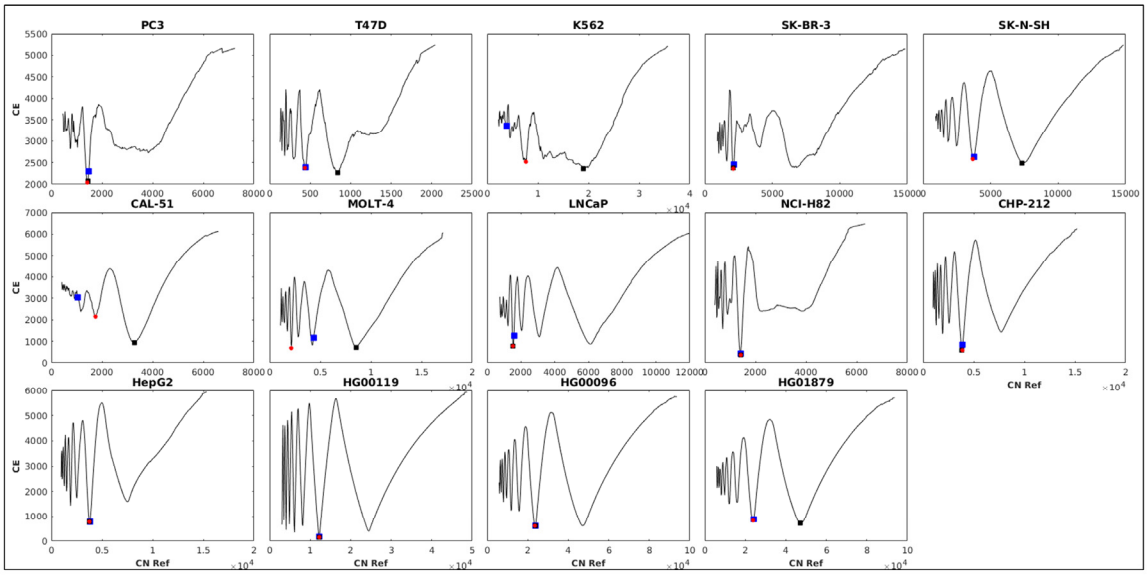
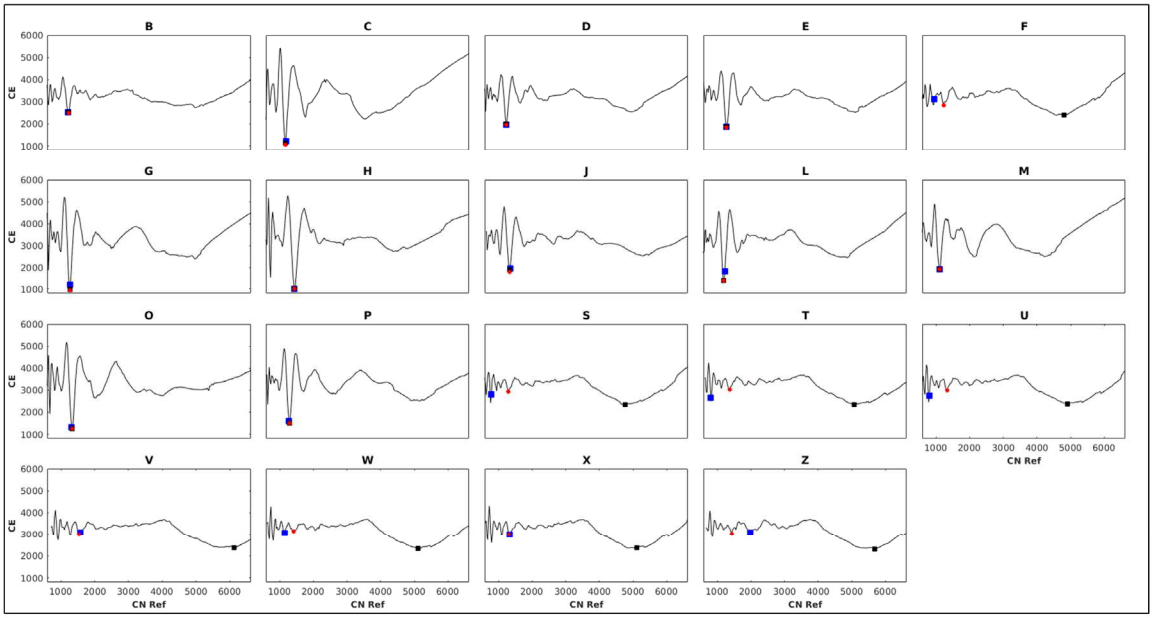

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
CE
CN Ref
0
2000
4000
6000
8000

Figure S7

Centralization error as a function of copy number reference in additional cancer cell lines/1000 Genomes Project samples (a) and MCF7 strains (b) not shown in the Fig. 3. The centralization error (CE) 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.



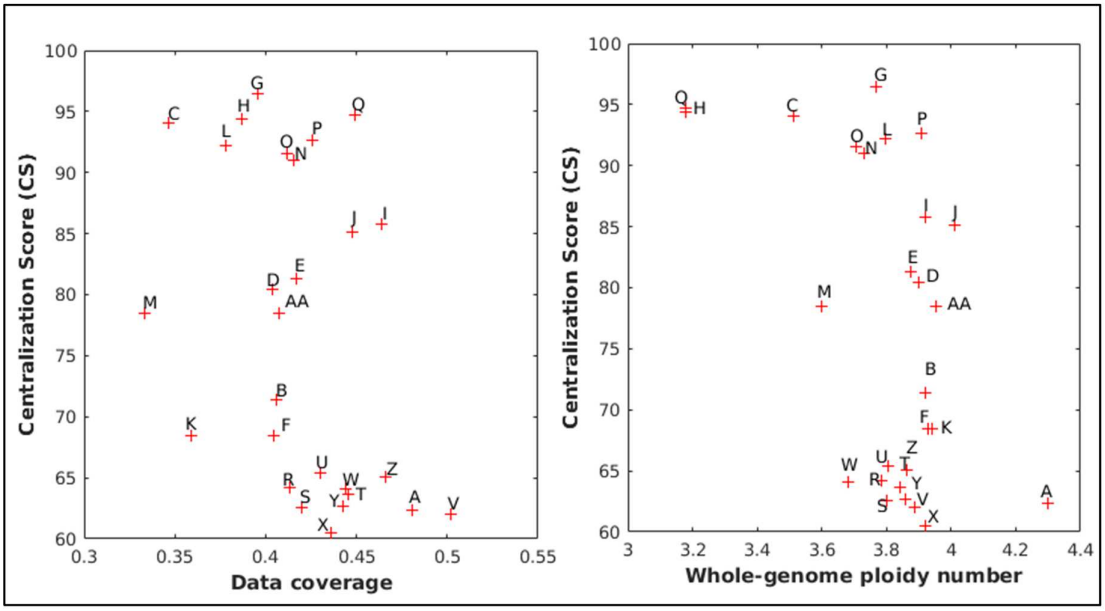
(a) Cancer cell lines/1000 Genomes Project samples



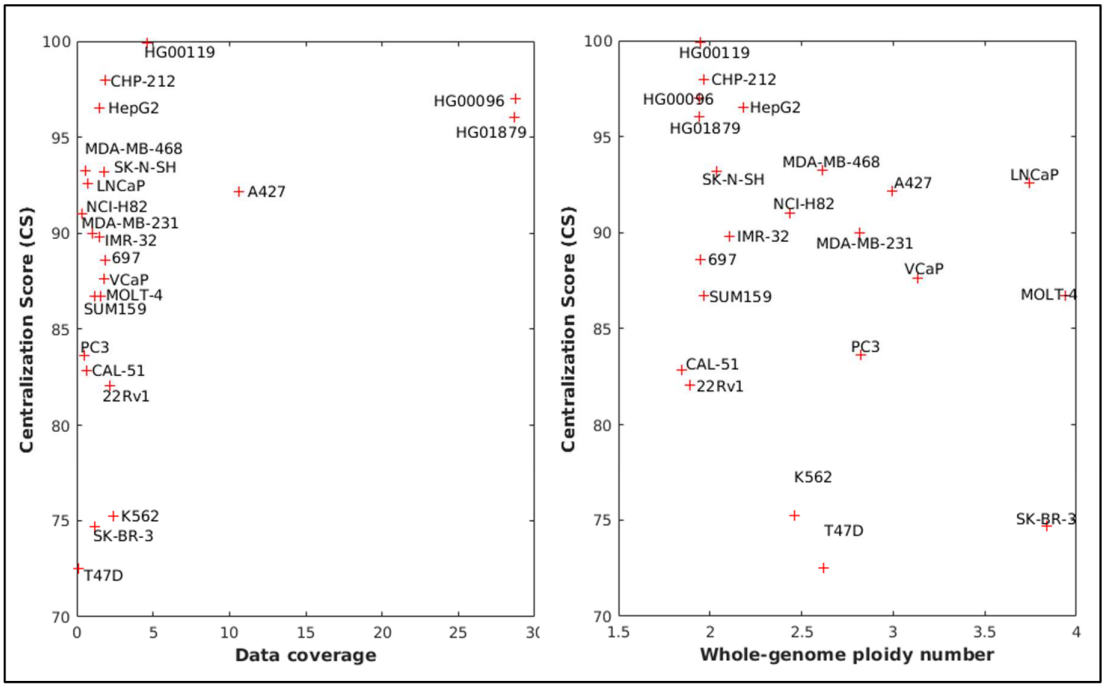
(b) MCF7 strains

Figure S8

Centralization score (CS) varies among different cell lines and different MCF7 strains. Dot plot showing the relationship of the CS with respect to the data coverage (left) and the whole-genome ploidy number (right) for MCF7 strains (a) and cancer cell lines/1000 Genomes Project samples (b).



(a) MCF7 strains



(b) Cancer cell lines/1000 Genomes Project samples

Figures

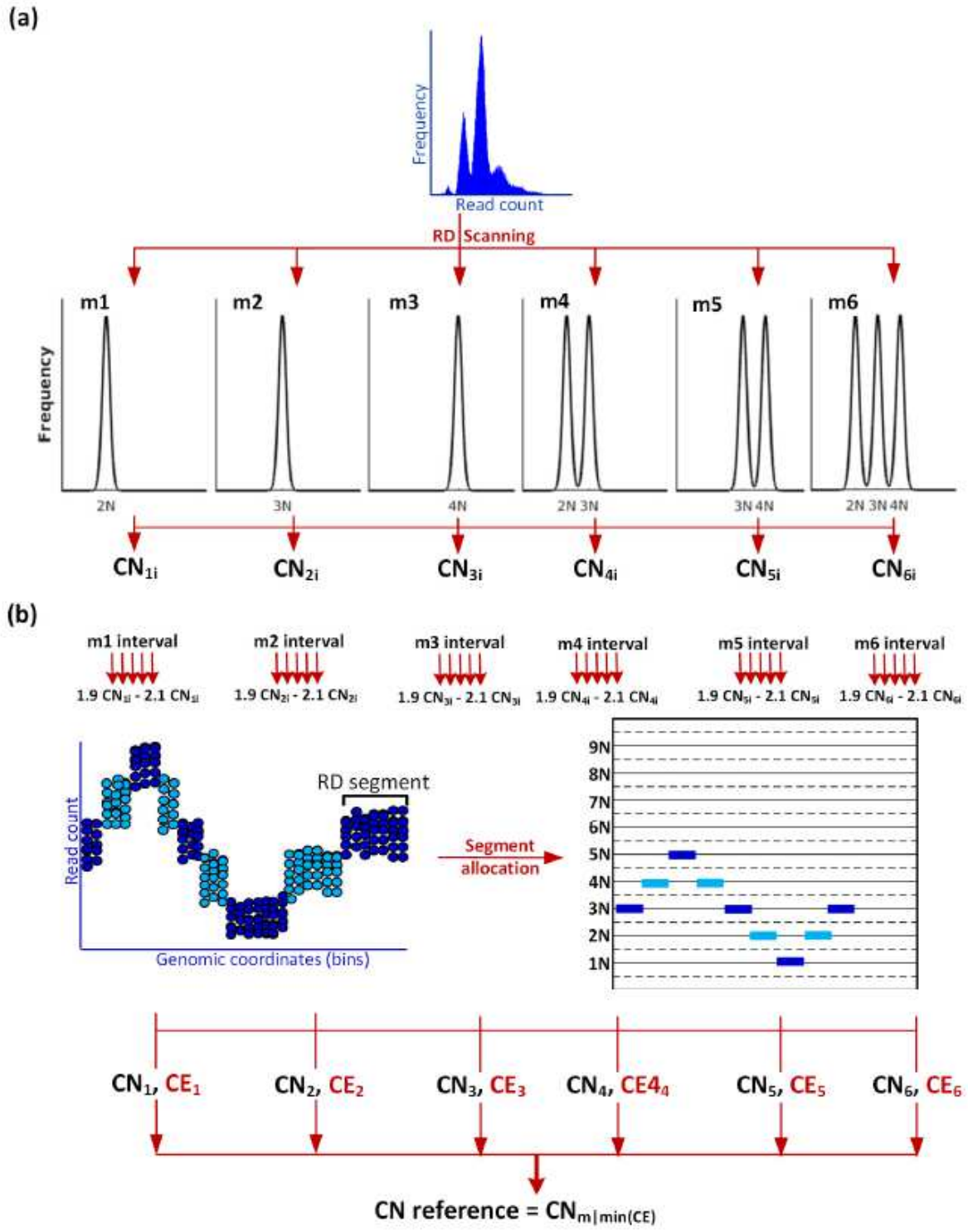


Figure 1

AStra framework. (a) RD frequency distribution, extracted from WGS reads, is scanned against six prospective models (m1 to m6) to identify the initial CN reference candidates (CN1i-6i). (b) RD segments are utilized for fine-tuning the initial CN reference candidates by searching their narrow intervals (1.9CN to

2.1CN). For each interval, the CN reference candidate (CN1-6) are the RD value that best allocates the RD segments around integer CN states. Final CN reference is selected with minimum CE.

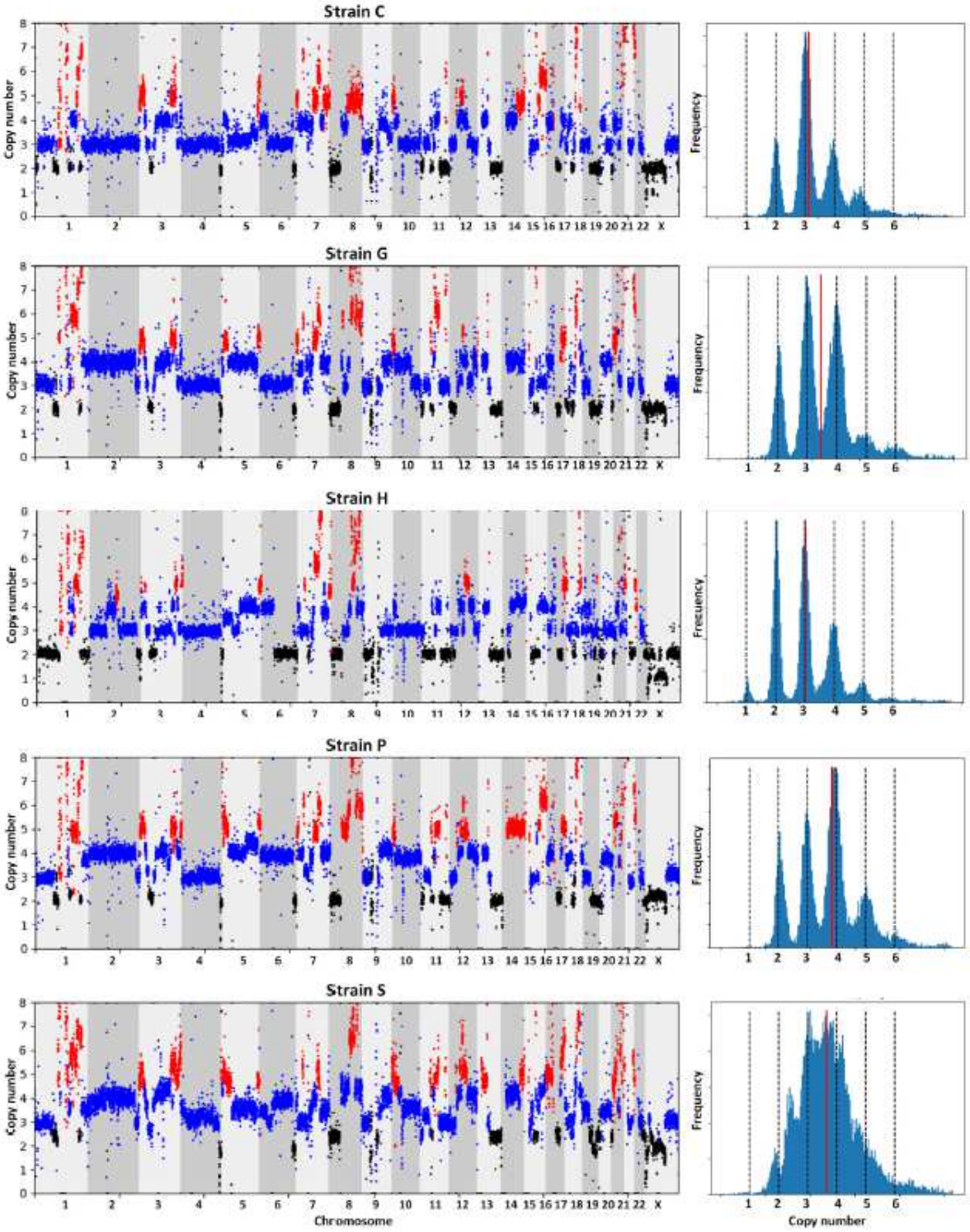


Figure 2

Aneuploidy signature revealed the genetic variations among the MCF7 strains. The genome-wide aneuploidy profile (left) and aneuploidy spectrum (right) for strain C, G, H, P and S are shown. Genomic loci are colored based on their copy numbers (CN ≤ 2: black, 3 ≤ CN ≤ 4: blue, CN > 4: 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.

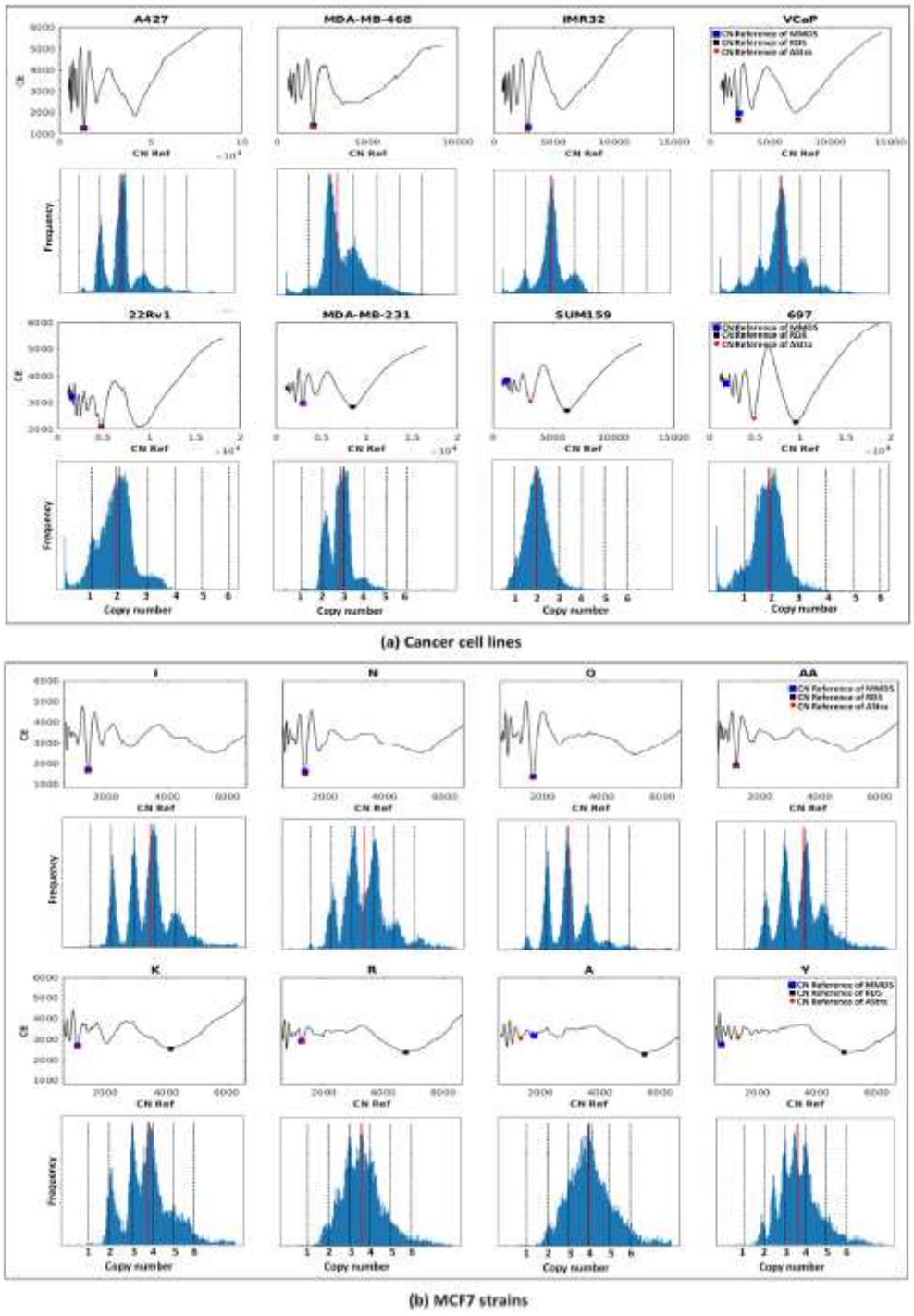


Figure 3

Centralization error as a function of copy number reference in cancer cells lines (a) and MCF7 strains (b). (Top) The centralization error 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) The corresponding aneuploidy spectrum of cancer cell lines/MCF7 strains are shown where the dotted black lines denote the CN states whereas the red line denotes the median RD signal.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1MCF7AneuploidySpectrum.xlsx
- S2DatasetAneuploidySpectrum.xlsx