Digital karyotyping of cancer cell lines from NGS data

Ahmed Ibrahim Samir Khalil

2020

Ahmed Ibrahim Samir Khalil. (2020). Digital karyotyping of cancer cell lines from NGS data.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/138176>

<https://doi.org/10.32657/10356/138176>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0
International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 22:08:39 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

DIGITAL KARYOTYPING OF CANCER CELL LINES FROM NGS DATA

AHMED IBRAHIM SAMIR KHALIL

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

2020

DIGITAL KARYOTYPING OF CANCER CELL LINES FROM NGS DATA

AHMED IBRAHIM SAMIR KHALIL

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy.

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarized materials, and has not been submitted for a higher degree to any other University or Institution.

Ahmed Ibrahim Samir Khalil

School of Computer Science and Engineering,
Nanyang Technological University, Singapore.

Date: 27 April, 2020

Supervision Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Dr.-Ing. Anupam Chattopadhyay
School of Computer Science and Engineering,
Nanyang Technological University, Singapore.
Date: 27 April, 2020

Authorship Attribution Statement

This thesis contains material from 1 paper published in peer-reviewed journal and 2 preprints where I am the first author.

The content of Chapter 3 is available as preprint — *Digital Karyotyping for Rapid Authentication of Cell Lines*. Ahmed Ibrahim Samir Khalil, Anupam Chattopadhyay, Amartya Sanyal.

<https://www.researchsquare.com/article/8fb6098b-3771-48e1-a06a-98bff5ed1c6b/v1>

- I designed the AStra algorithm, followed by developing the Python-based AStra software (<https://github.com/AISKhalil/AStra>), and performed all the analysis for benchmarking it.
- I and A. Sanyal wrote the manuscript, and A. Chattopadhyay edited it.
- A. Sanyal and A. Chattopadhyay guided the work.

The content of Chapter 4 has been published as — Khalil AIS, Khyriem C, Chattopadhyay A, Sanyal A. “Hierarchical discovery of large-scale and focal copy number alterations in low-coverage cancer genomes.” *BMC Bioinformatics*. 2020;21(1):147. Published 2020 Apr 16. doi:10.1186/s12859-020-3480-3. PMID: 32299346.

- I designed the CNATra algorithm, followed by developing the MATLAB-based CNATra software (<https://github.com/AISKhalil/CNATra>), and performed all the analysis for benchmarking it.
- C. Khyriem installed and applied the available CNA detection tools for performance evaluation.
- I and C. Khyriem developed and implemented the algorithm for generating the realistic simulated data.
- I and A. Sanyal wrote the manuscript; A. Chattopadhyay and C. Khyriem revised and edited it.

- A. Sanyal and A. Chattopadhyay guided the work.

The content of Chapter 5 is available as preprint — *Identification and Utilization of Copy Number Information for Correcting Hi-C Contact Map of Cancer Cell Line*. Ahmed Ibrahim Samir Khalil, Siti Rawaidah Muzaki, Anupam Chattopadhyay, Amartya Sanyal. I presented it in "Multiscale Modeling of Chromatin: Bridging Experiment with Theory, Les Houches 2019".

<https://www.biorxiv.org/content/10.1101/798710v1>

- I designed the HiCNAttra algorithm, followed by developing the MATLAB-based HiCNAttra software (<https://github.com/AISKhalil/HiCNAttra>), and performed all the analysis for benchmarking it.
- Siti Rawaidah Muzaki did the 3C-seq experiment for generating the 3C-seq data.
- I and A. Sanyal wrote the manuscript, and A. Chattopadhyay edited it.
- A. Sanyal and A. Chattopadhyay guided the work.

Ahmed Ibrahim Samir Khalil
School of Computer Science and Engineering,
Nanyang Technological University, Singapore.
Date: 27 April, 2020

Acknowledgments

In the name of Allah, the most merciful the most gracious; all thanks to Allah the Lord of the Heavens and Earth and peace be upon Mohamed and his companions.

I would like to express my indebtedness and gratefulness to my supervisors *Asst. Prof. Amartya Sanyal* and *Assoc. Prof. Anupam Chattopadhyay*. It has been a pleasure to work with them and learn from such extraordinary advisors. They have always made themselves available for help and advice with their boundless enthusiasm and positive thinking. Special thanks go to my colleagues at Nanyang Technological University; especially *Costerwell Khyriem* and *Yao Chen* for their continuous support and collaboration.

Lastly, I want to thank my family and my fiancée for their invaluable support during my arduous PhD journey; After Allah, without their support I could not have done it.

Thanks go to you all.

Table of Contents

Chapter 1	Introduction.....	29
1.1	Biological experiments	33
1.1.1	Conventional chromosome conformation capture (3C).....	33
1.1.2	Hi-C	35
1.1.3	3C-seq	35
1.2	Visualization of genomic alterations	35
1.3	Problem statement.....	41
1.4	Contributions.....	42
1.5	Thesis roadmap	44
Chapter 2	Literature Review.....	45
2.1	CNA detection through NGS data	46
2.1.1	Single-sample CNA detection tools.....	48
2.2	Ploidy estimation through NGS data	57
2.2.1	SNP-based methods	58
2.2.2	RD-based methods	59
2.3	Contact-map correction and CNA detection from Hi-C/3C-seq data.....	60
2.3.1	Contact map normalization	61
2.3.2	Hi-C copy number detection.....	65
2.4	Summary	66
Chapter 3	Astra.....	67
3.1	Methods	69

3.1.1 AStra framework.....	69
3.1.2 Data availability.....	73
3.2 Results.....	76
3.2.1 RD signal captures the strain-specific karyotype of MCF7 cell line	76
3.2.2 RD signal can be decoded into biologically-relevant aneuploidy information	77
3.2.3 AStra successfully identifies the aneuploidy signatures of cancer cell lines.....	80
3.2.4 AStra framework provides a pragmatic solution for computing CN reference	84
3.2.5 Centralization score for assessment of the quality of sequencing data for copy number analysis.....	89
3.2.6 AStra outperforms nQuire for estimating the whole-genome ploidy in low-coverage WGS data	91
3.2.7 Aneuploidy spectrum for calibration of single-sample CN detection tools	92
3.3 Discussion.....	93
Chapter 4 CNAttra.....	95
4.1 Methods	97
4.1.1 CNAttra framework.....	97
4.1.2 Estimation of CNAttra tuning parameters	105
4.1.3 Generating simulated data with artificial CNAs.....	108
4.1.4 Evaluation experiment	112
4.1.5 Data availability.....	113
4.2 Results.....	114
4.2.1 Accurate estimation of CN reference is essential for CNA calling in cancer genomes.....	114
4.2.2 Hierarchical framework enables CNAttra to detect and distinguish between LCV and FA.....	117
4.2.3 Coverage-based thresholding enables the detection of high confidence CNAs in low-coverage datasets.....	119

4.2.4 CNAttra detects experimentally validated LCVs and FAs across cell lines.....	121
4.2.5 CNAttra confirms the CNA profiles identified by SNP array in CCLE and COSMIC database using low-coverage NGS data.....	124
4.2.6 CNAttra successfully identifies copy number profiles of cancer patient samples with variable purity.....	126
4.2.7 CNAttra is the superior tool for detecting large-scale and focal alterations.....	130
4.2.8 Visualization and manual inspection of CNA calling exhibit CNAttra is best equipped to capture the complexity of cancer genomes.....	133
4.3 Discussion.....	137
Chapter 5 <i>HiCNAttra</i>	141
5.1 Methods	144
5.1.1 HiCNAttra framework.....	144
5.1.2 Data availability.....	152
5.2 Results.....	155
5.2.1 HiCNAttra computes high-resolution RD signal by utilizing all types of Hi-C reads	155
5.2.2 Entire fragment counting approach successfully extracts copy number-associated features of RD signal.....	156
5.2.3 HiCNAttra can extract high-resolution copy number events from Hi-C/3C-seq datasets.	160
5.2.4 HiCNAttra successfully ameliorates the effects of CNAs on the contact map.....	164
5.2.5 HiCNAttra better attenuates the 1D signal variation without distorting its inherent features	168
5.3 Discussion.....	178
Chapter 6 <i>Conclusion and Outlook</i>.....	181
6.1 CN state-driven coverage-based thresholding	183

6.2 Directions for accurate TAD identification.....	185
6.2.1 Two steps approach for calling accurate TADs	186
6.2.2 Epigenetics-driven robust TADs.....	188
<i>Bibliography</i>	191
<i>Appendix A: CNAttra Pseudocodes</i>	209
A.1 Pseudocode for L1 segmentation procedure	209
A.2 Pseudocode for IB assembly algorithm.....	210
<i>Appendix B: CNVnator and FREEC Calibration</i>	213
B.1 CNVnator results vary significantly with different bin size options.....	213
B.2 Effect of different ploidy assumptions on the CNA calling by FREEC illustrated using Chr 12 of NCI-H82	214

Table of Figures

Figure 1.1: Visualization of the copy number events in the RD signal.....	32
Figure 1.2: Conventional 3C protocol.....	34
Figure 1.3: Clonal evolution of cancer cell line over passaging time.....	37
Figure 1.4: Characterization of the RD signal of HG00119 normal cell line.....	39
Figure 1.5: Characterization of the RD signal of A427 cancer cell line.....	40
Figure 3.1: AStra framework.....	70
Figure 3.2: Visual illustration to compute centralization error (CE) and centralization score (CS).....	74
Figure 3.3: Aneuploidy signature revealed the genetic variations among the MCF7 strains.	
<i>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 ($CN \leq 2$: black, $3 \leq CN \leq 4$: 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.</i>	78
Figure 3.4: The genome-wide aneuploidy profile for all MCF7 strains (except C, G, H, P and S).	79
Figure 3.5: Aneuploidy signatures of neo-genomes generated using simulated data.....	81
Figure 3.6: Aneuploidy signatures of cancer cell lines.....	83
Figure 3.7: Flowchart of RDS and MMDS methods.....	85
Figure 3.8: Centralization error as a function of copy number reference in some simulated neo genomes.	87
Figure 3.9: Centralization error as a function of CN reference in cancer cells lines (a) and MCF7 strains (b).....	88
Figure 3.10: Centralization score (CS) varies among different cell lines and different MCF7 strains.....	90

Figure 4.1: CNATra framework	99
Figure 4.2: The Savitzky-Golay filter is critical for CNATra oversegmentation approach.	102
Figure 4.3: Procedure for estimating CNATra coverage-based resolution using HG00119 dataset.....	106
Figure 4.4: Estimation of CNATra parameters using exponential regression.	108
Figure 4.5: Generation of simulated CNA profile.....	110
Figure 4.6: An illustrative example of our realistic simulated data.	111
Figure 4.7: Visualization of copy number estimation using CNATra.....	115
Figure 4.8: Normality test for individual IBs using Q-Q plot and Kolmogorov-Smirnov test.	118
Figure 4.9: Coverage-based thresholding.	120
Figure 4.10: CNATra validation using the genetic profile of HepG2.....	122
Figure 4.11: CNATra validation using reported LCVs and FAs.....	123
Figure 4.12: Visual illustration of COSMIC and CCLE CNA calls (using SNP arrays) superimposed on the NGS-based coverage plot.....	125
Figure 4.13: CNATra calls using cancer patient samples.....	127
Figure 4.14: CNATra successfully identified the copy number events of tumor samples with purity >50%.	129
Figure 4.15: Performance evaluation of CNA detection tools on the simulated datasets.....	132
Figure 4.16: Visual comparison of CNA profiles generated by copy number detection tools on Chr 1 of CHP-212.....	135
Figure 4.17: Visual comparison of CNA profiles generated by copy number detection tools on Chr 3 of A427.	136
Figure 5.1: HiCNATra framework and RD signal computation.	145
Figure 5.2: Interval normalization successfully normalized the RD signal of normal/cancer Hi-C data from the systematic biases.....	148

Figure 5.3: Fragment length of dangling-end and extra dangling-end reads from Hi-C/3C-seq datasets.....	154
Figure 5.4: Evaluation of HiCNAta entire restriction-fragment counting approach for computing the RD signal.....	157
Figure 5.5: Visual comparison between genome-wide RD signal derived from Hi-C/3C-seq and WGS datasets (WGS data is taken from input control data of ChIP-seq experiment). ...	159
Figure 5.6: Numbers and widths of large-scale copy number variations (LCVs) and focal alterations (FAs) estimated from different cell lines.	162
Figure 5.7: Visual illustration of copy number events identified from Hi-C/3C-seq reads. ...	163
Figure 5.8: HiCNAta normalization successfully ameliorates the effects of CNAs on the chromatin contact maps.	165
Figure 5.9: Visual illustration of the performance of Hi-C correction tools (ICE, CAIC, OneD, HiCNAta) using MCF7 cancer cell line having abnormal karyotype.....	167
Figure 5.10: HiCNAta successfully eliminated the CNA-induced artifacts on the heatmaps of Hi-C/3C-seq datasets (LNCaP, PC3, NCI-H82, K562).....	168
Figure 5.11: HiCNAta approach successfully attenuates the 1D signal variation of MCF7 without distorting the signal trend.....	170
Figure 5.12: Visual comparison of the Hi-C correction tools (ICE, CAIC, OneD, HiCNAta) in attenuating the 1D signal variation using MCF7 cancer cell line.....	171
Figure 5.13: Visual comparison between HiCNAta simultaneous correction approach and OneD sequential correction approach after integrating HiCNAta-computed CNA track (modified OneD) using MCF7 cancer cell line.).	175
Figure 5.14: HiCNAta simultaneous approach is better than OneD sequential approach for correcting Hi-C/3C-seq cancer datasets.....	176
Figure 5.15: Spearman correlation between the biases and the raw, HiCNAta-corrected, modified OneD-corrected, and cis contact maps for cancer Hi-C/3C-seq datasets (MCF7, LNCaP, PC3, K562, NCI-H69).....	177

Figure 6.1: Illustration of the dispersion level at IBs of different CN states.	184
Figure 6.2: Cartoon illustration of on-diagonal TADs and off-diagonal inter-TAD (TAD-TAD) interactions.	187

Table of Tables

<i>Table 2.1: Single-sample RD-based copy number detection tools.....</i>	<i>56</i>
<i>Table 3.1: Datasets for evaluating the performance of AStra.</i>	<i>75</i>
<i>Table 4.1: Datasets for evaluating the performance of CNAttra.....</i>	<i>113</i>
<i>Table 4.2: Estimation of copy number reference (CNR) using CNAttra.....</i>	<i>116</i>
<i>Table 5.1: Datasets for evaluating the performance of HiCNAttra.....</i>	<i>153</i>
<i>Table 5.2: Spearman correlation between the raw 1D signal and corrected 1D signal of all MCF7 chromosomes by Hi-C correction tools (ICE, CAIC, OneD, HiCNAttra).....</i>	<i>173</i>

Abbreviations

3C	Chromosome Conformation Capture
3D	Three-Dimensional
BAM	Binary Alignment map
CBS	Circular Binary Segmentation
CCLE	Cancer Cell Line Encyclopedia
CE	Centralization Error
CGH	Comparative Genomic Hybridization
Chr	Chromosome
CN	Copy Number
CNA	Copy Number Alteration
CNR	Copy Number Reference
CNV	Copy Number Variation
CS	Centralization Score
COSIMIC	Catalogue of Somatic Mutations in Cancer
DOC	Depth of Coverage
ENCODE	Encyclopedia of DNA Elements
FA	Focal Alteration
FDR	False Discovery Rate
GAM	Generalized Additive Model
GLM	Generalized Linear Model
HMM	Hidden Markov Model
IB	Iso-copy number Block
LOH	Loss of Heterozygosity
Kb	Kilobase

LCV	Large-scale Copy number Variations
Mb	Megabase
MML	Maximal Molecule Length
NB	Negative Binomial
NGS	Next-generation Sequencing
PanNET	Pancreatic Neuroendocrine Tumor
PCR	Polymerase Chain Reaction
PE	Paired-End
RD	Read Depth
ROC	Receiver operating characteristic
SCNA	Somatic copy number alteration
SE	Single-End
SLM	Shifting Level Model
SNP	Single-nucleotide Polymorphism
TAD	Topologically Associating Domain
TCGA	The Cancer Genome Atlas
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
ZS	Zero Score

Abstract

Human immortalized and transformed cell lines are the workhorse of cancer research. They are widely used to study cancer biology and to test and screen anti-cancer compounds in order to improve the efficacy of cancer treatment. Detection of DNA copy number alterations (CNAs) is critical to understand genetic diversity, genome evolution and pathological conditions of cancer cells. Several computational tools have been developed to identify CNAs using the read depth (RD) from single-sample next generation sequencing (NGS) data. However, cancer genomes are plagued with widespread multi-level structural aberrations, such as large-scale copy number variations (LCVs) and focal alterations (FAs), of different length scales, and distinct biological origins and functions. Additionally, cancer cells undergo an adaptive evolutionary process even under controlled culture environment to generate clonal variability which includes changes in copy number. Moreover, genomic NGS data are prone to inherent biases such as GC content, low-mappability regions, experiment biases, and coverage-influenced overdispersion. All these attributes pose challenges for authentication of cancer cell lines as well as accurate modeling and interpretation of their biological phenomenon. Consequently, in case of cancer cell lines, these computational tools fail to identify the correct CNA profile and to distinguish between large-scale and focal alterations due to inaccurate modeling of cancer genomes. Additionally, at low-coverage ($\sim 1x-2x$), RD signal is affected by overdispersion-driven biases which significantly inflate false detection of CNA regions. For solving these problems, we have developed AStra and CNATra tools for digital karyotyping and discovery of CNAs of cancer genomes.

First, AStra has been developed for whole-genome sequencing (WGS)-derived digital karyotyping and authentication of cancer cells. AStra is a Python-based software for *de novo* estimation of the genome-wide aneuploidy profile from raw WGS reads without any prior information about exact chromosome number and aneuploidy levels. AStra identifies the best-fit aneuploidy profile with most genomic segments toward copy number states of positive integer values. It provides an

analytical and visual platform for cell authentication through rapid and easy comparison between different cell lines/strains. We demonstrated that aneuploidy profile offers a unique signature that can distinguish the clonal variations of a cell line. We evaluated our approach using simulated and cancer datasets and showed that cancer cell line exhibit distinct aneuploidy profile which corroborates well with the experimental observations reported earlier. Additionally, AStra provides CN-associated features such as the whole-genome ploidy level and number that can be used for tuning single-sample CNA detection tools.

Next, we have developed CNAttra to hierarchically discover and classify 'large-scale' and 'focal' copy number gain/loss from a single WGS sample. CNAttra first utilizes a multimodal-based distribution to estimate the copy number reference from the complex RD profile of the cancer genome. Savitzky-Golay smoothing filter and Modified Varri segmentation are then implemented to capture the change points of the RD signal. We then developed a CN state-driven merging algorithm to identify the large segments with distinct copy number. Next, focal alterations were identified in each large segment using coverage-based thresholding to mitigate the adverse effects of signal variations. Using cancer cell lines datasets and clinical samples from patients, we confirmed CNAttra's ability to detect and distinguish the segmental aneuploidies and focal alterations. We used realistic simulated data for benchmarking the performance of CNAttra against other single-sample detection tools, where we artificially introduced CNAs in the original cancer profiles. We found that CNAttra is superior in terms of precision, recall, and *f-measure*. CNAttra shows the highest sensitivity of 93% and 97% for detecting large-scale and focal alterations respectively. Visual inspection of CNAs revealed that CNAttra is the most robust detection tool for low-coverage cancer data.

The presence of CNAs in the genome can impact the interpretation of many genetic analyses. One of the examples is the analysis of genome-wide chromatin interactions. Apart from imaging techniques, high-throughput chromosome conformation capture (3C)-based techniques, such as Hi-C, have been extensively used to capture the spatial organization of chromatin, in the form of

contact map employing NGS reads. Normalization of Hi-C contact maps is essential for accurate modelling and interpretation of high-throughput chromatin conformation experiments. Most Hi-C correction methods are originally developed for normal cell lines and mainly target systematic biases either implicitly or explicitly. However, most Hi-C data were generated using cancer cell lines that carry multi-level copy number CNAs which render over- or under-representation of interaction frequencies compared to CN-neutral regions. Therefore, CNA-driven biases need to be corrected to generate euploid-equivalent chromatin contact maps in cell lines with abnormal karyotypes.

We developed HiCNATra framework that extracts RD signal from Hi-C or 3C-seq reads to generate the high-resolution CNA profile and use this information to correct the systematic biases in chromatin contact map. We introduce a novel “entire fragment” counting approach for better estimation of the RD signal and CNA profile. We demonstrated that RD signal calculated from Hi-C reads recapitulates the WGS-derived coverage signal of the same cell line. Utilizing this CNA information with other systematic biases, HiCNATra simultaneously estimate the contribution of each bias and explicitly correct the interaction matrix using Poisson regression. HiCNATra normalization results in removal of CNA-induced artifacts on contact map leading to a ‘homogeneous’ heatmap. Benchmarking against OneD and CNV-Adjusted Iterative Correction (CAIC) methods, which specifically targets CNA bias, as well as commonly-used iterative correction and eigenvector decomposition (ICE) method showed that HiCNATra correction results in the least 1D signal variations without deforming the inherent chromatin interaction landscape.

To sum up, our computational tools provide an analytical and visualization platform for digital karyotyping of hyperploid cancer cell lines. AStra provides the genome-wide snapshot of large-scale chromosomal alterations of the cancer genome such as whole-genome ploidy. CNATra and HiCNATra provide more-detailed karyotyping that includes the LCV and FA information from WGS

and chromatin interaction data respectively. They also provide platforms for visualization of the CNA profiles and chromatin contact maps.

Chapter 1

Introduction

Cancer is the abnormal cellular state that causes uncontrolled growth and division of cells. This typically happens due to the accumulation of various genetic and epigenetics abnormalities in DNA of the cells [1]. Cancer incidence and death are rapidly growing worldwide. According to recent estimates from Global Cancer Observatory (GLOBOCAN), there are 18.1 million new cancer cases and 9.6 million cancer deaths worldwide in 2018 [2]. Lung cancer and breast cancer are the leading reasons of cancer death for male and female, respectively [2]. In 2015, World Health Organization (WHO) estimates that cancer is the first or second leading cause of death before 70 years of age in 91 of 172 countries, and it ranks third or fourth in an additional 22 countries [3]. Therefore, advancement in cancer research is essential for timely diagnosis and treatment of cancer that can save millions of human lives. However, the diversity of tumours, with different kinds of variations, complicates the understanding of this disease and the creation of novel therapies.

Human cancer cell lines are the quintessential experimental model for cancer research [4]. Cancer studies are built on various models such as cancer cell lines [5], primary tumours [6] and xenografts [7]. Among them, cell line models have been widely used for genetic analysis and drug development since they are easy to culture, characterize and manipulate experimentally [8]. Studies based on cell lines have helped to understand the genetic, epigenetic and cellular pathways as well as the biological mechanisms involved in cancer [9]. Cell lines formed the foundation for developing and testing new drug therapies [10, 11]. However, the accurate characterization of cancer cell lines is still problematic due to the presence of many genetic and biological heterogeneity with different origins and functionalities.

Genetic aberrations- including point mutations, indels, and DNA copy number alterations (CNAs)- are common hallmarks of cancer cells [12]. CNA (also commonly referred to as copy number variation or CNV) is a generic term broadly used to define genetic variations that lead to the changes in the number of copies of genomic regions. CNA events are 'relative' gain or loss of DNA regions compared to the reference sample(s) or assembly that are 1 kb or larger in size [13-16]. Cancer genomes are usually plagued with CNAs at different length scales. For example, profiles of 47 ovarian cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) database showed that up to 74% of their genomes are altered [17]. The overwhelming extent of CNA size distribution resulted in their further classification into *microscopic level* large-scale and *sub-microscopic level* focal copy number alteration events [15, 16, 18].

Large-scale copy number variations (LCVs) concern chromosomal abnormalities at megabase (Mb) scale that can be cytogenetically detected. This includes whole-genome or segmental aneuploidy that results from chromosome missegregation and genome instability of cancer cells [16, 19]. LCVs may also represent polymorphic variations among individuals [20, 21]. Aneuploidy refers to the presence of abnormal number of chromosomes in a cell. The accumulation of whole chromosomal or segmental aneuploidies in cancer cells leads to a deviation from 'normal' diploid karyotype, with two copies of each chromosome, to abnormal karyotype (aneuploidy profile), with different number of each chromosome [22, 23]. As result, cancer cells frequently have abnormal karyotype with polyploidy (a cellular state with more than two paired sets of chromosomes) or hyperploidy (a cell containing one or more extra chromosome) profiles along with widespread segmental aneuploidies [24-26]. For example, analysis of the Mitelman database indicates 86% of solid tumors and 72% of blood cancers exhibit gain or loss of entire chromosome, while 26% and 6% of them are near-polyploid, respectively [27].

Focal alterations (FAs) range between several kilobases (kb) to a few Mbs in size containing a small number of genes [28] believed to harbor important oncotargets. They emerge as a consequence of adaptive selection events that facilitate selective growth advantage and evolution

of malignant cells during tumor development and drug resistance [15]. Focal amplifications generally have high-level gains of oncogenes or anti-apoptotic genes while focal deletions usually involve tumor suppressor or pro-apoptotic genes [29, 30].

Heterogeneity is another prevalent signature of cancer. It is evident in a single tumor of an individual (intra-tumor heterogeneity) and between cancers of different individuals (inter-tumor heterogeneity) [31]. Although cancer cell lines are usually derived from an individual, they still suffer from a degree of intra-tumor heterogeneity since it contains a heterogeneous population of cells (supportive cells, transformed cells, tumor-infiltrating cells) [32]. Cancer cells can remain stable with aneuploid karyotype, but sometime, they may undergo continuous genetic alterations over passaging time and culture conditions [33]. For example, it was observed that the same cell line cultured in different labs exhibits karyogram changes over time [34]. These facts suggest that cancer cells undergo an adaptive evolutionary process even under controlled culture environment to generate clonal variability at copy number, gene expression and phenotypic levels [4].

Taken together, all these alterations give rise to ‘mosaic’ karyotypes of cancer cell populations. As explained, cancer cell lines are associated with many concurrent biological phenomena at both *individual cell* level and *population* level [15, 35, 36]. The co-occurrence of these events posits a big challenge for accurate and complete genetic profiling of cancer cell lines that are essential for interpreting biological questions and making reliable conclusions [33, 37, 38]. The next-generation sequencing (NGS) techniques are DNA sequencing technologies that can parallelly sequence large numbers of segments of DNA [39]. Using NGS, the whole human genome can be sequenced within a single day [40]. The advent of NGS allowed high-resolution estimation of the complete repertoire of aneuploidy and copy number alterations at different structural levels [41-43]. Although several computational tools have been developed for aneuploidy estimation and copy number detection from single-sample NGS data [41-43], they failed to capture the variable and wide-range accumulation of genomic alterations in cancer cell lines due to different reasons.

Figure 1.1: Visualization of the copy number events in the RD signal.

Briefly, for copy number detection, most single-sample CNA detection tools were originally developed for modeling normal genomes or cancer patient genomes with relatively lower percentages of CNAs compared to cancer cell lines [41-43]. Therefore, they can be applied to only limited cases of cancer cell lines. Additionally, none of these detection tools can distinguish between the two phenomena of large structural variations and focal alterations. These tools utilize the read depth (RD) signal from WGS data from individual sequencing sample [44-51]. RD signal (also commonly referred to as depth of coverage (DOC) signal) contains the count of the mapped sequence reads at each reference base position. In other words, RD signal indicates the number of unique NGS reads mapped to a particular genomic locus. Sequencing data are available at different coverages. In the present thesis, we refer a WGS data as low coverage if the sequencing depth is $\sim 1x-2x$ and as high coverage if the depth is $\sim 10x$ or above. In absence of copy number events, reads are expected to be distributed uniformly over the genome with no change in the amplitude of the RD signal. Copy number events (amplification and deletions) result in an amplitude shift of the RD signal. Figure 1.1 provides a cartoon representation of the RD signal that illustrates the presence of different copy number events (amplifications and deletions). Majority of the genomic regions are CN neutral with same number of reads. The copy number event (CNA1) shows deleted region with lesser number of reads. Similarly, CNA2 indicates amplified region with higher number of reads. Similarly, most genome-wide ploidy tools compute ploidy by assessing the distribution of allele frequency for single-nucleotide polymorphisms

(SNPs) [51-53]. These SNP-based tools require very high-coverage sequencing data and therefore, they are less applicable for low-coverage WGS data.

To conclude, digital karyotyping of cancer cell lines is still an open and challenging problem. Cancer cell lines are afflicted with wide-spread chromosomal abnormalities and population-level heterogeneity. They complicate the accurate characterization of structural variation of cancer cell lines. Although many tools are available to address these problems, there is still a gap between the biological problems associated with genetic alterations and their available computational solutions.

1.1 Biological experiments

RD signal is the input for the digital karyotyping of cancer cell lines (copy number profile). RD signal can be extracted/estimated from the genome-wide sequencing reads that are resulted from any NGS-based experiment such as WGS, Hi-C, and 3C-seq. Whole genome sequencing (WGS) is a method for analyzing the entire genome of an individual [54]. Last decade, Hi-C and 3C-seq experiments have been developed for capturing the 3D genome conformation based on the 3C experiment [55]. In 2009, Lieberman-Aiden has developed Hi-C method, the first method to study the genome-wide interactions between all loci [56]. In addition, 3C-seq [57], a combination of 3C library preparation [55] followed by NGS, has been developed as a simplified protocol to study the genome-wide DNA interactions. These chromosome conformation capture (3C)-based methods produce paired-end genomic libraries in which each read is supposed to define an interaction between pair of genomic loci. Therefore, short sequencing reads of these experiments can be utilized differently for computing the RD signal.

1.1.1 Conventional chromosome conformation capture (3C)

In a conventional 3C experiment [55], illustrated at Figure 1.2, the interacting genomic regions are chemically cross-linked by formaldehyde, which can covalently link proteins to proteins as well as proteins to DNA. Thus, physically interacting genomic regions are cross-linked through

DNA-bound proteins. The amount of cross-linking is based on spatial proximity, suggesting a higher frequency of cross-linking of spatially close regions than that of spatially distant regions. To quantify the frequencies of cross-linking, a restriction enzyme that can produce single stranded DNA overhang is used to digest the cross-linked chromatin and the resulting restriction fragments are ligated by T4 DNA ligase. The ligation is carried out in dilute condition to favor the intramolecular ligation over intermolecular ligation. The number of ligation products formed between two segments represents the frequency with which they interact with each other. These ligation junctions contain the information of chromatin interaction and thus are the end products of a 3C experiment. To obtain the 3C library containing ligation products, the cross-linking is then reversed by Proteinase K treatment, and DNA is purified. In conventional 3C technique, polymerase chain reaction (PCR) is used for the detection of individual ligation products. Generally, the interaction frequency is inversely proportional to the linear genomic distance between two loci. Interaction between two genomic regions that occurs with a frequency significantly higher than expected is considered as long-range looping interaction.

Figure 1.2: Conventional 3C protocol.

Procedures in a conventional 3C experiment, including formaldehyde cross-linking, restriction enzyme digestion, intramolecular ligation, reverse cross-linking, DNA purification and PCR detection.

1.1.2 Hi-C

In Hi-C [\[56\]](#), after the restriction enzyme digestion, the sticky ends (5'- or 3'-overhangs) are filled with dNTPs, among which dCTP is labeled with biotin. This is followed by blunt-end ligation using T4 DNA ligase. After DNA purification, the library is sheared and the biotin-labeled DNA segments are pulled down with streptavidin beads via specific binding of streptavidin to biotin. This step allows the enrichment of digested and ligated segments and thus increases the percentage of informative segments in the Hi-C library. The library is then sequenced by high-throughput sequencing. In Hi-C, paired-end sequencing is used to obtain sequence information of both sides of the ligation products.

1.1.3 3C-seq

In 3C-seq [\[57\]](#), the library preparation is same as that in conventional 3C [\[55, 58\]](#). Instead of using PCR for the detection of pair-wise interactions, the entire library is sheared by sonication to a median size of 500 bp (range 300-800 bp) followed by NGS library preparation and paired-end sequencing. Absence of biotin labeling and pulling down steps, used in Hi-C protocol, decreases the loss of DNA in the 3C-seq protocol. Therefore, 3C-seq results in more genome-wide distributed reads that can be used for better estimation of the RD signal.

1.2 Visualization of genomic alterations

In this section, we provide a comprehensive visual illustration of the copy number alterations associated with cancer cell lines and the resulting RD signal, which forms the basis for digital karyotyping. As explained earlier, cancer cell lines evolve continuously during propagation in culture and they acquire various types of genomic alterations over time including DNA copy number changes.

These genomic alterations (CN events) can be classified into chromosomal duplications and deletions, segmental duplications and deletions based on their structure. Chromosomal

duplication and deletion are the gain and loss of one or multiple complete chromosome(s) respectively whereas segmental duplication and deletion are the gain and loss of a part of a chromosome. From another perspective, as explained earlier, these copy number events can be classified into large-scale copy number alterations [16, 19] and focal alteration based on the biological origins and functionalities,

To illustrate the temporal changes of genomic alterations, cartoon representation of cancer cell at different passage numbers is provided in Figure 1.3. In this example, we assume that cell line initially contains 7 cells (initial cells) with same set of genomic alterations (Figure 1.3 left). Each initial cell has three copies of chromosome 2 (chromosomal duplication) and a segmental duplication in chromosome 1 (H2 region). With continuous passaging in culture over time, each cancer cell keeps gaining or losing other CN events (Figure 1.3 bottom). First, 6 out of the 7 cells gain additional copies of chromosome 1. Then, 5 cells undergo segmental duplication of F3 region in both copies of chromosome 3. Finally, 4 cells lose H2 region (segmental deletion) from all three copies of chromosome 2. Overall, with passaging time there is gain of one additional copy of chromosome 1, loss of H2 region (segmental deletion) from all three copies of chromosome 2, and segmental duplication of F3 region in both copies of chromosome 3 (Figure 1.3 right).

The combination of these CN events defines the aneuploidy profile (LCVs and FAs) which adds to the intra-tumor heterogeneity of cancer cell line over time. In the previous example, the passaged cells are heterogeneous (Figure 1.3 right) while the initial cells are homogenous (Figure 1.3 left). Compared to normal cell lines, this phenomenon introduces an extra level of complexity for modeling the CN distribution of cancer cell lines using predetermined statistical distribution. This can be visually observed from genome-wide RD signal and distribution (Figure 1.4 and Figure 1.5).

Figure 1.3: Clonal evolution of cancer cell line over passaging time.

In this cartoon representation, one cancer cell from the early passage is zoomed (left side) to show the genomic alterations at whole-chromosome or segmental levels. Over time, continuous passaging (culturing) may lead to acquisition of additional genetic alterations (for example, one cell from late passage, shown in zoomed version on the right side, shows additional genomic aberrations) leading to intra-tumor heterogeneity and clonal evolution.

For illustrating the complexity of modeling cancer cell lines, compared to normal cell lines, we provided two examples using coverage plots from WGS data of HG00119 normal individual (1000 Genomes Project) (Figure 1.4) and A427 cancer (Figure 1.5) cell line, respectively. In both figures, each dot represents a bin. Top panels show the genome-wide RD signal (coverage plot) (Figure 1.4a and Figure 1.5a right) and the genome-wide RD distribution (Figure 1.4a and Figure 1.5a left). As explained earlier, CN events can be visualized as amplitude of the RD signal. For HG00119 cell line, genome-wide RD signal is almost free of any LCV (Figure 1.4a left) and most of the genome is copy number neutral (all chromosomes sharing similar statistical measures for RD signal). This results in a unimodal distribution of RD signal (Figure 1.4a left). Therefore, unimodal distribution, such as normal and Poisson, is reasonably applicable to ‘normal’ human genomes where structural variations are rare. In contrast, LCVs are pervasive as evidenced from the genome-wide RD signal of A427 cancer cell line (Figure 1.5a right) which results in the multimodal frequency distribution of RD signal (Figure 1.5a left).

A closer look at the RD signal of chromosome 2 of both cell lines (Figure 1.4b and Figure 1.5b) shows that only focal events such as *CCNT2* amplification are presented in the HG00119 (Figure 1.4b). In contrast, RD signal of A427 cell line is afflicted with LCVs and FAs (Figure 1.5b). Some focal events are interspersed within LCVs creating an intricate relationship between them. For example, focal amplifications containing *USP34* and *CCNT2* genes are part of different LCVs in 2p and 2q region respectively (Figure 1.5b). Therefore, there can be a complex scenario where a genomic segment may have a hemizygous segmental deletion (LCV) which in turn can contain a focal amplification. In the coverage plot, LCVs appear as ‘large’ segments whereas focal amplifications and deletions are represented as ‘sharp’ peaks and troughs respectively (Figure 1.4b and Figure 1.5b,c).

Notably, RD signal of A427 cell line are strongly affected by ‘wave artifacts’ (indicated as a blue curve in (Figure 1.5b,c). Wave artifacts are systematic biases due to deviation from equimolar coverage [59, 60]. On the other hand, wave artifacts are not clearly recognized at coverage plot of

HG00119 cell line (Figure 1.4b,c). This also complicates the accurate detection of focal alterations in case of cancer cell lines, especially from low-coverage data. The overdispersion of the RD signal is inversely proportional with the data coverage [61]. Therefore, at low-coverage, the overdispersion of the RD signal may magnify the impact of the wave artifacts. In addition, RD signals from WGS data are also prone to inherent biases associated with NGS owing to genome GC content, low-mappability regions, and other experiment-based biases [62, 63]. All these biases should be considered for accurate identification of the biological features of the cancer genomes.

Figure 1.4: Characterization of the RD signal of HG00119 normal cell line.

(a) The RD frequency distribution (left) and genome-wide coverage plot of all chromosomes (right) of normal genome (HG00119) showing the normal individual genome is usually free of LCVs. (b) Coverage plot of Chr 2 shows the presence of focal alterations. (c) Zoomed in view of Chr 2 (184-215 Mb) showing that wave artifacts are not affected the RD signal of the normal genomes.

Figure 1.5: Characterization of the RD signal of A427 cancer cell line.

(a) The RD frequency distribution (left) and genome-wide coverage plot of all chromosomes (right) of A427 cell line showing the presence of widespread large-scale segmental aneuploidies. (b) Coverage plot of Chr 2 showing LCVs of different copy number states. (c) Zoomed in view of Chr 2 (184-215 Mb) shows the presence of focal alterations (both amplification and deletions) inside the LCV. The blue curved lines in (b) and (c) denote the upper envelope of RD signal showing the regions affected by wave artifacts. Few focal amplifications and deletions are indicated.

1.3 Problem statement

RD signal of cancer data is usually plagued with biologically relevant features in addition to other sources of biases (NGS-driven such as GC-content and mappability, experiment-driven such as distribution of restriction enzymes in Hi-C/3C-seq experiments). Here, our main aim is to extract and distinguish the important features of cancer cell line including aneuploidy profile, large-scale copy number variations, and focal alterations. These features themselves are valuable information for answering many biological questions such as role of genetic alterations in cancer and chemotherapeutic drug response. Moreover, they can be utilized for studying other biological phenomena in an unbiased manner. For example, copy number events can be used for correcting Hi-C data for accurate estimation of chromatin interactions.

In this thesis, we formulate this *problem* as **computational karyotyping of individual cancer cell lines from low-coverage NGS data through read depth analysis**. More specifically, if any genome-wide NGS (low/high-coverage) data of a cancer cell line is provided, then the main goal is to identify its complete and accurate copy number profile from this data without any additional information or prior knowledge.

In other words, if one has WGS, Hi-C, or 3C-seq data from a cancer cell line, then how to authenticate these cell lines, to assess its reliability for downstream copy number analysis and to extract the maximum accurate copy number information, including LCVs and FAs, in the presence of the inherited biases of that data. More importantly, this should be done without making specific assumptions of the percentage alterations that vary much between cancer cell lines, or without using other information of that same cell line since they are evolved under passaging time and culture conditions.

1.4 Contributions

This thesis is mainly focused on digital karyotyping of cancer cell lines from single-sample NGS data, especially at low-coverage scenario. The key point for digital karyotyping is the accurate identification of copy number events, including both large-scale and focal copy number alterations. Therefore, to make our karyotyping more robust so that it is valid for any cancer cell line with different degree of variability in structural aberrations, we built our solutions based on the fundamental assumption that majority of genomic regions of any cell line population should have integer/near-integer copy number values.

Compared to the available CN detection tools, our successful contributions are mainly originated from the usage of multimodal distributions for modeling the RD signal and from defining copy number variations as large-scale and focal events based on their origin and functionalities. In contrast to most available tools that use a unimodal probability distribution (normal, Poisson, and negative binomial) for modeling the RD signal (explained in Chapter 2), we utilize multimodal probability distributions for representing the RD signal of cancer cell lines. This results in a more accurate modeling of hyperploid cell lines with abnormal karyotype (the chromosome count of an organism). In addition, we computationally introduce the notions of large-scale copy number variations and focal alterations. This enables the optimization of segmentation and CNA calling methods for identifying both LCVs and FAs separately, even from the low-coverage data. Consequently, we have developed three computational tools (AStra, CNATra, HiCNATra) for successively defining the aneuploidy spectrum and authentication of cancer cell lines, for detecting both LCVs and FAs, and for correcting the Hi-C/3C-seq chromatin interaction frequencies. Our digital karyotyping tools, AStra and CNATra, are also equipped with interactive platform to visualize and manually inspect the aneuploidy spectrum, complete CNA profiles, and CNA-associated information for further validation and interpretation. Similarly, HiCNATra also has platform for visualization of contact maps (raw and normalized) as well as all the features associated with copy number analysis from Hi-C/3C-seq data.

AStra (Aneuploidy Spectrum (detection) through read depth analysis), is a Python-based software for cell line authentication from the WGS data. AStra captures the complete aneuploidy profile of cancer genome without any prior knowledge of the input sequencing data, e.g. the genome-wide ploidy level of input cell line. As a result, AStra's aneuploidy profile can be easily used for cell line authentication. AStra also can infer the CN-associated features such as the centralization score (CS) and whole-genome ploidy level. The CS measures the percentage of genomic regions around CN states. They can be used for selecting and tuning the single-sample copy number detection tools.

CNAtra (Copy Number Alteration (detection) through read depth analysis), a MATLAB-based software for the sensitive and robust detection of LCVs and FAs. CNAtra empirically models the RD signal based on a multimodal distribution. This approach allows us to define the accurate 'interval' of copy number states which aids to hierarchically identify segmental aneuploidies and focal alterations in a robust manner largely unaffected by percentage of genomic aberrations and wave artifacts. Also, CNAtra introduced coverage-based thresholding to overcome the overdispersion problem of the RD signal from low-coverage datasets and detect only highly significant focal alterations.

HiCNAtra (Hi-C Normalization Approach through read depth analysis), a MATLAB-based software for identifying the copy number events from the Hi-C/3C-seq data and utilizing them to normalize the chromatin contact frequencies. HiCNAtra first utilize all the raw Hi-C reads (not only valid pairs) using a novel entire fragment approach for precise estimation of RD signal at high resolution, and detection of both LCVs and FAs based on CNAtra approach. Then, we utilize a generalized linear model (GLM) by incorporating the data-driven CNA tracks with other data-independent systematic biases to correct the interaction frequencies between genomic loci.

1.5 Thesis roadmap

The rest of this document is organized as follows. Chapter 2 includes a comprehensive up-to-date literature review of the current state-of-the-art computational tools for ploidy estimation, copy number detection and Hi-C/3C-seq data correction. Chapter 3 presents a detailed description of the framework and evaluation of *AStra software*. Chapter 4 demonstrates the CNAttra solutions for detecting LCVs and FAs from low-coverage WGS data. Chapter 5 presents the three modules of HiCNAttra for estimating the RD signal, detecting copy number events, and correcting the contact maps obtained from Hi-C/3C-seq data. Finally, Chapter 6 concludes the thesis with a summary of the research problems and solutions, and possible directions for future research.

Chapter 2

Literature Review

Genetic aberrations describe the changes in DNA sequences between individual. Some of these changes are tightly associated with changes in gene expression level [64]. In many cases, these aberrations influence phenotypic traits and/or cause diseases, such as cancer, autism, schizophrenia, and Alzheimer's disease [65-69]. These variations range from single to millions of nucleotide (base) changes which include single nucleotide variations, small insertions or deletions (indels) and copy number variations/alterations.

CNAs (also called CNVs) can be further classified, based on their sizes, into the submicroscopic level 'focal' alterations (between several kb to a few Mb) and microscopic level large-scale copy number alterations (Mb scale) with different origins and functionalities [15, 16, 18]. Usually genomes of normal cell lines are free of LCVs [70] whereas genomes of cancer cell lines are afflicted with both LCVs and FAs [17, 27]. Accumulation of LCVs may result in changing the aneuploidy spectrum and the whole-genome ploidy level [24-26].

Starting in 2003, many computational tools were developed for detecting CNAs, inferring the whole-genome ploidy level, or both [41-43, 71-73]. They attempted to identify CNAs by utilizing data from different technologies such as array-based comparative genomic hybridization (array CGH) and NGS-based approaches. Over the last decade, NGS [39] has developed as a standard strategy for genotyping and comprehensive profiling of CNAs by generating millions of short reads of the entire human genome in a single sequencing run [74]. Compared to array-based approaches, NGS-platforms have many advantages including higher data coverage, better estimation of RD signal, accurate detection of breakpoints, and identification of novel CNAs [75,

[76](#)]. Taking these advantages into account, numerous computational tools have been developed to detect CNAs based on different attributes that can be extracted from NGS data.

Here, we present **an overview of the single-sample computational tools for identifying and correcting CNA-associated features from NGS data**. Particularly, in the upcoming sections, we first provide a survey on the different algorithms for copy number detection from single-sample WGS data. Then, we analyzed the available methods for identifying copy number alterations from Hi-C/3C-seq datasets as well as for normalizing the CNA-driven biases on their genome-wide contact frequencies. In this chapter, we mainly review the applicability of these tools on NGS data of cancer cell lines. In Chapter 3, Chapter 4 and Chapter 5, **we will revisit the main techniques and analyze their performance in comparison with our proposed methods (AStra, CNATra, HiCNATra)**.

2.1 CNA detection through NGS data

Many tools have been developed for identifying the copy number events such as duplication, deletions, and insertions from NGS data. Generally, they detect these events by comparing the sequence and frequency of short reads versus the sequence of the reference genome in two ways (1) whole exome sequencing (WES) and (2) WGS [[41-43](#), [71](#)]. While WES-based methods focus only on the exome (coding DNA) regions which represent only 1% of the human genome [[77](#)], WGS-based methods examine the complete genome including the non-coding DNA. WGS is more reliable source than WES for copy number analysis and it can capture all genetic variations including LCVs and FAs [[78](#)]. WES methods cannot precisely detect the LCVs in cancer cells since they only examine the exome regions which may not include the changepoints of these LCVs. Additionally, they cannot capture the variations in gene's regulatory elements in non-coding regions of genome such as enhancers and repressors. In contrast, WGS-based techniques enable the detection of the complete repertoire of copy number events that include genes and all other

regulatory elements [78]. This copy number information can be effectively used for correcting the CN-driven bias on any genome-wide study such as Hi-C.

WGS-based CNA detection methods can be categorized into multiple strategies including (1) paired-end mapping (PEM), (2) split read (SR) and (3) read depth [41-43]. PEM identifies small CNAs from discordantly mapped paired-reads whose distances are significantly different from the average distance between the two-sides of reads from the same library/experiment [43]. Similarly, SR methods examine read pairs for identifying small CNAs [43]. They start from pairs in which one read from each pair is aligned to the reference genome uniquely while the other one fails to map or only partially maps to the genome. Those unmapped or partially mapped reads defines breaking points of these CNAs. It is important to note that PEM algorithms [79] are only applicable to paired-end sequencing whereas SR methods are for unique regions in the reference genome. Moreover, both techniques can only find the small deletions and insertions, but they cannot identify large segmental gain or loss [43]. Therefore, last decade, RD methods have become the dominant technique to calculate the CN gain/loss and find duplication and deletion regions due to the availability of WGS data at low cost. RD methods are built on the hypothesis that number of mapped reads in a genomic region is correlated with the copy number of the region [43]. This allows to identify CNAs of different sizes and estimate their exact CNs. These tools are in strong demand as they are central to CNA prediction from personalized genomes or individual tumor biopsy for clinical decision-making. RD techniques can find CNAs of different lengths in complex genomic areas since it is based on the hypothesis that copy number of a genomic region is correlated with the coverage depth of that regions [41]. However, the accuracy of currently available RD methods depends on data coverage, the availability of a matched sample, and the complexity of genomic CN profiles, etc.

Numerous computational tools have been developed for CNA detection using depth of coverage from WGS data [41-43]. They utilize the RD signal to identify 'absolute' copy number by modelling the RD signal either from an individual sample (single sample), or to discover 'relative' copy

number by taking advantage of matched normal sample (paired case-control sample) or using samples from healthy population [41]. Each of these RD-based approaches have been successfully used depending on the study design and data availability [43]. When matched control samples are available, they can be used as a reference genome to detect the CNA assuming this control sample is correctly representing the individual genome. Based on this assumption, many tools [80-84] were developed and achieved high accuracy and sensitivity. However, matched control from tumor-adjacent normal tissue is generally difficult to procure [85]. Hence, majority of tumor datasets do not have corresponding tumor-adjacent normal sample in genomic databases, such as The Cancer Genome Atlas (TCGA) [85]. As an alternative, whole blood is commonly used as matched control for solid tumors to control for genetic background, and sometimes non-invasively collected samples (e.g. saliva or buccal samples) are also employed for this purpose. However, the source of 'normal' samples greatly impacts the quality of genome analysis such as copy number detection outcome [86, 87]. Additionally, in case of cancer cell lines, most of them do not have the corresponding normal counterpart derived from the same individual. Therefore, single-sample-based computational tools, which do not rely on matched controls, are more applicable for both cancer cell lines as well as patient tumor samples.

2.1.1 Single-sample CNA detection tools

Several tools have been presented over the years that utilize a single sequencing sample for copy number detection [44-51]. They are built on different assumptions of the underlying probabilistic distribution and percentage of chromosome karyotype abnormality (Table 2.1). Generally, the single-sample CNA detection tools follow the subsequent approach. First, the entire genome is divided into consecutive non-overlapping bins of equal size to compute the RD signal (RD value/read count per bin). Then, RD signal is corrected to remove the effect of the systematic biases such as GC-content and mappability. Then, copy number reference (CNR) is estimated based on a genome-wide measure such as global RD mean or median. This CN reference represents the RD value (read counts per bin) corresponding to the copy number neutral state

(two copies of each chromosome for diploid individuals). In absence of CNA, all genome regions are supposed to have read counts per bin similar to the copy number reference. Therefore, CN amplified (deleted) region can be defined as any continuous region with a significant increase (decrease) of RD values compared to the copy number reference. Then, each chromosome is partitioned into segments of different RD values using various clustering techniques such as circular binary segmentation (CBS) [88, 89]. Finally, CNA regions are detected using a statistical test for finding the statistically significant segments assuming that the RD signal follows particular distribution such as normal, Poisson, or negative binomial (NB) distribution.

As demonstrated in Chapter 1, RD signal of cancer and normal genomes are significantly different due to the widespread presence of large-scale segmental/chromosomal aneuploidies in cancer genomes in addition to focal alterations. As a result, the RD frequency distribution of cancer genomes exhibits multimodal frequency distributions (Figure 1.4), compared to the unimodal distribution in case of normal genome (Figure 1.5). Additionally, the global statistical measures (median or mean) may not truly represent the copy number neutral state since different chromosomes have different local statistical measures and RD frequency distribution is no-longer unimodal distribution. These properties of cancer genomes may invalidate the assumptions of most single-sample CNA detection tools and lead to inaccurate copy number profiling of cancer genomes. First, the CN reference cannot be computed as the global median/mean of the RD signal due to the multimodality of RD frequency distribution. Second, disregarding the distinction of large-scale and focal events with different biological origins [15, 16, 18], these single-sample tools suffer from oversegmentation of LCVs and erroneous calling of FAs. For example, if one LCV accommodates one or more FAs (FAs inside LCV scenario), this LCV may be inaccurately divided into smaller LCVs and FAs. Therefore, none of the available tools can capture the entire spectrum of copy number aberrations in cancer cells and distinguish between the LCVs and FAs.

Additionally, many WGS datasets of cancer cell lines are available at low-coverage which pose a big challenge for accurate identification of CNAs, especially the focal alterations. In NGS, data

coverage refers to the average number that each nucleotide in the genome is sequenced. Dispersion of RD signal is inversely proportional to the data coverage. Therefore, at low-coverage, RD signal is prone to the overdispersion problem. Consequently, statistical-based segmentation/CNA-calling may result in many false calls of CNAs. To solve this problem, some CNA detection tools have the option to increase the bin size of RD signal to avoid false detection of CNAs. However, this approach may result in missing true FAs.

In the next subsection, we provide **a detailed description of the single-sample CNA detection tools from WGS data**. For each tool, we also discuss the impact of presence of LCVs and the overdispersion problem in its performance. In Chapter 4, **we analyze their performance in comparison to CNATra by applying them in both real cancer datasets and simulated data**.

2.1.1.1 ReadDepth

ReadDepth [49] uses a variable bin size for computing the RD signal. Based on the data coverage, they first estimate the appropriate bin size by modeling the RD signal using a negative binomial distribution. The mean and variance of the NB distribution are computed based on number of reads per bin n and the user-defined variance-to-mean ratio (default = 3). They calculated the mean number of reads per bin $\lambda = n * b/g$ where g and b are the genome size and bin size, respectively. They progressively test different bin size until they find the smallest bin size under the constraints of the user-defined false-discovery rate. Then, they utilized the circular binary segmentation (CBS) [88, 89] to partition the RD signal. CBS consider the RD signal as a continuous circle and divide it hierarchically into segments by iteratively locating the change points. These change points Z_c are defined as the points (i, j) which maximize the statistics Z_{ij} :

$$Z_c = \max_{1 < i < j < n} |Z_{ij}| \quad (2.1)$$

$$Z_{ij} = \frac{\frac{s_j - s_i}{j-i} - \frac{s_n - s_j + s_i}{n-j+i}}{\sqrt{\frac{1}{j-i} - \frac{1}{n-j+i}}} \quad (2.2)$$

where S_i, S_j, S_k are the summation of RD signal from bin 1 to bins i, j, k respectively. Finally, they identified the CNAs by examining the resulted segments against the NB distribution. This algorithm has two main drawbacks for CNA analysis of cancer cell lines. First, it uses the genome-wide RD mean as the copy number reference. However, as explained earlier, this may result in inaccurate CN estimation in case of cancer cell lines with abnormal karyotype and multimodal RD distribution. Second, they use large bin size for low-coverage data to reduce the false positive CNAs, which may result in missing many true focal alterations.

2.1.1.2 CNVnator

CNVnator [44] uses a user-defined bin size for computing the RD signal. First, *CNVnator* normalizes the RD signal for GC-content bias using the interval (scaling) normalization method. Then, it utilizes mean-shift technique for segmenting the RD signal. They formulate copy number variation problem as finding a probability distribution function which is a mixture of many distributions, each corresponding to a CN state. So, they use the mean-shift technique to divide the RD signal into clusters by iteratively shifting each point to this density maxima along the mean-shift vector. Then, they define the copy number reference as the average read counts per bin assuming that the RD signal follows a normal (Gaussian) distribution. Consequently, they use the statistical one sample t-test to identify the significant copy number events. For each genomic segment, they calculate its one sample t-test P-value to check if the segment mean RD value has a value of genomic average:

$$t = \frac{\overline{RD}_{global} - \overline{RD}_{segment}}{\sigma_{segment}} * \sqrt{n} \quad (2.3)$$

where $\overline{RD}_{global}$ is the genome-wide average RD signal, $\overline{RD}_{segment}$ is the segment average RD signal, $\sigma_{segment}$ is the segment standard deviation, and n is the number of bins within the segment. Finally, P-value is corrected assuming that 99% of the genome is CN neutral. Where *CNVnator* is widely used for copy number detection of normal genomes (1000 Genomes Project) with largely unimodal RD signal distribution, its assumptions may not be accurate in case of

cancer cell lines with abnormal karyotype. First, the genome-wide RD mean may not accurately represent the copy number neutral state. Second, in cancer cell lines, large sections of the genome can be affected by CNAs, not only 1% of the genome as assumed in CNVnator algorithm. For example, the analysis of 47 ovarian cancer cell lines showed that up to 74% of their genomes are altered [17].

2.1.1.3 FREEC

To address the problem of polyploidy, FREEC (control-FREE Copy number caller) [45] incorporates the ploidy information as an input. It uses the genome-wide ploidy level (diploid, triploid, and tetraploid) to define the number of copies of each chromosome (2 copies, 3 copies, and 4 copies, respectively). The FREEC pipeline starts by using a sliding-window technique to compute the RD signal into non-overlapping bins. Then, they fit a polynomial function for normalizing the effects of GC-content on the RD signal. This fitted model is built based on many assumptions: (1) the whole-genome ploidy level (P) of the sample should be provided, (2) RD signal can be modeled by a polynomial of GC contents, (3) the RD values per regions of different CN states are linearly proportional to each other, and (4) the genomic regions in the CN state of P (2 for diploid, 3 for triploid, and 4 for tetraploid) contain the complete interval of measured GC contents. Then, they use only the regions in the state of P to fit the polynomial function and normalize the RD signal. Next, they segment the RD signal using a LASSO-based algorithm [90]. Finally, they classify the resulted segments based on their copy number to identify the significant CNAs. FREEC algorithm has two main advantages for CNA detection of cancer cell lines. First, it identifies the CNAs based on their CN states (1, 2, 3, 4, ...), not distribution-based outliers. This allows FREEC to identify CNA from low-coverage data. Second, they can correct the CN values for triploid and tetraploid profiles as well as diploid profile. However, it may not be able to model the cell lines with hyperploid profiles such as hyperdiploid and hypertriploid. For example, in hyperdiploid karyotype, only few chromosomes have more than two chromosomes and the whole-genome ploidy number may be ~ 2.5 , not 2 or 3. Additionally, in this case, the genomic

regions in the nearest CN state of P (2 or 3) may not include the complete interval of measured GC contents.

2.1.1.4 CLImAT

CLImAT [51] has been developed for detecting CNAs of patient tumor samples. In general, tumor samples differ from cancer cell lines in two main fundamental points. First, although both cancer cell lines and patient tumor samples have abnormal karyotypes, their degree of abnormality differs. For example, aneuploidy of cancer cell lines can range from diploidy to tetraploidy (hyper- or hypo-diploid, hyper- or hypo-triploid, and so on) while patient tumor samples are usually hyper- or hypo-diploid. Additionally, tumor samples suffer from purity issue compared to clonally-propagated cancer cell lines which are 'largely' free of such issue. Moreover, unlike cancer cell lines, most of the WGS data of tumor samples from patients are available at high coverage. Therefore, for patient samples, *CLImAT* was tuned for high-coverage data (>30x) and utilized the SNP information in addition to the RD signal. In the *CLImAT* pipeline, the RD signal is first computed by counting the reads at windows of 1000 base pairs centered at each SNP. Additionally, it computes the B-allele frequency (BAF) signal that contains all known SNPs in dbSNP [91]. Then, it corrects the RD signal from the GC-content and mappability biases. Then, *CLImAT* utilizes an integrated Hidden Markov Model (HMM) for clustering the RD signal and defining the copy number states by integrating both RD signal and BAF signal. As the integrated HMM defines its states using the BAF (SNP information), its performance strongly depends on the data coverage of the input WGS data. Therefore, *CLImAT* may not be suitable for detecting CNAs from low-coverage data.

2.1.1.5 XCAVATOR

XCAVATOR [48] has been developed for detecting copy number events from short and long reads WGS experiments. In its evaluation, it mainly identified genomic aberrations from 1000 Genomes Projects [92] as well as patient tumor samples. In both cases, the majority of the genome are CN neutral and the genome-wide ploidy level is diploid/near-diploid. In *XCAVATOR* pipeline, it first

computes the RD signal by counting the number of reads that map to consecutive and non-overlapping windows of the reference genome. Then, similar to CNVnator and CLImAT, XCAVATOR corrects the RD signal from GC content and mappability using the scaling approach:

$$\overline{RD}_i = RD_i * \frac{m}{m_x} \quad (2.4)$$

where RD_i and $\overline{RD}_i$ are the raw and corrected RD values of the i window respectively, m is the genome-wide median of the RD signal, and m_x is the median of the RD signal of all the windows that have the same X value where X is either the GC content or the mappability score. Then, assuming diploid genome, it uses the genome-wide RD median as copy number reference (corresponds to two copies of each chromosome). Then, XCAVATOR utilizes the shifting level model (SLM) algorithm [93], a special class of HMM, for detecting the boundaries of RD segments. In SLM algorithm, they model the RD signal as summation of two normal (Gaussian) distributions for representing the RD value per window and the normally-distributed noise. Finally, they classify the alteration regions into five CN states (double deletion, single deletion, normal, duplication, multiple copy duplication) based on a mixture of normal distributions. Although XCAVATOR can model the RD profiles of normal genomes and to some extent the patient tumor samples, the widespread of prevalence of LCVs in cancer cell lines may invalidate its assumptions. First, the copy number reference may not be truly represented by the genome-wide RD median in case of cancer cell lines with abnormal karyotype. Second, the unimodal normal distribution can not fit the multimodal RD frequency distribution of cancer cell lines. Third, the classification of alteration regions in only five states (CN =0, CN =1, CN =2, CN =3, and CN >=4) can not differentiate between adjacent regions with CNs of 4 or higher.

2.1.1.6 AMYCNE

AMYCNE [46] estimates the copy numbers of input targeted-regions. In addition to the main two files that contains the coverage and GC-content of non-overlapping bins covering the whole genome, it requires a bed file that includes the target regions. They estimate the copy number of the targeted regions by dividing the read counts of each region by the average read counts of all

regions with same GC score. Then, AMYCNE integrates the input information of the ploidy of the organism in order to model the RD signal of aneuploid cell lines. They multiply the estimated copy number by the ploidy number. Although AMYCNE approach may be applicable for cancer cell lines, there are two points that may negatively impacts its performance. First, as it requires the targeted regions as input, it can only identify all genome-wide events if the genomic segments are accurately provided. However, this strongly depends on the approach that is used for RD segmentation. Additionally, it may limit the ability to detect novel CNAs. Second, for accurate CN estimation, the whole-genome ploidy number of cell line (not ploidy level such as diploid and triploid) should be computed by another ploidy estimation tool since cancer cells may undergo an adaptive evolutionary process and their established whole-genome ploidy number of the same cell line may be inaccurate [4].

2.1.1.7 SCNV

SCNV [50] have been developed for detecting copy number variations from single-cell data of patient tumor samples. These single-cell data contain the WGS data of many individual cells, each at low coverage. Moreover, single-cell data of tumor samples usually contain some normal cells in addition to the tumor cells based on sample purity. Therefore, SCNV can be considered as control/matched copy number detection tool, but from the same sample. In SCNV pipeline, reads of each cell are first binned in every 100 kb to compute the RD signal. The 100 kb bin is selected to preserve sufficient resolution for detecting change points of the RD signal. Then, SCNV applies a control-free segmentation procedure on each cell separately. It partitions the RD signal into segments using the CBS algorithm. This segmentation step is used for classifying the individual cells into normal or tumor cells (cell status). Then, they pool a group of normal cells as a control for the tumor cells. Finally, they adapt a control-based bulk population CNA detection [94] to identify the copy number events from tumor samples. As SCNV targets the copy number identification from single-cell data of tumor samples, it is less applicable for cancer cell lines. SCNV uses a control-based CNA detection approach assuming that the presence of a fraction of

normal cells among the single-cell sample. However, as explained before, matched control samples are difficult to procure for most cancer cell lines. Additionally, the bin size of 100 kb limits its ability to detect small-sized copy number alteration (focal alterations).

Table 2.1: Single-sample RD-based copy number detection tools.

Tool	Input	Language	Methodology
ReadDepth	Coverage bed file	R	ReadDepth uses Circular-Binary Segmentation for dividing the RD signal and fitting an NB distribution.
CNVnator	BAM file	C++	CNVnator uses the mean-shift technique for segmenting the RD signal and applies t-test to find statistically-significant alterations.
FREEC	BAM file	C++	FREEC fits a polynomial for normalizing the RD signal for GC-content and uses the lasso-based algorithm for segmentation.
CLImAT	BAM file + SNPs	MATLAB	CLImAT computes the RD signal at windows centralized at SNPs and utilizes integrated HMM for segmenting the CN states.
XCAVATOR	BAM file	Fortran, Perl, R	XCAVATOR segments the RD signal using the SLM algorithm and identifies the CNAs based on a mixture of truncated normal distributions.
AMYCNE	Coverage bed file + target regions	Python	AMYCNE estimates the copy number of the target regions and detects the large CNAs and chromosomal aneuploidies.
SCNV	BAM file	R	SCNV identifies the cell status (normal and tumor) and uses the normal cells as controls for tumor cells.

2.2 Ploidy estimation through NGS data

The whole-genome ploidy number and ploidy level (state) are **inherent features** associated with the copy number profiles. The whole-genome ploidy number is the average of the observed copy number across the entire genome, and the genome-wide ploidy level (state) is the CN state ('diploid', 'triploid', 'tetraploid') that harbors maximum percentage of genomic segments. Both features can be extracted accurately from the copy number profile that is computed by a CNA detection tool. However, most of the copy number detection tools for cancer cell lines with abnormal karyotype (such as FREEC [45], CLImAT [51], and AMYCNE [46]) either require the information of whole-genome ploidy as input or estimate it using the SNP information. Therefore, a bunch of computational tools has been developed for estimating the whole-genome ploidy level or number.

As explained in Chapter 1, deviation from 'normal' karyotype (two copies of each chromosome), is a common hallmark of cancer cells [22, 23] due to the accumulation of whole chromosomal or segmental aneuploidies. The chromosomal aneuploidy and microscopic-level segmental gain or loss are classically analyzed by karyotyping (imaging) technique [95]. Whole-genome ploidy level with modal chromosome number has been routinely characterized for most cancer cell lines using imaging techniques by individual labs and cell repositories [96-98]. However, as illustrated in Figure 1.3 earlier, cancer cells undergo clonal evolution over time and their karyotypes may change even under controlled culture environment [4]. Therefore, same cell line at different labs may exhibit different karyotypes over time and the established karyotype ploidy characterization may not be stable [4]. Moreover, karyotyping is a low-resolution technique. It can only detect genomic aberrations at microscopic level (e.g. whole chromosome or large segmental aneuploidy). Therefore, standard karyotype cannot distinguish between two genomes with normal chromosome complement but harboring different submicroscopic copy number amplifications or deletions.

The dynamics of these microscopic and submicroscopic copy number alterations represent timestamp 'barcode' of aneuploidy spectrum of cancer cells [22, 23]. The advent of NGS-based techniques allowed high-resolution estimation of the complete repertoire of aneuploidy at different structural levels [74, 75]. Most of the ploidy estimation tools encode (characterize) the aneuploidy information as the change of whole-genome ploidy level ('diploid', 'triploid', 'tetraploid') assuming that genome gain or loss a complete/near-complete set of chromosomes. However, most cancer cell lines have hyperploid profiles. In such situation, whole-genome ploidy level cannot represent or capture the complete aneuploidy information. For example, it cannot capture the percentages of genomic alterations at each copy number state. The majority of NGS-based tools for ploidy estimation utilize SNPs, except nQuire which uses the RD signal.

2.2.1 SNP-based methods

Most whole-genome ploidy tools compute ploidy levels by assessing the distribution of allele frequency at biallelic single-nucleotide polymorphisms [51-53]. They estimate the whole-genome ploidy-level by assuming different ratios of biallelic SNPs at different ploidy levels. For example, ConPADE [53] have been developed for the ploidy estimation in highly polyploid plant genomes. It utilizes a probabilistic method that estimates the ploidy based on the relative proportions of alleles in heterozygous positions. For a diploid region, all heterozygous positions will display the two alleles in a 1:1 ratio. Similarly, alleles in a triploid region can be present in 2:1 or 1:2 ratios. In general, the number of heterozygous possibilities of ploidy level (p) is $p - 1$. Based on these assumptions, it can estimate the whole-genome ploidy level, even for $p > 10$. Another SNP-based method, ploidyNGS [52] visually analyzes the distributions of SNPs at biallelic heterozygous sites to estimate the whole-genome ploidy level (p) up to heptaploid ($p = 7$). Although these tools can identify the whole-genome ploidy level of highly polyploid genomes, their assessment of changes at the single nucleotide level requires very high coverage data. As a result, they were successfully applied to small genomes like yeast but are less adaptable in the human scenario, especially for low-coverage sequencing data. Additionally, the majority of human cancer cell lines

usually have diploid, triploid, and tetraploid ploidy level. Therefore, there is an urgent need for estimating whole-genome ploidy level from the low-coverage data. In case of cancer cell lines, these tools can be built based on the assumption that whole-genome ploidy level ranges from diploid to tetraploid.

2.2.2 RD-based methods

Recently, **nQuire** [99] has been developed to estimate the whole-genome ploidy level using the RD signal. To the best of our knowledge, nQuire is the first tool that models the RD signal for ploidy estimation, independently from the SNP analysis. It was applied to estimate the ploidy level of small yeast genomes from very high-coverage data (> 200x). In nQuire pipeline, the RD signal is first computed from the short sequencing reads. Then, it models the RD frequency distribution as Gaussian Mixture Model (GMM). The nQuire utilizes three pre-determined distributions, with fixed contribution from each ploidy state, under diploid, triploid, and tetraploid assumptions. It computes the log-likelihoods between the RD frequency distribution and each model-based pre-determined distribution. Finally, the fixed model with the smallest log-likelihoods is chosen as the true ploidy level (diploid, triploid, or tetraploid). The methodology of ploidy-based distribution fitting is novel, and theoretically it can identify the whole-genome ploidy level from low-coverage data. However, ploidy assessment based on fitting to fixed models is not always true for cancer cell lines. As explained earlier, cancer cell lines harbor variable combinations of ploidy levels that can be modeled using only three distribution. Additionally, due to the overdispersion at low-coverage data, RD signal should be re-binned at large bin sizes before fitting the pre-defined distributions.

2.3 Contact-map correction and CNA detection from Hi-C/3C-seq data

Last 15 years has witnessed extensive use of high-throughput chromosome conformation capture (3C)-based techniques such as Hi-C [\[56\]](#) for understanding genome-wide three-dimensional (3D) chromatin topology by identifying chromatin contacts between genomic regions. These 3C-related methods are a collection of molecular techniques to analyze the three-dimensional (3D) chromatin interactions inside the cell nucleus [\[100\]](#). Briefly, the conventional 3C method [\[55\]](#) has been modified and combined with next-generation sequencing (NGS) to interrogate the chromatin interactions of genomic loci at different length scales [\[56, 57, 100\]](#). In 2009, the genome-wide high-throughput Hi-C technique was developed that confirmed the chromosome territory hypothesis and established the hierarchical organization of our genome into A/B compartments [\[56\]](#) and topologically-associating domains (TADs) [\[56, 101\]](#). TADs are linear units of chromatin that fold as discrete 3D neighborhood that organizes physical proximity between genomic loci such as enhancers and gene promoters to regulate their expression [\[102\]](#). Parallely, 3C-seq technique has been developed as a simple and straight-forward experimental protocol to study the genome-wide chromatin interactions. In this approach, a conventional 3C library is prepared first and then the library is sonicated, size-selected and subsequently sequenced using NGS platform [\[57, 103, 104\]](#).

Ideally, Hi-C/3C-seq library should include paired-end reads which belong to two different restriction fragments and each read (one side of the paired read) has to be near a restriction site. Also, all loci are assumed to be equally observed in the resultant Hi-C/3C-seq library [\[105\]](#). However, this is not the usual situation due to many reasons such as (1) the variability of restriction fragment lengths (effective length) which may affect the frequency of each locus in the library; (2) genome biases such as GC-content and mappability; (3) CNAs that change the interaction frequencies between loci; (4) non-specific cutting by restriction enzymes; and (5)

ligation-related issues which may produce either un-ligated fragment ("dangling end") or a circularized ligated fragment ("self-circle") [105]. These biases may result in inaccurate interpretation of 3D genome organization and identification of TADs. Therefore, computational filtering and normalization methods are applied at fragment and loci levels to correct the data before analyzing and interpreting Hi-C/3C-seq data.

In the next subsection, we provide **an overview of the available methods for correcting the Hi-C contact frequencies as well as for detecting CNA from Hi-C data**. For the contact-map normalization approaches, we mainly discuss the impact of copy number events on their performance. In Chapter 5, **we analyze their performance in comparison to HiCNAttra by applying them on cancer cell line datasets with abnormal karyotype**.

2.3.1 Contact map normalization

Many normalization procedures have been developed to correct the Hi-C contact frequencies from (1) sequencing biases such as GC-content and mappability, (2) experiment-associated effective length bias that represents the variable lengths of restriction fragments, and (3) CNA-associated bias. Correction of the Hi-C interaction matrix is the most challenging task in the Hi-C pipeline due to concurrent presense of all these biases, espically for cancer cell lines. The majority of Hi-C correction methods [106-114] are single-sample methods, except diffHic [115] that utilizes a scaling or non-linear normalization method to correct two Hi-C samples concurrently before identifying differential genomic interaction between them.

Most correction methods [106-111] are initially introduced to remove the impact of sequencing biases (GC-content and mappability) and effective length bias on the Hi-C interaction matrix of normal cell lines. They gave less attention to the CNA-driven bias since normal genomes are usually free from large-scale copy number alterations. Implicit methods correct the Hi-C contact map as matrix-balancing problem assuming that genomic regions of equal size should have similar coverage [106-109]. On the other hand, explicit methods normalize the contact matrix by

modelling the relationship between the contact frequencies and the known systematic biases introduced primarily by GC content, restriction fragment length and mappability biases [110, 111]. Starting in 2016, the chromosome-adjusted Iterative Correction Bias method (caICB) [112], OneD [113], and implicit-based CNV-Adjusted Iterative Correction (CAIC) [114] have been developed for correcting the CNA-driven bias as well as the inherited systematic biases (GC-content, mappability, and effective length). They mainly adapted the pipeline of the existing implicit or explicit tools [106, 111] by integrating a CNA correction module.

2.3.1.1 Implicit methods

Implicit methods usually model the Hi-C correction as a matrix-balancing problem [106-109]. They aim to balance the interaction matrix assuming that each bin should be observed approximately the same number of times in the experiment, regardless of the sources of biases. For instance, The iterative correction and eigenvector decomposition (ICE) method [106] has been commonly used to implicitly correct the contact maps of most Hi-C datasets due to its ease of application and low computational time. It applies an iterative correction (IC) approach, which does not require a priori knowledge of the sources of biases, to iteratively correct for all possible biases till it reaches a tolerance threshold. Finally, it generates a bias vector that converts the raw interaction matrix to a normalized matrix. Computationally, at high-resolution, the size of interaction matrices become huge and the interaction frequencies become sparse given the same number of NGS reads. So, ICE method requires a massive amount of memory space since it applies the IC approach on the whole matrix simultaneously. Therefore, Hi-Corrector [109] is proposed as a scalable version of iterative correction [106]. They divide the interaction matrix into n sub-matrices and apply the iterative correction method on these sub-matrices sequentially.

Theoretically, ICE-based methods may indirectly account for the effects of CNA-associated biases on the interaction matrix and they can be used for correcting Hi-C data of cancer cell lines. however, recent analyses demonstrated that ICE method is not suitable for cancer cells that are plagued with widespread CNAs. ICE has been shown to result in overcompensation of original

CNA bias leading to distortion of the interaction frequency signal [113, 114]. This can be also attributed to the usage of empirical conditions for filtering the contact frequencies without any strong biological basis. For example, hiclib [106] removes reads coming from the top 0.5% of the most frequently detected restriction fragments and removes the 1% poor bins with lowest abundance. Also, the convergence of their techniques depends on empirical conditions such as tolerance threshold. These thresholds may be reasonable for normal cell lines in the absence of structural variations. However, they cannot be used for cancer cell lines where the contact frequencies are influenced significantly by copy number variations. For example, higher abundance contact frequencies may result from copy number gain, not from sequencing/protocol systematic bias.

2.3.1.2 Explicit methods

In contrast to the implicit approaches which correct the contact frequencies without any prior knowledge of the sources of biases, explicit approaches [110, 111] aim to estimate and normalize the impacts of known systematic bias sources including the mappability, GC content, and restriction fragment length. In 2011, Yaffe and Tanay showed that these systematic biases dramatically affect the read counts in Hi-C interaction matrices [110]. To correct the effects of these systematic biases, they utilize a non-parametric probabilistic model that explicitly models the probability of having interaction between two genomic loci. In 2012, HiCNorm utilizes a generalized linear model with Poisson distribution to correct the systematic biases in Hi-C contact maps [111]. Poisson regression is used to capture the relationship between interaction frequencies and the systematic biases. Both explicit approaches (Yaffe and HiCNorm) resulted in similar normalized contact frequencies of normal cell lines, however, HiCnorm was able to reduce the computation cost significantly [111]. More importantly, both approaches do not account for the data-driven (data dependent) bias such as copy number alterations. CNAs are different for each cell line and they have to be estimated from the Hi-C data itself. This CNA bias will not impact the Hi-C contact matrix of normal cell lines significantly in the absence of large structural

variations. In contrast, cancer cell lines harbor widespread copy number change events that have significant impact on the interaction frequencies. Therefore, Hi-C data of cancer cell lines should be corrected for CNA bias as well as the systematic biases.

2.3.1.3 CNA-driven methods

A crude attempt to normalize CNA bias has been made by calCB [112]. The calCB applies a uniform copy number for individual chromosome ignoring the copious presence of segmental aneuploidy and focal CNAs. Two methods have been recently introduced to address the CNA bias effects on Hi-C contact maps. First, the explicit-based approach, OneD [113] normalizes the contact map sequentially for the sample-independent systematic biases (such as GC content, fragment length and mappability) followed by dividing the contact frequencies between two genomic loci by the product of their copy numbers. In OneD pipeline, it first utilizes an additive model based on the negative binomial distribution for normalizing the systematic biases. Then, CNAs are estimated by hidden Markov model (HMM) of eight states. Finally, it divides the contact frequencies between two genomic loci by the estimated CN of these loci assuming a perfect correlation between the CNAs and contact frequencies. However, any additive model cannot estimate the accurate contribution of each systematic bias on the interaction frequencies in case of the overwhelming presence of large-scale copy number variations. Briefly, the regression model for the variable (interaction frequency f) should include all predictors (GC-content g , mappability m , effective length e , and copy number c) that affect the variable's value: $f = a_1 * g + a_2 * m + a_3 * e + a_4 * c$, where a_1, a_2, a_3 and a_4 are the regression coefficients. By inclusion of all predictors, the regression coefficients can be estimated accurately. However, in OneD approach, they used the Generalized Additive model (GAM) based on three predictors only (GC-content g , mappability m , effective length e): $f = a_1 * g + a_2 * m + a_3 * e$. This leads in inaccurate estimation of the regression model coefficient ($a_1 - a_3$), especially in case of abnormal karyotype when copy number is the dominant bias source. Therefore, it cannot normalize the contact frequencies accurately and completely from both the CNA and other systematic biases.

Instead, CNA tracks should be integrated with other systematic-biases such as GC-content and mappability in the same additive model for accurate normalization of Hi-C interaction matrices. On the other hand, the implicit-based CAIC (CNV-Adjusted Iterative Correction) [114] method are developed based on the ICE approach. In CAIC pipeline, CNAs are first estimated using a pruned dynamic programming algorithm [116] Then, it performs a local iterative correction that assumes all bins of genomic regions with same copy number should be observed approximately the same number of times. Therefore, in this step, it aims to correct the impacts of the systematic biases only, not the CNA bias. Then, CAIC corrects the CNA-bias assuming the raw interaction count between two loci as product of their copy numbers. However, as an ICE-based approach, CAIC may lead to overcompensation of the CNA bias and distortion of the Hi-C interaction frequencies.

2.3.2 Hi-C copy number detection

While most of CNA-related tools for Hi-C data aim to correct the contact frequencies from the CNA bias, HiCnv [117], to the best of our knowledge, is the only available tool that have been developed for identifying copy number events from Hi-C data. In HiCnv pipeline, RD signal (30-kb bin) is first computed from the Hi-C interaction matrix using only the valid pairs (paired-end reads that are mapped to two different restriction fragments). Then, it utilizes HMM for segmenting the RD signal. Afterward, in case of diploid cell, it compares the mean of these segments to the average mean of all segments to identify the significant amplifications and deletions. For higher ploidy, it requires additional information of a reference diploid chromosome from WGS or karyotyping analysis to determine the true copy number of each chromosome and accurately infer the CN of each segment. However, as explained earlier, the karyotype of cancer cell lines may change over time and, thus, this information may not be always true. Additionally, HiCnv can only identify large-scale copy number variations (LCVs with size > 1 Mb) since it uses large bin size of 30 kb.

2.4 Summary

We presented a summary of the NGS-based CNA-associated computational tools for identifying and correcting CNA-associated features. They can be divided into two main categories (1) RD-based copy number detection from WGS data and (2) contact map correction from Hi-C data. For each tool, we analyzed its computational pipeline, methods, and assumptions with respect to the applicability on cancer cell lines with abnormal karyotype. In next chapters, we will describe and discuss our proposed methods (AStra, CNATra, HiCNATra) and evaluate their performance in comparison to these computational tools.

Chapter 3

AStra

Aneuploidy Spectrum (detection) through read depth analysis

Cancer cell lines are the cornerstone of cancer research and drug screening. Intriguingly, established cancer cell lines undergo continuous genetic alterations over passaging time and culture conditions to display significant clonal variations (strain differences) [4, 33, 34, 118, 119]. Additionally, in laboratory conditions cancer cell lines face the risk of contamination with other cell lines (intraspecies contamination) [37, 38]. 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 the same cell line [120].

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 SNP profiling [120]. 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 NGS technology, augmented by development of numerous computational methods, has paved the way for comprehensive characterization of human cell lines [121]. Recent

explosion in publicly-available genome-wide reads of cancer cell lines has provided powerful method to identify different strains. These genome-wide reads are available from different sources such as WGS data, ChIP-seq ‘input control’ data, and Hi-C/3C-seq data. The ChIP-seq ‘input control’ and Hi-C/3C-seq datasets can be utilized to analyze the entire genome similar to WGS. This WGS can address several quality assessment metrics for cell line authentication, such as intra- and interspecies contamination, chromosomal alterations, genetic drift, etc. in a genome-wide manner [120]. The availability of WGS data have been successfully used for CNA detection as well as ploidy, mutation and SNP analyses [34, 118, 122]. These WGS-derived genetic profiles can reproduce the results from 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 the aneuploidy profiling as WGS-based karyotyping (digital karyotyping) of cancer cells. Aneuploidy profile provides the copy number information of every locus of the genome. Additionally, the aneuploidy spectrum is the percentage representation of genomic segments with particular CN state. We have developed AStra (Aneuploidy Spectrum (detection) through read depth analysis), a Python-based software for de novo estimation of aneuploidy profile which does not require any control/reference sample or any prior information about the cell line. AStra evaluation of 27 strains of MCF7 revealed strain-specific differences in their aneuploidy signatures. Similarly, analysis of additional 19 cancer cell lines revealed that aneuploidy signatures are distinct for every cell line and represent their characteristic feature. Therefore, aneuploidy profile and spectrum are rapid in silico methods to authenticate the cell lines solely from NGS-based experimental data. Comparison of aneuploidy profiles from NGS data produced by different labs will provide a quick quality control check for genetic drift or strain differences. Additionally, AStra extracts the CN-associated features such as the whole-genome ploidy level and whole-genome ploidy number that can be used as inputs for single-sample CNA detection tools.

3.1 Methods

3.1.1 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 can robustly estimate 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 prior knowledge of karyotype information of most cancer cell lines.

AStra first scans the RD frequency distribution to approximately identify the candidate CN reference. For that, we employ 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 (Figure 3.1a). We utilize each model (m) separately to find the model-specific CN reference (CN_m) in two 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 [123], to find the model-associated CN reference (CN_m) by scanning the CN reference interval (1.9N to 2.1N) around the CN_{mi} (Figure 3.1b). 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 (CN₁₋₆), that yields the lowest CE (Figure 3.1b). At the end, we obtain a collection 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. Finally, we extract the CN-associated features such as whole-genome ploidy level and ploidy number.

(a)

(b)

Figure 3.1: 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-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_{1-6}) are the RD value that best allocates the RD segments around integer CN states. Final CN reference is selected with minimum CE.

3.1.1.1 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 (Figure 3.1a):

$$f(x) = \sum_i c * \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right), x \geq 0 \quad (3.1)$$

such as

$$\int_0^\infty f(x).dx = 1 \quad (3.2)$$

$$\int_0^\infty \sum_i c * \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right).dx = 1 \quad (3.3)$$

$$\sum_i c * \int_0^\infty \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right).dx = 1 \quad (3.4)$$

$$\sum_i c = 1 \quad (3.5)$$

$$c = \frac{1}{\sum_i 1}, \quad (3.6)$$

where i is the common CN state (2, 3, 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).

3.1.1.2 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 $f(x)$ (Figure 3.1a). The RD frequency distribution of A427 cell line is shown in Figure 3.1a as an example of the input RD frequency distribution $r(x)$. 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:

$$R_k = \sum_{j=s}^e r(j) * f_k(j), \quad k \in [s, e] \quad (3.7)$$

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

3.1.1.3 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 (Figure 3.2a):

$$CN_i = RD_i * \frac{2}{CNR} \quad (3.8)$$

$$S_i = \text{round-to-nearest}(CN_i) \quad (3.9)$$

$$E_i = |S_i - CN_i| \quad (3.10)$$

$$CE_j = \sum_i E_i * W_i, \quad S_i == j \quad (3.11)$$

$$CE = \sum_{j=1}^n CE_j, \quad (3.12)$$

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), W_i is the width of the segment i , and CE_j is the centralization error of segments of CN state j . Figure 3.2a shows an illustrative cartoon example to compute the CE. 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 denotes the error weight of each segment (E_i) (the difference between the CN of that segment and the nearest integer CN state). For each candidate CN reference, the centralization error is computed as the summation of the product of each RD segment width (W_i) and its error (E_i).

3.1.1.4 Characterization of the aneuploidy spectrum

The aneuploidy profile, with minimum CE, is decoded into the aneuploidy spectrum which contains the percentage of genome segments in each CN state/interval. 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 (Figure 3.2b):

$$CS = \frac{\sum_i W_i, |S_i - CN_i| \leq 0.25}{\sum_i W_i} \quad (3.13)$$

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 . An illustrative example of calculating the CS is provided in Figure 3.2b. 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. Sequencing data with high CS are more reliable for CNA analysis. Second, we compute the whole-genome ploidy number as average of the observed copy number across the entire genome. Third, we infer the genome-wide ploidy level ('diploid', 'triploid', 'tetraploid') as the CN state harboring maximum percentage of genomic segments.

3.1.2 Data availability

In addition to the simulated data that represent genomes of various levels of aneuploidy, we validated AStra approach using 27 strains of MCF7 breast cancer cell line with reported copy number profiles [4]. Similarly, we evaluated the accuracy of AStra using additional datasets of 19 cancer cell lines as well as three WGS data from the 1000 Genomes Project (Table 3.1). In this table, GEO accession number indicates the sequencing data source of the cell line in the Gene Expression Omnibus (GEO) database (a public functional genomics data). American Type Culture Collection (ATCC) modal chromosome number is number of chromosomes reported by karyotyping for each cell line. Whole-genome ploidy level is the karyotype of the same cell line, reported by published study. Coverage is the average number that each nucleotide in the genome is sequenced. All these data were aligned using Bowtie 2 software [124] to human genome assembly GRCh37 (hg19) and their BAM-formatted files were used as input for AStra.

(a) Centralization error

(b) Centralization score

Figure 3.2: Visual illustration to compute centralization error (CE) and centralization score (CS).

(a) Illustration to compute centralization error which allows to obtain the best-fit aneuploidy profile. Each blue line represents a genomic segment while the red arrow is the error weight of this segment (the difference between the CN of that segment and the nearest CN state). (c) Illustration to calculate the centralization score (CS). Each blue line represents a segment with estimated CN within $0.25N$ from the nearest CN state whereas each red line represents a segment with estimated CN $>0.25N$ from the nearest CN state.

Table 3.1: Datasets for evaluating the performance of AStra.

Cell Line	GEO Accession number	ATCC modal Chromosome number	Whole-genome ploidy level	Coverage (x)
PC3	GSM1917731, GSM1917732 (GSE74351)	62	Near-triploid [97]	0.478754
T47D	GSM589247	65	Near-triploid [96]	0.09760
MDA-MB-468	GSM2258900, GSM2258901 (GSE85158)	64 (range = 60 to 67)	Near-triploid [98]	0.602257
MDA-MB-231	GSM2258864, GSM2258865 (GSE85158)	64 (range = 52 to 68)	Near-triploid [125]	1.018764
K562	GSM2054698	NA	Hypertriploid [126]	2.3680593
SK-BR-3	GSM2258810, GSM2258811 (GSE85158)	84	hypertriploid to hypotetraploid [96]	1.202806
22Rv1	ENCSR933UGQ, ENCFF745ZXO, ENCFF868MWO	NA	Hyperdiploid [127]	2.19145
697	GSM2107713	46 (range = 45 to 48)	Pseudodiploid [128]	1.89361
SK-N-SH	GSM2664362	47	Hyperdiploid [129]	1.81939
CAL-51	GSM1693027	NA	Diploid [98]	0.63860
VCaP	ENCFF273KUS, ENCFF466WDC	NA	near-triploid [127]	1.784808
MOLT-4	GSM2090932 (GSE79288)	95	Hypotetraploid [130]	1.524776
LNCaP	GSM1902617	84	Hypotetraploid [131]	0.733167
SUM159	GSM1553140	NA	Diploid [98]	1.205006
NCI-H82	GSM1526702	58	Near-triploid [132]	0.334848
IMR-32	GSM2664328	49	Near-diploid [133]	1.482567
CHP-212	GSM2664356	NA	Diploid [134]	1.855291
HepG2	ENCSR055XHIN	55 (range = 50 to 60)	Near-diploid [135]	1.505710
A427	DRX015190	NA	Hypotriploid to hypertriploid [24]	10.59587
HG00119	1000 Genomes Project	NA	Diploid [92]	4.601165
HG00096	1000 Genomes Project	NA	Diploid [92]	28.7330
HG01879	1000 Genomes Project	NA	Diploid [92]	28.67377

3.2 Results

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

Aneuploidy is associated with cell proliferation defects as well as uncontrolled cellular growth and cancer [136-139]. Aneuploidy is a hallmark of human cancers reported in 90% of solid tumors and 75% of blood cancers [140-142]. Chromosome missegregation and instability have been implicated for manifestation of complex genetic makeup of cancer cells and their evolution [19, 143, 144]. Consequently, the cancer cell lines show widespread aneuploidy at whole-chromosome and segment levels. Interestingly, cancer cell lines spontaneously acquire new chromosomal alterations at a high rate even under controlled culture condition [145] resulting in their phenotypic evolution and strain difference [4]. Hence, same cell line cultured in different labs can exhibit changes in karyogram and mutational landscape [4, 34, 118, 119].

Currently, NGS-based experimentation improved upon the conventional genetic techniques for unbiased, reliable and high-resolution results. The explosion of NGS-based data 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 available WGS data of 27 strains of the common breast cancer cell line, MCF7 [4] 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 estimated from low-coverage (<1x) data.

3.2.2 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). Aneuploidy spectrum, on the other hand, is the normalized RD signal frequency distribution that summarizes the percentage contribution of genomic bins with different CN states. For that, AStra provides *de novo* computation of the aneuploidy profile and aneuploidy spectrum without the aid of control/reference sample or prior knowledge of karyotype of input sample.

We evaluated the AStra aneuploidy profiles of the 27 MCF7 strains against their CNA profiles reported earlier using a panel of normal samples as reference [4]. AStra successfully identified the correct aneuploidy profiles of all strains as interpreted in the original study [4] without inputting any additional information or reference control. Visually, both the aneuploidy profile and aneuploidy spectrum demonstrated remarkable differences between MCF7 strains (Figure 3.3 and Figure 3.4). 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 (Figure 3.3). 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.

Figure 3.3: 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 ($CN \leq 2$: black, $3 \leq CN \leq 4$: 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 3.4: The genome-wide aneuploidy profile for all MCF7 strains (except C, G, H, P and S).

Notably, the cytogenetic analysis of MCF7 metaphase chromosome spreads showed that their genome-wide ploidy level is hypertriploid to hypotetraploid [146]. 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 3.3 and Figure 3.4). 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 also emphasized that these 27 MCF7 strains have different gene expression profiles and displayed differential sensitivity to cancer drug treatments [4]. Their phenotypic differences can be attributed to the variations in aneuploidy profiles [145]. Therefore, we recommend aneuploidy profiling for *in silico* digital karyotyping and cell line authentication. This will allow reliable integration of different datasets generated by different laboratories derived from same cell line and their downstream interpretation.

3.2.3 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 (Table 3.1). For the simulated data, we manipulate the WGS reads of HG00119 (1000 Genomes Project sample of diploid male) to artificially introduce large-scale copy number gain/loss, preserving the inherent systematic biases of the WGS data. **The detailed explanation of the pipeline for simulated data is provided in Chapter 4.** 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. The original reads and the spiked-in artificial reads were merged into a new BAM file and used as input for AStra.

Figure 3.5: 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 ≤ 1 : black, CN = 2: blue, CN ≥ 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 manipulation, are estimated correctly from all neo-genomes.

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 replaced the original chr 2 only with artificial chr2 keeping the rest of the chromosomes of HG00119 intact. 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 (Figure 3.5). 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 19 cancer cell line data and three 1000 Genomes Project samples of varying genome-wide ploidy levels that were experimentally verified (Table 3.1). As demonstrated for MCF7 strains, Astra successfully identified the accurate aneuploidy spectrum of all cell lines fulfilling two conditions. First, CN states are detected as peaks of input RD signal distribution (Figure 3.6; frequency distribution histograms). Second, majority of genomic segments (Figure 3.6) have CN around the reported genome-wide ploidy level of these cells (Table 3.1).

Figure 3.6: Aneuploidy signatures of cancer cell lines.

The genome-wide aneuploidy profile (left) and aneuploidy spectrum (right) for a variety of cancer cell lines 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.

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 relevant genome-wide NGS datasets without additional cost of performing targeted WGS. Additionally, as a general rule, in the aneuploidy spectrum of any cell line, CN states should be detected as peaks of input RD signal distribution. Therefore, in most cases, user can evaluate Astra’s results by visually exploring the aneuploidy spectrum without the knowledge of whole-genome ploidy level of the input cell line. 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 of low-coverage (<3x) data in less than 3 minutes and high-coverage (~28x) in about 15 minutes.

3.2.4 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 [147]. 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 (Figure 3.7). 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 (multimodal distribution and input RD signal distribution). 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.

Figure 3.7: 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 achieves the maximum overlap between the input RD frequency distribution and the multimodal distribution.

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 [148] and COSMIC [149] database information. Therefore, Astra devised 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 underlying reason for choosing 6 prospective models (M1-M6) (Figure 3.1a).

To illustrate the advantage of Astra approach over RDS and MMDS methods, we applied all these methods 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 (Figure 3.8 and Figure 3.9). In general, we observed that CN reference computed by Astra, RDS and MMDS methods are identical for simulated datasets (Figure 3.8). In contrast, CN reference varies considerably across methods in many cases for cancer datasets and MCF7 strains (Figure 3.9). 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). 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). Interestingly, we found that CE is a non-convex function and has many local minima (Figure 3.8 and Figure 3.9). Moreover, the CN reference corresponding to global minimum of the CE may not be always correct. This strongly confirmed our hypothesis that scanning of the complete RD interval may result in inaccurate CN reference, especially in cases where the RD frequency distribution shows overlapped peaks. Therefore, the CN reference may be assigned wrongly as the global minimum or any other inaccurate local minima.

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 and allocated toward the CN states (e.g. in case of simulated data in Figure 3.8), all these three methods can accurately estimate the CNs of these segments. On the other hand, the discordance between three methods, in case of real cancer datasets, may be attributed to the differences in their centralization scores that measure the localization of genomic segments around CN states, which are reflected as the degree of separation between different copy number states (peaks) in RD frequency distribution.

Figure 3.8: Centralization error as a function of copy number reference in some simulated neo genomes.

The CE for all CN candidates are plotted by scanning the entire range of values of the RD signal. 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. All three methods accurately computed the CN reference.

(a) Cancer cell lines

(b) MCF7 Strains

Figure 3.9: Centralization error as a function of CN 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.

3.2.5 Centralization score for assessment of the quality of sequencing data for copy number analysis

Presence of hyperploidy and segmental aneuploidy in a cancer cell line result in a CN profile which contains a collection of genomic segments at different CN states. Ideally, in case of high-coverage data of a homogeneous cancer cell line, these genomic segments are expected to be well-separated and centralized at integer/near integer CN states. However, this condition depends on the other source of biases in cancer sequencing data. The main bias originates from the purity of the sample, clonal variations, and heterogeneity of the cancer cells/samples which adversely affects peak separation. In addition, coverage signal of cancer cells may be affected by wave artifacts which may contribute to signal variations and peak separation [60]. In low-coverage data, the effects of heterogeneity and wave artifacts are magnified due to the sparseness of reads representing each region of the genome. Consequently, we introduced the centralization score to evaluate the degree of separation between different copy number states of the cancer data.

We analyzed the CS of the 27 MCF7 strains and 19 cancer cell lines. Although all 27 strains have similar coverage ($\sim 0.5x$), their CS vary remarkably (Figure 3.10a). Interestingly, we noticed that strains (A, F, S, T, U, W, Y and Z) whose CN references were wrongly identified by both RDS and MMDS methods generally have lower CS (Figure 3.10a). We observed similar trend in case of T47D, K562, 22Rv1 and CAL-51 cancer cell lines (Figure 3.10b). 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 data coverage (Figure 3.10). We found that CS is weakly correlated with the coverage of the data (Spearman correlation = -0.45 for MCF7 strains and 0.39 for other cancer cell lines). Similarly, CS is weakly correlated with the whole-genome ploidy number (Spearman correlation = -0.38 for MCF7 strains and -0.24 for other cancer cell lines). These weak correlations suggest that data coverage and whole-genome ploidy number are not the only keys for allocating the RD segments around integer CN states.

(a) MCF7 Strains

(b) Cancer cell lines/1000 Genomes Project samples

Figure 3.10: 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).

Alternatively, CS provides an empirical measure of the separation of peaks of RD frequency distribution as well as the centralization of genomic segments around CN states. We found that sequencing data with higher CS are less affected by the biases of cancer genomes. For example, MCF7 strain S with low CS (62.5%) shows low separation of RD distribution peaks as demonstrated by their coverage plots and frequency distributions (Figure 3.3). Its RD frequency distribution show overlapped peaks and its coverage plots are affected by wave artifacts. In contrast, the RD frequency distributions of strains C, G, H, and P with high CS of 94.5%, 96.4%, 94.4%, and 92.6% respectively (Figure 3.3), show separable and distinct peaks centralized at the CN states. Overall, the CS reflects a unique feature of the WGS data. We believe that CS may hold clue for assessment of the quality of the sample where low CS may be attributed to sample- or sequencing-driven bias such as sample heterogeneity, inter-tumor contamination, overdispersion, etc.

3.2.6 AStra outperforms nQuire for estimating the whole-genome ploidy in low-coverage WGS data

We have also benchmarked our tool against nQuire by applying it on 10 datasets of cancer cell lines with known karyotype information and three 1000 Genomes Project datasets (SK-N-SH, CAL-51, SUM159, IMR-32, CHP-212, HepG2, HG00119, HG00096, HG01879, VCaP, A427, MOLT-4, and LNCaP) (Table 3.1). As explained in Chapter 2, nQuire is the only ploidy tool that models the RD signal for whole-genome ploidy level estimation which makes it more applicable for low-coverage data. The nQuire employs three RD frequency distributions to model the RD signal under diploid, triploid, and tetraploid assumptions only. Therefore, for a fair comparison, we keep only cancer cell lines with whole-genome ploidy level of diploid/near-diploid, triploid/near-triploid, and tetraploid/near-tetraploid. Their RD frequency distributions have a major peak at the ploidy CN state, a ratio between genomic segments at major CN state and any other states is at least 2, and CS > 80%. This resulted in 9 diploid (SK-N-SH, CAL-51, SUM159, IMR-32, CHP-212, HepG2, HG00119, HG00096, HG01879), 2 triploid (VCaP and A427) and 2 tetraploid (MOLT-4

and LNCaP) datasets. We found that nQuire estimated the genome-wide ploidy level of all cells as tetraploid. In comparison, AStra correctly estimates the whole-genome ploidy level based on the CN state with the highest percentage of genomic segments.

Although both tools utilize multimodal distribution for estimating the whole-genome aneuploidy, AStra performed better than nQuire. This may be attributed to the fact that AStra evaluates the aneuploidy profile based on the allocation of genomic segments at CN states. AStra's models are used for narrowing down the interval of CN reference candidates, followed by searching for the CN reference that best allocates genomic segments around CN states. This allow AStra to count for any percentage of segmental aneuploidy, and therefore they can identify hyperploid profile from low-coverage data of cancer cells with complex genetic makeup. In contrast, nQuire evaluates the aneuploidy profiles based on the fitting of the RD signal to pre-defined distributions under each ploidy assumption. It assumes unimodal distribution for diploid cells ignoring the existence of any segmental aneuploidy change. Similarly, nQuire triploid distribution assumes that the data distribution can have only two peaks $2N$ and $3N$, and so on.

We attempted to apply PloidyNGS, the most-recent SNP-based ploidy estimation tool, on same datasets, however, it failed to execute due to memory outage (RAM = 32GB). We checked the GitHub repository of PloidyNGS [<https://github.com/diriano/ploidyNGS>] and found that this tool is mainly applicable for small genomes such as yeast.

3.2.7 Aneuploidy spectrum for calibration of single-sample CN detection tools

Characterization of the aneuploidy profile forms the basis of CNA detection for single-sample cancer data [41]. As demonstrated in Chapter 2, most single-sample CN detection tools, which do not rely on matched/control reference, are built on some underlying assumptions for modeling the RD signal. Additionally, they require additional information based on the prior knowledge of the cell line. For example, FREEC [45] and AMYCNE [46] use the whole-genome ploidy number to define the CN states while ReadDepth [49] uses gain/loss percentage to adjust the underlying

Poisson/negative binomial distribution. Similarly, CNVnator [44] assumes that 99% of the genome is CNA-free. However, this information changes across cell lines and between same cell line data obtained from different labs owing to clonal variability as presented in this study (Figure 3.3 and Figure 3.6). Therefore, knowledge of aneuploidy spectrum is fundamental for selecting and calibrating CN detection tools. Any CNA-associated feature can be easily extracted from the aneuploidy spectrum of the input WGS data. Consequently, AStra provides the whole-genome ploidy number and the percentage of genomic regions per CN state, which can be used as inputs for copy number detection tools.

3.3 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 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 cancer cell lines undergo genetic drift with continuous passaging. Therefore, these 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 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 [113, 114, 117]. Therefore, the free access of NGS raw reads from hundreds of cancer cell lines has paved the way for easy

sharing and as well as integrative analysis of data. However, before such analysis caution should be employed that the data are generated from authenticated cell lines and are of good quality.

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 there is an urgent need to report cell line authentication data for every study, unfortunately, 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/clones 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.

Chapter 4

CNAtra

Copy Number Alteration (detection) through read depth aalysis

DNA copy number alterations (also commonly referred to as copy number variation) are genetic aberrations that lead to the change in the number of copies of DNA in comparison to the reference genome. Cancer genomes are usually plagued with copy number alterations (CNAs) at different length scales. These CNAs can be further classified into microscopic level large-scale copy number variations (LCVs) at Mb-size scale or submicroscopic level focal alterations (FAs) at kb-size scale [\[15, 16, 18\]](#). Both LCVs and FAs are pervasive in cancer cell lines, which serve as important pre-clinical models for cancer research, drug screening and discovery [\[150\]](#). Naturally, accurate detection of both these alteration phenomena is crucial for gaining insights on their origin and biological context. Because of their variable size distribution, the current copy number detection tools generally target a specific range of CNA size [\[43\]](#). Therefore, detection methodology needs to be tuned to identify the complete spectrum of CNAs (large-scale and focal events) and should include procedures to distinguish them.

Many computational tools utilize the RD signal to identify CNAs using a single sample of WGS data independent of the requirement of a control sample [\[41-43\]](#). Single-sample approaches are more suitable for copy number profiling of cancer cell lines since they usually do not have a matched normal sample [\[44-51\]](#). However, as explained in Chapter 2, these tools cannot capture the complete profile of copy number alterations for different reasons. First, they usually utilize a unimodal probabilistic distribution to fit the RD signal of cancer cell lines ignoring their multimodal feature due to the extensive presence of LCVs. This usually results in false calling and

inaccurate CN estimation of alteration events. Second, they do not tend to distinguish between large-scale and focal alteration events. This leads to oversegmentation of LCVs and erroneous calling of FAs. Third, most RD-based detection tools are adversely affected by low-coverage data which result in overdispersion and short-term variations such as wave artifacts [59, 63, 81]. In such scenarios, statistics-based segmentation [44, 49, 88] and CNA calling lead to either in false segmentation or missing the focal alterations. Nevertheless, even the low-coverage NGS data is a valuable resource for CNA profiling with higher sensitivity compared to array CGH data [151]. Therefore, additional challenges associated with low-coverage noisy data raise additional challenges that need to be circumvented for accurate copy number detection.

Therefore, we have developed CNAtra, a MATLAB-based hierarchical computational framework for the sensitive and robust detection of LCVs and FAs from a single sequencing sample. CNAtra is built on the fundamental assumption that most genomic regions of any cell are centralized toward copy number states of positive integer values. CNAtra empirically models the RD signal based on a multimodal distribution and estimate the CN reference. This approach allows us to define the accurate ‘interval’ of CN states which aids in identifying segmental aneuploidies (and FAs within them) in a robust manner largely unaffected by coverage, percentage of karyotype abnormality and wave artifacts. For this, we first applied a robust signal-processing technique of univariate time series to identify significant change points of the RD signal. These change points are used for assembling large segments based on their CN states (we called them iso-copy number blocks or IBs) and for identifying the candidate FAs within these IBs. In addition, for handling the overdispersion problem of low-coverage data, we incorporated coverage-based thresholding parameters beside the conventional statistical test to identify the significant FAs. CNAtra also provides an interactive platform to visualize and manually inspect the complete (genome-wide) copy number profile and accessory information for further validation, interpretation and downstream application of CNA calls. We successfully verified CNAtra results using experimentally-validated segmental aneuploidies and focal amplifications/deletions across

several cancer cell lines. We also applied CNATra to patient tumor samples and showed that CNATra successfully detected the copy number changes in a single WGS data sample without using in the absence of any matched control samples. We then benchmarked the performance of CNATra against five single-sample CN detection tools using realistic simulated data by randomly introducing copy number events in the original cancer genome. The evaluation showed the ability of CNATra to resolve complex CNA profiles into LCVs and FAs with the highest f-measure. Manual review and visualization also verified the advantage of CNATra over other single-sample tools.

4.1 Methods

From the methodology perspective, CNATra is superior for detecting both LCVs and FAs from low-coverage data due to four proposed solutions. First, CNATra utilizes multimodal distributions for computing the copy number reference and the other copy number states regardless of the extent (percentage) of chromosomal and segmental aneuploidies in the genome. Second, CNATra is the first tool to computationally introduce the notions of large-scale copy number variations and focal alterations. This divides the CNA detection problem into two optimization problems (LCV detection and FA detection). Third, we developed CN-based merging algorithm that assembles the small neighboring/adjoining segments, within same copy number intervals, into iso-copy number blocks (IBs). IBs are long contiguous segments with distinct CN state. Fourth, CNATra incorporates coverage-based thresholds that allow detection of significant FAs from the low-coverage data. Also, detection of FAs separately allows the user to set/tune the thresholding parameters to identify the focal amplifications and deletions for more specific analysis.

4.1.1 CNATra framework

CNATra has been developed to detect LCVs and FAs in cancer genomes hierarchically. CNATra is a MATLAB-based single-sample CNA discovery tool particularly adapted for low-coverage cancer genomes. CNATra comprises two modules – RD calculator and CNA caller (Figure 4.1a). In the RD

calculator, we compute the RD signal as base count frequency at 1-kb bin from the input WGS data after initial read filtering steps. One-kb bin allows us to fine-tune the tool to precisely define the boundaries for the two copy number events in case of low-coverage data. We then correct the RD signal for systematic biases due to GC content (isochore normalization) and low-mappability regions. The CNA caller module constitutes the hierarchical framework to delineate the multi-level alterations in the cancer genomes. We first compute the CN reference by fitting a multimodal distribution over the RD frequency histogram (Figure 4.1b). Then, we use a multi-step framework for detecting both LCVs and FAs (Figure 4.1c-e). This includes smoothing of RD signal using Savitzky-Golay filter (Figure 4.1c black line) and segmenting the RD signal using Modified Varri technique, followed by detection of IBs with distinct CN states (Figure 4.1d). Finally, we discover candidate peaks of ‘focal’ amplification and deletions in each CN-defined large segment/IB (Figure 4.1e).

4.1.1.1 Binning the short reads

CNAttra accepts BAM/SAM files as input sequence alignment data for RD calculation. It can handle single-end (SE), or paired-end (PE) reads. A MATLAB BioMap object is constructed from the input sequences to store the mapped short reads and compute the RD signal. These reads are filtered based on their mapping quality (MAPQ) to remove PCR-duplicated reads, partially mapped PE reads and non-uniquely mapped reads. We also remove reads mapped to chromosome Y and chromosome M (mitochondrial genome) from the downstream analysis. Subsequently, reads that pass the filtering steps are used to calculate base (nucleotide) count at 1-kb non-overlapping bin (RD signal per bin). For SE reads, each read is considered as a sequencing fragment. For PE reads, sequencing fragment is reconstructed from the start of side 1 to the end of side 2. Genomic data coverage is computed as the ratio between the summation of lengths of these sequencing fragments and the genome size.

Figure 4.1: CNAttra framework.

(a) Flowchart of CNAttra pipeline. (b) Schematic representation of multimodal distribution under different ploidy assumptions. Dotted red lines denote the boundaries of CN states (CN intervals). (c-e) A hierarchical framework for detecting both LCVs and FAs. The CNAttra approach includes smoothing of RD signal using Savitzky-Golay filter (black line) (c), followed by detection of IBs with distinct CN states (d) and then identification of FAs inside each IB (e).

4.1.1.2 Bin correction of the RD signal

CNAttra corrects the RD signal for GC-content [152] and mappability [62] biases. Based on the read length, we compute the mappability score as the fraction of unique mappable nucleotides per bin. For that, we used the Umap tracks of uniquely mappable regions [62]. RD signal is first normalized for mappability effects. The base counts per bin are divided by the mappability score of that bin. Also, bins with low-mappability scores (< minimum mappability threshold) are filtered out to prevent overcorrection. User can set the minimum mappability threshold from 0 to 1 (default 0.5). Next, we correct the RD signal for GC-content using the CNVnator approach [44]. GC ratio per bin can be calculated from the input reads, or they can be downloaded from the precomputed GC-tracks (default option) for different read-lengths [49].

4.1.1.3 Estimation of copy number reference

We utilize a multimodal distribution for computing CN reference (2N) and all other CN states (1N, 3N, 4N, ...) based on the user-defined whole-genome ploidy level of a given input data (default = 'free'). Our multimodal distribution is defined as the summation of normal distributions of different probabilities centered at different CN states (Figure 4.1b):

$$f(x) = \sum_{i=2}^6 c * w_i * \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right), x \geq 0 \quad (4.1)$$

such as

$$\int_0^{\infty} f(x).dx = 1 \quad (4.2)$$

$$\int_0^{\infty} \sum_{i=2}^6 c * w_i * \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right).dx = 1 \quad (4.3)$$

$$\sum_{i=2}^6 c * w_i * \int_0^{\infty} \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} * e^{-\frac{(x-i)^2}{2\sigma^2}} \right).dx = 1 \quad (4.4)$$

$$\sum_{i=2}^6 c * w_i = 1 \quad (4.5)$$

$$c = \frac{1}{\sum_{i=2}^6 w_i}, \quad (4.6)$$

where i is the copy-number state (2, 3, 4, 5, 6), w_i is the weight of the Gaussian distribution at state i and c is a constant for normalization of the probability distribution function. The “free

model" (Figure 4.1b) assumes that all CN states have the same weight to be presented ($w_i = 1$). For other model ('diploid', 'triploid', 'tetraploid') (Figure 4.1b), the weights w_i are $2^{|i-n|}$ where i is the CN state and n is the main ploidy state assuming that the majority of segments have CNs near this ploidy state.

Based on the user-defined ploidy level ('diploid', 'triploid', 'tetraploid' or 'free'), the CN reference is defined as the RD value that achieves the maximum matching/overlap between the RD frequency distribution $r(x)$ and the multimodal distribution $f(x)$ of this ploidy assumption. For computing the CN reference of a given cell line, RD signal is re-binned at 100 kb to make the RD frequency distribution $r(x)$ centered around or near CN states. Then, the interval of RD values $[s, e]$ is divided into m equally-spaced candidates of a CN reference. At each candidate value i , a multimodal distribution $f_i(x)$ is generated assuming that i is the CN reference (corresponding to $x = 2$ in the multimodal distribution). A rank R_i is computed as the convolution between the RD frequency distribution $r(x)$ and the multimodal distribution $f(x)$:

$$R_i = \sum_{j=s}^e r(j) * f_i(j), \quad i \in [s, e] \quad (4.7)$$

Finally, candidate value i with the maximum rank R_i is chosen as copy number reference and is used to normalize the RD signal.

4.1.1.4 Detection of candidate segments for identifying IBs/FAs

We utilized signal processing methods (Savitzky-Golay smoothing filter [153]) and Modified Varri method [154]) for over-segmenting the RD signal and detecting all change points. Savitzky-Golay smoothing filter [153] is applied to eliminate short-term variations and wave artifacts without affecting the 'sharp' signal change points. Savitzky-Golay filter is a method of data smoothing using a local least-squares polynomial approximation. Savitzky-Golay filter has advantages over other smoothing filters (such as moving average) since it tends to preserve features of the data such as sharp edges. Therefore, the RD signal can be smoothed without losing the locations of change points (Figure 4.1c). Also, compared to the wavelet filter [155], Savitzky-

Golay filter does not suffer from the shifting effect which is an essential characteristic to detect the accurate change points [156]. We then adapted the Modified Varri method [154] to detect the amplitude-shift points of the RD signal that define the boundary of primary segments. The Modified Varri method with moving average showed high ability for correctly segmenting the noise time-series data and has better accuracy comparing with moving average filter [156]. Therefore, our oversegmentation approach depends on the usage of both Savitzky-Golay filter and Modified-Varri segmentation together. To illustrate this, we applied CNAttra with and without smoothing (Savitzky-Golay filter) on the simulated datasets. Our results show that true positives decrease significantly in case of FAs in the absence of Savitzky-Golay filter (from 79-93% to 31-41%) in 4 experiments using simulated datasets (Figure 4.2) while the numbers of LCVs remain unchanged (detailed explanation of the generation of simulated data is provided in section 4.1.3). Therefore, combining Savitzky-Golay filter with Modified Varri segmentation enables the robust identification of true positives with low false discovery rate.

Figure 4.2: The Savitzky-Golay filter is critical for CNAttra oversegmentation approach.

CNAttra is applied to simulated CNA datasets (#1 to #4) with (SG) or without (No-SG) smoothing the RD signal using Savitzky-Golay filter. The bar plot showed the number of true-positive (green) and false positives (purple) FAs in different experiments using simulated datasets.

The pseudo code of the L1 segmentation procedure for detecting the candidate segments of the RD signal is provided in Appendix A.1. First, the RD signal is smoothed to attenuate its fluctuation using the Savitzky-Golay filter (Figure 4.1c) [157]. Then, it is oversegmented to capture all sharp-edge (change-points) by adapting the modified Varri method [156] in two steps: 1) finding the peaks and troughs of the smoothed RD signal, 2) computing the change points as the mid-points between each consecutive peak and trough. Finally, the change points are selected based on target size of IBs and FAs. We apply the L1 segmentation procedure twice with two Savitzky-Golay frame lengths to identify the large and small segments. Large segments are chosen expecting one segment per 1 Mb region (the minimum IB width parameter) and these large segments are used to build the iso-copy number blocks (IBs). Short segments are used to identify the FAs. They are selected expecting one segment per $10 \times \text{CN}_{\text{Atra}}$ resolution assuming that at least 90% of the genome is free of FAs.

4.1.1.5 Assembly of iso-copy number blocks

A subsequent CN state-based merging process combines adjacent initial segments into large contiguous segments. We termed these merged contiguous segments with distinct CN state as IBs (Figure 4.1d). IBs with CN state different from CN reference are considered as segmental aneuploidy or LCVs. The size range of IBs (LCVs) are generally in the megabase (Mb) scale (default value $\geq 1\text{Mb}$, which can be tunable by the user). Each IB is used as a population of bins for the discovery of FAs.

The pseudo code of our IB assembly module is provided in Appendix A.2. Our IB assembly approach is a heuristic algorithm to build the IBs (Figure 4.1d) from the primary segments that resulted from the L1 segmentation procedure. We define an IB as a large segment ($>$ user-defined minimum IB width) with distinct copy number from its adjacent segments, and it is not a part of a larger IB ($> 5\%$ of its embedding IB). The primary segments are then classified based on segment width, segment copy number ($\text{CN}_{\text{segment}}$), and segment state/interval (I_{segment}). First,

subsequent primary segments with the same CN state are merged to form the initial segments. These initial segments are again merged iteratively to form the final IBs (Figure 4.1d). In each iteration, long initial segments are labeled first. Then consecutive long initial segments, with the same CN interval, are assigned to the same label. Then, if two long initial segments are interrupted by short initial segment(s), these short segments are assigned to a label of one of the surrounding long segments or to a new label to maximize intra-segment similarity based on their copy numbers.

4.1.1.6 Calling of candidate FAs

Focal alterations (Figure 4.1e) are detected by examining the candidate regions (L1 short segments) versus the IBs that are resulted from the IB assembly algorithm. We consider each IB as a population of bins and find its significant amplification/deletion regions. Each IB is used as a local reference for its candidate regions and modeled using a normal distribution. They are classified by the statistical t-test and the coverage-based thresholding test to find the initial significant FAs. Class 1 FAs pass the t-test while Class 2 FAs pass the thresholding test.

The t-test may result in many false positives for large regions especially in low-coverage data. Therefore, we integrated the random sampling approach to apply the t-test on segments of fixed reasonable sample size (m) and find the statistically significant candidates (Class 1 FAs). If the region size (n) is greater than the sample size (default = 35 bins), the t-test is applied on k random-samples and the maximum P_{value} is chosen as the P_{value} of the region:

$$t_{\text{region}} = \begin{cases} \max \left(\frac{CN_{\text{sample}_i(\text{region})} - CN_{\text{IB}}}{S_{\text{sample}_i(\text{region})}} * \sqrt{m} \right), i \in [1, k] & \text{if } n > m, \\ \frac{CN_{\text{region}} - CN_{\text{IB}}}{S_{\text{region}}} * \sqrt{n}, & \text{otherwise,} \end{cases} \quad (4.8)$$

where CN is the estimated copy number, S is the standard deviation, m is the size of the sampled region. P_{value} is then corrected for multiple hypotheses testing assuming that at least 90% of each IB are not focal CNAs:

$$P_{\text{value}} = \frac{P_{\text{value}}(t_{\text{region}}) * \text{Width}_{IB} * 0.90}{\text{Width}_{\text{region}}} \quad (4.9)$$

$$\text{statistical condition} = \text{sign}(\text{abs}(P_{\text{value}} - \alpha)), \quad (4.10)$$

where α is the threshold for a statistically significant region (default value $\alpha = 0.05$). Also, we use a thresholding method to define *high confidence* candidate regions (Class 2 FAs):

$$\text{coverage condition} = \text{sign}(\text{abs}(\text{abs}(\text{CN}_{\text{region}} - \text{CN}_{\text{segment}}) - \lambda)), \quad (4.11)$$

where λ is the coverage-based amplification or deletion threshold of highly significant regions. Then, adjacent significant regions of the same class and same type are merged.

4.1.1.7 Filtering the candidate FAs

Candidate FAs are filtered based on coverage-based parameters and user-defined flags to keep only FAs with a high degree of confidence. First, we filtered out candidate FAs around the low-quality and low-abundance regions of the genome (dark bins). Low-quality regions are the bins which are located around centromeres, telomeres and blacklisted regions [158]. Low-abundance regions are mostly associated with gap regions, repeated regions, and low-mappability regions. User can set a threshold of the maximum-allowed dark bins per candidate FAs to keep (default = 0.5). Also, we remove small regions (width < CNAttra resolution).

4.1.2 Estimation of CNAttra tuning parameters

We used relatively high-coverage datasets (10x-14x), available from 1000 Genomes Projects (HG00119, HG01879, HG00096) and A427 cell line, for estimating the relationship between the data coverage and CNAttra parameters including resolution, amplification and deletion thresholds. For each dataset, we generated subsamples of the original data using Picard (<http://broadinstitute.github.io/picard/>) and SAMtools [159]. Then, we computed the optimum values of the analysis parameters of the original and subsamples. These values and their corresponding data coverages were used for fitting exponential regression models (Figure 4.9)

which showed best-fitting compared to other regression models such as polynomial, power-decay, linear models.

Resolution is defined as the minimum width of FAs that can be detected to achieve a low FDR ≤ 0.05 . It is used as the threshold to filter the alteration regions at FA filtering step. Same resolution is also used as the Savitzky-Golay frame length to normalize the RD signal at L1 segmentation stage. Amplification and deletion thresholds are used to define the Class 2 focal amplification and deletion regions respectively (Fig. 2d). These thresholds are the minimum differences between the copy number of amplified/deleted regions and the copy number of their IB.

Figure 4.3: Procedure for estimating CNAttra coverage-based resolution using HG00119 dataset.

Larger bin size results in low FDR. FDR is 28.5% at bin-size of 1-kb (a), 10.7% at bin-size of 5-kb (b), FDR = 6.64% at bin-size of 10-kb (c). The black, blue, and red lines are the 5%, 90%, and 50%-reads normal distributions corresponding to copy numbers of 1N, 2N, and 3N respectively.

For the resolution analysis, at each genomic coverage (x), 1-kb binned RD signal was first modeled using a normal distribution (μ, σ). Assuming that maximally 10 % of the genomic regions can be FAs, we generated random RD values for 5%, 90%, and 5% of bins from normal distributions corresponding to copy numbers of 1N, 2N, and 3N respectively (Figure 4.3). False discovery rate (FDR) was computed as the fraction of false CN-assigned bins to the total number of bins. We repeated these steps by incrementing the bin size. Resolution is computed as the minimum bin-size that achieves FDR of ≤ 0.05 . Figure 4.3 shows an example for estimating the CNAttra coverage-based resolution from the HG00119 dataset. First, at 1-kb binning, we compute the mean (2N) and the standard deviation (σ) of the RD signal. Then, we generate random RD values from normal distributions with means of 1N (Figure 4.3a black color curve), 2N (Figure 4.3a blue color curve) and 3N (Figure 4.3a red color curve), and standard deviation (σ). We compute the false positives as bins with CN states that are not equal the CNs of the underlying distributions. At 1-kb binning, the false discovery rate (FDR) is 28.5%. We re-binned the RD signal at 5-kb (Figure 4.3b) and 10-kb (Figure 4.3c), and repeated this methodology. This showed improvement of the FDR (10.7% at 5-kb and 6.64% at 10-kb bins). In our approach, we keep incrementing the bin size till achieving FDR of ≤ 0.05 to define the CNAttra resolution.

For the focal amplification (or deletion) thresholding studies, we used approximately 0.5N as the amplification (or deletion) threshold of the original high-coverage dataset, and we computed the average amplitude shift between the CN of the amplification (or deletion) region and the IB (reference amplitude shift). Then, for each subsample, we computed the new average amplitude shift between the copy numbers of these regions and the IB copy number (sample amplitude shift). The amplification (or deletion) threshold was adjusted by adding the difference between the sample amplitude shift and the reference amplitude shift.

The reason to utilize exponential regression models for estimating coverage-based parameters is as follows. We have tested exponential regression model as well as other regression models for modeling the CNAttra parameters. For that, we used MATLAB 'fit' function with different

underlying models ('ploy2': $Y = p1 * x^2 + p2 * x + p3$, 'weibull': $Y = a * b * x^{b-1} * e^{-a*x^b}$, 'exp2': $Y = a * e^{bx} + c * e^{dx}$, 'power2': $Y = a * x^b + c$). Then, we applied the two-sample Kolmogorov-Smirnov test (KS test) to validate each model assumption. Among these models, the KS test for both exponential regression model 'exp2' and power-decay model 'power2' didn't reject the Null-hypothesis with asymptotic p-value > 80% for all the parameters. Therefore, theoretically both can be used for modeling the relationship between the data coverage and parameters. However, based on the manual inspection, we found that exponential model shows better fitting (Figure 4.4).

Figure 4.4: Estimation of CNAttra parameters using exponential regression.

Exponential regression model 'exp2' and power-decay model 'power2' can be utilized for modeling of the relationship between the data coverage and CNAttra parameters.

4.1.3 Generating simulated data with artificial CNAs

We derived simulated CNA profile from real cancer cell line datasets. Starting from NGS reads (BAM file) of a cancer cell line data, we artificially introduced both LCVs and FAs of random copy number and width into cherry-picked chromosomes devoid of any observable large-scale copy number changes. These spiked-in CNAs (LCVs and FAs) provided us with *de facto* ground truth for benchmarking CNAttra against five other single-sample CNA detection tools.

The pipeline of artificial CNA generation contains two steps – 1) random selection of candidate location, and 2) artificial read spike-in. For each chromosome, we simulate M number of IBs and N number of FAs. For this, we first divide the chromosome into M contiguous large segments randomly which represents IBs. LCVs are IBs with CN not equal to 2. Similarly for focal alterations, we randomly choose N non-overlapping small regions (range 50 to 200 kb) which satisfy the following conditions: 1) regions must not overlap IB boundary or any blacklisted/gap regions and 2) each region should have number of reads similar to the median of the selected chromosome based on threshold cut-off:

$$\text{Cut-off} = \left(|Med_{IB} - RC_{FAR}| / Med_{IB} \right) \leq 5\% \quad (4.12)$$

where Med_{IB} is the median read counts of the chromosome, and RC_{IB} is read count of the N regions. After the selection of coordinates of LCVs (IBs with CN not equal 2) and FAs, we next 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 $\left\lceil X * \left(\frac{C2}{C1} - 1 \right) \right\rceil$ artificial reads, where X is the initial number of reads of that region. This number of 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 CNA evaluation.

As an illustrative example, in Figure 4.5, we demonstrate our methodology to artificially introduce 3 IBs and 7 focal alterations in Chr 2 of a hypothetical cancer genome. We first selected Chr 2 which is free of any observable LCVs (Figure 4.5a). Then, Chr 2 is divided randomly into 3 IBs with random CN (1,2,3,4). We generated artificial reads for IB2 (LCV with CN of 3) and IB3 (LCV with CN of 4) based on their resulted CN (blue rectangles in Figure 4.5b). Similarly, we spiked-in 7 focal alterations with random coordinates and CNs (e.g. 4 focal amplifications and 3 focal deletions in Figure 4.5c). For the focal amplification, we generated reads based on the

difference between their CNs and the CN of their parent IBs (red rectangles in Figure 4.5c). For the homozygous deletion, we removed all the reads from that region.

(a) Cherry-picked chromosome : Chr2

(b) Number of IBs : 3

(c) Number of FAs : 7

Figure 4.5: Generation of simulated CNA profile.

Illustrative example shows artificial introducing of 3 IBs and 7 FAs in Chr 2.

Another illustrative example of the resulted coverage plot of simulated data, after applying our approach and introducing both LCVs and FAs, is provided in Figure 4.6a (top panel). This is derived from the RD signal of CHP-212 Chr 12 (Figure 4.6a bottom panel). Upon introducing the CNAs, the new simulated RD signal maintains the multimodal characteristic of the cancer cells (Figure 4.6b).

Figure 4.6: An illustrative example of our realistic simulated data.

4.1.4 Evaluation experiment

We generated four versions of the simulated datasets with different widths, frequencies, and copy numbers. The first dataset is generated with 40 FAs of 50-100 kb width per chromosome, and a minimum difference between LCV and FA copy numbers (CN shift) of 2N. It is used for the detailed performance evaluation of each tool. The second dataset is created with 30 FAs of 50-200 kb width per chromosome, and CN shift of 2N. The third dataset is generated with 30 FAs of 50-200 kb width per chromosome, and CN shift of 1N. The fourth dataset is generated with 40 FAs of 50-100 kb width per chromosome, and CN shift of 1N. Using these artificial datasets, we performed comparative evaluation of the performance of CNATra, ReadDepth, CNVnator, FREEEC, CLImAT, and AMYCNE. We could not pursue XCAVATOR [48] and SCNV [50] due to their complicated installation procedure. We optimized the performance of each tool to handle the overdispersion and multimodality of the low-coverage cancer datasets (Appendix B:1).

We found that CNVnator results change dramatically with the input bin size. At 1-kb binning, CNVnator detects many false positives. At 10-kb binning, false positives are decreased significantly. However, it misses some true focal alterations (true positives) that are detected using the 1-kb resolution as illustrated in Appendix B:1. Therefore, we generated CNVnator results using bin sizes of 1-kb and 10-kb. The main parameter of FREEEC is the ploidy level of the cell line. For example, the organization of segments differs considerably at different ploidy levels as illustrated using the 12p region of NCI-H82 (Appendix B:2). It uses the ploidy information to set the copy number reference (2N) based on the median of the RD signal. This affects the segmentation and calling of CNAs. However, this tool cannot compute the copy number reference accurately for hyperdiploid, near-triploid or higher ploidy cells. We generated FREEEC results of the artificial datasets under diploid and triploid assumptions for comparison. For CLImAT, we tried smaller values of the minDepth parameter (default 10) to adapt to the low-coverage data. We found that minDepth of 0.1 gives the best segmentation and calling results. For ReadDepth, there are many tuning parameters such as overdispersion (default 3), percCNGain (default 0.05),

and percCNLoss (default 0.05). We tried higher values of overdispersion. Also, we set the percCNGain and percCNLoss based on the multimodal distribution of simulated data. However, the results did not show notable enhancement. We believed that ReadDepth failed to model the multimodal distribution and to correctly estimate the copy number reference. For AMYCNE, we use its default parameters since results showed a balance between true positives and false positives.

4.1.5 Data availability

All the datasets used in this study are publicly available (Table 4.1). We used CNA profiles of cancer cell lines from CCLE [148] and COSMIC [160] databases for performance evaluation of CNATra.

Table 4.1: Datasets for evaluating the performance of CNATra.

Cell Line	GEO Accession	Coverage (x)	Read length	Whole-genome ploidy level
HG00096	1000 Genomes Project	13.94	250	Diploid [92]
HG01879	1000 Genomes Project	14.34	250	Diploid [92]
HG00119	1000 Genomes Project	10.087	76	Diploid [92]
CHP-212	GSM2664356	1.752	101	Diploid [134]
IMR-32	GSM2664328	1.4001	101	Near-diploid [133]
HepG2	ENCSR055XHN	1.422	76	Near-diploid [135]
A427	DRX015190	10.67	101	Hypertriploid [24]
NCI-H82	GSM1526702	0.3162	40	Near-triploid [132]
MCF7	GSM2258738	0.293	36	Hypertetraploid [125]
LNCaP	GSM1902617	0.6299	50	Hypertetraploid [131]
Case#6	TCRBOA6-T-WGS.lane1	20.2071	101	Diploid: Texas Cancer research Biobank (http://txcrb.org/data.html)
Case#7	TCRBOA7-T-WGS.lane4	19.2884	101	

4.2 Results

4.2.1 Accurate estimation of CN reference is essential for CNA calling in cancer genomes

As demonstrated in Chapter 3, human cancers frequently have hyperdiploid, near-triploid or higher genome-wide ploidy levels [24-26]. All these anomalies manifest as multimodal distribution of the RD signal. We analyzed the RD frequency distribution of the datasets that are used for CNAtra evaluation [24, 92, 125, 131-135] (Table 4.1). Cancer genomes (IMR-32, A427, MCF7, CHP-212, HepG2, NCI-H82, LNCaP) exhibited a complex *multimodal* distribution (Figure 4.7a) as opposed to normal genomes (1000 Genomes Project datasets) which follow a unimodal distribution (e.g. HG00119).

Based on AStra principle, we utilized three multimodal distributions for estimating the CN reference (2N) under different ploidy assumptions. Each distribution is a summation of normal distributions of different probability centralized at CN states (Figure 4.1b). While most single-sample tools assume unimodal distribution of the RD signal and use the global median as CN reference (2N), CNAtra multimodal allows accurate estimation of the CN reference. As shown in Figure 4.7a, our estimated CN reference (2N; solid blue line) and higher CN states (dashed blue line) are coinciding with peaks of the RD signal. Therefore, we denote the ‘actual’ CN reference as the estimated CN reference by CNAtra when the true whole-genome ploidy level is provided. In contrast, the median-based CN reference (solid black line) deviates from the ‘actual’ CN reference (solid blue line) by 2.5-87% depending on the percentage of LCVs in the cancer cells as opposed to 0.34-0.56% for normal genomes. Similarly, in contrast to normal genomes (Figure 4.7b), we also found that statistical measures (mean and median) of RD signal are highly inconsistent for cancer cells (Figure 4.7c) and cannot capture the CN reference estimated by CNAtra using multimodal modeling.

Figure 4.7: Visualization of copy number estimation using CNAttra.

Nevertheless, karyotype or ploidy information is not always available for cell lines. In that case, our free model (Figure 4.1b) can still predict the CN reference with a maximum error of 0.4% compared to the ‘actual’ CN reference given the ploidy information (Table 4.2). Therefore, our multimodal approach can be applied universally regardless of the knowledge of the underlying ploidy level or karyotype. Defining the CN reference forms the basis to define the segmental aneuploidies and estimate the copy number states accurately.

Table 4.2: Estimation of copy number reference (CNR) using CNATra.

The actual CNR is the estimated CNR by CNATra given the true whole-genome ploidy level as input. RD global median is the median of the RD signal that is used by some CNA detection tools as CNR. The CNATra free CNR is the copy number reference estimated by CNATra using the ‘free’ model.

Cell line	Actual CNR	RD global median	Actual CNR – RD global median (%)	CNATra free CNR	Actual CNR – free CNR (%)
HG00119	11158.07638	11117.41164	0.36444	11158.07638	0
HG00096	15201.7051	15116.5035	0.56047	15201.7051	0
HG01879	15574.3865	15521.3532	0.34051	15574.3865	0
CHP-212	2007.338554	1956.111938	2.551966962	2007.338554	0
IMR-32	1449.388925	1520.469956	4.904206838	1449.388925	0
HepG2	1483.53594	1578.920375	6.429533198	1483.53594	0
A427	7619.16942	11186.58037	46.82152019	7638.036473	0.24701444
NCI-H82	311.6544291	370.6617928	18.93358741	312.2339631	0.185608896
MCF7	179.7474664	323.7492603	80.11339287	180.4597787	0.394720794
LNCaP	383.6634819	719.2432633	87.46721991	384.635724	0.252769585

4.2.2 Hierarchical framework enables CNAttra to detect and distinguish between LCV and FA

We have taken a pragmatic approach for solving the two major problems associated with cancer genomes- 1) presence of LCVs and 2) systematic biases such as overdispersion and wave artifacts which are pronounced in low-coverage data. We used Savitzky–Golay filter (a weighted moving average filter that smooth out short-term variations and preserves inherent RD features without shifting effect) to successfully attenuate the wave artifacts and signal variability in order to identify the primary segments by Modified Varri segmentation (Figure 4.1c; black line). However, copy number detection methods based solely on segmentation may suffer from false segmentation or oversegmentation. For example, neighboring segments may represent copy number altered regions belonging to the same CN state which has no biological basis (false segmentation). In addition, a LCV region can be falsely ‘oversegmented’ into several regions with the same copy number due to the presence of focal amplifications and deletions within it. This may lead to failure in capturing the entire LCV as a single event. We solved this problem using an assembly algorithm which successfully merges primary segments with the same copy number to define the iso-copy number blocks (IBs) (Figure 4.1d). Each IB represents distinct (unimodal) peak in the multimodal RD signal distribution centered on/near a CN state. The IB with CN state different from CN reference is considered as segmental aneuploidy or LCV.

After defining the IBs, we found that RD signal from each IB follows the normal distribution using Q-Q plot and Kolmogorov–Smirnov test with a reasonably good approximation (5% significance level) (Figure 4.8). In each subfigure, 1 Mb continuous sample from an IB is randomly selected, and Q-Q plots are generated from cancer cell lines (A427, CHP-212, HepG2, IMR-32 and NCI-H82). Therefore, we use each IB as a population of bins for the discovery of *statistically significant* FAs (**Class 1**) using the t-test. However, due to overdispersion of RD signal in cancer genomes, statistical tests may reject the null hypothesis of long segments with a small mean difference (particularly in the presence of wave artifacts) resulting in many false positives. Therefore, we

employed coverage-based thresholding to define *high confidence* (Class 2) FAs based on the local CN reference of the parent IB (Figure 4.1e).

Figure 4.8: Normality test for individual IBs using Q-Q plot and Kolmogorov-Smirnov test.

4.2.3 Coverage-based thresholding enables the detection of high confidence CNAs in low-coverage datasets

Dispersion of RD signal is inversely proportional to the depth of coverage. Therefore, CNA detection tools usually recommend using large bins (coverage-driven bin-size approach) to avoid false detection of CNAs in low-coverage datasets. However, this approach may result in missing significant focal alterations. As an alternative, we propose to adapt the coverage-driven thresholding approach instead of increasing the bin-size. This allows the identification of focal CNAs without increasing the false positives.

Theoretically, any 'candidate' focal alteration with amplitude-shift $> 0.5N$ (threshold) from the CN of its parent IB can be identified as significant focal alteration since it belongs to another copy number state (Figure 4.9a). For example, in a diploid genome, the majority of bins have copy number of $2N$ (blue circles in Figure 4.9a). In such situation, any segment with $CN > 2.5N$ can be called as focal amplifications (red circles in Figure 4.9a) such as FA1 and FA2. Similarly, any segment with $CN < 1.5N$ can be called as focal deletions (grey circles in Figure 4.9a) such as FA3 and FA4. However, low-coverage data suffer from higher variability resulting in false-calling of focal CNAs with the same threshold. Therefore, we define the coverage-based parameters (resolution, amplification and deletion thresholds) to adapt with this overdispersion in low-coverage data (Figure 4.9b). CNAttra is calibrated using several WGS datasets of high coverage. We utilized negative exponential regression for modeling the relation between sequencing coverage and the coverage-based parameters. Our thresholding parameter enables a user to strike a balance between false positives and false negatives. For example, using the same threshold of $0.5N$, the subsampling of A427 data to 3.2x coverage yields more false positives compared to the original 10.67x coverage (Figure 4.9c top and middle panel). Increasing the threshold gets rid of these false positives (Figure 4.9c bottom panel). Therefore, the advantage of

coverage-based parameter tuning makes the CNAttra results more robust at different data coverage.

(a)

(b)

(c)

Figure 4.9: Coverage-based thresholding.

4.2.4 CNAttra detects experimentally validated LCVs and FAs across cell lines

We evaluated the ability of CNAttra to detect and distinguish both large-scale and focal events using validated data as ‘ground truth’ from multiple cancer datasets. First, CNAttra confirms the complete genetic profile of LCVs of HepG2 reported earlier using array CGH analysis [135]. CNAttra correctly identified the whole chromosome (2, 16 and 20) and segmental gains involving Chr 6, 14, and 17 (Figure 4.10). The bars below the ideogram indicate the segmental amplification (red), copy number neutral (white) and centromere (black) regions. CNAttra detected additional IB6 in Chr 14 using WGS data of HepG2. This difference can be attributed either to real biological difference (different sublines, different passage number or differences in culture condition of HepG2 used for the two experiments) or due to the difference in the detection platforms used (array CGH versus NGS based detection).

Moreover, CNAttra precisely identified the well-known monoallelic 1p deletion in neuroblastoma cell lines (IMR-32 and CHP-212) [161] (Figure 4.11a). In both these cells, CNAttra called 1p deleted region as a single LCV event with correct CN estimation (CN=1). Concurrently, CNAttra also successfully detected the *well-known* focal amplifications of *MYC* (NCI-H82) [162] and *MYCN* (CHP-212, IMR-32) [163] loci as well as homozygous focal deletions of *BANK1/4q24* (NCI-H82) [164], *LKB1/STK11* (A427) [165], and *p16INK4a/CDKN2A* (A427) [166] loci (Figure 4.11b).

Figure 4.10: CNAttra validation using the genetic profile of HepG2.

(a)

(b)

Figure 4.11: CNAtlas validation using reported LCVs and FAs.

4.2.5 CNAtra confirms the CNA profiles identified by SNP array in CCLE and COSMIC database using low-coverage NGS data

In order to estimate our performance in a global manner, we used genome amplification data of CHP-212 and NCI-H82 cells, identified using SNP array, which are available from both CCLE [148] and COSMIC [160] databases. In the COSMIC database, CNAs were called using PICNIC software [167] while CCLE calls were based on CBS approach [88]. We applied CNAtra on the low-coverage NGS data of CHP212 (1.4x) and NCI-H82 (0.31x) and compared our calls. In COSMIC data, the CNAs have been called in a gene-centered manner where each region is associated with one or more genes. COSMIC data has 6 and 13 amplified regions for the CHP-212 and NCI-H82 cells respectively. We found that out of 6 amplified regions in CHP-212, CNAtra identified two as focal CNAs and other four regions inside a single LCV (Figure 4.12a). Similarly, for NCI-H82, 8 out of 13 CNAs have been detected as focal events, and three are embedded in two LCVs. In comparison, we found that CCLE detected 11 and 99 amplified regions in CHP-212 and NCI-H82 cells respectively. In CHP-212, CNAtra detected 10 out of 11 amplified regions as four focal events, three LCVs and, the rest three are embedded in one LCV. Similarly, 78 out of 99 amplified regions were detected by CNAtra for NCI-H82. Out of 78, CNAtra identified 11 as FAs, 7 LCVs and rest 60 are part of 19 LCVs. This suggests that CCLE calls are segmenting LCVs into smaller CNA segments (Figure 4.12b). Taken together, CNAtra detected 89.5% COSMIC and 80% of CCLE calls. It is important to note here that these differences may also be due to different approaches (SNP array versus NGS based) of identifying CNAs. Surprisingly, there are only two CHP-212 and four NCI-H82 consensus amplification regions that were detected by both COSMIC and CCLE suggesting poor concordance of CNA results. All these six consensus regions are detected by CNAtra. However, the number of consensus regions is scanty for a robust assessment of performance. Also, COSMIC and CCLE CNAs do not distinguish between LCVs and FAs. The absence of a detailed and verified cancer CNA profile with complete repertoire of LCVs and FAs encouraged us to use realistic simulated data as ground truths for performance evaluation.

Figure 4.12: Visual illustration of COSMIC and CCLE CNA calls (using SNP arrays) superimposed on the NGS-based coverage plot.

4.2.6 CNAtra successfully identifies copy number profiles of cancer patient samples with variable purity

Apart from abnormal karyotype of cancer genome, tumour purity adversely affects the genomic analyses of patient samples. The proportion of non-cancer cell types in the tumor sample imparts a strong influence on the outcome and biological interpretation of copy number estimation. This problem is generally countered by using matched control samples by some computational tools for the purity estimation and copy number detection [147, 168]. The flagship cancer project, TCGA had initially set a minimum of 80% tumor purity as a quality threshold for inclusion of tumor samples in their study which was later changed to 60% [169]. Using 21 cancer types data, the median purity level of tumor samples available from TCGA is estimated to be around 75% [169].

We next applied CNAtra to two publicly-available WGS samples of cancer patients: Case#6 pancreatic neuroendocrine tumor (PanNET) and Case#7 B-cell follicular lymphoma (Table 4.1) with tumor cellularity of 80% and 90% respectively. The copy number information for these two samples are also available from the same database which were identified by VarScan 2 [36], a computation tool which uses paired case-control sample for copy number detection. RD signals from tumor/cancer samples generally showed that majority of the genome is copy-number neutral resulting in a RD frequency distribution with a prominent peak at 2N which better fits our multimodal distribution with diploid assumption (Figure 4.1b). Therefore, for patient samples, it is recommended to model the RD signal distributions using the diploid assumption.

Figure 4.13: CNAttra calls using cancer patient samples.

(a) CNAttra detection of genome-wide CNAs in Case#6 (top) with pancreatic neuroendocrine tumor (PanNET) and Case#7 (bottom) with B-cell follicular lymphoma (Supplementary Table 1) with tumor cellularity of 80% and 90% respectively. (b) CNAttra detects focal gain affecting 6p22.2 (top), segmental loss of 15p region (middle), and segmental gain of 17p region (bottom) in Case#6 patient with PanNET.

For Case#6, CNAttra accurately detected all the chromosome-level aneuploidies (as detected by VarScan 2) using solely the WGS data of the cancer (case) sample (Figure 4.13a). For example, CNAttra identified the characteristic loss of heterozygosity (LOH) of chromosome 3 and 11 as well as the copy number gain of 17q region in Case#6 that are frequently observed in patients with PanNET [170, 171] (Figure 4.13b). We also observed a focal gain affecting 6p22.2 in Case#6 which is a recurrent feature of PanNETs [172] (Figure 4.13b). In Case#7, chromosome-level alterations were largely absent. However, we easily interpreted Case#7 to be a male subject as the X chromosome has only one copy in contrast to the Case#6 (female subject) who has two copies of X chromosomes (Figure 4.13a). Our evaluation with two cancer patient data suggests that CNAttra can be successfully applied to clinical samples with $\geq 80\%$ purity.

To further evaluate the impact of tumor purity on CNAttra calls, we used the publicly available simulated tumor samples with purity from 60% to 90% [https://www.yfish.org/data/singleclone_2x/]. CNAttra successfully detected all copy number “events” correctly from samples with purity of 60%, 70%, 80% and 90 as visually illustrated by the CNA profile of chromosome 2 at different purity level (Figure 4.14a). However, in some cases/events, tumor purity information can be used to correctly estimate the exact copy number values. The success of CNAttra to detect CNA events in tumor samples, without relying on matched control samples, can be attributed to the utilization of CN interval and CN state for detecting CNA. To illustrate that, Figure 4.14a shows a cartoonic explanation of CNAttra calls in a hypothetical sample with tumor purity of 60%. For example, the copy number values of the red segment (originally CN=1N in tumor cells) and blue segment (originally CN=3N in tumor cells) will change to 1.4 and 2.6 respectively due to the ‘contamination’ of tumor sample with 40% non-cancerous (normal) cells. CNAttra can estimate their copy number correctly as 1N and 3N respectively based on the ‘interval’ definition of CN state. For example, 1N state ranges from copy number values of 0.5 to 1.5, and 3N state ranges from 2.5 to 3.5. This allows CNAttra to accurately estimate the

change in copy number state up to 1N with tumor purity >50% and up to 2N with purity >75% and up to 3N with purity >90%.

(a) Simulated tumor data: Chr2

(b) Cartoonic illustration of CNAttra CN intervals

Figure 4.14: CNAttra successfully identified the copy number events of tumor samples with purity >50%.

4.2.7 CNATra is the superior tool for detecting large-scale and focal alterations

In the absence of extensively validated datasets of both LCVs and FAs, we relied on the simulated CNA profile. However, a simple distribution-based simulated data may not capture the inherent biases of the RD signal of cancer genomes. Therefore, we developed a novel approach to manipulate the original WGS reads of cancer genome to randomly introduce FAs embedded within the LCVs maintaining the inherent CNA features and complexities of the RD signal (Figure 4.6a). We generated simulated CNA profiles using low-coverage (< 2x) CHP-212, IMR-32, and HepG2 data (Figure 4.6b). We introduced LCVs and FAs in 5 chromosomes per cell line. For each chromosome, we have incorporated an average of 1-4 LCVs and then introduced 40 focal amplifications and deletions (size ranging between 50 and 100 kb) within these LCVs resulting in 600 FAs encompassing three cell lines.

Using this simulated data, we benchmarked the performance of CNATra against five RD-based single-sample CN detection tools which include ReadDepth [49], CNVnator [44], FREEC [45], CLImAT [51] and AMYCNE [46]. We analyzed the performance of these tools after optimizing their parameters for low-coverage data (section 4.1.4 Evaluation experiment). We set the criteria of > 75% overlap between the spiked FAs and the FA calls to be considered as a true call, and similarly, we set 90% overlap as criteria for true calls in case of LCVs. For a fair evaluation, we used the false amplifications only as false positives since many deletions may be originally present at these cherry-picked chromosomes, however, we cannot distinguish them from the spiked-in deletions.

Our evaluation showed that CNATra could overall detect 93.3% (560/600) of the ‘introduced’ focal alterations (IMR-32 91.5%; CHP-212 97% and HepG2 91.5%) (Figure 4.15a). In comparison, the second-best tool, FREEC can detect FAs with 63.5% (381/600) accuracy under the diploid assumption. All other tools (CNVnator, AMYCNE, ReadDepth, and CLImAT) can detect FAs with 0-

50% accuracy (0-284 focal CNAs/600). We also found that CNATra has the highest accuracy for estimating copy number correctly with average CN difference of 0.251N followed by CNVnator with 10-kb binning (0.3642N) and FREEC (0.4357N) (Figure 4.15b). To evaluate the detection power, we computed precision and recall (sensitivity) of each tool and found that CNATra outperforms all other tools (Figure 4.15c). Although ReadDepth showed the highest precision with zero false positives among all the tools, it can only detect 34/600 (5.6%) FAs with 33 deletions and one amplification event with wrong estimated CN. Therefore, we compute *f-measure* to estimate detection accuracy which balances the precision and the recall values. CNATra showed the highest *f-measure* value of 87.77% followed by FREEC which showed 56.8% (Figure 4.15c). Next, we plot the ROC (receiver operating characteristic) curve for evaluating the performance of the tools under the assumption that the maximum cumulative focal alteration locus length is 10% of the genome (for estimation of true negatives). The ROC curve clearly shows that CNATra is superior in detecting focal CNAs in terms of true-positive and false-positive rates (Figure 4.15d). When we compare their performance to detect LCVs, CNATra again emerges at the top. CNATra could detect 31/32 LCVs (96.8%) while CLImAT can detect 18/32 LCVs (56.25%). Rest of the tools failed to detect any LCV event. We repeated this procedure to generate three additional simulated CNA profiles using different widths, frequencies and copy numbers. All the analysis showed similar relative performance between the tools.

In addition, we found that CNATra is robust in estimating the CN reference regardless of the presence of LCVs. We stated earlier that LCVs could adversely affect local median that in turn can affect the copy number estimation. For example, after spiking the RD signal with LCVs, the global median changes by 3-7% (Figure 4.15e) which may lead to the wrong estimation of CN reference. Despite this, CNATra can correctly estimate the CN reference (Figure 4.15e). We also analyzed the computation time of all copy number detection tools. We have compared only the processing time for CNA calling modules since different tools have different input formats. We found that for low-

coverage datasets, CNAtra, ReadDepth and CLImAT take the shortest computation time (average 35, 45 and 41 seconds respectively).

Figure 4.15: Performance evaluation of CNA detection tools on the simulated datasets.

4.2.8 Visualization and manual inspection of CNA calling exhibit CNATra is best equipped to capture the complexity of cancer genomes

Review of CNA calls necessitates post-processing procedures which include visual inspection and curation of the results. Visualization and manual review of CNA profile in terms of copy number, size and structure can help to fine-tune the tool parameters as well as refinement and curation of the results for downstream applications. Therefore, CNATra provides an interactive visualization platform for the user to inspect and authenticate its results manually. We utilized this visual inspection approach to comprehensively understand the advantages and limitations of all tools using cancer datasets (Figure 4.16 and Figure 4.17). In both figures, the red, black, and grey dots are bins belonging to focally amplified, deleted, and CN-neutral regions respectively. The blue line represents the copy number of each CNA. Any amplitude transition indicates a new CNA region. Top panel shows the IBs identified by CNATra (each red box represents one IB). CNATra examines each IB to detect the FAs (second panel). Rest of the panels show CNA output by other detection tools.

We found that CNATra is the only tool to comprehensively detect both LCVs and FAs (Figure 4.16 and Figure 4.17). Moreover, we found that most tools are affected by imperfect segmentation of the large segments. For example, all tools except CNATra have wrongly attributed the monoallelic 1p deletion in CHP-212 neuroblastoma cells into several segments. Only CLImAT identified this 1p deletion as a single event; however, they fail to correctly determine the right boundary of the segment (Figure 4.16). Additionally, the focal amplification (FA1) inside the monoallelic 1p segmental deletion, which harbors enhancer region based on ENCODE ChromHMM [173], is detected by CNATra, CNVnator and AMYCNE (Figure 4.16). Also, 1q segmental amplification (correctly detected by CNATra and CLImAT only) harbors many focal deletions. This confirms that focal amplification can be a part of the monoallelic segmental deletion and similarly focal deletions are present inside segmental amplification. None of the currently available tools

addressed the coexistence of large and focal CNAs in cancer genomes. Therefore, they cannot distinguish between these phenomena and tend to favor either of them. In addition, we found that all the tools suffer from the false estimation of copy number based on the proportion of LCVs in the genome. This results in misidentification of 1p and 1q regions with the same copy number by ReadDepth (Figure 4.16). This effect is more evident in the A427 triploid cell line [24]. As illustrated in Figure 4.17, IB2 with CN 3 is misclassified as normal region and IB3 with CN 2 is wrongly identified as a deletion event since the median of the RD signal is matching the 3N (black line in Figure 4.7a A427) and not the correct CN reference (2N) (blue line in Figure 4.7a A427). All the tools are affected by overdispersion in low-coverage data which may result in false positives and false negatives as estimated using simulated data. CNAtra circumvents this problem using thresholding parameters, however, the user can apply higher stringency thresholding to curate the CNA data manually.

Figure 4.16: Visual comparison of CNA profiles generated by copy number detection tools on Chr 1 of CHP-212.

Figure 4.17: Visual comparison of CNA profiles generated by copy number detection tools on Chr 3 of A427.

4.3 Discussion

Hyperploidy and pervasive genetic alterations are the hallmark of cancer genomes. In the current study, we analyzed several cancer cell lines with different levels of aneuploidy which showed a complex multimodal distribution due to widespread LCVs and FAs. This is in stark contrast to unimodal RD signal distribution of normal human genomes (such as 1000 Genomes Project samples) which are mostly devoid of segmental aneuploidy. Current single-sample CN detection tools have limited ability to handle cancer genomes due to their assumption of unimodal probabilistic distributions of RD signal. Erroneous modeling of the RD signal distribution may lead to incorrect estimation of CN reference and false segmentation which adversely affect the final CNA results. Consequently, cancer CNA profiles available from public databases (*viz.* COSMIC, CCLE) face the same problem. CNAttra successfully utilized a multimodal distribution to estimate the CN reference and then employed a CN-based merging algorithm to detect the large segments. Then each CN-designated segment formed the basis for detecting FAs where the copy number of the parental segment is used as the local reference.

High-coverage WGS datasets are generally used for CNA profiling; however, they are not available for many cancer cell lines. As an alternative solution, the NGS reads from 'input control' of histone modification ChIP-seq dataset can be effectively used which are publicly available for many cell lines. These input control reads are generated from sonicated crosslinked chromatin and typically used to normalize and peak calling of the ChIP-seq data. These input data contain the genome-wide reads but they are generally sequenced at low-coverage ($<2\times$). These data can be judiciously utilized to compute the RD signal for CNA analysis. However, any low-coverage dataset is afflicted with overdispersion of the RD signal which facilitates detection of a high number of false positives. CNAttra bypasses this problem by using coverage-based thresholding to detect high confidence FAs minimizing the number of false positives, which can be further tuned by the user for more stringent CNA calling. We have estimated the relationship between coverage and CNAttra

parameters using subsampling of a limited number of high-coverage WGS datasets including one cancer cell line data. Our heuristic approach of determining coverage-based thresholding parameters leaves space for future improvements by incorporating additional high-coverage datasets. For example, the dispersion levels vary based on the CN state of each LCV. Therefore, training using additional datasets will provide better estimation of thresholding parameters at different CN states. Also, presence of poor-mappability regions (bins) can lead to false detection of homozygous focal deletions surrounding these bins. CNAtra provides the option to filter these focal deletions based on the percentage of low-mappability bins.

We have also successfully applied CNAtra to patient samples with variable tumor purity. Using simulated tumor data, we demonstrated the robustness of CNAtra calls across different tumor purity levels. Both these analyses were performed using only the test (case) samples, without utilizing any matched control sample. Generally, tumor contains an individual's *germline* (inherited) CN variants as well as *somatic* copy number alterations (SCNAs). Distinguishing these two phenomena typically requires the paired case (tumor) and the 'normal' sample from the same individual. CNAtra uses a single sequencing sample at a time resulting in a CNA profile that contains both germline and somatic copy number variants. Nonetheless, we envisage that germline and somatic CN events can be effectively distinguished if both tumor and the matched 'normal' sample are processed separately using CNAtra followed by comparative analysis of CN profiles from both samples.

One of the major limitations of the performance evaluation of CN detection tools is the non-availability of a complete repertoire of experimentally-validated cancer CN profile. Therefore, we generated realistic simulated data using the available cancer cell line data as input to maintain the signal variability and features of cancer RD signal. Then we introduced CNAs of different length scales randomly as 'artificial' ground truths by manipulating the short sequence reads in the BAM file. Read manipulation helps to capture the natural variations encountered in WGS data

in an unsupervised manner. These simulated data with spiked-in CNAs can provide *de facto* ground truth for performance evaluation of CNA detection tools in terms of CN estimation, accuracy, and precision.

Chapter 5

HiCNAttra

Hi-C Normalization Approach through read depth analysis

Revealing 3D genome conformation may explain epigenetic-driven gene-regulatory processes such as enhancer-driven gene activation. Distal regulatory elements such as enhancers are known to control gene transcription through long-range physical interactions with gene promoters by forming chromatin loops [102, 174]. After the advent of NGS, many molecular biology-based methods have been developed to study the chromatin interactions of genomic loci at different length scales [100]. In the 15 years, conventional 3C technique [55] has been modified and combined with NGS for understanding genome-wide 3D chromatin conformation [56, 57]. In 2009, Lieberman-Aiden *et. al.* developed Hi-C protocol, the first method to study the genome-wide chromatin interactions [56]. In addition, 3C-seq has been developed as a simplified alternative approach, to study the genome-wide physical interactions, where first conventional 3C library is made and then the library is sonicated and sequenced using NGS platform [57, 103, 104].

Majority of available Hi-C datasets were generated using cancer cell lines. These cancer genomes harbor large-scale and focal copy number alterations [19, 21, 27, 28, 175] with pathological consequences [176]. CNAs can result in the rewiring of chromatin connectivity that may lead to alterations in the long-range control of gene expression [177]. They can also modulate gene-regulatory mechanisms by altering the copy number of regulatory elements or by modifying the higher-order chromatin structure [178]. For example, compared to normal prostate epithelial cells, the prostate cancer cell lines have smaller-sized TADs with new TAD boundaries coinciding with the CNA regions [179]. Additionally, in multiple myeloma (MM) cells with whole/partial

chromosomal gains/losses, the CN-amplified regions have higher interaction frequencies than CN-neutral (normal) regions [180]. This is expected as regions with copy number gains or losses will have greater or lower relative chances of being captured during Hi-C pull-down step respectively which will affect their interaction profiles [177, 180, 181]. The Wu *et. al.* study then identified the TADs by insulation score method [103] using ICE-normalized interaction matrices [106] and found that the numbers of TADs increased by 25% and the average size of TADs decreases significantly compared to the normal B cells. These observations suggest that chromatin interaction signal is greatly impacted by CNAs in case of cancer cells with abnormal karyotype. Therefore, while interpreting the chromatin interactions of CNA regions with respect to the CN-neutral regions, the apparent CNA effects must be masked/corrected first to obtain a euploid-equivalent contact map.

For correcting the CN effect, it is primary to obtain the CNA information of that cell line. Hi-C sequencing reads can be effectively utilized to estimate the RD signal and discovery of CNAs without additional cost of performing WGS. For example, the HiCnv tool has been developed to identify CNAs from Hi-C reads [117], however, it can only identify large CNAs (> 1Mb size) which leave a lot of room for improvement in identifying small-sized CNAs (focal alterations).

Hi-C contact maps are affected by different kinds of systematic biases as well as the CNA-induced bias [110, 113, 182]. Correction of the raw contact frequencies is an essential step before any downstream analysis of the chromatin interaction data. Many normalization procedures have been developed over the years to correct the effects of these biases. Implicit methods correct the Hi-C contact map as matrix-balancing problem assuming that genomic regions of equal size should have similar coverage [106-109, 114]. On the other hand, explicit methods normalize the contact matrix by modelling the relationship between the contact frequencies and the known systematic biases introduced primarily by GC content, restriction fragment length and mappability biases [110, 111]. Recently, OneD [113], and implicit-based CAIC (CNV-Adjusted Iterative Correction) [114] have been developed for correcting the CNA-driven bias as well as the

inherited systematic biases (GC-content, mappability, and effective length). They mainly adapted the pipeline of the existing implicit or explicit tools [106, 111] to address the CNA bias effects on Hi-C contact maps. First, the explicit-based approach, OneD [113] normalizes the contact map sequentially for the sample-independent systematic biases (such as GC content, restriction fragment length and mappability) followed by dividing the contact frequencies between two genomic loci by the product of their copy numbers (estimated by a hidden Markov model). On the other hand, the implicit-based CAIC (CNV-Adjusted Iterative Correction) [114] method first performs a local iterative correction followed by CNA-bias normalization assuming the raw interaction count between two CNA loci as product of their copy numbers (estimated using a pruned dynamic programming algorithm [116]). However, our analysis showed that OneD- and CAIC-corrected Hi-C contact maps of cancer cell line do not show high concordance. We noticed that OneD can only partially remove the CNA-induced artifacts of the contact map while CAIC suffers from signal inversion effects in corrected 1D signal.

Here, we developed HiCNAttra (Hi-C Normalization Approach through read depth analysis) framework for copy number estimation and contact map normalization of Hi-C/3C-seq data . HiCNAttra utilizes all types of raw Hi-C reads using a novel “entire fragment” counting approach. This enables estimation of RD signal at high resolution that allows precise detection of both large-scale and focal alterations based on CNAttra approach [183]. For 3C-seq data, we generate the RD signal and CNA profiles exclusively from the genomic reads. Our RD-computing approach from Hi-C/3C-seq data from cancer cell lines successfully recapitulates the RD signal derived from WGS data of the corresponding cell lines. We then utilize a generalized linear model (GLM) to correct the interaction matrix from the sample-driven CNA bias along with sample-independent systematic biases. We applied HiCNAttra on Hi-C/3C-seq data of five cancer cell lines that demonstrated its ability to ameliorate the effects of CNA-induced bias on contact maps. We benchmarked the performance of HiCNAttra against ICE (iterative correction and eigenvector decomposition), OneD, and CAIC using MCF7 breast cancer cell line. Comparative results showed

that HiCNAta-normalized contact maps achieved the least 1D variations with no signal inversion effect and produced a heatmap with homogeneous signal.

5.1 Methods

From the methodology perspective, HiCNAta is superior for detecting both LCVs and FAs, and for correcting the contact frequencies of cancer cell lines due to three main reasons. First, HiCNAta proposes the ‘entire fragment’ counting approach and utilizes all types of Hi-C reads for computing the RD signal from Hi-C libraries. This allows HiCNAta for best estimation of the RD signal at higher resolution. Second, HiCNAta identifies the copy number events using CNAta approach to detect both LCVs and FAs. Third, HiCNAta normalizes the chromatin contact maps for all the systematic biases simultaneously by integrating CNA track with other systematic biases in the GLM. This enables HiCNAta to accurately estimate the contribution of each bias on the interaction frequencies and normalize their effects

5.1.1 HiCNAta framework

HiCNAta utilizes Hi-C/3C-seq reads for both identifying their CNA profiles and normalizing 3D genome contact frequencies. HiCNAta pipeline is divided into three modules (Figure 5.1a): 1) computation of the RD signal from Hi-C or 3C-seq reads (*RD calculator*), 2) RD-based detection of copy number events (*CNA caller*) and 3) bias correction of interaction matrix (*Contact map normalization*). Briefly, HiCNAta pipeline starts with computation of RD signal from the Hi-C/3C-seq reads. Next, we use CNAta [183] approach to identify large-scale and focal alterations from the RD signal and integrate them to generate the CNA track that is used as an explicit bias source for correcting interaction matrix. Finally, we utilize Poisson-based GLM to simultaneously normalize the contact map for GC content, mappability, and fragment length biases as well as biases introduced by copy number gains/losses.

a) HiCNAta pipeline

b) Read depth calculation

Figure 5.1: HiCNAta framework and RD signal computation.

(a) Block diagram of HiCNAta pipeline. **(b)** Cartoon illustrations of the RD signal computation method from Hi-C (left) and 3C-seq (right) reads. For Hi-C, informative and non-informative (only dangling-end shown) reads **(i)** that were mapped within the restriction fragment-end windows **(ii)** are used for RD calculation. Then, number of reads is counted for each window **(iii)**. Base counts are then estimated for each restriction fragment as the summation of the counts of its two fragment-end windows **(iv)**. For 3C-seq libraries **(v)**, genomic reads **(vi)** are exclusively used to compute the RD signal in an unbiased manner similar to WGS paired-end reads **(vii)**.

5.1.1.1 Computing the RD signal from Hi-C/3C-seq NGS reads

Hi-C/3C-seq datasets comprise different read types such as valid pairs, dangling-end, extra dangling-end, self-circle, and single-sided reads as well as genomic reads [105]. These reads can be mainly categorized into (1) ‘informative’ reads containing valid pairs that represent interactions between genomic loci, and (2) ‘non-informative’ reads that include all other types of reads including dangling-end (un-ligated DNA fragment), extra dangling-end (un-ligated DNA fragment that includes a restriction enzyme), self-circle (circularized ligated DNA fragment), single-sided (Hi-C read with only one-side aligned to the genome), and genomic reads [105]. In case of Hi-C, valid pairs generally comprise a maximum of ~70% of the total mapped reads and these reads are solely used to generate contact map. HiCNATra integrates other non-informative reads (dangling-end, extra dangling-end, self-circle, and single-sided reads) along with valid pairs to compute the RD signal at higher resolution. On the other hand, in the absence of biotin labeling and pull-down steps, the majority of 3C-seq reads are contributed by genomic reads. Therefore, for 3C-seq datasets, we compute the RD signal exclusively from these genomic reads similar to RD signal computation from paired-end WGS reads. Depending on the Hi-C or 3C-seq experiment, reads are utilized differently for computing the RD signal (default bin size = 5 kb).

For Hi-C datasets, HiCNATra computes the RD signal from all Hi-C reads (Figure 5.1b(i)) using the “entire (restriction) fragment” counting approach. For this, we first retain only those reads that are located within the restriction fragment-end windows – region of maximal molecule length (MML) adjacent to the restriction site (Figure 5.1b(ii)). The MML is the maximum length of the DNA molecule in the Hi-C NGS library. Therefore, reads that align more than the MML away from the closest restriction enzyme site are results of either non-canonical enzymatic activity (such as star activity resulting in alteration of the specificity of restriction enzyme) or non-enzymatic physical breakage of the chromatin [106]. Second, we count the reads for each restriction fragment based on the assumption that each continuous DNA-sequence read represents a particular restriction fragment and contribute to the abundance of that fragment. Therefore, we

count each non-informative read as a single count whereas for valid pairs, side1 and side2 are counted separately (Figure 5.1b(iii)). So for a particular restriction fragment, the fragment count is calculated as the sum of the number of reads located in both fragment-end windows (Figure 5.1b(iv)). For example, the fragment count of restriction fragment 1 (rFrag1) is the sum of read counts of fragment-end windows, w2 and w3 (Fig. 1b(iv)). Finally, we assign the fragment count of a restriction fragment to all of its nucleotides (bases) and compute the RD signal.

For 3C-seq datasets, genomic reads are distributed uniformly along the genome and represent the majority of reads (Figure 5.1b(v)). Therefore, we use these paired-end genomic reads (Figure 5.1b(vi)) to compute the RD signal in an unbiased manner (Figure 5.1b(vii)).

5.1.1.2 Extracting the bias features of RD signal

After computing the RD signal, we corrected against all systematic biases such as GC-content, mappability score, and length of restriction fragments. For that, we computed these biases from Miller's GC-track [49], the unique mappability track [184], and the reference genome sequence respectively. For 3C-seq data, GC-content and mappability are calculated using CNATra approach [183]. For Hi-C data, we define these tracks based on the fragment-end windows using the same method that is used for computing the RD signal of Hi-C data (Figure 5.1b). First, the mappability and GC scores are computed for each window. Scores are then calculated for each restriction-fragment as the summation of the scores of its two fragment-end windows, resulting in Hi-C mappability and GC tracks. Finally, these tracks are binned to find the mappability and GC scores per bin. In addition, effective (restriction fragment) length feature is defined as the number of nucleotides in the fragment-end windows next to restriction sites per bin.

5.1.1.3 Normalization of the RD signal

RD signal is first filtered to remove the uncertainty areas and low-abundance areas of the genome (described in Chapter 4). Then, we normalize it for the systematic biases by employing the interval (scaling) normalization method [44, 49, 51]. This method is widely-used by CNA

detection methods (CNVnator [44], ReadDepth [49], CLImAT [51]) for correcting the RD signal since it successfully eliminates the effects of the systematic biases without distorting the copy number events. These systematic biases include the GC content (GC-score per bin), mappability, and effective length bias that represents the variable lengths of restriction fragments.

Figure 5.2: Interval normalization successfully normalized the RD signal of normal/cancer Hi-C data from the systematic biases.

Bar chart showing the mean read counts before (blue) and after (red) normalization for each systematic bias (effective length, mappability, and GC-score). For each bias source, the interval of bias values is divided into 20 equally spaced sub-interval and we compute the average RD signals over all bins within each sub-interval. The bar charts are plotted for GM12878 (a), IMR90 (b), MCF7 (c) and LNCaP (d) cell lines.

For each bias source, the interval of bias values is divided into 100 equally spaced sub-interval. A correction factor is computed for each sub-interval as the ratio between the genome-wide mean of the RD signal to the average RD signal of bins with bias value in this sub-interval. Then, these correction factors are used for computing the normalized RD signal.

$$RD_{corrected}^{i,j,k} = RD_{raw}^{i,j,k} \times G^i \times M^j \times E^k \quad (5.1)$$

$$G^i = \frac{\overline{RD}_{genome}}{\overline{RD}_{GC(i)}} \quad (5.2)$$

$$M^j = \frac{\overline{RD}_{genome}}{\overline{RD}_{mappability(j)}} \quad (5.3)$$

$$E^k = \frac{\overline{RD}_{genome}}{\overline{RD}_{effective\ length(k)}}, \quad (5.4)$$

where i, j, k are the sub-intervals indices of GC-content, mappability, effective length biases; respectively. $RD_{raw}^{i,j,k}$ is the raw RD signal for a bin, $RD_{corrected}^{i,j,k}$ is the corrected RD signal of this bin, and $\overline{RD}_{genome}$ is the genome-wide mean of the RD signal. $\overline{RD}_{gc(i)}$, $\overline{RD}_{mappability(j)}$, and $\overline{RD}_{effective\ length(k)}$ are the average RD signal over all bins with the GC-score, mappability, and effective length as in sub-intervals i, j and k , respectively. G^i, M^j , and E^k are the correction factors for a bin with the GC-score, mappability, and effective length in sub-intervals i, j and k , respectively. We evaluated our interval normalization method on Hi-C datasets of different normal and cancer cell lines (Figure 5.2). While the raw RD-signals of different Hi-C datasets shows high correlation with the systematic biases, the corrected RD-signals become uncorrelated with the systematic biases (Figure 5.2).

5.1.1.4 CNA identification

The CNA calling module of HiCNatra is based on CNatra CNA calling module (described in Chapter 4). CNatra CNA caller constitutes the hierarchical framework to delineate the multi-level copy number alterations in the cancer genomes (Figure 4.1a). Briefly, we first define the CN reference (CN=2) by fitting the RD signal to a multimodal distribution. Then we utilize a multi-step framework to first identify large genomic segments with distinct CN state. Segments with CN

state other than 2 ($CN \neq 2$) are considered as large-scale copy number variations (LCVs) or segmental aneuploidies. Next, focal amplifications and focal deletions in each CN-defined segment are detected based on coverage-based thresholding. HiCNAttra then computes the CNA tracks by merging both LCVs and focal alterations (FAs) for Hi-C correction. Hi-C RD signal is usually influenced by the length of restriction fragments, and its maximum resolution is limited by the restriction enzyme. Therefore, we tuned CNAttra parameter to define the significant FA with a minimum width of 100 kb and a minimum CN change of 1N from its parent IB (amplification/deletion threshold = 1N).

5.1.1.5 Computing the chromatin contact map

We use the valid pairs of Hi-C/3C-seq for computing the interaction frequency between all loci of the genome and constructing the contact maps (genome-wide heatmap). Valid pairs are filtered to keep only read pairs within the MML from the restriction sites. Then, we bin the filtered reads at a user-defined resolution (default bin size = 500 kb). Reads are assigned to bins based on the midpoint of the restriction fragments [106]. Interaction frequency per bin is influenced by GC-content, mappability, restriction fragment length, and copy number variations. Therefore, we compute these four bias features for normalizing the contact map. The GC, mappability, and effective-length scores are computed following the HiCNorm method [111].

5.1.1.6 Correction of the chromatin contact map

We built our contact-map correction approach based on the HiCNorm approach [111]. HiCNorm utilizes a GLM for normalizing the interaction frequencies from the systematic biases (GC-content, mappability, and effective length) assuming that they follow a Poisson distribution or negative binomial distribution. In the HiCNorm study, both distributions showed similar bias reductions [111]. Since Poisson regression model is mathematically simpler, it is adapted for the Hi-C normalization.

HiCNorm approach successfully corrected the contact frequency of normal cell lines from the systematic biases [111]. Therefore, for cancer cell lines, we follow the same approach by incorporating the CNA track as a data-associated bias with other systematic biases. Inspired by HiCNorm, we employ GLMs to normalize the contact frequencies against all biases (GC-content, mappability, effective length, and CNA features). Two GLM with Poisson distribution are used to fit the *cis* (intrachromosomal) and *trans* (interchromosomal) contact maps separately. By default, we apply our normalization approach in a genome-wide manner for better estimation of the GLM parameters.

Normalization of *cis* contact map. Let $F = \{f_{kl}^i\}_{1 \leq k, l \leq n_i}$ represent the $n_i \times n_i$ *cis* interaction frequencies for chromosome i , where n_i is the number of bins in chromosome i . Each element f_{kl}^i represents the number of interactions between genome loci from bin k and bin l in chromosome i . Let x_k^i, y_k^i, z_k^i , and w_k^i represent the GC content, mappability, effective length, CNA features of bin k in chromosome i , respectively; whereas x_l^i, y_l^i, z_l^i , and w_l^i represent the features of bin l in chromosome i , respectively. The interaction frequency f_{kl}^i follows a Poisson distribution with rate λ_{jk} :

$$\log(\lambda_{jk}) = \alpha_0^i + \alpha_1^i * \log(x_k^i x_l^i) + \alpha_2^i * \log(y_k^i y_l^i) + \alpha_3^i * \log(z_k^i z_l^i) + \alpha_4^i * \log(w_k^i w_l^i), \quad (5.5)$$

where α_{0-4}^i are the coefficients of the GLM. We fit this model and use the coefficient estimates $\hat{\alpha}_{0-4}^i$ for computing the estimated Poisson rates:

$$\hat{\lambda}_{kl} = \exp [\hat{\alpha}_0^i + \hat{\alpha}_1^i * \log(x_k^i x_l^i) + \hat{\alpha}_2^i * \log(y_k^i y_l^i) + \hat{\alpha}_3^i * \log(z_k^i z_l^i) + \hat{\alpha}_4^i * \log(w_k^i w_l^i)]. \quad (5.6)$$

Then, $\hat{\lambda}_{kl}$ is used to compute the normalized interaction frequency $\hat{f}_{kl}^i$ between bin k and bin l in chromosome i of bin:

$$\hat{f}_{kl}^i = \frac{f_{kl}^i}{\hat{\lambda}_{kl}} \quad (5.7)$$

Normalization of *trans* contact map. Let $F = \{f_{kl}^{ij}\}_{1 \leq k \leq n_i, 1 \leq l \leq n_j}$ represent the $n_i \times n_j$ *trans* interaction frequencies between chromosome i and chromosome j , where n_i and n_j are the

number of bins in chromosomes i and j , respectively. Each element f_{kl}^{ij} represent the number of interactions between genome loci of bins k and l from chromosomes i and j , respectively. Let x_k^i , y_k^i , z_k^i , and w_k^i represent the GC content, mappability, effective length, CNA features of bin k in chromosome i , respectively; whereas x_l^j , y_l^j , z_l^j , and w_l^j represent the features of bin l in chromosome j , respectively. The interaction frequency f_{kl}^{ij} follows a Poisson distribution with rate λ_{jk} :

$$\log(\lambda_{jk}) = \alpha_0^{ij} + \alpha_1^{ij} * \log(x_k^i x_l^j) + \alpha_2^{ij} * \log(y_k^i y_l^j) + \alpha_3^{ij} * \log(z_k^i z_l^j) + \alpha_4^{ij} * \log(w_k^i w_l^j), \quad (5.8)$$

where α_{0-4}^{ij} are the coefficients of the GLM. We fit this GLM model and use the coefficient estimates $\hat{\alpha}_{0-4}$ for computing the estimated Poisson rates:

$$\hat{\lambda}_{kl} = \exp [\hat{\alpha}_0^{ij} + \hat{\alpha}_1^{ij} * \log(x_k^i x_l^j) + \hat{\alpha}_2^{ij} * \log(y_k^i y_l^j) + \hat{\alpha}_3^{ij} * \log(z_k^i z_l^j) + \hat{\alpha}_4^{ij} * \log(w_k^i w_l^j)]. \quad (5.9)$$

Then, $\hat{\lambda}_{kl}$ is used to compute the normalized interaction frequency $\hat{f}_{kl}^{ij}$ between bins k and l from chromosomes i and j , respectively:

$$\hat{f}_{kl}^{ij} = \frac{f_{kl}^{ij}}{\hat{\lambda}_{kl}} \quad (5.10)$$

5.1.2 Data availability

We used six publicly available Hi-C datasets (GM12878, IMR90, PrEC, MCF7, LNCaP and PC3) and two in-house generated 3C-seq datasets (K562 and NCI-H69). We used input control reads of ChIP-seq experiments of MCF7, LNCaP and K562 cells to generate WGS-derived RD signal to validate HiCNAta RD signal (Table 5.1). For Hi-C, the data coverage is approximately computed as $\frac{\text{number of reads} * \text{read length}}{\text{genome length}}$. For 3C-seq and WGS, the data coverage is computed as the total count of base pairs divided by the genome length. We also compute the coverage gain that represents the percentages of all non-valid pair reads to the total reads from Hi-C data. This reflects the advantage of integrating the non-informative reads.

The iterative mapping has been shown to recover the maximum numbers of aligned reads compared to fixed length mapping [106]. Therefore, for both Hi-C and 3C-seq datasets, we applied

the iterative-mapping technique of hiclib using default parameters for aligning short sequence reads to human reference genome GRCh37 (hg19) [106]. The iterative mapping output HDF5 files are used as input for HiCNAta tool. Classification of read types (valid pairs, dangling-end, extra dangling-end, self-circle) depends on the maximum molecule length (MML) which is not always mentioned for the public available data. Therefore, for accurate analysis, we tuned it based on the fragment length distribution of dangling-end/extra dangling-end reads (Table 5.1, Figure 5.3). For each dataset, we approximately set the MML as the size (in hundreds of bp) that is greater than the fragment length (side1-start to side2-end of a paired-end read) of 99% of dangling-end and extra-dangling end reads (Figure 5.3).

Table 5.1: Datasets for evaluating the performance of HiCNAta.

Cell Line	GEO Accession	Data	RD coverage (x)	Coverage gain (%)	MML
MCF7	GSM1631185	Hi-C	12.89	30.308	400
LNCaP	GSM1902607-14	Hi-C	25.329	37.0834	600
PC3	GSM1902605-06	Hi-C	6.9794	29.42	800
PrEC	GSM1902602-04	Hi-C	19.95	25.5577	500
GM12878	GSM1181867	Hi-C	19.727	22	600
IMR90	GSM1055800	Hi-C	4.07	43.49788	600
K562	In-house (Sanyal lab, NTU)	3C-seq	14.26		1000
NCI-H69	In-house (Sanyal lab, NTU)	3C-seq	12.374		1000
MCF7	GSM2258738	ChIP-seq Input control	0.293		
LNCaP	GSM1902617	ChIP-seq Input control	0.6298		
K562	ENCFF-350QUB,- 466QHR, -851QFR, - 922JGL	ChIP-seq Input controls (merged)	2.2927

Figure 5.3: Fragment length of dangling-end and extra dangling-end reads from Hi-C/3C-seq datasets.

5.2 Results

5.2.1 HiCNAta computes high-resolution RD signal by utilizing all types of Hi-C reads

Data coverage (number of reads) is an important factor for detecting the CN events, especially the FAs. Hi-C reads comprise different read types such as valid pairs and other reads which contain dangling-end, extra dangling-end, and single-sided mapped reads [105]. Although valid pairs are the main target of Hi-C/3C-seq protocols for capturing the chromatin interactions, other non-informative reads can be additionally utilized for estimation of the RD signal at a higher resolution. Based on the six Hi-C datasets we analyzed, Hi-C valid pairs represent 39-64% of the total mapped reads (Table 5.1). Other reads, which represent a significant chunk (36-61% of Hi-C total mapped reads), are usually filtered out during Hi-C downstream analysis. However, these reads are still ‘informative’ for estimation of the genome-wide RD signal at a higher resolution. Therefore, we took advantage of these ‘unutilized’ reads and computed RD signal using both valid pairs and non-informative reads. Addition of these non-informative reads increases the resolution of the RD signal by 22-44% for different Hi-C datasets (Table 5.1). The validity of incorporating non-informative reads to estimate the RD signal is evaluated by comparing “read counts per restriction fragment” from all Hi-C reads versus valid pairs only. We found that read counts are highly correlated (Spearman correlation 80-91%) confirming the validity of using non-informative Hi-C reads in RD calculation which in turn results in RD signal computation at high resolution. This conferred advantage to HiCNAta over other tools (HiCnv, OneD, CAIC) which use only the valid pair reads for estimating the copy numbers [113, 114, 117].

On the other hand, 3C-seq technique, which combines 3C library with NGS sequencing step, results in a higher percentage of genomic reads among the double-sided mapped reads (66% for K562 and 77% for NCI-H69). Therefore, similar to WGS datasets, RD signals can be conveniently computed from these genomic reads. Each paired-end genomic read is used to calculate the

fragment length (side1-start to side2-end). This results in computing RD signal at high coverage (14.26x for K562 with 117 million genomic reads and 12.37x for NCI-H69 with 120 million reads). This demonstrates that genomic reads from the 3C-seq datasets can be effectively used for estimating the RD signal in an unbiased manner without any additional cost of WGS data. Therefore, 3C-seq contact frequencies can be self-normalized using its genomic reads.

5.2.2 Entire fragment counting approach successfully extracts copy number-associated features of RD signal

Accurate estimation of the RD signal from Hi-C data is the primary key for identification of copy number events, LCVs and FAs. Generally, most Hi-C analysis tools [[106-111](#), [113](#), [114](#), [117](#)] assign the Hi-C reads to the midpoint of corresponding restriction fragment (midpoint approach). Alternatively, exact cut site of the read can be used (exact-cut approach) [[182](#)]. These approaches are suitable for contact frequency calculation since Hi-C matrices are usually generated using large bin size compared to the restriction fragment length. Usually, this bin size is at least one order of magnitude larger than the fragment length. However, smaller bins are essential for effective identification of change points at high-resolution for discovery of both LCVs and FAs. Therefore, we introduce “entire (restriction) fragment” counting approach, for RD calculation, where RD signal counts all base pairs (bps) of a restriction fragment if any read is mapped to this fragment (Figure [5.1b](#)). This allowed us to compute the RD signal at high resolution.

We compared HiCNAta RD calculation method with exact-cut and midpoint approaches using MCF7 Hi-C data generated using *HindIII* enzyme (Figure [5.4](#)). We calculated the RD signal at 5-kb bin using Hi-C reads which is of the same order of magnitude as the experimental resolution (average 4096 bases for 6-bp cutter).

Figure 5.4: Evaluation of HiCNatra entire restriction-fragment counting approach for computing the RD signal.

(a) The coverage plot (left) at 5-kb bin and the RD frequency distribution (right) of MCF7 Chr16:56-88Mb region extracted from Hi-C data using exact-cut mapping (top), midpoint mapping (middle), and HiCNatra entire restriction-fragment counting (bottom) approaches. Each red dot represents copy number per bin. (b) Box plot of centralization scores (5-kb and 100-kb bin) and zero scores (5-kb bin) computed from the Hi-C-derived RD signal of 6 cell lines (MCF7, LNCaP, PC3, GM12878, IMR90, PrEC) using entire fragment, exact-cut, and midpoint counting approaches at 5-kb and 100-kb bin sizes.

Interestingly, we found that our approach can effectively capture LCVs as evidenced by two peaks in the RD signal distribution (bottom panel in Figure 5.4a right). In addition, FAs are conspicuous using this approach viz. FA1 and FA2 (bottom panel in Figure 5.4a left). In comparison, the RD signal generated using other approaches show overdispersion and make the distribution more skewed toward bins with low reads (top and middle panels in Figure 5.4a). Consequently, their RD frequency distributions are more skewed toward bins with low reads (bottom panel in Figure 5.4a left).

For a more quantitative comparison, we computed the centralization score and zero score for all the six Hi-C datasets. Centralization score (CS) is computed as the percentage of bins that are near their integer CN states (described in Chapter 3). CS can be used as a measure of distribution of genomic loci into distinct CN states. Next, we defined zero score (ZS) as the percentage of bins with CN=0. ZS can be used as a measure of the sparseness of the RD signal. We found that entire fragment counting approach achieved the maximum CS (for both 5-kb and 100-kb bin) and minimum ZS in all Hi-C datasets, compared to the exact-cut and midpoint approaches (Figure 5.4b). This is also confirmed by the visual inspection of coverage plot and RD frequency distribution derived from these three approaches (Figure 5.4a). High CS implies that RD signal is distributed near distinct CN states which aids in the identification of CNAs. Low ZS is essential to avoid false detection of deleted regions, especially in large-sized restriction fragments.

Figure 5.5: Visual comparison between genome-wide RD signal derived from Hi-C/3C-seq and WGS datasets (WGS data is taken from input control data of ChIP-seq experiment).

(a) The genome-wide coverage plot (chr1-chrX) of MCF7 cancer cell computed from Hi-C data (top) and WGS data (bottom). (b) The genome-wide coverage plot (chr1-chrX) of LNCaP cancer cell computed from Hi-C data (top) and WGS data (bottom). (c) The genome-wide coverage plot (chr1-chrX) of K562 cancer cell computed from 3C-seq data (top) and WGS data (bottom). Each dot represents the RD value per bin computed from Hi-C/3C-seq data (red) or WGS data (grey). WGS data is obtained from “input control” of ChIP-seq experiment.

Next, we validated that the entire fragment counting approach extract the RD signal correctly from Hi-C reads. We compared the RD signal derived from Hi-C reads of MCF7 and LNCaP with the WGS-derived coverage signal of these cell lines. First, we found that HiCNAttra RD signal has the highest correlation with the WGS-derived RD signal for both MCF7 (Spearman's $\rho = 0.714$) and LNCaP (Spearman's $\rho = 0.55$) cells. In comparison, exact-cut and midpoint approaches have the Spearman's ρ of 0.32 and 0.31 for MCF7, and 0.26 and 0.24 for LNCaP cells, respectively. This confirmed that the HiCNAttra counting approach is the best estimator of the RD signal. Similarly, we computed the RD signal (5-kb bin) from genomic reads of K562 3C-seq dataset. As expected, RD signal extracted from K562 3C-seq data correlated well with WGS-derived coverage signal of K562 cells (Spearman's $\rho = 0.573$). This confirmed that the RD signal can be reproducibly derived from Hi-C/3C-seq datasets as visually illustrated in Figure 5.5. We plot the genome-wide RD signal from Hi-C/3C-seq data (Figure 5.5 top panels) versus the RD signal from computed from input control data of ChIP-seq experiments (Figure 5.5 bottom panels) of MCF7, LNCaP and K562 cell lines. Red dots are bins of the RD signal from Hi-C/3C-seq data and grey dots represent bins of the RD signal from the input control data of ChIP-seq experiment (Figure 5.5).

5.2.3 HiCNAttra can extract high-resolution copy number events from Hi-C/3C-seq datasets.

Successful cloning of the WGS-derived RD signal from Hi-C/3C-seq reads enables HiCNAttra to accurately detect CNA regions. To the best of our knowledge, HiCnv is the only tool available to detect CNA from Hi-C data [117]. However, HiCnv computes the RD signal exclusively from the valid pairs using midpoint approach. As shown before, valid pairs represent 39-64% of total mapped-Hi-C reads. Therefore, HiCnv sacrifices a large percentage of reads which can be effectively used for calculating the RD signal based on HiCNAttra approach. Moreover, the midpoint approach limits the power of HiCnv to detect only large CNAs (size >1 Mb) [117]. As an alternative solution, we adapted the hierarchical CNA detection approach of CNAttra (described in Chapter 4) in HiCNAttra CNA caller module to identify both LCVs and FAs.

We analyzed the CNA profiles of all six Hi-C and two 3C-seq datasets using HiCNAttra (Table 5.1). Results show that all cancer cell lines (MCF7, LNCaP, PC3, K562, NCI-H69) are enriched for LCVs (41-122 regions) with a median width of 5.7-45.7 Mb, whereas normal cell lines (GM12878, IMR90 and PrEC) are almost free of LCVs (1-2 regions) (Figure 5.6 top panel). On the other hand, FAs are pervasive in both cancer (162-238 regions) and normal cell lines (25-107 regions) with median width ranging from 157 to 225 kb (Figure 5.6 top panel). FAs are divided into focal amplifications (5-161 regions), homozygous deletions (CN=0) (10-72 regions), and hemizygous deletions (CN=1) (7-95 regions). Visual inspection of the CNA profiles generated from MCF7 Hi-C data and NCI-H69 3C-seq data clearly illustrated that chromosomes are interspersed with both large-scale and focal alteration events (Figure 5.7). In this figure, the blue and grey dots are bins belonging to focally amplified/deleted regions, and CN neutral regions respectively. The red line represents the copy number of each CNA (any amplitude transition indicates a new CNA region) whereas the black line represents the copy number of IBs/LCVs only (any amplitude transition indicates a new CNA region). The green vertical lines denote the boundaries of centromeres and telomeres. Detection of the complete CNA profile allows HiCNAttra to filter and normalize the contact frequencies of Hi-C/3C-seq datasets. First, we use the homozygous deletions besides the mappability- and GC-tracks for filtering out bins belonging to uncertainty and low-abundance genomic loci. Then, we integrate the LCVs, focal amplifications and hemizygous deletions to form the CNA tracks that are used for correcting the contact frequencies.

(a) Large-scale copy number variations

(b) Focal alterations

Figure 5.6: Numbers and widths of large-scale copy number variations (LCVs) and focal alterations (FAs) estimated from different cell lines.

Box plot of the width/length (in Mb) of LCVs (a) and focal alterations (b) estimated from Hi-C/3C-seq data are shown. The numbers of LCVs and FAs are denoted at the bottom of the box plot for each cell line. The LCVs detected in the normal cell lines (GM12878 and IMR90) are located at the vicinity (boundary) of the telomeres.

Figure 5.7: Visual illustration of copy number events identified from Hi-C/3C-seq reads.

Copy number profiles computed from MCF7 Hi-C data (a) and NCI-H69 3C-seq data (b) show that cancer cell lines carry both large-scale copy number variations (LCVs) and focal alterations (FAs). The red line represents the CNA track. Any amplitude transition indicates a new CNA region. The grey dots represent RD signal per bin whereas the blue bins indicate focally amplified/deleted regions. The green vertical lines denote the boundaries of centromere and telomeres.

5.2.4 HiCNAttra successfully ameliorates the effects of CNAs on the contact map

Copy number alteration is a hallmark of cancer cells. It is well-established that chromatin contact map is highly influenced by the copy number events [179, 180]. The copy number change may lead to apparent increment or decline in interaction frequencies creating visible artifacts on contact maps. This results in fallacious interpretation of the chromatin conformation data at different levels. Hence, we developed HiCNAttra which utilizes GLM with Poisson distribution to normalize the chromatin contact frequencies versus all bias sources simultaneously including CNA-induced bias.

In order to measure the impact of CNA bias on contact frequencies, we computed the Spearman correlation between different biases and the interaction frequencies of the raw (uncorrected) and HiCNAttra-corrected contact maps. As previously reported, our analysis using five cancer cell lines also demonstrated that raw interaction frequencies are positively correlated with the copy number track derived from Hi-C/3C-seq data of the same cell line. In fact, we found that CNA and effective (restriction fragment) length impart the dominant biases on the raw contact map of cell lines with abnormal karyotypes (Figure 5.8a). For example, Spearman correlation of genome-wide interaction frequencies with CNA is 0.47 for MCF7 and 0.51 for LNCaP, whereas Spearman correlation with restriction fragment length is 0.34 for MCF7 and 0.41 for LNCaP. In contrast, we observed that HiCNAttra-corrected interaction frequencies became virtually uncorrelated with the CNA tracks (Spearman correlation of genome-wide interaction frequencies with CNA is 0.1 for MCF7 and 0.13 for LNCaP) as well as with other systematic biases across cell lines (Figure 5.8a).

Figure 5.8: HiCNAttra normalization successfully ameliorates the effects of CNAs on the chromatin contact maps.

(a) Bean plot showing Spearman correlations between *cis* contact frequencies and sample-dependent (CNA) and sample-independent (effective length, GC-content and mappability) biases in raw/uncorrected (orange) and HiCNAta-corrected (blue) contact maps. Correlations were calculated across chromosomes of five cancer datasets (MCF7, LNCaP, PC3, K562, NCI-H69). **(b)** Raw (top) and HiCNAta-normalized (bottom) interaction heatmaps (500-kb bin) of MCF7 chr 1. Copy number track of chr1 computed by HiCNAta from MCF7 Hi-C data was shown on top. Each grey dot represents the copy number of a bin. The red line represents the copy number track where any amplitude transition indicates a new CNA region. The numbers on the axes of the heatmap represent the bin numbers. Two amplified regions CNA1 and CNA2 were shown that resulted in two off-diagonal block patterns (visible artifacts). **(c)** Post-normalized Hi-C interaction heatmaps (500-kb bin) of MCF7 chr1 using ICE (top), CAIC (middle), and OneD (bottom) tools. Copy number track was shown on top. For the OneD tool, *OneD+CN* normalization module was used.

As a complementary approach, visualization and manual inspection of contact heatmaps are simple and effective ways to qualitatively ascertain the effects of CNAs on interaction frequencies pre- and post-normalization. Cancer cell lines, which typically harbor amplifications and deletions, create visible artifacts on the raw heatmaps. These alteration regions in the CNA tracks show concomitant changes in contact frequencies in the raw (uncorrected) heatmap (Figure 5.8b, Figure 5.9, and Figure 5.10). For instance, in MCF7 chromosome 1, highly-amplified regions CNA1 and CNA2 resulted in two off-diagonal block patterns (artifacts) in the heatmap with signal higher than their surrounding regions (Figure 5.8b, shown by arrowhead). These blocks indicate the overrepresented *cis* interactions between these CNA regions with other genomic regions. As expected, similar artifacts are also observed on the raw heatmaps of chromosomes 3, 9, 13 and 15 of MCF7 (Figure 5.9) as well as different chromosomes of other cancer cell lines (LNCaP chr3, PC3 chr14, NCI-H69 chr2, and K562 chr15) (Figure 5.10). Interestingly, in all chromosomes, HiCNAta ameliorates CNA-induced artifacts as evidenced by removal of off-diagonal lines (artifacts) from the corrected heatmap without distorting the inherent chromatin landscape such as domain structures (Figure 5.8b, Figure 5.9, and Figure 5.10). Collectively, these data suggest that HiCNAta approach can successfully diminish the effects of CNA and all other systematic biases effectively.

Figure 5.9: Visual illustration of the performance of Hi-C correction tools (ICE, CAIC, OneD, HiCNAttra) using MCF7 cancer cell line having abnormal karyotype.

For each chromosome (chr3, chr9, chr13, and chr15), the top panel shows the copy number (CNA) track computed by HiCNAttra. Each grey dot represents the copy number of a bin. The black line represents the copy number track where any amplitude transition indicates a new CNA region. The second panel shows the uncorrected contact map (raw). The next four panels show the post-corrected contact-map using ICE, CAIC, OneD and HiCNAttra tools.

Figure 5.10: HiCNATra successfully eliminated the CNA-induced artifacts on the heatmaps of Hi-C/3C-seq datasets (LNCaP, PC3, NCI-H82, K562).

For each chromosome (LNCaP chr3, PC3 chr14, NCI-H69 chr2, and K562 chr15), the top panel shows the copy number (CNA) track computed by HiCNATra. Each grey dot represents the copy number of a bin. The black line represents the copy number track where any amplitude transition indicates a new CNA region. The second and fourth panels show the contact map before correction (raw) and post-correction using HiCNATra, respectively. The third and fifth panels show the 1D signal before correction (raw) and post-correction using HiCNATra, respectively.

5.2.5 HiCNATra better attenuates the 1D signal variation without distorting its inherent features

We evaluated the performance of HiCNATra against widely-used ICE method and CNA-targeted Hi-C correction tools, OneD and CAIC. As a qualitative comparison, we visually inspected the corrected heatmaps resulted from these tools and interpret their abilities to ameliorate CNA-induced artifacts. Though OneD-corrected heatmap showed improvement over the raw heatmap, the footprints of off-diagonal blocks (artifacts) of weaker strength are still visible after bias

correction (Figure 5.8c). We witnessed similar observations while inspecting other regions of MCF7 for OneD (Figure 5.9). In comparison, to the naked eye, iterative correction-based methods- ICE and CAIC- performed similarly as HiCNAttra in ameliorating the CNA artifacts (Figure 5.8c and Figure 5.9). This further strengthens the fact that HiCNAttra's explicit approach is fairly adequate for correcting contact maps of cancer cell lines with abnormal karyotypes.

Nonetheless, the visual interpretation is not sufficient to evaluate the performance of Hi-C normalization techniques. The main criteria for any normalization should be to attenuate the signal variations without changing the inherent behavior of the data. As defined by implicit-based approach, the ultimate target of Hi-C normalization is to achieve near-equal visibility of all genomic regions. It is self-evident that this normalization should also preserve the inherent behavior of the chromatin interaction landscape. Therefore, we compared the 1D tracks of the corrected heatmaps generated by all tools to evaluate the quality of their normalization in attenuating the overall biases. For this comparative analyses, we have used MCF7 Hi-C data as this breast cancer cell line have abnormal karyotypes with extensive copy number gain/loss [4]. Additionally, both OneD and CAIC used this cell line for performance evaluation. Next, we plotted 1D signal (the summation of each column of the contact matrix) of raw and corrected heatmaps generated by ICE, CAIC, OneD and HiCNAttra (Figure 5.11a,b and Figure 5.12). Visually, the HiCNAttra-corrected 1D tracks showed the minimum overall variation compared to the other tools (Figure 5.11a,b and Figure 5.12). This is substantiated by the violin plots and cumulative plot of standard deviations per chromosome (Figure 5.11c,d). All other tools diminish the 1D signal variation compared to the raw (uncorrected) signal with varying degrees. The iterative-correction-based methods, CAIC and ICE, achieve the second-least and third-least variation levels respectively (Figure 5.11c,d).

Figure 5.11: HiCNAttra approach successfully attenuates the 1D signal variation of MCF7 without distorting the signal trend.

(a, b) Visual comparison of 1D signal of pre-normalized (raw) and normalized contact map of MCF7 chr1 **(a)** and chr9 **(b)** using different Hi-C correction tools. For each chromosome, the top panel shows the HiCNAttra-computed copy number track where each grey dot represents the copy number of a bin and the black line represents the copy number track. The next five panels show the raw and post-corrected 1D signal (red dots) using ICE, CAIC, OneD and HiCNAttra tools, respectively. The blue shows the moving average of the 1D signal. The black lines with arrows represent a region showing signal inversion effects. For each post-normalized 1D signal, the Spearman correlation (ρ) and standard deviation (s) are indicated next to each panel. **(c)** Bean plot of the standard deviations (std) per chromosome of the *cis* contact map corrected by ICE (green), CAIC (purple), OneD (red) and HiCNAttra (blue) tools. **(d)** Cumulative standard deviation per chromosome of the *cis* interaction frequencies normalized by ICE (green), CAIC (purple), OneD (red) and HiCNAttra (blue) tools. For the OneD tool, *OneD+CN* normalization module was used in all the analyses.

Figure 5.12: Visual comparison of the Hi-C correction tools (ICE, CAIC, OneD, HiCNAttra) in attenuating the 1D signal variation using MCF7 cancer cell line.

For each chromosome, the top panel shows the copy number (CNA) track computed by HiCNAttra. Each grey dot represents the copy number of a bin. The black line represents the copy number track where any amplitude transition indicates a new CNA region. The second panel shows the raw 1D signal without correction (raw). The next four panels show the post-correction 1D signal using ICE, CAIC, OneD and HiCNAttra tools.

Intriguingly, upon close inspection, we observed the ‘flipping’ (signal inversion effect) of the signal in ICE-corrected 1D trajectory relative to the raw 1D signal (Figure 5.11a,b, shown by black lines). We witnessed similar ‘flipping’ effects for other chromosomes of MCF7 (Figure 5.12). This is further substantiated by the fact that the ICE-corrected 1D signal of all MCF7 chromosomes showed negative correlations compared to the raw 1D signal (Table 5.2). This signal inversion (flipping) effect contradicts the general idea of normalization which tend to attenuate the bias effect without distorting the main feature of the signal. This also confirms the previous studies which demonstrated ICE method is not suitable for normalizing Hi-C contact map of cancer cells with abnormal karyotype [113, 114]. These studies illustrated that ICE results in overcompensation of original CNA bias leading to distortion of the interaction frequency signal. Though CAIC attempted to overcome this problem by utilizing CNA information in their matrix correction, their 1D signal still carry the footprints of signal inversion effects (Figure 5.11a,b, shown by black lines) as evidenced by negative correlation of their 1D signal relative to the uncorrected ones in 16 chromosomes (Table 5.2). It is noteworthy to mention here that CAIC also utilizes local iterative correction in their approach. This clearly suggests that iterative correction-based approach may not be suitable for normalizing Hi-C data of cells with abnormal karyotypes.

It is also intriguing to us that OneD method, which is based on similar explicit-based approach as HiCNAttra, has not performed to its potential. Briefly, OneD first utilizes a GAM with negative binomial (NB) distribution to correct the contact frequencies from the effective fragment length, mappability, and GC-content. Then, it normalizes the resulted interaction frequencies again for the CNA bias, estimated by a hidden Markov model (sequential bias correction). Therefore, the sub-optimal performance of OneD against HiCNAttra in correcting contact maps can be attributed to two aspects: 1) HMM-based 8-states CNA calling of OneD versus CNAttra-based CNA calling of HiCNAttra and/or 2) *sequential* bias correction of OneD versus *simultaneous* bias correction of HiCNAttra.

Table 5.2: Spearman correlation between the raw 1D signal and corrected 1D signal of all MCF7 chromosomes by Hi-C correction tools (ICE, CAIC, OneD, HiCNAttra).

Chromosome	ICE	CAIC	OneD	HiCNAttra
chr1	-0.73064	-0.130231	0.063536	0.103533
chr2	-0.619891	-0.477564	0.735285	0.566515
chr3	-0.697526	-0.341217	0.508838	0.423114
chr4	-0.403214	-0.360016	0.609564	0.733345
chr5	-0.690545	-0.0412825	0.657112	0.1272
chr6	-0.427902	-0.297994	0.622519	0.587857
chr7	-0.594512	-0.233437	0.530939	0.477461
chr8	-0.862631	0.286152	0.796497	-0.0672376
chr9	-0.257246	-0.507427	0.530248	0.536933
chr10	-0.405442	-0.332628	0.359638	0.204658
chr11	-0.547111	-0.338925	0.10694	0.333631
chr12	-0.576321	0.0177084	0.125575	0.232727
chr13	-0.769818	-0.0540486	-0.00683562	0.326046
chr14	-0.489251	-0.544295	-0.211464	0.22315
chr15	-0.566855	0.0324998	0.689149	0.216919
chr16	-0.508687	-0.421141	0.517179	0.248787
chr17	-0.640297	0.234448	0.326991	-0.130521
chr18	-0.717771	-0.357175	0.456625	0.780741
chr19	-0.324659	0.359993	0.09225	0.531433
chr20	-0.794497	-0.113305	0.414559	-0.445632
chr21	-0.055987	-0.421802	0.597194	0.405315
chr22	-0.0906334	0.151441	0.437842	0.344997

In order to test the impact of CNA calling method on Hi-C normalization, we implemented the ‘*modified OneD*’ method by incorporating HiCNAta-generated CNA tracks instead of their default HMM-based CNA information. We found utilization of HiCNAta CNA information alone enhanced the performance of *modified OneD*. The *modified OneD*-corrected heatmap showed less variation in 1D signal compared to the default OneD method (Figure 5.13 and Figure 5.14). This suggests that the estimation of CNA profiles from Hi-C data plays an important role in contact map normalization of cancer cell lines with abnormal karyotypes. Next, we applied HiCNAta and *modified OneD* methods to five Hi-C/3C-seq contact maps of cancer cell lines. We observed that *modified OneD*-normalized heatmap still harbors the footprint of off-diagonal blocks (artifacts) of weaker strength (Figure 5.13). Next, we plotted 1D signal of raw and corrected heatmaps generated by *modified OneD* and HiCNAta (Figure 5.13). Visually, the HiCNAta-corrected 1D tracks showed less variation compared to the *modified OneD* (Figure 5.13). This is substantiated by the violin plots and cumulative plot of standard deviations per chromosome (Figure 5.14a,c).

Additionally, the HiCNAta-corrected interaction frequencies achieved the minimum correlation with CNA tracks across cell lines (Figure 5.14b and Figure 5.15). For example, MCF7 genome-wide interaction frequencies with CNA is 0.1 for HiCNAta and 0.24 for *modified OneD*. Similarly, for LNCaP, Spearman correlation is 0.13 for HiCNAta and 0.33 for *modified OneD*. Similarly, the HiCNAta-corrected interaction frequencies showed similar or lesser correlations with other systematic biases, compared to *modified OneD*-corrected interaction frequencies, except for GC-score in K562 and NCI-H69 datasets. However, HiCNAta-corrected interaction frequencies of K562 and NCI-H69 datasets are uncorrelated with GC-score (Spearman correlation of ~ 0.1) (Figure 5.14b and Figure 5.15).

Taking together, we speculate that the advantage of HiCNAta is due to improved RD calculation and CNA calling as well as implementing *simultaneous* bias correction as opposed to *sequential* bias correction of OneD. Moreover, HiCNAta approach is fundamentally same as HiCNorm [111] approach in case of normal cells with copy-number neutral karyotype. Therefore, we envisage

that HiCNAttra will be equally applicable for correcting Hi-C contact maps of normal as well as cancer cells.

Figure 5.13: Visual comparison between HiCNAttra simultaneous correction approach and OneD sequential correction approach after integrating HiCNAttra-computed CNA track (modified OneD) using MCF7 cancer cell line.).

For each chromosome (chr1, chr3, chr13, chr15), the top panel shows the CNA track computed by HiCNAttra. Each grey dot represents the copy number of a bin. The black line represents the copy number track where any amplitude transition indicates a new CNA region. The second, fourth and sixth panels show the raw, *modified OneD*-corrected, and HiCNAttra-corrected contact maps respectively. Similarly, the third, fifth, and seventh panels show the raw and *modified OneD*- and HiCNAttra-corrected 1D signal respectively.

Figure 5.14: HiCNAttra simultaneous approach is better than OneD sequential approach for correcting Hi-C/3C-seq cancer datasets.

(a) Bean plot of the standard deviations per chromosome of the contact map corrected by OneD, *modified OneD*, and HiCNAttra methods. **(b)** Spearman correlation between the biases and the corrected *cis* contact maps using *modified OneD* and HiCNAttra methods. **(c)** Cumulative standard deviation per chromosome of the raw, OneD-corrected, *modified OneD*-corrected, and HiCNAttra-corrected *cis* contact maps.

Figure 5.15: Spearman correlation between the biases and the raw, HiCNAttra-corrected, *modified OneD*-corrected, and cis contact maps for cancer Hi-C/3C-seq datasets (MCF7, LNCaP, PC3, K562, NCI-H69).

5.3 Discussion

Cancer genomes carry pervasive large-scale CNAs and focal alterations. Amplification and deletion events will naturally result in over- and under-representation of chromatin interactions respectively in Hi-C/3C-seq contact matrices. Previous studies have demonstrated the adverse effects of CNA regions on TAD sizes and structures in cancer cells. Our analysis also pointed out a positive correlation between the copy number and the strength of the interaction frequencies of genomic loci. This clearly suggests that CNA imparts a dominant explicit bias in case of cancer cell lines with abnormal karyotypes. Therefore, CNAs can greatly influence the interpretation of chromatin contact maps and may lead to false identification of genomic regions with over- or under-represented interaction strengths. The effect of CNA bias should be compensated for the amplified/deleted regions along with other systematic biases for accurate interpretation of interaction frequency with respect to the CN-neutral regions. Therefore, we employed a two-pronged approach for correcting contact map using 1) high-resolution CNA tracks derived from Hi-C/3C-seq reads and 2) normalization of contact frequencies by simultaneously integrating all the systematic biases including CNA bias.

High-resolution extraction of RD signal from Hi-C/3C-seq datasets is central to discover copy number events of smaller sizes. We propose two solutions to better estimate RD signal from Hi-C datasets. First, HiCNATra utilized all types of Hi-C reads resulting in high coverage RD signal unlike other tools which use only valid pairs. Second, HiCNATra employs the “entire fragment” counting approach to create the RD signal track which closely matches the WGS-derived RD profile of the corresponding cell line.

Utilizing a Poisson regression model, HiCNATra aims to learn the effects of CNAs on the contact map in conjunction with other systematic biases. Therefore, HiCNATra can estimate the contributions of each bias simultaneously instead of merely assuming a pre-defined specific effect of the CNAs on the interaction frequencies. HiCNATra thus utilized the GLM to normalize the

contact frequencies (the variable) versus CNA, effective fragment length, mappability, and GC-content (the predictors) simultaneously. In principle, HiCNAttra can also include newer sources of biases in the same method and therefore it is adaptable to new biases. HiCNAttra-corrected contact maps alleviate the CNA-induced visual artifacts as evidenced by near uncorrelation of contact frequencies with the CNA regions.

Our benchmarking with ICE, OneD and CAIC methods clearly demonstrated that HiCNAttra is superior to other existing tools in eliminating the CNA effects on contact matrices. The inversion effect on 1D signal exhibited by matrix-balancing approaches clearly suggests that explicit-based Hi-C normalization methods are more suitable and reliable to overcome the CNA bias. OneD normalization using HiCNAttra-calculated CNA track as input (modified OneD) pointed out the importance of capturing of high-resolution CNA events from Hi-C/3C-seq data for effective normalization of contact maps from CNA bias.

In conclusion, our results suggest that HiCNAttra provides a better solution for a) computing high-coverage RD signal and detecting large-scale and focal CNAs from Hi-C/3C-seq datasets, and b) for explicitly correcting the chromatin contact frequencies from systematic biases as well as data-associated bias that is introduced by CNA events in cell lines with abnormal karyotypes.

Chapter 6

Conclusion and Outlook

In this thesis, we aimed for comprehensive multi-level karyotyping of cancer cell lines beyond what has been reached by the state-of-the-art single-sample NGS-based computational tools. In particular, we were interested in the computational methods that utilize the RD signal for identifying the copy number from single-sample WGS, Hi-C, or 3C-seq data. In other words, the main question, we addressed, was how to maximally utilize the RD signal of single-sample datasets for copy number profiling of cancer cell lines.

In Chapter 1, we explained the biological features and inherent biases associated with RD signal of cancer cell lines that complicate their accurate modeling and genetic characterization, especially from low coverage NGS data. So, we developed computational solutions to address them in order to improve the accuracy of copy number profiling. Next, we surveyed the literature in Chapter 2 to point out the limitations of available computational solutions that mainly targeted the problem of copy number detection from individual WGS or Hi-C data, especially at low-coverage. Additionally, we reviewed the computational tools for correcting the contact frequencies from the Hi-C data.

In Chapter 3, we presented AStra framework that employs the RD signal for aneuploidy profiling which can be used effectively for cell line authentication. Briefly, AStra searches the RD frequency distribution to find the CN reference that best allocates genomic segments into integer CN states. Then, it computes the aneuploidy profile that includes the copy number of each genomic locus. Therefore, 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 of the aneuploidy profile between different cell lines/strains.

In Chapter 4, we presented CNATra approach that utilizes the RD signal from WGS data for sensitive and robust detection of both large-scale copy number variations and focal alterations from low-coverage complex cancer profiles. Briefly, we proposed a novel hierarchical approach to first identify the iso-copy number blocks (IBs), and subsequently discover the focal alterations inside them. CNATra also incorporates coverage-based thresholds that allow detecting significant focal alterations, inside these IBs, from the low-coverage data. Therefore, CNATra is a single-sample CNA detection tool that provides an analytical and visualization framework for DNA copy number profiling without relying on any reference control. It can detect chromosome-level segmental aneuploidies and high-confidence focal alterations, even from low-coverage data. We also demonstrated the feasibility of applying CNATra on cancer patient data with variable purity.

Also, in Chapter 4, we proposed a novel method for generating realistic simulated-data driven from real cancer data to maintain the inherent CNA features and complexities of the RD signal. To achieve that, we manipulate the original WGS reads of cancer genome to randomly introduce copy number events. This helps for accurate benchmarking of CNA detection tools in the absence of extensively validated datasets of both LCVs and FAs.

In Chapter 5, we presented HiCNATra tool for copy number profiling of cancer cell lines from Hi-C/3C-seq data and use this information to correct their chromatin contact matrices from systematic biases including bias introduced by CNA events. Briefly, HiCNATra utilizes the Hi-C/3C-seq reads to estimate the high-resolution RD signal, to identify both LCVs and FAs, and to correct the Hi-C/3C-seq contact frequencies from the CNA-associated bias as well as other systematic biases. We proposed the ‘entire fragment’ RD computing approach for Hi-C data that allows HiCNATra to reproduce the RD profile generated from WGS data of the same cell line. Then, we apply the CNA detection method of CNATra for detecting the LCVs and FAs from RD signal derived

from Hi-C data. Can you say one line for 3C-seq RD calculation. Finally, we utilize a GLM to correct the contact frequencies of Hi-C/3C-seq data from all biases together. This model showed excellent ability for eliminating the CNA-associated bias as well as other systematic biases. Therefore, HiCNAta is a chromatin contact map correction tool that provides an analytical and visualization framework for DNA copy number profiling from Hi-C/3C-seq data and chromatin contact map normalization.

For the remainder of this chapter, we present an overview of potential future research directions in this field as well as alternative research trends that are related to the line of research pursued in this thesis.

6.1 CN state-driven coverage-based thresholding

As demonstrated in Chapter 3 and Chapter 4, data coverage is an important key for accurate detection of focal copy number gains and losses. Generally, data sparseness is inversely proportional to the data coverage. This results in overdispersion of the RD signal from low-coverage data. To overcome this, in this thesis, we introduced the coverage-based thresholding that enables the detection of focal alterations from low-coverage data. We utilized an exponential regression models that learned the relationship between coverage and the amplification/deletion thresholds using subsampling of four high-coverage data. However, our heuristic approach of determining coverage-based thresholding parameters leaves space for future improvements by incorporating more high-coverage cancer datasets of different complexities.

For example, CNAta and HiCNAta use the same thresholds for identifying amplifications and deletions from all IBs regardless of the copy number states of these IBs. However, we observe that the dispersion levels vary between IBs belonging to different CN states (Figure 6.1). In this figure, each grey dot represents a bin from a chromosome. The horizontal black line represents the copy number of each IB. Any amplitude transition in this line indicates a new IB (3 IBs for Chr

17 of CHP-212 cell line and 7 IBs for Chr 3 of A427 cell line). The green vertical lines denote the boundaries of centromeres and telomeres. Within each IB, height of the red arrow represents the dispersion level of that IB. Notably, the height of red arrow in an IB is positively correlated with the CN of that IB (Figure 6.1).

Figure 6.1: Illustration of the dispersion level at IBs of different CN states.

Therefore, we recommend developing *CN state-driven coverage-based thresholds* for further enhancement of the sensitivity and precision of CNAttra and HiCNAttra tools. Amplification and deletion thresholds can be modeled better as a function of both data coverage and CN state of the

IB. For this study, first, more cancer datasets of complex copy number profiles, at high coverage, should be incorporated to capture the main features of cancer data. Secondly, due to the absence of complete repertoire of validated datasets that distinguish LCVs and FAs, our realistic-simulated data generator can be utilized for making different scenarios of complex CNA profiles as 'ground truth' from these high-coverage data. Finally, these realistic data are subsampled for accurate estimation of the relationship between our thresholding parameters, CN states of each IB, and the data coverage of the sample.

6.2 Directions for accurate TAD identification

Using the corrected Hi-C contact maps, different analysis methods [105, 185] are applied for interpreting the 3D genome structure and addressing various research questions such as the enhancer-driven gene activation process. Genome-wide Hi-C analysis revealed that the genome is divided into TADs that represent higher-order chromatin folding [101, 186]. TADs may represent a group of interacting enhancers and gene-promoters that controls the gene regulation in metazoans. These interactions often take place at sub-Mb distances. Therefore, TADs are believed to be the fundamental units of the genome organization.

Many computational tools have been developed to identify TADs from the Hi-C interaction matrices by utilizing various clustering and optimization techniques [101, 103, 108, 187-190]. For example, Dixon *et. al.* utilized HMM to infer TADs based on a directionality index [101]. As TADs create unexpectedly low numbers of contacts crossing the boundary regions, Dixon *et. al.* used this feature to infer a directional index for each restriction fragment, which is high if the fragment insulates its upstream regions from the downstream, and to identify the boundaries of TADs [101]. Rao *et. al.* used the arrowhead transformation for detecting TADs based on the observation that there is a very rapid decline in contact probability as moving away from the diagonal of a contact map [108]. Therefore, they detect the TADs by identifying the corners of the domain that have rapid decline in interaction frequencies. First, they apply the arrowhead

transformation to normalize the interaction frequency using the predicted interaction frequency between genomic loci. Then, they identify the domain corners that have higher normalized interaction frequency. Considering sub-TADs and hierarchical property of association domains, Filippova *et. al.* uses optimization functions to find optimal and near-optimal domains for a given resolution [190]. Then, they identify a consensus set of domains that are consistent across different resolutions. They use the consensus domains and the resolution-specific domains as TAD calls.

However, various methodologies result in different boundaries and sizes of TADs [191, 192]. Also, tuning parameters and data resolution (number of Hi-C short reads) have significant impacts on the TAD calling by different tools [193]. In other words, TADs are defined in different ways using various methodologies and parameters resulting in lack of consensus structural domains that are agreeable with each other. In the following sections, we propose two solutions for the identification of accurate and robust TADs.

6.2.1 Two steps approach for calling accurate TADs

Most TAD calling tools identify TADs by examining the changes in the interaction frequencies only between the adjacent genomic loci and does not consider any interaction between distal loci. In other words, they mainly search along the diagonal of the Hi-C interaction matrix to find significant declines in the interaction frequencies that define the boundaries of TADs. Therefore, they cannot detect the chromatin loops between distal loci that are physically connected (interacting). Moreover, TADs usually contain multiple hierarchies of overlapping block-like structures, called as sub-TADs, but it is difficult to distinguish between TADs and sub-TADs based only on a single variable (enhanced interaction frequencies between adjacent regions). As a result, TADs called by different tools are inconsistent [191-193].

Figure 6.2: Cartoon illustration of on-diagonal TADs and off-diagonal inter-TAD (TAD-TAD) interactions.

Here, we introduce a new definition of TAD as contiguous genomic loci that have higher self-interaction frequencies and their boundaries accurately define the inter-TADs (off-diagonal blocks that are formed by the interaction between two or more underlying TADs). Figure 6.2 provides a cartoon representation of the Hi-C heatmap that illustrates the difference between TADs and inter-TAD interactions. TADs are triangle-shaped blocks that are located on the diagonal of Hi-C heatmap, whereas inter-TAD interactions are shown as rectangle/square-shaped off-diagonal blocks on the Hi-C heatmap (Figure 6.2). Similar definition is introduced [194] for off-diagonal pairwise TAD-TAD interaction as TAD clique. Paulsen et al. showed that TAD cliques represent spatial assemblies of large heterochromatic domains [194].

By this definition, we have two variables (on-diagonal interaction frequency and off-diagonal inter-TAD interaction frequency) which can be optimized for identifying accurate TADs in different ways. For example, the interaction matrix can be visualized as 2D-image. Therefore, we suggest a two-step approach based on image-segmentation for calling both TADs and TAD-TAD interactions accurately. First, the image-segmentation techniques can be applied to find the primary small building units of TADs (on-diagonal blocks) and inter-TAD interactions (off-diagonal blocks) from the Hi-C interaction matrix. Then, these blocks are assembled to find the 'robust' TADs that satisfy these two features. First, genomic loci inside a TAD should have more interaction with each other than with loci from any other TAD. Second, any inter-TAD interaction should reflect a physical TAD-TAD interaction.

6.2.2 Epigenetics-driven robust TADs

Genome-wide epigenetic datasets are available from many large-scale studies such as the Encyclopedia of DNA Elements (ENCODE) [195] and Roadmap Epigenomics [196]. These studies suggest the influence of epigenetics such as histone post-translational modifications, and binding patterns of chromatin-binding proteins in chromatin function. For example, it is recognized that enhancers can be identified based on epigenetic markers by detecting higher ratio of H3K4me1 compared to H3K4me3 and CREB-binding proteins [197]. Also, H3K27ac is considered as an active enhancer mark [198]. Therefore, many supervised learning frameworks have been developed to predict the gene expression [199] and identify the regulatory elements [174, 200] using the epigenetic markers.

Recently, Hi-C study correlated CTCF binding at TAD boundaries [108]. Moreover, ChIA-PET analysis found that CTCF is involved in defining the interface between condensed and open compartments for structural regulations and the majority of chromatin interaction loops are associated with CTCF-binding sites [201]. Additionally, the CTCF and the cohesin protein complex, such as RAD21 subunit, have been frequently reported at TAD boundaries [191, 202].

This suggests that epigenetic information can be efficiently utilized for better detection of TADs. This is also supported by the fact that epigenetic data are available at higher resolution for many cell lines, compared to the resolution of Hi-C data.

Consequently, Huang et al. [203] have developed a computational model to predict the chromatin interaction hubs, the genomic loci with frequent chromatin interactions, and TAD boundaries from ChIP-seq data of histone modifications. Results showed that both features can be predicted accurately using ChIP-seq of CTCF and 9 histone post-translational modification marks for IMR90 and GM12878 which are immortalized cell lines. They also reported that H3K4me1 and H3K27me3 are the main histone marks for identifying the chromatin interaction hubs whereas CTCF and H3K4me1 were the informative marks for predicting TAD boundaries. In 2017, Pierro et al. [204] extended the idea of ChIP-seq for predicting 3D genome organization. They developed a computational tool for predicting the human genome architecture in terms of the Hi-C interaction using 106 epigenetic marks. They trained their model using the odd chromosomes of IMR90 normal cell line and tested it on the even chromosomes. Results showed they can successfully predict the contact frequencies. However, both methods require relatively large number of epigenetic markers that are not readily available for all cell lines. Additionally, they applied their techniques using only normal cell lines which are expected to have similar TAD structure since TADs are highly conserved among cell types, and even across different species [205]. For example, Rao et al. showed that 55-75% of the frequent chromatin loops, that link promoters and enhancers, found in any given cell type were also found in GM12878 [108]. Also, Dixon et al. reported that 53.8% of human TAD boundaries are boundaries in mouse and 75.9% of mouse TAD boundaries are boundaries in humans [101].

As an alternative solution, we propose to use epigenetic marks along with Hi-C data for accurate detection of TADs. First, the candidate genomic loci of TAD boundaries can be identified using the available epigenetic markers. For example, candidate genomic regions are expected to be

enriched with CTCF and RAD21 as well as other epigenetic markers (e.g. H3K4me1). Therefore, the genome can be first examined to find a set of candidate genomic loci of TAD boundaries that are associated with peaks of the epigenetic markers. Then, previously-introduced techniques such as directionality index [\[101\]](#) or arrowhead transformation [\[108\]](#) can be applied to these candidate loci only. This will result in diminishing of the genome-wide noise while comparing these genomic loci and detecting the robust TADs.

Bibliography

1. Vogelstein, B., et al., *Cancer genome landscapes*. science, 2013. **339**(6127): p. 1546-1558.
2. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA: a cancer journal for clinicians, 2018. **68**(6): p. 394-424.
3. Ferlay, J., et al., *Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods*. International journal of cancer, 2019. **144**(8): p. 1941-1953.
4. Ben-David, U., et al., *Genetic and transcriptional evolution alters cancer cell line drug response*. Nature, 2018. **560**(7718): p. 325-330.
5. van Staveren, W.C., et al., *Human cancer cell lines: Experimental models for cancer cells in situ? For cancer stem cells?* Biochim Biophys Acta, 2009. **1795**(2): p. 92-103.
6. Louzada, S., F. Adegá, and R. Chaves, *Defining the sister rat mammary tumor cell lines HH-16 cl.2/1 and HH-16.cl.4 as an in vitro cell model for Erbb2*. PLoS One, 2012. **7**(1): p. e29923.
7. Vargo-Gogola, T. and J.M. Rosen, *Modelling breast cancer: one size does not fit all*. Nat Rev Cancer, 2007. **7**(9): p. 659-72.
8. Ferreira, D., F. Adegá, and R. Chaves, *The importance of cancer cell lines as in vitro models in cancer methylome analysis and anticancer drugs testing*, in *Oncogenomics and cancer proteomics-novel approaches in biomarkers discovery and therapeutic targets in cancer*. 2013, IntechOpen.
9. Gillet, J.-P., S. Varma, and M.M. Gottesman, *The clinical relevance of cancer cell lines*. Journal of the National Cancer Institute, 2013. **105**(7): p. 452-458.
10. Nakatsu, N., et al., *Chemosensitivity profile of cancer cell lines and identification of genes determining chemosensitivity by an integrated bioinformatical approach using cDNA arrays*. Mol Cancer Ther, 2005. **4**(3): p. 399-412.

11. Gazdar, A.F., et al., *Lung cancer cell lines as tools for biomedical discovery and research.* J Natl Cancer Inst, 2010. **102**(17): p. 1310-21.
12. Dayton, J.B. and S.R. Piccolo, *Classifying cancer genome aberrations by their mutually exclusive effects on transcription.* BMC Med Genomics, 2017. **10**(Suppl 4): p. 66.
13. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Res, 2014. **42**(Database issue): p. D986-92.
14. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.
15. Mishra, S. and J.R. Whetstine, *Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive.* Mol Cell Biol, 2016. **36**(7): p. 1050-63.
16. Tang, Y.C. and A. Amon, *Gene copy-number alterations: a cost-benefit analysis.* Cell, 2013. **152**(3): p. 394-405.
17. Domcke, S., et al., *Evaluating cell lines as tumour models by comparison of genomic profiles.* Nat Commun, 2013. **4**: p. 2126.
18. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.
19. Gordon, D.J., B. Resio, and D. Pellman, *Causes and consequences of aneuploidy in cancer.* Nat Rev Genet, 2012. **13**(3): p. 189-203.
20. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.
21. Sansregret, L. and C. Swanton, *The Role of Aneuploidy in Cancer Evolution.* Cold Spring Harb Perspect Med, 2017. **7**(1).
22. Bakhoum, S.F. and C. Swanton, *Chromosomal instability, aneuploidy, and cancer.* Front Oncol, 2014. **4**: p. 161.
23. Giam, M. and G. Rancati, *Aneuploidy and chromosomal instability in cancer: a jackpot to chaos.* Cell Div, 2015. **10**: p. 3.

24. Speicher, M.R., et al., *Analysis of chromosomal alterations in non-small cell lung cancer by multiplex-FISH, comparative genomic hybridization, and multicolor bar coding*. Lab Invest, 2000. **80**(7): p. 1031-41.
25. Schulze, S. and I. Petersen, *Gender and ploidy in cancer survival*. Cell Oncol (Dordr), 2011. **34**(3): p. 199-208.
26. Lazarevic, V., et al., *Prognostic significance of high hyperdiploid and triploid/tetraploid adult acute myeloid leukemia*. Am J Hematol, 2015. **90**(9): p. 800-5.
27. Zasadil, L.M., E.M. Britigan, and B.A. Weaver, *2n or not 2n: Aneuploidy, polyploidy and chromosomal instability in primary and tumor cells*. Semin Cell Dev Biol, 2013. **24**(4): p. 370-9.
28. Brosens, R.P., et al., *Candidate driver genes in focal chromosomal aberrations of stage II colon cancer*. J Pathol, 2010. **221**(4): p. 411-24.
29. Leary, R.J., et al., *Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers*. Proc Natl Acad Sci U S A, 2008. **105**(42): p. 16224-9.
30. Bignell, G.R., et al., *Signatures of mutation and selection in the cancer genome*. Nature, 2010. **463**(7283): p. 893-8.
31. Prasetyanti, P.R. and J.P. Medema, *Intra-tumor heterogeneity from a cancer stem cell perspective*. Mol Cancer, 2017. **16**(1): p. 41.
32. Ellsworth, R.E., et al., *Molecular heterogeneity in breast cancer: State of the science and implications for patient care*. Semin Cell Dev Biol, 2017. **64**: p. 65-72.
33. Fisher, R., L. Pusztai, and C. Swanton, *Cancer heterogeneity: implications for targeted therapeutics*. Br J Cancer, 2013. **108**(3): p. 479-85.
34. Thompson, S.L. and D.A. Compton, *Chromosomes and cancer cells*. Chromosome Res, 2011. **19**(3): p. 433-44.

35. Turashvili, G. and E. Brogi, *Tumor Heterogeneity in Breast Cancer*. Front Med (Lausanne), 2017. **4**: p. 227.
36. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing*. Genome Res, 2012. **22**(3): p. 568-76.
37. Dawson, M.A. and T. Kouzarides, *Cancer epigenetics: from mechanism to therapy*. Cell, 2012. **150**(1): p. 12-27.
38. Altschuler, S.J. and L.F. Wu, *Cellular heterogeneity: do differences make a difference?* Cell, 2010. **141**(4): p. 559-63.
39. Metzker, M.L., *Sequencing technologies - the next generation*. Nat Rev Genet, 2010. **11**(1): p. 31-46.
40. Behjati, S. and P.S. Tarpey, *What is next generation sequencing?* Arch Dis Child Educ Pract Ed, 2013. **98**(6): p. 236-8.
41. Alkods, 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.
42. Duan, J., et al., *Comparative studies of copy number variation detection methods for next-generation sequencing technologies*. PLoS One, 2013. **8**(3): p. e59128.
43. 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.
44. 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.
45. 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.

46. Eisefeldt, J., et al., *AMYCNE: Confident copy number assessment using whole genome sequencing data*. PLoS One, 2018. **13**(3): p. e0189710.
47. Eisefeldt, J., et al., *TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data*. F1000Res, 2017. **6**: p. 664.
48. Magi, A., T. Pippucci, and C. Sidore, *XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments*. BMC Genomics, 2017. **18**(1): p. 747.
49. 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.
50. Wang, X., H. Chen, and N.R. Zhang, *DNA copy number profiling using single-cell sequencing*. Brief Bioinform, 2018. **19**(5): p. 731-736.
51. Yu, Z., et al., *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**(18): p. 2576-83.
52. Augusto Correa Dos Santos, R., G.H. Goldman, and D.M. Riano-Pachon, *ploidyNGS: visually exploring ploidy with Next Generation Sequencing data*. Bioinformatics, 2017. **33**(16): p. 2575-2576.
53. Margarido, G.R. and D. Heckerman, *ConPADE: genome assembly ploidy estimation from next-generation sequencing data*. PLoS Comput Biol, 2015. **11**(4): p. e1004229.
54. Bick, D. and D. Dimmock, *Whole exome and whole genome sequencing*. Curr Opin Pediatr, 2011. **23**(6): p. 594-600.
55. Dekker, J., et al., *Capturing chromosome conformation*. Science, 2002. **295**(5558): p. 1306-11.
56. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome*. Science, 2009. **326**(5950): p. 289-93.

57. Sexton, T., et al., *Three-dimensional folding and functional organization principles of the Drosophila genome*. Cell, 2012. **148**(3): p. 458-72.
58. Naumova, N., et al., *Analysis of long-range chromatin interactions using Chromosome Conformation Capture*. Methods, 2012. **58**(3): p. 192-203.
59. van de Wiel, M.A., et al., *Smoothing waves in array CGH tumor profiles*. Bioinformatics, 2009. **25**(9): p. 1099-104.
60. van Heesch, S., et al., *Systematic biases in DNA copy number originate from isolation procedures*. Genome Biol, 2013. **14**(4): p. R33.
61. Payne, E.H., et al., *An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data*. Commun Stat Simul Comput, 2018. **47**(6): p. 1722-1738.
62. Karimzadeh, M., et al., *Umap and Bimap: quantifying genome and methylome mappability*. Nucleic Acids Res, 2018.
63. Ai, N., et al., *CNARA: reliability assessment for genomic copy number profiles*. BMC Genomics, 2016. **17**(1): p. 799.
64. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping*. Nat Rev Genet, 2011. **12**(5): p. 363-76.
65. Swaminathan, S., et al., *Analysis of copy number variation in Alzheimer's disease: the NIALOAD/ NCRAD Family Study*. Curr Alzheimer Res, 2012. **9**(7): p. 801-14.
66. Shlien, A. and D. Malkin, *Copy number variations and cancer*. Genome Med, 2009. **1**(6): p. 62.
67. Marshall, C.R. and S.W. Scherer, *Detection and characterization of copy number variation in autism spectrum disorder*. Methods Mol Biol, 2012. **838**: p. 115-35.
68. Cuccaro, D., et al., *Copy Number Variants in Alzheimer's Disease*. J Alzheimers Dis, 2017. **55**(1): p. 37-52.

69. Bassett, A.S., S.W. Scherer, and L.M. Brzustowicz, *Copy number variations in schizophrenia: critical review and new perspectives on concepts of genetics and disease*. Am J Psychiatry, 2010. **167**(8): p. 899-914.
70. Genomes Project, C., et al., *A global reference for human genetic variation*. Nature, 2015. **526**(7571): p. 68-74.
71. Pierre-Jean, M., G. Rigai, and P. Neuvial, *Performance evaluation of DNA copy number segmentation methods*. Brief Bioinform, 2015. **16**(4): p. 600-15.
72. Karimpour-Fard, A., et al., *A survey of analysis software for array-comparative genomic hybridisation studies to detect copy number variation*. Hum Genomics, 2010. **4**(6): p. 421-7.
73. Chari, R., W.W. Lockwood, and W.L. Lam, *Computational methods for the analysis of array comparative genomic hybridization*. Cancer Inform, 2007. **2**: p. 48-58.
74. Zhang, J., et al., *The impact of next-generation sequencing on genomics*. J Genet Genomics, 2011. **38**(3): p. 95-109.
75. Koboldt, D.C., et al., *Massively parallel sequencing approaches for characterization of structural variation*. Methods Mol Biol, 2012. **838**: p. 369-84.
76. Russo, C.D., et al., *Comparative study of aCGH and Next Generation Sequencing (NGS) for chromosomal microdeletion and microduplication screening*. J Prenat Med, 2014. **8**(3-4): p. 57-69.
77. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
78. Meienberg, J., et al., *Clinical sequencing: is WGS the better WES?* Hum Genet, 2016. **135**(3): p. 359-62.
79. Korbel, J.O., et al., *Paired-end mapping reveals extensive structural variation in the human genome*. Science, 2007. **318**(5849): p. 420-6.

80. Xie, C. and M.T. Tammi, *CNV-seq, a new method to detect copy number variation using high-throughput sequencing*. BMC Bioinformatics, 2009. **10**: p. 80.
81. Gusnanto, A., et al., *Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data*. Bioinformatics, 2012. **28**(1): p. 40-7.
82. Chiang, D.Y., et al., *High-resolution mapping of copy-number alterations with massively parallel sequencing*. Nat Methods, 2009. **6**(1): p. 99-103.
83. Xi, R., et al., *Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion*. Proc Natl Acad Sci U S A, 2011. **108**(46): p. E1128-36.
84. Ivakhno, S., et al., *CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data*. Bioinformatics, 2010. **26**(24): p. 3051-8.
85. Zeng, W.Z.D., et al., *Selecting precise reference normal tissue samples for cancer research using a deep learning approach*. BMC Med Genomics, 2019. **12**(Suppl 1): p. 21.
86. Trost, B., et al., *Impact of DNA source on genetic variant detection from human whole-genome sequencing data*. J Med Genet, 2019. **56**(12): p. 809-817.
87. Mao, R., et al., *Whole genome sequencing of matched tumor, adjacent non-tumor tissues and corresponding normal blood samples of hepatocellular carcinoma patients revealed dynamic changes of the mutations profiles during hepatocarcinogenesis*. Oncotarget, 2017. **8**(16): p. 26185-26199.
88. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data*. Biostatistics, 2004. **5**(4): p. 557-72.
89. Venkatraman, E.S. and A.B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array CGH data*. Bioinformatics, 2007. **23**(6): p. 657-63.
90. Levy-leduc, C. and Z. Harchaoui. *Catching change-points with lasso*. in *Advances in neural information processing systems*. 2008.

91. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Res, 2001. **29**(1): p. 308-11.
92. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing*. Nature, 2010. **467**(7319): p. 1061-73.
93. Magi, A., et al., *A shifting level model algorithm that identifies aberrations in array-CGH data*. Biostatistics, 2010. **11**(2): p. 265-80.
94. Shen, J.J. and N.R. Zhang, *Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing*. The Annals of Applied Statistics, 2012. **6**(2): p. 476-496.
95. Bates, S.E., *Classical cytogenetics: karyotyping techniques*. Methods Mol Biol, 2011. **767**: p. 177-90.
96. Rondon-Lagos, M., et al., *Effect of low doses of estradiol and tamoxifen on breast cancer cell karyotypes*. Endocr Relat Cancer, 2016. **23**(8): p. 635-50.
97. Ohnuki, Y., et al., *Chromosomal analysis of human prostatic adenocarcinoma cell lines*. Cancer Res, 1980. **40**(3): p. 524-34.
98. Gorringe, K.L., et al., *Evidence that both genetic instability and selection contribute to the accumulation of chromosome alterations in cancer*. Carcinogenesis, 2005. **26**(5): p. 923-30.
99. Weiss, C.L., et al., *nQuire: a statistical framework for ploidy estimation using next generation sequencing*. BMC Bioinformatics, 2018. **19**(1): p. 122.
100. de Wit, E. and W. de Laat, *A decade of 3C technologies: insights into nuclear organization*. Genes Dev, 2012. **26**(1): p. 11-24.
101. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions*. Nature, 2012. **485**(7398): p. 376-80.
102. Sanyal, A., et al., *The long-range interaction landscape of gene promoters*. Nature, 2012. **489**(7414): p. 109-13.

103. Crane, E., et al., *Condensin-driven remodelling of X chromosome topology during dosage compensation*. Nature, 2015. **523**(7559): p. 240-4.
104. Krijger, P.H., et al., *Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming*. Cell Stem Cell, 2016. **18**(5): p. 597-610.
105. Lajoie, B.R., J. Dekker, and N. Kaplan, *The Hitchhiker's guide to Hi-C analysis: practical guidelines*. Methods, 2015. **72**: p. 65-75.
106. Imakaev, M., et al., *Iterative correction of Hi-C data reveals hallmarks of chromosome organization*. Nat Methods, 2012. **9**(10): p. 999-1003.
107. Knight, P.A. and D. Ruiz, *A fast algorithm for matrix balancing*. IMA Journal of Numerical Analysis, 2013. **33**(3): p. 1029-1047.
108. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping*. Cell, 2014. **159**(7): p. 1665-80.
109. Li, W., et al., *Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data*. Bioinformatics, 2015. **31**(6): p. 960-2.
110. Yaffe, E. and A. Tanay, *Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture*. Nat Genet, 2011. **43**(11): p. 1059-65.
111. Hu, M., et al., *HiCNorm: removing biases in Hi-C data via Poisson regression*. Bioinformatics, 2012. **28**(23): p. 3131-3.
112. Wu, H.J. and F. Michor, *A computational strategy to adjust for copy number in tumor Hi-C data*. Bioinformatics, 2016. **32**(24): p. 3695-3701.
113. Vidal, E., et al., *OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes*. Nucleic Acids Res, 2018. **46**(8): p. e49.
114. Servant, N., et al., *Effective normalization for copy number variation in Hi-C data*. BMC Bioinformatics, 2018. **19**(1): p. 313.

115. Lun, A.T. and G.K. Smyth, *diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data*. BMC Bioinformatics, 2015. **16**: p. 258.
116. Picard, F., et al., *Joint segmentation, calling, and normalization of multiple CGH profiles*. Biostatistics, 2011. **12**(3): p. 413-28.
117. Chakraborty, A. and F. Ay, *Identification of copy number variations and translocations in cancer cells from Hi-C data*. Bioinformatics, 2017.
118. Spans, L., et al., *Variations in the exome of the LNCaP prostate cancer cell line*. Prostate, 2012. **72**(12): p. 1317-27.
119. Kleivi, K., et al., *Genome signatures of colon carcinoma cell lines*. Cancer Genet Cytogenet, 2004. **155**(2): p. 119-31.
120. Almeida, J.L., K.D. Cole, and A.L. Plant, *Standards for Cell Line Authentication and Beyond*. PLoS Biol, 2016. **14**(6): p. e1002476.
121. 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.
122. 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.
123. 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.
124. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. Nat Methods, 2012. **9**(4): p. 357-9.
125. Goodison, S., et al., *The interrelationship between DRIM gene expression and cytogenetic and phenotypic characteristics in human breast tumor cell lines*. BMC Genomics, 2003. **4**(1): p. 39.

126. Naumann, S., et al., *Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization*. Leuk Res, 2001. **25**(4): p. 313-22.
127. van Bokhoven, A., et al., *Spectral karyotype (SKY) analysis of human prostate carcinoma cell lines*. Prostate, 2003. **57**(3): p. 226-44.
128. Findley, H.W., Jr., et al., *Two new acute lymphoblastic leukemia cell lines with early B-cell phenotypes*. Blood, 1982. **60**(6): p. 1305-9.
129. Kim, G.J., et al., *Chromosomal aberrations in neuroblastoma cell lines identified by cross species color banding and chromosome painting*. Cancer Genet Cytogenet, 2001. **129**(1): p. 10-6.
130. Greenberg, J.M., et al., *Immunophenotypic and cytogenetic analysis of Molt-3 and Molt-4: human T-lymphoid cell lines with rearrangement of chromosome 7*. Blood, 1988. **72**(5): p. 1755-60.
131. Pan, Y., et al., *Characterization of chromosomal abnormalities in prostate cancer cell lines by spectral karyotyping*. Cytogenet Cell Genet, 1999. **87**(3-4): p. 225-32.
132. Grigorova, M., et al., *Chromosome abnormalities in 10 lung cancer cell lines of the NCI-H series analyzed with spectral karyotyping*. Cancer Genet Cytogenet, 2005. **162**(1): p. 1-9.
133. Hettmer, S., S. Ladisch, and K. Kaucic, *Low complex ganglioside expression characterizes human neuroblastoma cell lines*. Cancer Lett, 2005. **225**(1): p. 141-9.
134. Kiessling, M.K., et al., *Identification of oncogenic driver mutations by genome-wide CRISPR-Cas9 dropout screening*. BMC Genomics, 2016. **17**(1): p. 723.
135. Lopez-Terrada, D., et al., *Hep G2 is a hepatoblastoma-derived cell line*. Hum Pathol, 2009. **40**(10): p. 1512-5.
136. Weaver, B.A. and D.W. Cleveland, *The aneuploidy paradox in cell growth and tumorigenesis*. Cancer Cell, 2008. **14**(6): p. 431-3.

137. Weaver, B.A., et al., *Aneuploidy acts both oncogenically and as a tumor suppressor*. Cancer Cell, 2007. **11**(1): p. 25-36.
138. Sotillo, R., et al., *Mad2 overexpression promotes aneuploidy and tumorigenesis in mice*. Cancer Cell, 2007. **11**(1): p. 9-23.
139. Williams, B.R., et al., *Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells*. Science, 2008. **322**(5902): p. 703-9.
140. Weaver, B.A. and D.W. Cleveland, *Does aneuploidy cause cancer?* Curr Opin Cell Biol, 2006. **18**(6): p. 658-67.
141. Ganem, N.J., Z. Storchova, and D. Pellman, *Tetraploidy, aneuploidy and cancer*. Curr Opin Genet Dev, 2007. **17**(2): p. 157-62.
142. Albertson, D.G., et al., *Chromosome aberrations in solid tumors*. Nat Genet, 2003. **34**(4): p. 369-76.
143. 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).
144. 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.
145. 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.
146. 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.
147. Carter, S.L., et al., *Absolute quantification of somatic DNA alterations in human cancer*. Nat Biotechnol, 2012. **30**(5): p. 413-21.
148. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity*. Nature, 2012. **483**(7391): p. 603-7.
149. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution*. Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.

150. Wilding, J.L. and W.F. Bodmer, *Cancer cell lines for drug discovery and development*. Cancer Res, 2014. **74**(9): p. 2377-84.
151. Zhou, B., et al., *Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis*. J Med Genet, 2018. **55**(11): p. 735-743.
152. Benjamini, Y. and T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing*. Nucleic Acids Res, 2012. **40**(10): p. e72.
153. Savitzky, A. and M.J. Golay, *Smoothing and differentiation of data by simplified least squares procedures*. Analytical chemistry, 1964. **36**(8): p. 1627-1639.
154. Krajca, V., et al., *Automatic identification of significant graphoelements in multichannel EEG recordings by adaptive segmentation and fuzzy clustering*. Int J Biomed Comput, 1991. **28**(1-2): p. 71-89.
155. Fadili, M.J. and E.T. Bullmore, *A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps*. Neuroimage, 2004. **23**(3): p. 1112-28.
156. Azami, H., K. Mohammadi, and B. Bozorgtabar, *An improved signal segmentation using moving average and Savitzky-Golay filter*. Journal of Signal and Information Processing, 2012. **3**(1).
157. Steinier, J., Y. Termonia, and J. Deltour, *Smoothing and differentiation of data by simplified least square procedure*. Anal Chem, 1972. **44**(11): p. 1906-9.
158. Kundaje, A., *A comprehensive collection of signal artifact blacklist regions in the human genome*. ENCODE.[hg19-blacklist-README. doc-EBI]. Available online at, 2013.
159. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
160. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer*. Nucleic Acids Res, 2019. **47**(D1): p. D941-D947.

161. Kaghad, M., et al., *Monoallelically expressed gene related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers*. Cell, 1997. **90**(4): p. 809-19.
162. Okazaki, T., et al., *Detection of amplified genomic sequences in human small-cell lung carcinoma cells by arbitrarily primed-PCR genomic fingerprinting*. Hum Genet, 1996. **98**(3): p. 253-8.
163. Dassi, E., et al., *Multi-omic profiling of MYCN-amplified neuroblastoma cell-lines*. Genom Data, 2015. **6**: p. 285-7.
164. Nagayama, K., et al., *Homozygous deletion scanning of the lung cancer genome at a 100-kb resolution*. Genes Chromosomes Cancer, 2007. **46**(11): p. 1000-10.
165. Matsumoto, S., et al., *Prevalence and specificity of LKB1 genetic alterations in lung cancers*. Oncogene, 2007. **26**(40): p. 5911-8.
166. Shapiro, G.I., et al., *Multiple mechanisms of p16INK4A inactivation in non-small cell lung cancer cell lines*. Cancer Res, 1995. **55**(24): p. 6200-9.
167. Greenman, C.D., et al., *PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data*. Biostatistics, 2010. **11**(1): p. 164-75.
168. Luo, Z., et al., *Accurity: accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants*. Bioinformatics, 2018. **34**(12): p. 2004-2011.
169. Aran, D., M. Sirota, and A.J. Butte, *Systematic pan-cancer analysis of tumour purity*. Nat Commun, 2015. **6**: p. 8971.
170. Nagano, Y., et al., *Allelic alterations in pancreatic endocrine tumors identified by genome-wide single nucleotide polymorphism analysis*. Endocr Relat Cancer, 2007. **14**(2): p. 483-92.
171. Guo, S.S., et al., *Frequent deletion of chromosome 3 in malignant sporadic pancreatic endocrine tumors*. Mol Cell Endocrinol, 2002. **190**(1-2): p. 109-14.

-
172. Gebauer, N., et al., *Genomic landscape of pancreatic neuroendocrine tumors*. World J Gastroenterol, 2014. **20**(46): p. 17498-506.
 173. Ernst, J. and M. Kellis, *Chromatin-state discovery and genome annotation with ChromHMM*. Nat Protoc, 2017. **12**(12): p. 2478-2492.
 174. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome*. Nat Genet, 2007. **39**(3): p. 311-8.
 175. Krijgsman, O., et al., *Focal chromosomal copy number aberrations in cancer-Needles in a genome haystack*. Biochim Biophys Acta, 2014. **1843**(11): p. 2698-2704.
 176. Feuk, L., et al., *Structural variants: changing the landscape of chromosomes and design of disease studies*. Hum Mol Genet, 2006. **15 Spec No 1**: p. R57-66.
 177. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions*. Cell, 2015. **161**(5): p. 1012-1025.
 178. Beroukhi, R., X. Zhang, and M. Meyerson, *Copy number alterations unmasked as enhancer hijackers*. Nat Genet, 2016. **49**(1): p. 5-6.
 179. Taberlay, P.C., et al., *Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations*. Genome Res, 2016. **26**(6): p. 719-31.
 180. Wu, P., et al., *3D genome of multiple myeloma reveals spatial genome disorganization associated with copy number variations*. Nature communications, 2017. **8**(1): p. 1937.
 181. Franke, M., et al., *Formation of new chromatin domains determines pathogenicity of genomic duplications*. Nature, 2016. **538**(7624): p. 265-269.
 182. Gollosi, R., J.T. Sanders, and R.P. McCord, *Iteratively improving Hi-C experiments one step at a time*. Methods, 2018. **142**: p. 47-58.
 183. Khalil, A., et al., *Hierarchical Discovery of Large-scale and Focal Copy Number Alterations in Low-coverage Cancer Genomes*. bioRxiv, 2019: p. 639294.

184. Karimzadeh, M., Ernst, C., Kundaje, A., & Hoffman, M. M., *Umap and Bimap: quantifying genome and methylome mappability*. bioRxiv, 2016: p. 095463.
185. Ay, F. and W.S. Noble, *Analysis methods for studying the 3D architecture of the genome*. Genome Biol, 2015. **16**: p. 183.
186. Nora, E.P., et al., *Spatial partitioning of the regulatory landscape of the X-inactivation centre*. Nature, 2012. **485**(7398): p. 381-5.
187. Weinreb, C. and B.J. Raphael, *Identification of hierarchical chromatin domains*. Bioinformatics, 2016. **32**(11): p. 1601-9.
188. Levy-Leduc, C., et al., *Two-dimensional segmentation for analyzing Hi-C data*. Bioinformatics, 2014. **30**(17): p. i386-92.
189. Serra, F., et al., *Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors*. PLoS Comput Biol, 2017. **13**(7): p. e1005665.
190. Filippova, D., et al., *Identification of alternative topological domains in chromatin*. Algorithms Mol Biol, 2014. **9**: p. 14.
191. Zufferey, M., et al., *Comparison of computational methods for the identification of topologically associating domains*. Genome Biol, 2018. **19**(1): p. 217.
192. Forcato, M., et al., *Comparison of computational methods for Hi-C data analysis*. Nat Methods, 2017. **14**(7): p. 679-685.
193. Dali, R. and M. Blanchette, *A critical assessment of topologically associating domain prediction tools*. Nucleic Acids Res, 2017. **45**(6): p. 2994-3005.
194. Paulsen, J., et al., *Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation*. Nat Genet, 2019. **51**(5): p. 835-843.
195. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57-74.

196. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium*. Nat Biotechnol, 2010. **28**(10): p. 1045-8.
197. Li, W., D. Notani, and M.G. Rosenfeld, *Enhancers as non-coding RNA transcription units: recent insights and future perspectives*. Nat Rev Genet, 2016. **17**(4): p. 207-23.
198. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state*. Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.
199. Karlic, R., et al., *Histone modification levels are predictive for gene expression*. Proc Natl Acad Sci U S A, 2010. **107**(7): p. 2926-31.
200. Kundaje, A., et al., *Integrative analysis of 111 reference human epigenomes*. Nature, 2015. **518**(7539): p. 317-330.
201. Tang, Z., et al., *CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription*. Cell, 2015. **163**(7): p. 1611-27.
202. Hansen, A.S., et al., *CTCF and cohesin regulate chromatin loop stability with distinct dynamics*. Elife, 2017. **6**.
203. Huang, J., et al., *Predicting chromatin organization using histone marks*. Genome Biol, 2015. **16**: p. 162.
204. Di Pierro, M., et al., *De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture*. Proc Natl Acad Sci U S A, 2017. **114**(46): p. 12126-12131.
205. Krefting, J., M.A. Andrade-Navarro, and J. Ibn-Salem, *Evolutionary stability of topologically associating domains is associated with conserved gene regulation*. BMC Biol, 2018. **16**(1): p. 87.

Appendix A: CNatra Pseudocodes

Following is the Pseudocodes of L1 segmentation and IB merging algorithms that we developed in the CNatra framework.

A.1 Pseudocode for L1 segmentation procedure

L1 Segmentation Procedure

Input: **F**, RD signal of a chromosome

Output: **E**, change points (starting and ending points of segments).

1. Smooth the RD signal using Savitzky-Golay Filter with window length (**w**):

$$F = SG_Filter(F, k=7, w)$$

$$F = SG_Filter(F, k=1, w)$$

2. Find the location of the peaks (**X**) and bottoms (**Y**) of the signal **F**.

3. Find the corresponding RD values of **X** and **Y**:

$$F_x = F(X)$$

$$F_y = F(Y)$$

4. Merge and sort the peaks and bottoms:

$$Z = [X, Y]$$

$$F_z = [F_x, F_y]$$

$$[Z, I] = sort(Z)$$

$$F_z = F_z(I)$$

5. Compute the location of edges (**E**) and their weights (**E_w**):

$$E(i) = \frac{Z(i) + Z(i+1)}{2}$$

$$E_w(i) = abs(F_z(i) - F_z(i + 1)), \text{ where } i \text{ is the edge number}$$

6. Filter edges to define segments boundaries based on their weights

$$E = E(top(E_w))$$

A.2 Pseudocode for IB assembly algorithm

IB Assembly Algorithm

Input:

F: RD signal of a chromosome

[S_s, S_e]: starting and ending bins of each segment of a segment set (S)**Parameters:**d_{min}: user-defined minimum IB width (default = 1Mb)**Output:**[FS_s, FS_e]: starting and ending bins of each long segment

1. Compute the attributes of each segment such as the CN and the state/interval (I) of each segment (j):

$$RD(j) = \text{median}(F[S_s(j), S_e(j)])$$

$$W(j) = S_e(j) - S_s(j) + 1$$

$$CN(j) = RD(j) \times \frac{2}{CNR}$$

$$I(j) = \text{RoundNearestInteger}(CN(j))$$

2. Filter out homogenous deletion ($I = 0$).

3. Merge the adjacent segments of same copy number state to build the initial segments **S1**.

4. Merge the initial segments (**S1**) iteratively to maximize the intra-segment similarity and build the segments **FS**:

(i) Merging the small segments to one of its long surroundings:

- a) Divide segments into two groups: long segments (**LS**) with ($W \geq d_{min}$) and short segments (**SS**) with ($W < d_{min}$).

- b) Assign a merging label (**L**), starting from 1, to long segments: consecutive long segments with the same CN interval number (**I**) have the same label.

- For j in LS

If ($I(j) == I(j+1)$)

$L(j+1) = L(j)$

Else

$L(j+1) = L(j) + 1$

End

End

c) Short segments between consecutive long segments are merged into one of their surrounding segments or grouped to make a new segment based on their copy numbers.

- **While (#unlabeled SS $\neq 0$)**
 - **Find a group of unlabeled segments (U) between two labeled LS (x, y).**
 - **Assign the group to a label of their neighbor long segment L(x) or L(y) or to a new label L_n .**

- **CN(u) = majority (CN(i)) where $i \in U$**

- **If(CN(x) == CN(y))**

- If(CN(x) == CN(u))**

- L(u) = L(y) = L(x)**

- Else**

- L(u) = L_n**

- End**

- Else**

- If(CN(u) == CN(x))**

- L(u) = L(x)**

- Else if (CN(u) == CN(y))**

- L(u) = L(y)**

- Else**

- L(u) = L_n**

- End**

- End**

d) Merge segments having the same label to form the new segments (S2).

(ii) Merge relatively short segment if it falls between two long segments with similar CN:

- **For j in S2**

- Find previous segment (x) and next segment (y)**

- If (I(x) == I(y)) & ($\frac{W(j)}{W(x)+W(y)} < 0.05$)**

- L(j) = L(y) = L(x)**

- End**

- **Merge segments with the same label to construct the final segments**
-

 $[FS_s, FS_e].$

Appendix B: CNVnator and FREEC

Calibration

Following is the illustration examples of the effects of the bin size on CNVnator results and the effects of the whole-genome ploidy level on FREEC results.

B.1 CNVnator results vary significantly with different bin size options

CNVnator results of Chr 17 at different bin sizes in IMR-32 are plotted. Red, black and grey dots are bins belonging to focally amplified, deleted and CN-neutral regions. The blue line represents the copy number of each CNA where any amplitude transition indicates a new CNA region. At 1-kb binning, CNVnator can detect FA1, but it also results in oversegmentation of LCV into many CNAs with similar copy numbers due to short-term variations. At 10-kb binning, the effects of wave artifacts have lessened resulting in the detection of larger CNAs. However, this results in missing some focal amplifications such as FA1 which has been detected at 1-kb binning.

B.2 Effect of different ploidy assumptions on the CNA calling by FREEC illustrated using Chr 12 of NCI-H82

FREEC output is dependent on user-defined ploidy level of the cell line. Different ploidy assumptions result in different segments of the RD signal. For example, some large segments at 2N ploidy assumption are divided into smaller segments of similar copy number under higher ploidy assumptions. This results in a false assessment of these segments. Red, black and grey dots are bins belonging to focally amplified, deleted and CN-neutral regions. The blue line represents the copy number of each CNA where any amplitude transition indicates a new CNA region.