Tumor subclone structure reconstruction with genomic variation data

Author: Yi Qiao

Boston College

The Graduate School of Arts and Sciences

Department of Biology

TUMOR SUBCLONE STRUCTURE RECONSTRUCTION WITH GENOMIC VARIATION DATA

a dissertation

by

YI QIAO

submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

August 2014

Tumor subclone structure reconstruction with genomic variation data

ABSTRACT

YI QIAO

DISSERTATION ADVISOR: GABOR T. MARTH

Unlike normal tissue cells, which contain identical copies of the same genome, tumors are composed of genetically divergent cell subpopulations, or subclones. The abilities to identify the number of subclones, their frequencies within the tumor mass, and the evolutionary relationships among them are crucial in understanding the basis of tumorigenesis, drug response, relapse, and metastasis. It is also essential information for informed, personalized therapeutic decisions. Studies have attempted to reconstruct subclone structure by identifying distinct allele frequency distribution modes at a handful of somatic single nucleotide variant loci, but this question was not adequately addressed with computational means at the start of this dissertation work, and recent efforts either enforce certain assumptions or resort to statistical procedure which cannot guarantee the complete landscape of solution space.

This dissertation present a computational framework that examines somatic variation events, such as copy number changes, loss of heterozygosity, or point mutations, in order to identify the underlying subclone structure. Chapter 2 discuss the presence of intra-tumoral heterogeneity, and for historical interest, a method to reconstruct the parsimonious solution based on simplifying assumptions in tumor micro-evolution process. Analysis results on clinical datasets concerning Ovarian Serious Carcinoma and Intracranal Germ Cell Tumor based on this method, which confirmed the genomic complexity, are also presented.

Due to the reason that the linkage information i.e. whether two mutations are co-

localizing in the same cancer cell is lost during tissue homogenization and DNA fragmentation, common sample preparation steps used in whole genome profiling techniques, often there are more than one subclone model capable of explaining the observation. Chapter 3 describes an extended method that is able to search for all models consistent with the observation. Consequently, the solution to a specific input dataset is then a set of possible subclone structures. The method then trim this solution space in the case that more than one sample from the same patient are available, such as the primary and relapse tumor pairs. Furthermore, a statistical framework is developed that, when further trimming is not possible, predicts whether two mutations are co-localizing in the same subclone. The formal definition on the problem of subclone structure reconstruction, as well as techniques to pre-process various types of genomic variation data are given given here as well. Results on the analysis of published and novel datasets, ranging from cancer types including Acute Myeloid Leukemia, Sinonasal Undifferenciated Carcinoma and Ovarian Serious Carcinoma, and data types including whole genome sequencing, copy number array, single nucleotide polymorphism array and single nucleotide variant calls with deep sequencing are also included. They show that the method is applicable to these wide range of cancer and data types, able to independently replicate the published conclusion based on manual reasoning, and gain novel insights into the pattern of tumor recurrence and chemoresistance. It also shows that the method can be valuable in prioritizing variants for function study.

Chapter 4 summarizes the entire work, and provide future prospects in subclonality research.

(This page is left blank intentionally . . .)

This work is dedicated to my wife, Xiaomeng Huang, who is kind, smart, gentle, and absolutely beautiful; who is diligent, when I am weary; who remains strong, when I am down; who contributes ideas, when I am stuck; who keeps me accompanied on my journey, and makes living ever so more enjoyable, everyday. This work would have not been possible, had she not been part of my life. I shall be eternally grateful for the grace she bestows upon me.

This work is dedicated to my parents, Xiaocui Zhao and Shaoping Qiao, who strive to provide me with love, care, support, and inspiration, sometimes at their own sacrifice. This work would have not been possible, had they not been as supportive and understanding in many of my decisions.

This work is dedicated to all who suffer from cancer, a horrible disease that torments both the patients and their family. It is to my uttermost hope that this work can be of even the slightest help towards the ultimate goal of defeating this common enemy of man-kind.

Contents

List of Figures

List of Tables

(This page is left blank intentionally . . .)

Acknowledgments

It is towards my advisor, Prof. Gábor T. Marth, that I owe my highest gratitude, in the completion of this work. He has provided invaluable guidance, as well as great flexibility in my conduct of research. He made sure that the lab was resourceful, and was never stingy about it. He showed much understanding and support to my unreasonable request to sometimes work from 800 miles away. And he is above all a good friend to hang out with.

I would also like to express my gratefulness to my Thesis Advisory Committee members Prof. Michelle Meyer and Prof. Peter Clote, my current and former colleagues Alistair Ward, Amit Indap, Andrew Farrell, Brain D'Astous, Chase Miller, Chip Steward, Deniz Kural, Derek Barnett, Erik Garrison, Jiantao Wu, Krzysztof Grzęda, Mengyao Zhao and Wan-Ping Li, and collaborators Aaron Quinlan, David Wheeler and Roeland Verhaak, who all participated in the discussion of problems involved in this work and contributed ultimately to the final solution.

Lastly, I would like to extend my appreciation to my friends who introduced me to the great American culture. My period of being a graduate student may be one of my many life experiences, but from my first driving lesson, to my first road trip, from my first raspberry, and peanut butter jelly sandwich, to my first family dinner party, you turned it truly into an experience of a life time.

1

Introduction

TO OUR BEST KNOWLEDGE, cancer arises due to the accumulation of somatic mutations that one acquires throughout his or her life span. It is the outcome of a Darwinian evolution process among cell populations in the micro-environment provided by different tissues of an organ [1] that some cells acquire alleles in "cancer-causing genes" [2, 3]. In the past thirty years, cancer-causing genes have been categorized into two main types: **oncogenes**, which are inactivated in normal cells and has the potential to cause cancer [4], and **tumor suppressor genes**, which are normally activated and protect the cell from progressing towards cancer [5]. Mutation occurs for many different reasons, including exposure to mutagenic chemi-

cals, intrinsic error of the DNA repair mechanism, inherited mutation from fertilized egg ("germline") that confers to genome instability ("susceptibility"), exogenous DNA materials from bacterial or viral infections, epigenetic changes and so on. Mutation also comes in many different forms, which were defined, based on the scope of the effect, as Single Nucleotide Variation (SNV), Insertion or Deletion (INDEL), chromosomal rearrangements and Copy Number Variation (CNV) [1, 6]. Disruption of the normal function of oncogenes and tumor suppressors leads to cellular phenotype such as self sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis[1], and metastasis [7].

In this chapter, I will give a brief introduction to our current knowledge regarding the complexity of cancer genome, the methods to study genomic variations, and latest studies which revealed intra-tumoral heterogeneity.

1.1 Complexity of cancer genome

With the understanding of human genome, and the waves of technologies that have come into exist especially the Next Generation Sequencing (NGS), the genome of cancer has been systematically studied through many efforts. Following the launching of the Human Genome Project (HGP) in 1990, which could be considered as the first step to unravel the puzzle of cancer genome, three major organizations, the Wellcome Trust Sanger Institute, the National Cancer Institute (NCI), and the International Cancer Genome Consortium (ICGC) have been leading the efforts to generate high-quality -omic data on more than 25,000 tumors for up to 50 types of cancers [8]. NGS not only increased the resolution of genotype profiling, but also enabled the discovery of novel mutations in new cancer-causing genes. Original assumption of a single uniform background mutation rate ($\sim$1/Mb) turned out to be overly simplified. A recent study initiated by the Broad Institute (BI) revealed that the mutation frequencies varied across

[1] The development of new blood vessels

cancer types and across patients within a cancer type through the analysis of 27 cancer types [9]. The same study also showed that the heterogeneity in the mutational spectrum of tumors. For example, a cluster consisting of samples dominated by C>G or C>T mutation in the context of TpC contains mostly cervical, bladder, breast and some head and neck cancer patients. Consistent with this result, Nik-Zainal *et al.* [10] reported that a cluster substitutions like C>T, C>A or C>G in TpCpX trinucleotides was found to be associated with ER-positive[2] breast tumors. In addition, two new types of mutations were reported recently, chromothripsis[3] [11, 12] and kataegis[4] [10, 13, 14]. Chromothripsis, which refers to a catastrophic phenomenon that the chromosomes appear to be shattered and then stitched back together, was identified by the use of paired-end NGS across multiple cancer samples [12, 15]. This process occurs in $2\% - 3\%$ of human cancer [16, 17]. The mechanism was proposed as that erroneous chromosome escapes from pulverization of chromosomal segments and undergoes aberrant reassembly through non-homologous end-joining [12, 18]. Kataegis operates locally, generating a hypermutation region characterized by multiple base substitutions. Though the mechanism remains unclear, the activation-induced deaminase (AID) and apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (APOBEC) protein families are likely to be involved [10, 19].

There have been several very detailed review articles summarizing the studies of global cancer genome features [19–21]. Here we focus more on new studies (mainly in late 2013 to 2014) that are not covered by these articles, summarizing the landscape of somatic mutations in different types of cancer.

[2]Endocrine receptor (estrogen or progesterone receptor) positive

[3]a catastrophic phenomenon that the chromosomes appear to be shattered and then stitched back together

[4]a hypermutation region characterized by multiple base substitutions

1.1.1 Colorectal and gastric cancer

Genetic aberration were detected in colorectal carcinoma (CRC) by the The Cancer Genome Atlas (TCGA) through the analysis of whole exome sequencing, copy number, promoter methylation, and mRNA & microRNA expressions in 276 samples as well as low coverage whole genome sequencing data in 97 samples [22, 23]. Hypermutated tumors (defined as > 12 mutations per mega-base) were observed in 16% of the CRC samples, three quarters of which had high levels of micro-satellite instability frequently caused by the silencing of DNA mismatch-repair pathway gene *MLH1* due to the hypermethylation on its promoter region. 24 genes that were found significantly mutated involved in several pathways such as the *WNT* signaling (*APC, CTNNB1, SOX9, TP53, FAM123B*), *PI3K* pathway (*IGF2, IRS2, PTEN, PIK3CA*), the transforming growth factor-β (*SMAD2* and *SMAD4*), the *RTK-RTS* signaling pathway (*KRAS, BRAF* and *ERBB* family), and chromatin remodeling (*ARID1A*).

Wang *et al.* [24, 25] carried out two studies to characterize genetic features of gastric cancer. The earlier study performed exome sequencing on a small cohort in 2011 and discovered frequent *ARID1A* mutation in the MSI and Epsterin-Bar virus (EBV) subgroups. The recent one, based on whole genome sequencing of 100 paired tumor and normal samples, identified new driver mutations (*MUC6, CTNNA2, GLI3, RNF43, ZIC4* and others), in addition to previously known *TP53*, *ARID1A*, and *CDH1* [25]. Specifically, authors found 6 *RHOA* mutations with the whole genome sequencing data, and another 7 in a cohort of 67 diffuse-type tumors, which were recurrent hot spot mutations and caused defective *RHOA* signaling.

1.1.2 Lung cancer

There are two histological subtypes of lung cancer: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC). NSCLC if further classified into squamous cell carcinoma (SCC), adenocarcinoma and large-cell carcinoma subtypes, with adenocarci-

noma being the most common subtype of NSCLC. Exome and genome sequencing of 183 tumor/normal pairs revealed high somatic mutation rate as 12 mutations per megabase [26]. Mutations including recurrent somatic mutations in the splicing factor gene *U2AF1*, *RBM10*, and *ARID1A*, as well as exonic alteration within *EGFR* and *SIK2* kinases were identified, and may be therapeutically targeted [26]. Another independent study carried out in Korea consisted of RNA sequencing of 200 lung adenocarcinoma and identified novel driver mutations such as *LMTK2, ARID1A, NOTCH2,* and *SMARCA4* [27]. A comprehensive genetic profile was also generated from 178 lung SCC samples. A study by the Broad Institute found 260 exonic mutations, 165 genomic rearrangements, and 323 segments of copy number alteration per tumor [28]. Further more, Vignot *et al.* [29] carried out targeted NGS assay on primary and matched metastatic tumor pairs NSCLC samples from 15 patients (8 with adenocarcinoma, 2 with large-cell carcinoma, 2 with basaloid and 3 with SCC), in which 63 known recurrent and 248 novel (likely passenger) mutations were discovered.

SCLC is an aggressive lung tumor subtype with frequent metastasis and early death. Rudin *et al.* [30] and Peifer *et al.* [31] independently reported exome, whole genome, transcriptome and copy number alteration data from a total of more than 100 primary SCLC tumors. Both studies found frequent inactivation of *TP53* and *RB1* [32]. Rudin *et al.* identified 22 significantly mutated genes including genes encoding kinases (i.e. *STK38, LRRK2, PRKD3*, and *CDK14*), *Ras* family regulators (i.e. *RAB37, RASGRF1,* and *RASGRF2*) and chromatin-modifying proteins or transcriptional regulators (i.e. *EP300, DMBX1, MLL2, MED12L, TRRAP,* and *RUNX1T1*). Peifer *et al.* [31] found recurrent mutations in the *PTEN, SLIT2, EPHA7, CREBBP, EP300,* and *MLL* genes as well as *FGFR1* amplifications.

1.1.3 Adrenocortical carcinomas

Adrenocortical carcinomass (ACCs) are rare and progressive cancers originating in the cortex of adrenal gland. Assié *et al.* [33] reported alterations in known driver genes (*CTNNB1, TP53, CDKN2A, RB1,* and *MEN1*) [34, 35] as well as new ones (*ZNRF3, DAXX, TERT,* and *MED12*) by exome sequencing and Single Nucleotide Polymorphism (SNP) array analysis of 45 ACCs samples. Specifically, a cell surface *E3* ubiquitin ligase gene, *ZNRF3*, was frequently mutated and is a potentially new tumor suppressor gene involved in β-catenin pathway.

1.1.4 Esophageal squamous cell carcinoma

Esophageal squamous cell carcinoma (ESCC) is a subtype of esophageal cancer and particularly common in China. Lin *et al.* [36] identified several new mutated genes such as *FAT1, FAT2, ZNF750,* and *KMT2D* in addition to those already known such as *TP53, PIK3CA,* and *NOTCH1* by whole exome or targeted deep sequencing of 139 paired ESCC along with CNVs of over 180 ESCC samples.

1.1.5 Urothelial bladder carcinoma

Urothelial bladder carcinoma is the most common type of bladder cancer and so far no molecularly targeted agents have been approved for the treatment of this disease. As part of TCGA project, a very recent study (March 2014, [37]) analyzed 131 high-grade muscle-invasive urothelial carcinomas to characterize genetic alterations with data including DNA copy number, somatic mutations, mRNA and microRNA expression, protein and phosphorylated protein expression, DNA methylation, transcript splice variation, gene fusion, viral infection, pathway perturbation, clinical correlates and histopathology. 32 recurrent mutated genes were identified, which involve in cell-cycle regulation (e.g. *CDKN1A*), epigenetic regulation (e.g. *MLL2, ARID1A, KDM6A,* and *EP300*), and kinase signaling pathways (e.g. *PIK3CA*). Out of these 32 genes, 9 of them

with > 5% frequency have not been reported as significantly mutated in any other TCGA cancer types, which are *CDKN1A, ERCC2, RXRA, ELF3, KLF5, FOXQ1, RHOB, PAIP1,* and *BTG2*.

1.1.6 Uterine leiomyomas

Uterine leiomyomas is a benign smooth muscle neoplasm but affects the health of women. Mehine *et al.* [38] performed whole genome sequencing and gene expression profiling on 38 uterine leiomyomas from 30 women and investigated clonal origin of tumors from different patients, driver events in Complex Chromosomal Rearrangements (CCRs) and candidate targets of chromosome 7q deletion. Identical or shared variants among separate tumor nodules suggested the same origin of those nodules and additional rearrangements could suggest the relation of clonal evolution among nodules. Surprisingly, interconnected CCR resembling chromothripsis, which is often associated with advanced stage of cancers and a poor prognosis in other studies [39], was observed frequently among these benign nodules. Finally, authors proposed a potential mechanism model for leiomyoma development including events such as *MED12* mutation, biallelic loss of *FH*, translocation of the *HMGA2* and *RAD51B* loci and aberrations at the *COL4A5/COL4A6* locus.

1.1.7 Mantle cell lymphoma

Mantle cell lymphoma (MCL) is an aggressive subtype of non-Hodgkin lymphoma[5]. Whole transcriptome sequencing (RNAseq) of 18 primary tissue and 2 cell line, and exome sequencing of 56 primary tissue samples respectively in two studies have found novel recurrent mutations in *NOTCH1, RB1, WHSC1, POT1,* and *SMARCA4* in addition to *ATM, CCND1, MLL2,* and *TP53* [38, 40]. Zhang *et al.* [38] further carried out chromatin structure and epigenetic profiling of normal B cells and MCLs and found that

[5]Any of a large group of cancers of lymphocytes (http://www.cancer.gov/cancertopics/types/non-hodgkin)

frequent somatic mutations were associated with open chromatin.

1.1.8 Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL), like other types of leukemia, has heterogeneous clinical and biological behavior. Previously three independent whole exome and whole genome sequencing studies [41–43] identified several mutations including *TP53, ATM, NOTCH1, MYD88* and splicing factor *SF3B1*. A recent study showed transcriptional profile by performing deep RNA sequencing in different subpopulations of normal B-lymphocytes and CLL cells from a cohort of 98 patients [44]. Higher expression of genes involved in metabolic pathways and lower expression of genes related to sliceosome, proteasome and ribosome were observed in CLL samples. B-cell receptor (BCR), *JAK-STAT* signaling and the cytosolic DNA-sensing pathways were shown to be particularly enriched in CLL.

1.2 Genomic profiling methods

In the 1970s, banding patterns of a person's chromosomes, or the karyotype, has been the primary tool for the clinical assessment of patients with a variety of genomic abnormalities. Later, methods such as fluorescence *in situ* hybridization (FISH) and its derivative methods, spectral karyotyping (SKY) or multiplex-FISH (M-FISH) [45, 46] were used to map DNA sequences to specific regions of human genomes, which allowed a higher resolution than the standard G-banding approaches. FISH can also be used to identify gene copy number variations by more or fewer fluorescent dots in somatic cells. Comparative Genomic Hybridization (CGH), a more advanced, FISH-based technique, was developed to study the gain and loss of chromosomal regions. Briefly, genomic DNA are isolated and fragmented from both control subject and experimental subject, and labeled with green and red fluorescence respectively. Then, two DNA samples are pooled and hybridized with normal chromosomes (known as probes). As a result, yel-

low fluorescence represents no alteration in experimental subject, whereas red or green represents copy number gain and loss. By using this method, Shayeteh *et al.* [47] discovered the amplification of *PIK3CA* in ovarian cancer. Useful as they have proven to be, these methods are not suitable for large-scale, high-resolution mapping of the entire genome.

1.2.1 Array comparative genome hybridization

Array comparative genome hybridization (aCGH) is a method that combines the principles of CGH and microarray [48]. The probe chromosomes are immobilized on a glass slide in an ordered fashion. The size of the probes can vary from tens to thousands of base pairs based on the areas of interest. The DNA fragmentation from a test sample and a reference sample are directly comparable to the standard CGH procedure. After applying two genomic DNA to the microarray, digital imaging systems are used to capture and quantify the relative fluorescence intensities of the labeled DNA. Not only did aCGH enable the discovery of more genetic alteration types, such as sub-telomeric rearrangements [49] and peri-centromeric rearrangements [50], but also made the analysis of large number of samples (e.g. 8789 clinical cases [50]) possible.

1.2.2 Sanger sequencing

Developed by Sanger *et al.* in 1977 [51, 52], the first chain-terminating, by-synthesis sequencing method has been widely adopted, and put in heavy use, until only recently with the rise of NGS. Sanger sequencing relies on the addition of dideoxynucleotides (ddNTPs), along with normal deoxynucleotides (dNTPs), so that the DNA polymerization process is halted randomly after incorporating a ddNTP which lacks the 3'-OH group to form the phosphodiester bond with the next base. Traditionally, the sequencing process is divided into four separate experiments, with each experiment containing all the dNTPs and one specific ddNTP. After the synthesize is complete, the product of

the experiments are loaded into separate lanes for gel electrophoresis to sort the DNA fragments by their length. Finally, the sequence is determined by reading the bands on the electrophoresis from short to long, based on the lane they are in. For example, a fragment of length 10bp in the lane that represents the experiment in which ddCTP is added indicates that the nucleotide on the 10th bp in the original sequence is G.

Based on the same principle, automated procedures [53, 54] are developed that, by using florescent labeled ddNTPs, the DNA synthesize is carried out in one experiment, followed by capillary electrophoresis and automated florescence color readout. The automated procedure saw great adoption in the vast majority of sequencing projects.

Although Sanger sequencing, and its derivative methods, has the disadvantage of high cost and low throughput, it produces significant longer sequencing reads that is essential for certain genomic applications (e.g. *de novo* assembly), and is often used as a validation method to verify the genomic variations discovered through NGS technologies.

1.2.3 Next generation sequencing

Next generation sequencing is a collection of methods that increase the throughput of traditional Sanger sequencing by simultaneously sequence several hundred millions of DNA fragments in parallel, utilizing either cyclic reversible terminators (Illumina, Helicos) to reversibly terminate the process of DNA polymerization followed by image capturing, quantitatively detecting the release of the pyrophosphate released by the incorporation of dNTP (Roche/454), or use ligation instead of nucleotide addition (SOLiD).

There are generally four steps to generate NGS data: template preparation, sequencing, imaging and data analysis [55]. DNA fragments are either clonally amplified with emPCR[6] followed by fixation onto an amino-coated glass slide by chemical cross-link

[6] emulsion-based PCR

(Life/APG SOLID) [56] or fixed onto glass slide first, and then followed by solid phase amplification (Illumina). Either way, the procedure results in millions of spatially separated DNA molecule template clusters.

1.2.3.1 ILLUMINA

In the sequencing step of Illumina platform, fluorescent labeled dNTP, with chemical modification on the 3'-OH group to inhibit further dNTP incorporation, are used to proceed the DNA polymerization in single steps. Since four dNTPs are labeled with different colors, the current nucleotide can be decided for every clonal template cluster. A cleavage step is followed to remove the fluorescent group as well as the blocking group at 3' to enable further polymerization and sequencing [57]. In an ideal situation, all the templates in a clonal cluster should be sequenced in synchrony. However, de-phasing will occur if more (leading-strand de-phasing) or less (lagging-strand de-phasing) nucleotides are incorporated onto some of the templates and introduce noise to the florescence signal. Such noise is used to calculate a probability ("base quality") that a specific base call is incorrect [55]. As a consequence, de-phasing will result in limited read length due to the low quality bases aggregating at the 3' end.

1.2.3.2 ROACH/454 PYROSEQUENCING

Pyrosequencing is another "sequence-by-synthesis" method that, instead of relying on chain termination (Section 1.2.2, Sanger Sequencing), works by detecting the pyrophosphate (PPi) released during dNTP incorporation [58–60]. The experiment consists of cycles that only one of the four dNTPs are added into the system. If the particular dNTP is complementary incorporated onto the synthesizing strand, PPi will be released, which is quantitatively converted to light signal through luciferase-catalyzed reaction, allowing the system to determine the sequence.

454 Life Sciences, which has been acquired by Roach Diagnostics, developed a par-

allelized version utilizing the same principle as pyrosequencing, taking advantage of emPCR to clonally amplify single DNA template within water droplets suspended in an oil solution. These clonal colonies are then transferred onto plates with picoliter-volume wells, and standard pyrosequencing is carried out within each well [61]

1.2.3.3 SOLiD

With SOLiD sequencing strategy, the DNA sequences are not determined by single addition of nucleotide, but by ligation of a short DNA probe that recognizes 2 consecutive bases on the template strand [56]. Different florescent colors are assigned to two bases instead of one. The sequencing process consists of the incorporation of 8-mer florescent labeled probes by DNA ligase with the first two bases complementary to the template sequence, color imaging, cleavage of the last 3 bases of the 8-mer, and incorporation of the next 8-mer. An entire cycle of such procedure will give raise to a color sequence representing the base change of every 2 bases separated by 3 bases. To cover the entire sequence, the procedure will need to be repeated 3 more times, each time with a +1 base shift. As a result, every base is interrogated twice, and SNV will be easy to identify since substitution of one base will result in color changes of two consecutive positions. Sequencing errors are also easier to identify due to the fact that color changes with different inner bases are invalid.

1.3 NGS DATA PROCESSING AND VARIANT DISCOVERY

A common pattern of the current in-production NGS technologies is that they produce giga base-pairs of genomic sequencing data, but each individual sequencing read is relatively short [55]. Because the sequencing reads do not retail the knowledge as in where in the genome did they originated from, a common practice is to map all the reads to a reference genome scaffold to identify their location of origin. Heng *et al.* [62] reviewed the latest software packages designed to tackle this issue, with various

algorithmic approaches and adaptations to specific sequencing technologies.

Many types of genomic variations can then be identified from the alignments, either by measuring the amount of reads covering a certain region (read coverage) to identify CNV events [63–68], interrogating the raw sequences to discover SNV events [69, 70], or utilizing properties of the alignment reads (e.g. fragment length or orientation with paired-end reads) to uncover CNV and structural variations [66–68].

An important aspect of variant discovery regarding cancer samples is that, due to the existence of normal tissue mix-in and intratumoral heterogeneity, certain assumptions made for germline variation discovery may no longer hold true. For example, when only normal tissue is concerned, the allele frequency (AF) of a mutant allele at a specific locus is either 0% (when the subject does not have the variation), 50% (when the subject is heterozygous mutant) or 100% (when the subject is homozygous mutant). As a result of heterogeneity or copy number alteration, somatic events in cancer samples do not follow this trimodal distribution necessarily [71–73]. Methods have been developed to tailor for these specific conditions [74–81]. Xu *et al.* [82] reviewed the performance of these methods, and Kim *et al.* [83] proposed a strategy of combining multiple tools to increase overall call quality.

1.4 Studies reveal intra-tumor genomic heterogeneity

Cancer has been known to be heterogeneous long before the arrival of high throughput genomic profiling methods [84–105]. NGS, and more recently single cell sequencing (SCS), enabled the elucidation of subclone structure into unprecedented level. Here I briefly summarize the latest studies, utilizing NGS on bulk sample or SCS, that identified intratumoral heterogeneity in various types of cancer.

1.4.1 Studies based on NGS

1.4.1.1 Myeloma

Three studies [106–108] performed genome-wide analysis for the clonal landscape representing the heterogeneity in multiple myeloma (MM). Keats *et al.* [106] traced genetic changes over the entire disease course in a high-risk patient at 7 time points and identified 2 competing subclones. A Vk*MYC transgenic mouse was used to model the competition of these two subclones. An important lesson learned in this animal model was that the eradication a sensitive clone will probably resulted in the dominance of the other refractory clone. Thus the authors suggested that combination therapies targeting all co-existing subclones in the tumor would be more beneficial than sequential single-agent therapy. Egan *et al.* [107] performed whole genome sequencing on 4 time-points samples over tumor progression: diagnosis, first relapse, second relapse and end-stage secondary plasma cell leukemia (sPLC) in a t(4; 14) MM patient. Results showed that diagnostic and second relapse clones shared most SNVs, while the first relapse and the sPLC clones have some unique SNVs, which "suggested greater evolutionary divergence over time and disease aggressiveness". Walker *et al.* [108] compared a group of MM patient samples with t(4; 14). Only 3% of mutations, including driver mutations in *RAS/MAPK* pathway, were shared by both group and, in addition, *RAS* pathway mutations were not always present in the dominant clone, but instead in minor subclones in half of the samples.

Recently, two more studies looked at the heterogeneity of genomic evolution in MM [109, 110]. Bolli *et al.* analyzed 84 myeloma samples by whole exome sequencing and copy number profiling. Melchor *et al.* also used whole exome sequencing in addition to single cell qPCR[7]. Both studies identified linear and branching phylogenies, which contained 5 to 6 subclones. A very important observation was made that in some of the

[7] quantitative real time PCR

parallel evolution situations, two subclones independently "activated the *RAS/MAPK* pathway through *RAS* mutations", which resulted in distinct subclonal lineages [110].

1.4.1.2 Breast Cancer

Nik-Zainal *et al.* [111] applied novel algorithms developed by the same group to 21 breast tumors. Dominant subclonal linage (defined as more than 50% of tumor cells) was observed in every tumor and authors reasoned that hundreds and thousands of mutations were accumulated in one cell lineage before it expanded into the dominant subclone with the acquisition of "driver mutations". In another study, Shah *et al.* [112] discovered, with 104 triple-negative breast cancer (TNBC) cases and deep re-sequencing of 2,414 somatic mutations, that TNBC clonal structures vary drastically from case to case, and concluded that "understanding the biology and therapeutic responses of patients with TNBC will require the determination of individual tumor clonal genotypes".

1.4.1.3 Leukemia

Relapse caused the death of most acute myeloid leukemia (AML) patients [113]. By comparing of genomes from primary tumor and relapse, two main clonal evolution patterns were associated with relapse: either the major primary clones, or a minor surviving subclone in primary from initial chemotherapy gained mutations and evolved into the relapse [113]. Another group independently investigated clonal evolution of preleukemic hematopoietic stem cells (HSCs). By using targeted exome sequencing, Jan *et al.* [114] identified cellular and genomic path from HSCs to the dominant presenting leukemic clone.

1.4.1.4 Chronic Lymphocytic Leukemia

Schuh *et al.* [115] monitored the disease progression in 3 CLL patients by sampling at 5 time points over up to 7 years. Whole genome sequencing results on the collected

samples showed that each sample had up to 5 distinct subpopulations. The mutation profiles at different time points revealed the clonal evolution process, which was represented by the dynamics of subclones that declined or expanded over time.

Another intratumoral heterogeneity study was conducted by Landau *et al.* in 149 CLL cases [73]. Cell frequencies of somatic mutations were generated from whole exome sequencing and copy number analysis. Some driver mutations, such as *MYD88*, trisomy[8] 12 and del(13q) were found to be predominantly clonal, signaling early acquisition, whereas others such as *SF3B1* and *TP53* were found to be subclonal, representing later events.

1.4.1.5 Myeloproliferative neoplasms

Lundberg *et al.* [116] did comprehensive analysis in a cohort of 197 myeloproliferative neoplasm (MPN) patients by targeted NGS of 104 genes. A strong correlation was found between the total number of somatic mutations and survival and risk of leukemia transformation. Clonal analysis was carried out by genotyping DNA from signal colonies grown in methylellulose and genes focused in this study were epigenetic modifiers. Mutation profiles of *TET2, DNMT3A, JAK2, V617F, ASXL1, EZH2,* and *IDH1* revealed 8 types of clonal structures.

1.4.1.6 Mantle cell lymphoma

Beá *et al.* [117] reported a whole genome and/or exome sequencing study on 29 MCL and normal tissue pairs, and identified some recurrent mutations which then were investigated by targeted NGS in an independent 172 MCL cases. Sequencing data of two tumor samples were obtained from two different topographic sites or at two time points from each patient were used to establish 4 types of subclonal architectures in MCL.

[8] having three instances of a particular chromosome, instead of the normal two

1.4.1.7 Renal carcinomas

Gerlinger *et al.* [118] performed exome sequencing, chromosome aberration analysis, and ploidy[9] profiling on "multiple spatially separated" primary and associated metastatic renal carcinoma samples, and revealed branching evolution pattern in tumor growth. Tumor suppressor genes, such as *SETD2, PTEN,* and *KDM5C* exhibited convergence evolution, "underwent multiple distinct and spatially separated inactivating mutations within a single tumor". Ploidy heterogeneity was also observed in two of four tumors.

1.4.1.8 Pancreatic Cancer

Yachida *et al.* [119] carried out genomic sequencing of primary and metastatic cancers to assess their clonal relationships in 7 pancreatic cancer patients. They found that distant metastasis clones were originally within the primary carcinoma, which were non-metastatic clones. The authors also generated a time-line of metastasis initiation and occurring that at least 10 years for the tumor initiation, 5 years for the arise of parental metastatic clones, and then about 2 more years till decease. Another study was conducted by Compbell *et al.* [120] with parallel paired-end sequencing on 13 pancreatic adenocarcinoma patients. Besides identifying somatic mutations and rearrangements, they also investigated phylogenetic relationships, and confirmed that certain clones in the primary tumor had the ability to initiate metastasis. In addition, they discovered that organ-specific branching patterns of phylogenetic trees. The authors suggested two explanations: particular genotypes might drive metastasis to a particular organ; or, metastatic clones may expand in a stepwise process.

[9]The number of copies of a complete genome in a cell. Normal cells have two copies, thus diploid. Tumor cells with copy number variation could potentially contain three copies, or triploid, or more.

1.4.2 Studies based on single cell sequencing

1.4.2.1 Breast cancer

A method of combining flow-sorted nuclei, whole genome amplification, and NGS was able to accurately quantify genomic copy number changes within an individual nucleus [121]. By utilizing this method, Zik-Zainal *et al.* analyzed two sets of 100 single cells, and identified three distinct clonal subpopulations in a poly-genomic tumor and a single clonal expansion forming the primary tumor and seeded the metastasis in a mono-genomic primary tumor and its liver metastasis [121].

1.4.2.2 Myeloproliferative neoplasms

A high-throughput whole genome single cell sequencing method was developed by Hou *et al.* [122]. This method was of high sensitivity and had a distinct genomic distribution from tissue sequencing that GC extremely enriched regions had lower amplification efficiency. This method was used to sequence 90 cells from a *JAK2*-negative MPN patient, and identified a monoclonal evolution pattern in this patient sample.

1.5 Challenges

As it was made clear by the studies mentioned above, tumor samples are highly heterogeneous in terms of the genomic profiles of the constituting cancer cells. The heterogeneity itself contributes to the complexity of tumor genome, and hinders the investigation of mechanisms, such as tumorigenesis or metastasis, through traditional means. To make things worse, the problem of delineating each individual genomic profile within a tumor mass is fundamentally different from phylogenetic tree constructing since each individual genome cannot be observed separately. Thus it calls for novel computational methods to elucidate the number of different genomes, and their specific profiles, from various signals that represent the mixture of them all.

2

Linear evolution model

ONE OF THE CHALLENGES in analyzing cancer sequencing data was that the tumor sample used for sequencing would often have normal cells mixed in, or "normal contamination". For example, when a tumor is surgically removed, the surgeon will also remove the surrounding normal tissues as well to ensure maximum removal of tumor cells. Macrophage invasion and blood can also be sources for normal contamination. We started to design computational methods to estimate the level of contamination by looking for signals that separate the normal "clone" that all share the germline genome from tumor clone that all have the mutated somatic genome. Soon, however, we realized that the clonal structure is far more complex than

this binary segregation.

2.1 TUMOR IS HETEROGENEOUS

As part of a collaboration with Baylor College of Medicine, we gained access to the The Cancer Genome Atlas (TCGA) Ovarian Cancer dataset primary alignment data in the format of BAM[1] files. Although many studies had been published on the topic of estimating normal contamination, as well as extracting Copy Number Variation (CNV) features, with Single Nucleotide Polymorphism (SNP) array data [123–130], only a few methods were able to utilize Next Generation Sequencing (NGS) data [131–133], and none of which was considering normal contamination. We performed our own copy number analysis procedure on whole genome sequencing samples that consisted of the following steps to estimate normal contamination:

1. The read depth (RD), number of sequencing reads starting within a region, was scanned through the entire genome by a 10kb non-overlapping moving window (Figure 2.1.1 A, B).

2. The read depth in normal and tumor were separately normalized by the total number of reads in each sample.

3. The read depth ratio (RDR), with the definition that $RDR = \frac{RD_T}{RD_N}$, was calculated for every corresponding window, and a histogram was generated (Figure 2.1.1 C). This step also effectively filtered out any germline (inherited) events to allow the subsequent steps to only consider somatic (acquired) events.

4. The histogram envelope signal was extracted using Fast Fourier Transform (FFT), and low pass filtered to reduce high frequency noises (Figure 2.1.2 A, B, note that the sample is different from Figure 2.1.1).

[1] A binary file format widely used for storing sequencing reads alignments.

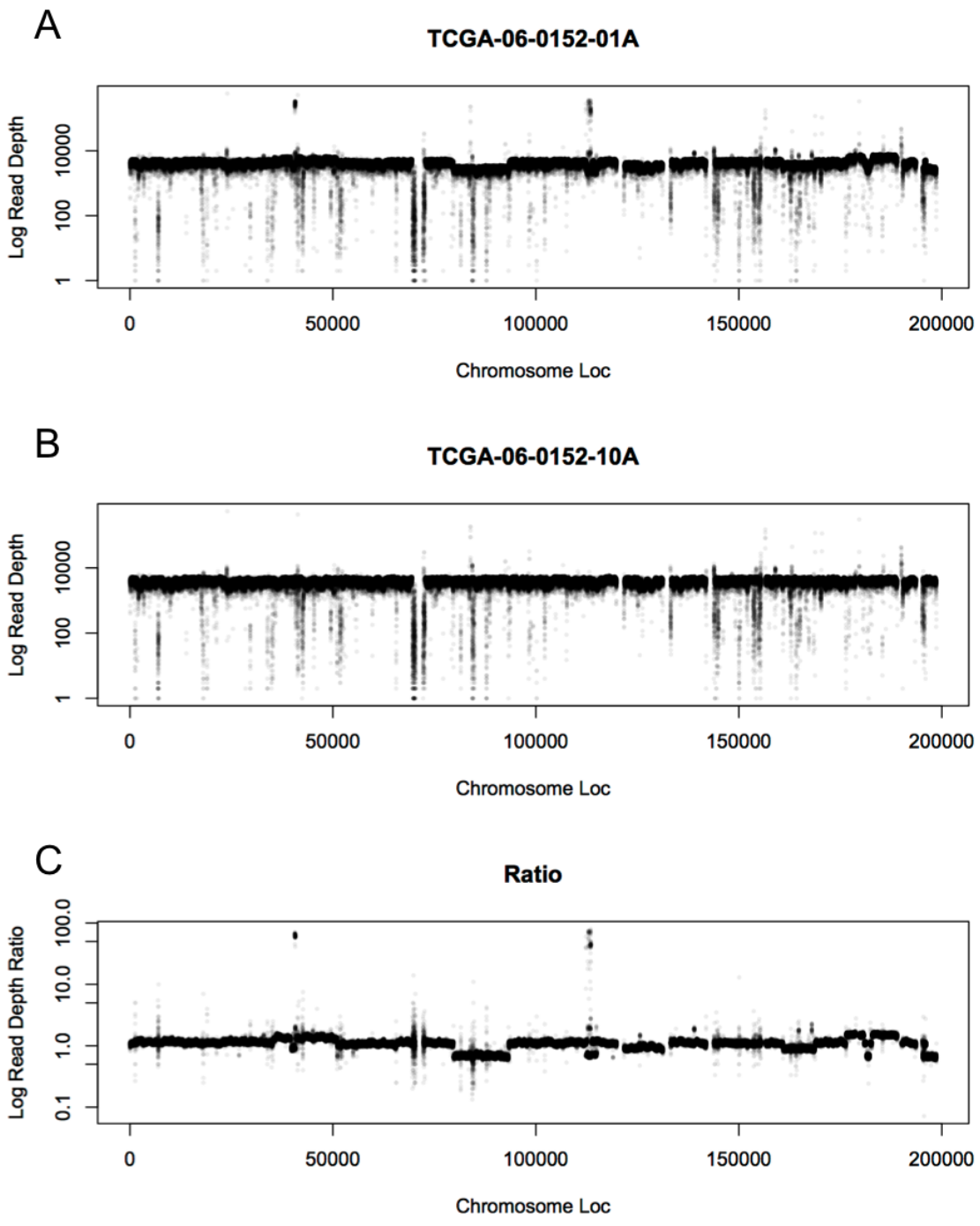


Figure 2.1.1: The procedure to calculate read depth ratio, *RDR*, from paired tumor-normal samples. A) Read depth measured in a 10kb moving window on the whole genome sequencing data of patient TCGA-06-0152 primary tumor sample. B) Read depth measured in a 10kb moving window on the whole genome sequencing data of patient TCGA-06-0152 normal sample. C) Read depth ratio between the tumor and normal sample of patient TCGA-06-0152.

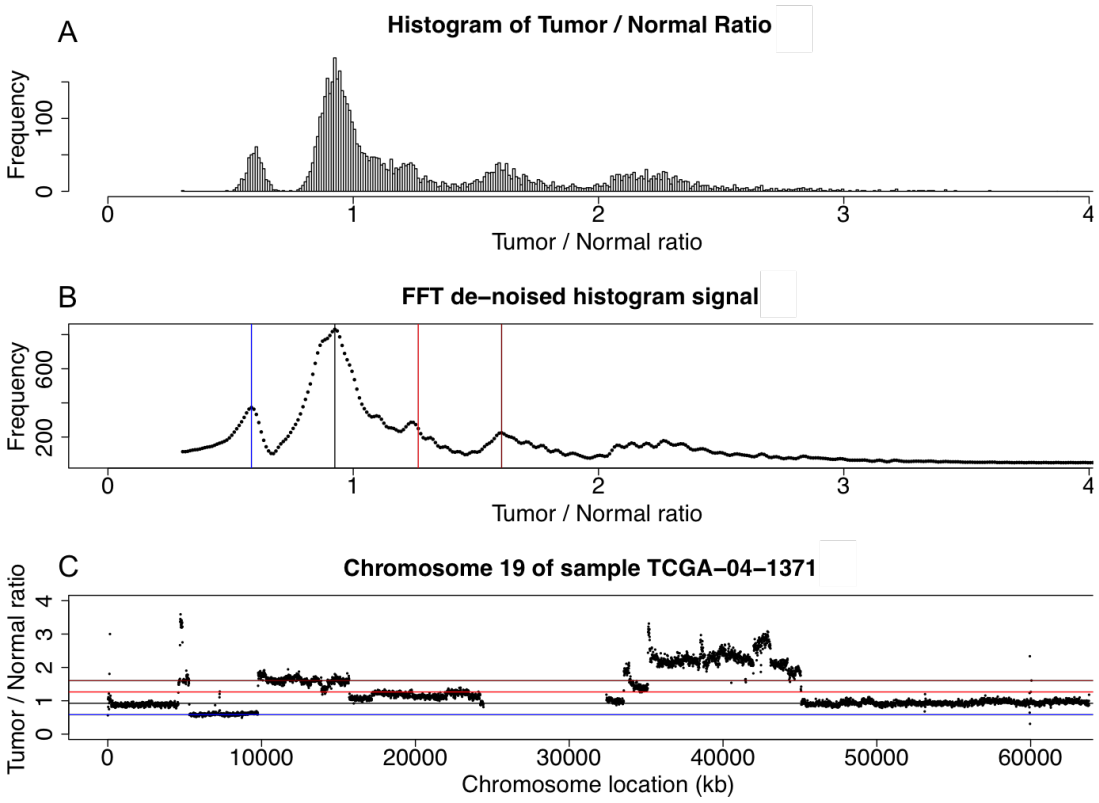


Figure 2.1.2: The procedure to estimate. A) The histogram of *RDR*. Peak shape is very obvious B) FFT de-noised histogram envelope. Black line represents identified copy number 2 peak, while blue line represents identified copy number 1 peak. Based on these two lines, location of copy number 3 and 4 are estimated with the knowledge of contamination (red and dark red lines). C) Plot of *RDR* along chromosome locations with identified and estimated ratio of specific copy numbers drawn as horizontal lines (definition is the same as in B)

5. Global maximum was identified and assumed to be copy number 2 with corresponding ratio denoted as R_2 (Figure 2.1.2 B, black line).

6. The first local maximum on the left side of R_2 was identified, and assumed to be copy number 1 and denoted as R_1 (Figure 2.1.2 B, blue line). In a simple model where only two genomes, tumor and mixed-in normal, are considered, the contamination level α could then be estimated by these two values

$$\frac{1 \times (1-\alpha) + 2 \times \alpha}{2} = \frac{R_1}{R_2}$$
$$\alpha = 2 \times \frac{R_1}{R_2} - 1 \tag{2.1}$$

7. With α estimated, the corresponding RDR of copy number $3, 4, \ldots, n$ can be calculated with Equation 2.2 (Figure 2.1.2 B, red and dark red lines).

$$\frac{R_n}{R_2} = [n(1-\alpha) + 2\alpha]/2 \tag{2.2}$$

As shown in Figure 2.1.2 this approach worked relatively well within the boundary of a single chromosome for the specific individual TCGA-04-1371. Using the estimated contamination level α, the RD peak for copy number 3 and 4 were predicted accurately. However, once we looked at all the chromosomes in the entire genome, it turned out that different chromosome resulted in different α estimation (Figure 2.1.3 A). This did not make sense because the normal contamination should reflect the amount of normal cell mixed within the tumor sample, which should be identical across the entire genome. More interestingly, the estimated α values for all the chromosomes seemed to cluster into three distinct groups. The only way to rationalize the observation was that (Figure 2.1.3 B)

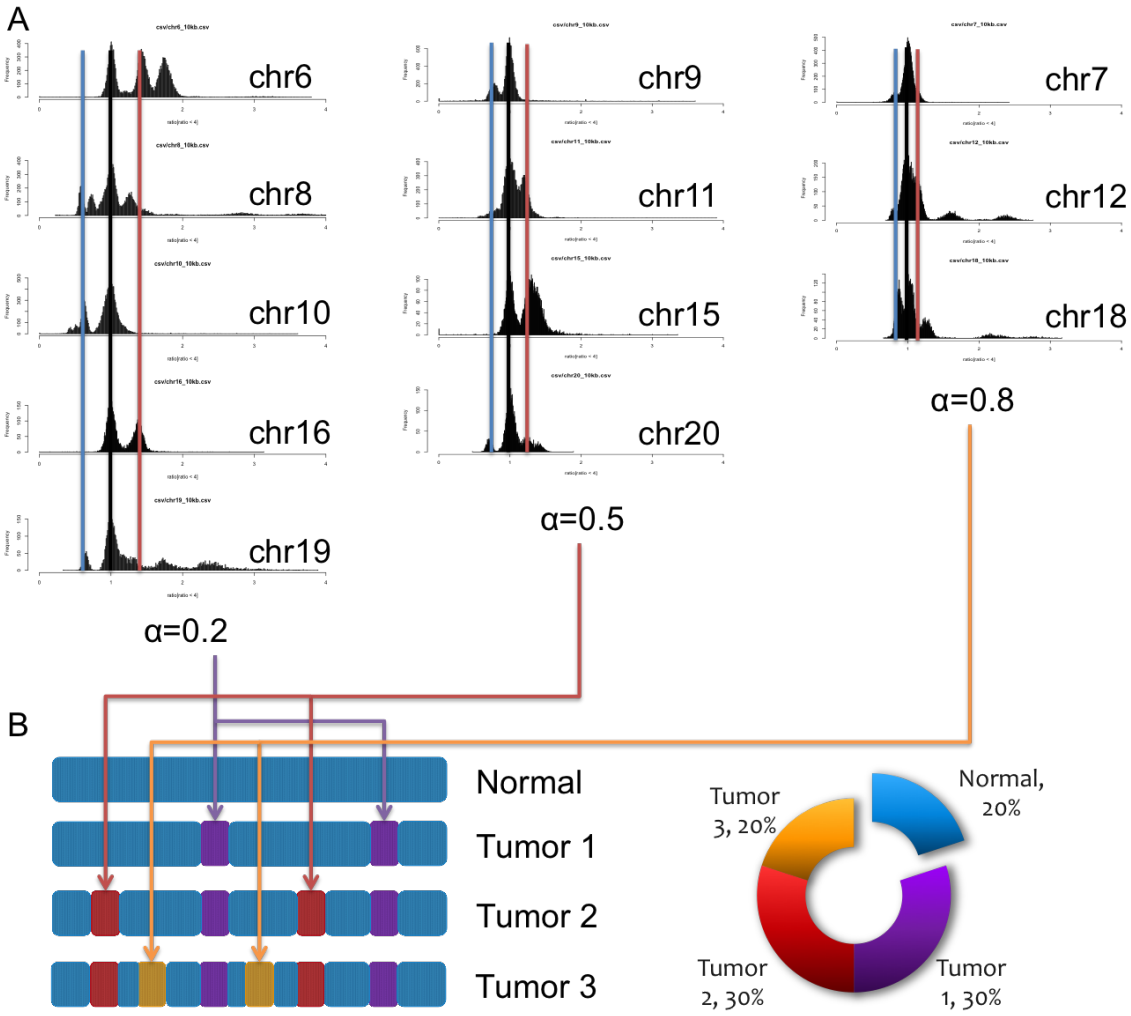


Figure 2.1.3: Normal contamination estimation based on all chromosomes in the primary tumor sample in patient TCGA-04-1371, and its parsimonious subclone structure. A) The estimation values from different chromosomes are grouping into three clusters, with $\alpha = 0.2, 0.5,$ and 0.8 respectively. B) The parsimonious subclone structure that explains the data. The events that resulted in lower α estimation, which represents higher tumor purity, exist in more subclones.

- The tumor cells, instead of sharing the same genome, actually were comprised of different subgroups, or subclones.
- The cells within the same subclone share the same genome.
- Genomic variations exist in at least one, but possibly more subclones. The subgroups which do not contain a specific variation would "act" like normal cells, and contribute to the estimated α, at the location of that variation.

2.2 PARSIMONIOUS SOLUTION BASED ON LINEAR EVOLUTION MODEL

Based on the observation we made in Section 2.1, we rationalized that, one way to model the multi-level α estimation based on the copy number data is that, for n different levels of distinct α, there exist at least n tumor subclones (hence parsimonious) plus 1 contaminating normal clone. The tumor subclones follow a linear evolutionary model, in which the events resulting in the lowest α estimation emerged in tumor tissue cells the earliest, and expanded into the initial tumor subclone population. One of the cell in the tumor population further gained mutations, and developed into another subpopulation. But because the events acquired later only exist in a subset of the entire tumor sample, when being interrogated alone, they will result in a higher α estimation because more cells (normal tissue plus those tumor cells that contain the initial events, but not the later events) would appear to be normal tissue at those specific genomic locations. New subclone always emerges from the most mutated, existing subclone, inheriting all the existing events, and in addition contain their own set of events.

We then developed an algorithmic procedure to reconstruct the parsimonious subclone structure based on this model, to which the input is a list of CNV events, and their associated RDR. The somatic RDR is then converted to Cell Prevalence (CP), that describes the fraction of cells in the sequenced tumor sample that harbors a particular event. In order to be able to compute CP, the absolute copy number (ACN) state of

an event, i.e. the exact number of copy of DNA at the location of the event, is necessary, which is often a non-trivial task to estimate. We estimate the ACN assuming that CP follows a Uniform Distribution $U(0,1)$, using a Maximum Likelihood method that $\mathscr{L}(RDR|ACN)$ is maximized. Given an overall ploidy[2] p, which can be estimated by methods such as ASCAT [134] and ABSOLUTE[135], or assumed to be 2 if no other information is available, the method will result in a closed form:

$$ACN = \arg\max_i \frac{i}{p} < RDR, i = 0, 1, \ldots \tag{2.3}$$

The method can be implemented in a iterative fashion in the following steps (Figure 2.2.1)

1. Initialize the subclone structure with a single subclone that represents the normal tissue mixture. This subclone contains no event, and its frequency $f = 1$

2. Identify the events in the event list that have the lowest $CP = CP'$; create a new subclone which contains all the events in the event list; set its frequency $f = CP'$, and subtract CP' from the frequency of the "normal" clone; remove the events considered in this step from the event list; subtract the CP of all the events in the event list by CP'

3. If the event list still has events in it, repeat step 2; Otherwise, return the subclone structure

Formally, the method is designed on the following definitions:

Definition 2.1. *A chromosomal location, L, is defined as* $L = \{chromosome, position\}$*, which describes a location on the genome.*

Definition 2.2. *A chromosomal segment, S, is defined as* $S = \{L, length\}$*, which describes a continuous region on the genome.*

[2]The number of copies of a complete genome in a cell. Normal cells have two copies, thus diploid. Tumor cells with copy number variation could potentially contain three copies, or triploid, or more.

Definition 2.3. *S contains L iff* $L.chromosome = S.L.chromosome$, $L.position \geq S.L.position$, *and* $L.position < S.L.position + S.L.length$. *We denote this as* $L \in S$

Definition 2.4. *Two chromosomal segments,* S *and* S', *overlaps iff* $S'.L \in S$, *assuming without loss of generality that* $S.L.position \leq S'.L.position$

Definition 2.5. *A segmental somatic CNV event (henceforth referred to as "event" if without specification), e, is defined as* $e = \{S, ACN\}$ *for a segment on the genome specified by S, with the absolute copy number state of* ACN.

Definition 2.6. *An observed somatic CNV event (henceforth referred to as "observed event" if without specification), oe, is defined as* $oe = \{e, CP\}$ *for an event e observed in* $CP > \epsilon_d$ *fraction of the total cells, with some detection sensitivity* $\epsilon_d > 0$.

Definition 2.7. *An observation, O, is defined as* $O = \{oe_1, oe_2, \ldots, oe_n\}$ *for a segmented tumor genome profile with n segments of non-modal RDR (The following discussion assume a modal copy number being 2, thus a modal RDR* $= 1$*). The chromosomal segments, S, of the events, e, in* oe_i *and* oe_{i+1} *need not be continuous, since locations of the genome can potentially be masked out. Observed events are identified with genomic segmentation algorithms, which will result in non-overlapping segments.*

Definition 2.8. *The complete events set, E, is defined as* $E = \{oe.e | oe \in O\}$.

Lemma 2.1. *For any given genomic location L, there exist at most 1 observed event oe so that* $L \in oe.e.S$

Proof. Suppose that, for a given L, there exist two observed events oe and oe', so that $L \in oe.e.S$ and $L \in oe'.e.S$. We denote, for simplicity, $oe.e.S$ as S and $oe'.e.S$ as S'. We also assume, without loss of generality, that $S'.L.\text{position} \geq S.L.\text{position}$

$$L \in S' \implies L.pos \geq S'.L.pos$$

$$L \in S \implies L.pos < S.L.pos + S.\text{length}$$

$$\implies S'.L.pos < S.L.pos + S.\text{length}$$

combined with the assumption that $S'.L.\text{position} \geq S.L.\text{position}$

$$\implies S'.L \in S$$

$$\implies S \text{ and } S' \text{overlaps}$$

This contradicts with Definition 2.7 □

Corollary 2.1. *For any given genomic location L, there exist at most 1 event $e \in E$ so that $L \in e.S$.*

Proof. Suppose that, for a given L, there exist two events $e \in E$ and $e' \in E$, so that $L \in e.S$ and $L \in e'.S$. Due to Definition 2.8, there must exist two observed events $oe \in O \wedge oe.e = e \implies L \in oe.e.S$ and $oe' \in O \wedge oe'.e = e \implies L \in oe'.e.S$. This contradicts with Lemma 2.1 □

Definition 2.9. *A subclone profile, C, is defined as $C^j = \{G, f\}^j, j = 0..m, f \geq 0$, in which $C^j.G \subseteq E$ is a set of events the j-th subclone contains. The 0-th subclone is a special one representing the normal tissue component, thus $C^0.G = \{\}$. $C^j.f$ represents the fraction the j-th subclone occupies over the entire cell population, or subclone frequency (SF), and that $\sum_{j=0}^{m} C^j.f = 1$*

Definition 2.10. *An actual (observed) genomic profile, A, is defined as $A = \{A_l\}, l = 1, 2, \ldots$ for each unique location L_l on the genome. Since at most 1 oe exists so that*

$L_l \in oe.e.S$ (Lemma 2.1), A_l is calculated as following

$$A_l = \begin{cases} \frac{oe.e.ACN}{2} \times oe.CP + (1 - oe.CP) & \exists! oe \in O : L_l \in oe.e.S \\ 1 & otherwise \end{cases}$$

which can be simplified as

$$A_l = \begin{cases} 1 + \frac{oe.e.ACN - 2}{2} \times oe.CP & \exists! oe \in O : L_l \in oe.e.S \\ 1 & otherwise \end{cases} \tag{2.4}$$

Definition 2.11. *A model genomic profile, M, is defined as $M = \{M_l\}, l = 1, 2, \ldots$ for each unique location L_l on the genome. Since for any given subclone C^j, at most 1 e exists for $C^j.G$ so that $L_l \in e.S$ (Corollary 2.1), M_l is calculated as following*

$$M_l = \sum_{j=0}^{m} \begin{cases} \frac{e.ACN}{2} \times C^j.f & \exists! e \in C^j.G : L_l \in e.S \\ 1 \times C^j.f & otherwise \end{cases} \tag{2.5}$$

Definition 2.12. *A model fitness score, fit, is defined as in Equation 2.6, which calculates the difference between the model and actual genomic profiles.*

$$fit = \sum_{l=1}^{genomic\ length} |M_l - A_l| \tag{2.6}$$

Definition 2.13. *If, in a given subclone profile C with $m+1$ subclones, the condition that $\forall j < m : C^j.G \subseteq C^{j+1}.G$ is satisfied, the subclone profile is said to be according to a linear evolution model. We say that a profile is parsimonious when the following conditions are met*

$$\forall j < m : C^j.G \subset C^{j+1}.G$$

$$\forall j \in [1, m] : C^j.f > \epsilon_f \text{ for a given error margin } \epsilon_f \geq 0$$

Definition 2.14. *The problem of subclone structure reconstruction with linear evolution model, is that given an observation O, find a subclone profile C, in which* $C^j.G \subset C^{j+1}.G, j = 0..m-1$*, that minimizes* fit*. We say that a subclone profile C is a solution to an observation O if, for a given error margin* $\epsilon_{fit} \geq 0$*,* $fit \leq \epsilon_{fit}$*.*

Theorem 2.1. *For any given subclone structure C that is a solution to an observation O and the implied complete event set E*

$$\bigcup_{j=1}^{m} C^j.G = E$$

Proof. Assume that

$$\bigcup_{j=1}^{m} C^j.G = E' \neq E$$

Due to the fact that $\forall j \leq m : C^j.G \subseteq E$ (Definition 2.9), we have

$$\begin{aligned} & E' \subset E \\ \Longrightarrow & \exists e \in E : e \notin E' \\ \Longrightarrow & \forall L_l \in e.S : M_l = 1 \text{ (Equation 2.5)} \end{aligned}$$

yet

$$
\begin{aligned}
& e \in E \\
\implies & \exists oe \in O : oe.e = e \\
\implies & \forall L_l \in e.S : A_l = 1 + \frac{oe.e.ACN - 2}{2} \times oe.CP \text{ (Equation 2.4)} \\
\implies & \forall L_l \in e.S : |M_l - A_l| = |1 - (1 + \frac{oe.e.ACN - 2}{2} \times oe.CP)| \\
\implies & fit \geq |\frac{2 - oe.e.ACN}{2} \times oe.CP| \times e.S.\text{length (Equation 2.6)}
\end{aligned}
$$

thus, for any given $\epsilon_{fit} < |\frac{2-oe.e.ACN}{2} \times oe.CP| \times e.S.\text{length}$, C cannot be a solution to the observation O, contradicting with the starting condition. □

Theorem 2.2. *For any parsimonious linear subclone profile (Definition 2.13) C with $m+1$ subclones that is also a solution to an observation O with n observed events, $m \leq n$*

Proof. Note that O is with n observed events implies that E is with n events (Definition 2.8). Assume that $m > n$, because C is parsimonious,

$$
\begin{aligned}
& \forall j < m : C^j.G \subset C^{j+1}.G \\
\implies & |C^{j+1}.G| > |C^j.G| \\
\implies & \forall j \leq m : |C^j.G| \geq j \\
\implies & \forall j \text{ that } m \geq j > n : |C^j.G| \geq j > n
\end{aligned}
$$

This contradicts with Definition 2.9, that $\forall j \leq m : C^j.G \subseteq E$.

□

Definition 2.15. *A set of event clusters, P, is defined as a partition over an observation O, that, for a given error margin $\epsilon_P \geq 0$, satisfies*

$$\forall p \in P :(\forall oe \in p, oe' \in p : |oe.CP - oe'.CP| \leq \epsilon_p)$$

$$and$$

$$\forall p \in P, p' \in P, p \neq p' :(\forall oe \in p, oe' \in p' : |oe.CP - oe'.CP| > \epsilon_p)$$

Each element $p \in P$ is called an event cluster. We denote $p.CP = \frac{\sum_{oe \in p} oe.CP}{|p|}$ as the cluster centroid.

We further impose, without loss of generality, that P is a sorted set, with respect to the cluster centroids, in descending order.

$$\forall i \in [1, n'], j \in [1, n'], i < j : p_i.CP > p_j.CP$$

Theorem 2.3. *For any parsimonious linear subclone profile (Definition 2.13) C with $m+1$ subclones that is a solution to an observation O with n observed events being partitioned as P with $n' \leq n$ clusters, $m \leq n'$*

Proof. P is a partition over $O \implies |P| \leq |O|$.

If $|P| = |O| \implies n' = n$, with Theorem 2.2, we have $m \leq n'$; otherwise ($|P| < |O|$), assume that $m > n'$

$$\bigcup_{j=1}^{m} C^j.G = E \text{ (Theorem 2.1)}$$
$$\Rightarrow |\bigcup_{j=1}^{m} C^j.G| > |P|$$

Due to pigeonhole principle,

$$\exists p \in P, |p| > 1 : \{$$
$$\exists j \leq m, j' \leq m, 0 < j < j' : [$$
$$\exists oe.e \in C^j.G, oe'.e \in (C^{j'}.G - C^j.G) : oe \in p \wedge oe' \in p$$
$$]\}$$

Let

$$f_{oe} = \sum_{k=j}^{m} C^k.f$$
$$f_{oe'} = \sum_{k=j'}^{m} C^k.f$$

Thus we have

$$oe \in C^j.G \wedge oe \in C^{j'}.G$$

$$oe' \notin C^j.G \wedge oe' \in C^{j'}.G$$

$$\Longrightarrow$$

$$\forall L_l \in oe.e.S : |M_l - A_l|$$

$$= |(\sum_{k=0}^{j-1} C^k.f + \sum_{k=j}^{m} \frac{oe.e.ACN}{2} \times C^k.f) - (1 + \frac{oe.e.ACN - 2}{2} \times oe.CP)|$$

$$= |(1 - f_{oe}) + \frac{oe.e.ACN}{2} \times (f_{oe} - oe.CP) - (1 - oe.CP)$$

$$= |(\frac{oe.e.ACN}{2} - 1) \times (f_{oe} - oe.CP)| \tag{2.7}$$

$$\forall L_{l'} \in oe'.e.S : |M_{l'} - A_{l'}|$$

$$= |(\sum_{k=0}^{j'-1} C^k.f + \sum_{k=j'}^{m} \frac{oe'.e.ACN}{2} \times C^k.f) - (1 + \frac{oe'.e.ACN - 2}{2} \times oe'.CP)|$$

$$= |(1 - f_{oe'}) + \frac{oe'.e.ACN}{2} \times (f_{oe'} - oe'.CP) - (1 - oe'.CP)$$

$$= |(\frac{oe'.e.ACN}{2} - 1) \times (f_{oe'} - oe'.CP)| \tag{2.8}$$

Assume that, at all other genomic locations, $|M - A| = 0$, for some small value $\epsilon_1 \geq 0$, the following conditions must hold

$$|\frac{oe.e.ACN}{2} - 1| \times |f_{oe} - oe.CP| < \epsilon_1$$
$$|\frac{oe'.e.ACN}{2} - 1| \times |f_{oe'} - oe'.CP| < \epsilon_1$$

in order for C to be a solution of O, with $\epsilon_{fit} > (oe.e.S.\text{length} + oe'.e.S.\text{length}) \times \epsilon_1$. Because $\forall oe : oe.e.ACN \in N \wedge oe.e.ACN \neq 2$ (Definition 2.7), $|\frac{oe.e.ACN}{2} - 1| \geq 0.5$, the above can be rewritten, with $\epsilon' = 2\epsilon_1$, as

$$|f_{oe} - oe.CP| < \epsilon'$$

$$|f_{oe'} - oe'.CP| < \epsilon'$$

combined with the fact that $|oe.CP - oe'.CP| < \epsilon_P$, under the worst case scenario that $f_{oe} > oe.CP$, $oe.CP > oe'.CP$, and $f_{oe'} < oe'.CP$, with $\epsilon_1 > \epsilon_P$, we have

$$\begin{aligned} f_{oe} - oe.CP &< \epsilon' \\ oe'.CP - f_{oe'} &< \epsilon' \\ oe.CP - oe'.CP &< \epsilon' \end{aligned}$$

when added together, we have

$$f_{oe} - f_{oe'} < 3\epsilon'$$

Recall that

$$f_{oe} = \sum_{k=j}^{m} C^k.f$$

$$f_{oe'} = \sum_{k=j'}^{m} C^k.f$$

$$0 < j < j' \leq m$$

We have

$$\sum_{k=j}^{j'-1} C^k.f < 3\epsilon'$$

However, this cannot be by Definition 2.13, when $\epsilon_f \geq \frac{3}{j'-j}\epsilon'$.

□

Theorem 2.4. *For any parsimonious linear subclone profile (Definition 2.13) C with $m+1$ subclones that is a solution to an observation O with n observed events being partitioned as P with $n' \leq n$ clusters, $m \geq n'$*

Proof. Assuming that there exists one subclone profile C with $m < n'$ that is a solution to O, due to pigeonhole principle, $\exists p \in P, p' \in P, \exists j \leq m, \exists oe \in p, oe' \in p'$, that

$$oe.e \notin C^{j-1}.G \wedge oe'.e \notin C^{j-1}.G$$

$$oe.e \in C^{j}.G \wedge oe'.e \in C^{j}.G$$

Similar to the proof of Theorem 2.3, if we let

$$f_{oe} = \sum_{k=j}^{m} C^k.f$$

$$f_{oe'} = f_{oe}$$

we have

$$\forall L_l \in oe.e.S : |M_l - A_l|$$
$$= |(\frac{oe.e.ACN}{2} - 1) \times (f_{oe} - oe.CP)| \tag{2.9}$$
$$\forall L_{l'} \in oe'.e.S : |M_{l'} - A_{l'}|$$
$$= |(\frac{oe'.e.ACN}{2} - 1) \times (f_{oe'} - oe'.CP)|$$
$$= |(\frac{oe'.e.ACN}{2} - 1) \times (f_{oe} - oe'.CP)| \tag{2.10}$$

Assume that, at all other genomic locations, $|M - A| = 0$, for some small value $\epsilon_2 \geq 0$, the following conditions must hold

$$|\frac{oe.e.ACN}{2} - 1| \times |f_{oe} - oe.CP| < \epsilon_2$$
$$|\frac{oe'.e.ACN}{2} - 1| \times |f_{oe} - oe'.CP| < \epsilon_2$$

in order for C to be a solution of O, with $\epsilon_{fit} > (oe.e.S.\text{length} + oe'.e.S.\text{length}) \times \epsilon_2$. Because $\forall oe : oe.e.ACN \in N \wedge oe.e.ACN \neq 2$ (Definition 2.7), $|\frac{oe.e.ACN}{2} - 1| \geq 0.5$, the above can be rewritten, with $\epsilon' = 2\epsilon_2$, as

$$|f_{oe} - oe.CP| < \epsilon'$$
$$|f_{oe} - oe'.CP| < \epsilon'$$

under the worst case scenario that $f_{oe} > oe.CP$, $f_{oe} < oe'.CP$, with $\epsilon_2 < \frac{1}{4}\epsilon_P$, we have

$$f_{oe} - oe.CP \quad < \epsilon'$$
$$oe'.CP - f_{oe} \quad < \epsilon'$$

when added together, we have

$$oe'.CP - oe.CP < 2\epsilon' < 4\epsilon_2 < \epsilon_P$$

However this contradicts with Definition 2.15, that

$$\forall p \in P, p' \in P, p \neq p' : (\forall oe \in p, oe' \in p' : |oe.CP - oe'.CP| > \epsilon_P)$$

□

Corollary 2.2. *If there exists a parsimonious, linear subclone profile C, which is also a solution to an observation O with a clustering partition P having n' clusters, there are exactly $n'+1$ subclones in C (Theorem 2.3 and Theorem 2.4) with the following relationships among the error margins, for some constant $C_1 > 0, C_2 > 0, C_3 > 0$ and some small value $\epsilon_1 \geq 0, \epsilon_2 \geq 0$*

$$C_1\epsilon_{fit} > \epsilon_1$$
$$C_2\epsilon_{fit} > \epsilon_2$$
$$C_3\epsilon_f > \epsilon_1$$
$$\epsilon_1 > \epsilon_p > 4\epsilon_2$$

Theorem 2.5. *A parsimonious, linear subclone profile C with exactly $n'+1$ subclones that is a solution to an observation O with a clustering partition P having n' clusters always*

exists, and can be constructed as the following

$$\forall j \leq n' : C^j.G = \begin{cases} \{\} & j = 0 \\ \{oe.e | oe \in \bigcup_{i=1}^{j} p_i\} & otherwise \end{cases} \tag{2.11}$$

$$\forall j \leq n' : C^j.f = \begin{cases} 1 - p_1.CP & j = 0 \\ p_j.CP - p_{j+1}.CP & 0 < j < n' \\ p_j.CP & j = n' \end{cases} \tag{2.12}$$

We then prove that the subclone profile C is linear, parsimonious, and a solution to O.

Proof. C is linear

$$\begin{aligned} &\forall j < n' : C^j.G = \{oe.e | oe \in \bigcup_{i=1}^{j} p_i\}, C^{j+1}.G = \{oe.e | oe \in \bigcup_{i=1}^{j+1} p_i\} \\ \implies & C^{j+1}.G - C^j.G = \{oe.e | oe \in p_{i+1}\} \\ & \because P \text{ is a partition over } O \\ & \therefore p_{i+1} \neq \emptyset \\ \implies & C^j.G \subset C^{j+1}.G \\ \implies & C^j.G \subseteq C^{j+1}.G \end{aligned} \tag{2.13}$$

□

Proof. C is parsimonious

$\forall j < n' : C^j.G \subset C^{j+1}.G$ has already been proven by Equation 2.13.

When $j = n'$, $C^j.f = p_j.CP > \epsilon_d$ (Definition 2.6);

When $0 < j < n'$,

$$\begin{aligned} C^j.f &= p_j.CP - p_{j+1}.CP \\ &> \epsilon_p \text{ (Definition 2.15)} \end{aligned}$$

Thus $\forall j \in [1, n'] : C^j.f > \epsilon_f$ as long as $\epsilon_f < \min(\epsilon_d, \epsilon_p)$ □

Proof. C is a solution to the observation O.

$\forall L_l$ over the entire genome, one of the two following things can happen

- $\forall oe \in O : L_l \notin oe.e.S$ (case 1)
- $\exists oe \in O : L_l \in oe.e.S$ (case 2)

In case 1, $M_l = \sum_{j=0}^{n'} C^j.f = 1$ (Definition 2.9); $A_l = 1$ (Equation 2.4); $|M_l - A_l| = 0 \leq \epsilon_{fit}$ for any $\epsilon \leq 0$.

In case 2, $\exists! j \leq n' : oe \in p_j \implies oe.e \in C^j.G, C^{j+1}.G, \ldots, C^{n'}.G$.

$$\begin{aligned} M_l &= \sum_{k=0}^{j-1} 1 \times C^k.f + \sum_{k=j}^{n'} \frac{oe.e.ACN}{2} \times C^k.f \\ &= (1 - \sum_{k=j}^{n'} C^k.f) + \frac{oe.e.ACN}{2} \times \sum_{k=j}^{n'} C^k.f \\ A_l &= \frac{oe.e.ACN}{2} \times oe.CP + (1 - oe.CP) \\ &= (1 - oe.CP) + \frac{oe.e.ACN}{2} \times oe.CP \end{aligned}$$

Consequently

$$|M_l - A_l| = |(\frac{oe.e.ACN}{2} - 1) \times [\sum_{k=j}^{n'} C^k.f - oe.CP]|$$
$$= |(\frac{oe.e.ACN}{2} - 1)| \times |\sum_{k=j}^{n'} C^k.f - oe.CP|$$

Because

$$\begin{aligned}\sum_{k=j}^{n'} C^k.f =& (p_j.CP - p_{j+1}.CP) \\ &+ (p_{j+1}.CP - p_{j+2}.CP) \\ &+ \cdots \\ &+ (p_{n'-1}.CP - p_{n'}.CP) + p_{n'}.CP \\ =& p_k.CP\end{aligned}$$

We have

$$|M_l - A_l| = |(\frac{oe.e.ACN}{2} - 1)| \times |p_j.CP - oe.CP|$$

Combine case 1 and case 2 together, over the entire genome, we have

$$
\begin{aligned}
fit &= \sum_{l}^{\text{genome length}} |M_l - A_l| \\
&= \sum_{i=1}^{n} [oe_i.e.S.\text{length} \times |(\frac{oe_i.e.ACN}{2} - 1)| \times |p_{i'}.CP - oe_i.CP|] \text{ (in which } oe_i \in p_{i'}) \\
&\leq \sum_{i=1}^{n} [oe_i.e.S.\text{length} \times |(\frac{oe_i.e.ACN}{2} - 1)| \times \epsilon_3]
\end{aligned}
$$

in which $\epsilon_3 \geq 0$ is the largest difference between any cluster centroid and the CP value of their member observed events, or, formally

$$\forall p \in P : \forall oe \in p : \epsilon_3 \geq |p.CP - oe.CP|$$

$$\therefore \text{for any } \epsilon_{fit} \geq \sum_{i=1}^{n} [oe_i.e.S.\text{length} \times |(\frac{oe_i.e.ACN}{2} - 1)| \times \epsilon_3], \text{ we have } fit \leq \epsilon_{fit}. \quad \square$$

The method, which is outlined in Listing 2.1, implements Equation 2.11 and Equation 2.12 described in Theorem 2.5, albeit starting with the event cluster that has the lowest cluster centroid. It modifies P and O so that the SF of the newly introduced subclone is always determined by the third case in Equation 2.12. A toy example is shown in Figure 2.2.1, in which those nodes that have a genotype and frequency label besides them are the subclones in the model.

2.3 Analysis of Ovarian Serious Carcinoma copy number variation dataset

2.3.1 Introduction

High grade serous ovarian cancer is a highly aggressive and lethal disease. While most patients achieve an initial clinical remission, approximately 80% of patients recur within a five year period. Continuing advancements in sequencing technologies allow stud-

```
Input: P = P[1], P[2], ..., P[n']
Input: O = O[1], O[2], ..., O[n]

Initialize C with one subclone, C[0]
C[0].f = 1; C[0].G = {}

for i = n' .. 1 :
    newC.f = P[i].CP
    newC.G = { oe.e for oe in O }
    C[0].f -= newC.f
    insert newC after C[0]

    // update P and O due to the introduction of newC
    P = P - {P[i]}
    n'--
    for j = 1 .. n' :
        P[j].CP -= newC.f
    end-for

    O = O - {oe for oe in P[i]}
    n = n - size(P[i])
    for j = 1 .. n :
        O[i].CP -= newC.f
    end-for
end-for
```

Listing 2.1: Pseudo code of the linear model algorithm.

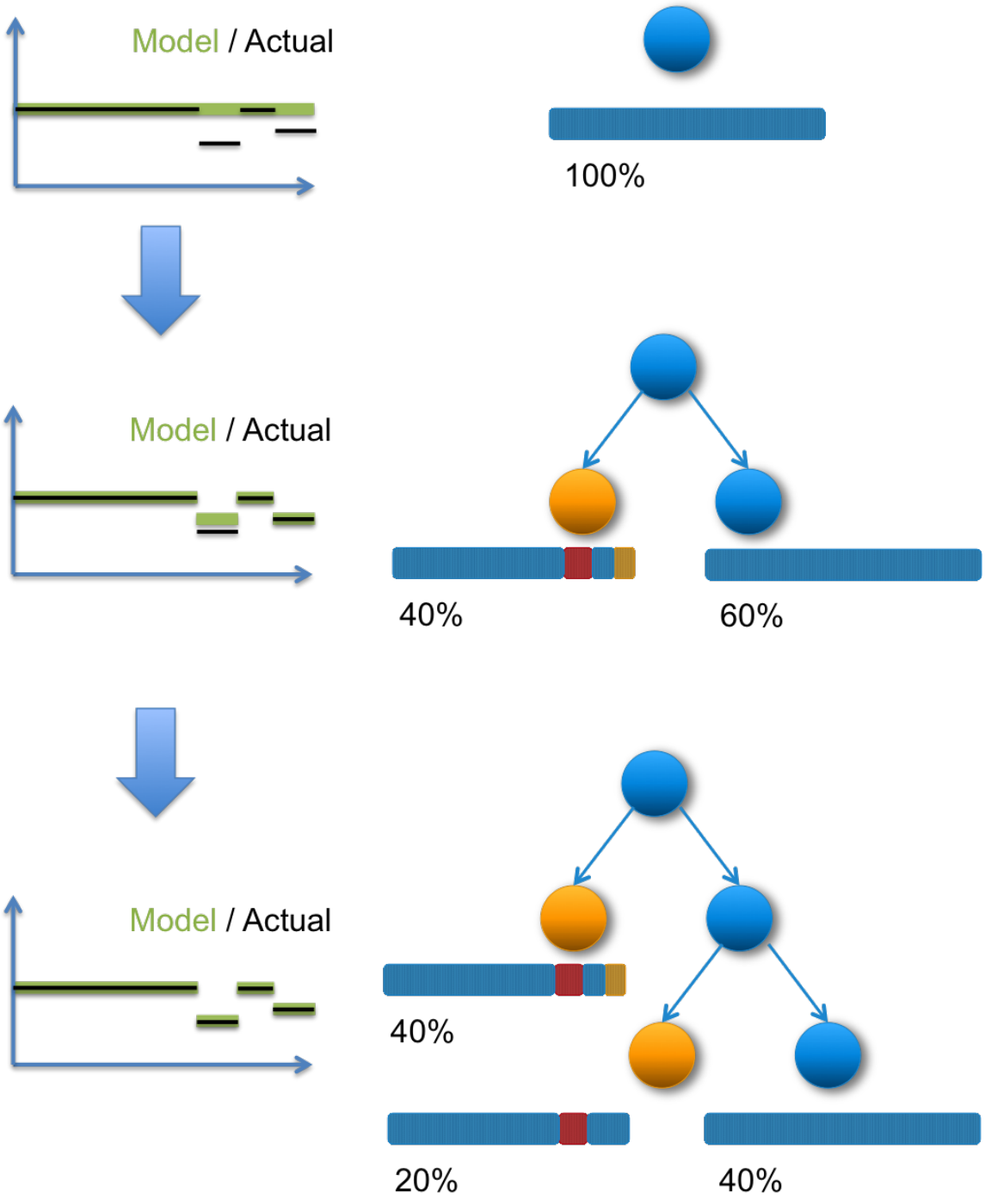


Figure 2.2.1: Illustration of the parsimonious method to reconstruct the subclone structure with a linear heritage model. New subclones are introduced into the model stepwise, each time explaining the least prevalent events completely, until the observation (actual) can be entirely explained by the model.

ies of the tumor genome in ever increasing detail. Improving knowledge on intratumoral heterogeneity and the identification of the clonal structure of tumor samples, by sequencing of single cells or sequencing of tumors at very high coverage levels, may provide important new insights into mechanisms of tumor evolution and progression [89, 121, 136, 137].

To gain insight into the mechanisms used by tumor cells to evade the cytotoxic effects of chemotherapy, and taking advantage of the highly standardized nature of treatment regimens for ovarian serous carcinoma, we participated in a joint effort to characterize the genomes of the primary and relapse tumors of seventeen patients with ovarian carcinoma, using a combination of whole genome sequencing, whole exome sequencing, DNA copy number, methylation and gene expression profile. We applied the method described in Section 2.2 in an attempt to provide insights into the mechanism of tumor relapse in ovarian cancer.

Please refer to Table 4.0.1 for a summary of input data types and major conclusions.

2.3.2 Results

2.3.2.1 Clonal structure reconstruction

First, we used the precise and linear measurements of copy number level established by whole genome sequencing to reconstruct each of the ten whole gnome samples into a set of subclones. Next we assessed the distance between each primary and matching relapse subclone to derive a model of tumor progression in each patient.

To establish the clonal structure of a tumor sample, we removed all loci of copy number gain, and kept only segments with equal to or less than the modal copy number. Next, all remaining copy number segments were clustered into discrete levels, with the requirement for each cluster to contain DNA segments covering at least 10 mega-bases. In a simplified model of a diploid[4] tumor genome, three possible ACN categories exist

[4]Having two copies of the complete genome.

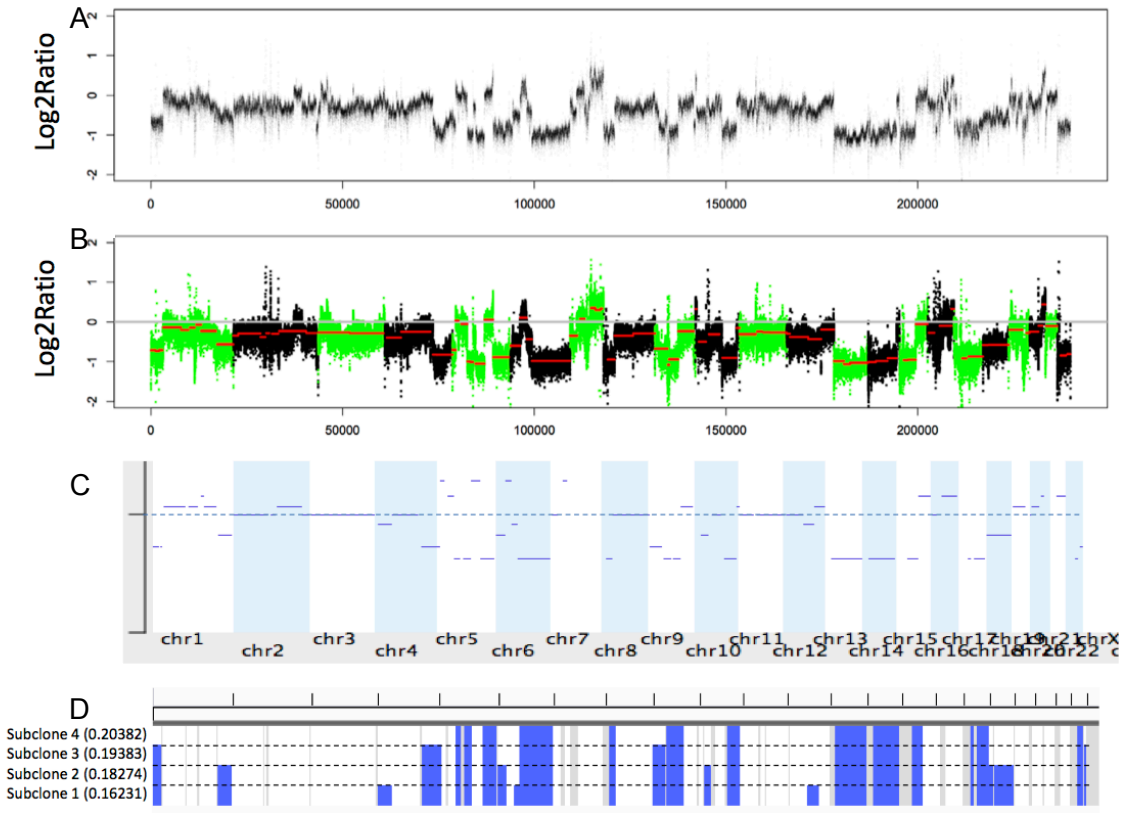


Figure 2.3.1: Inferring subclones through clustering of copy number levels. A) Whole genome copy number profile is quantified by calculating, for each 10kb non-overlapping moving window, the log 2 ratio of the number of reads initiating within a specific window (RD) in the tumor sample relative to the read depth in the paired-normal sample. B) The log 2 ratio is subjected to circular binary segmentation [138] for identifying continuous regions with the same underlying copy number state. C) The identified segments are clustered to find discrete levels. D) A parsimonious subclone structure is generated with a biologically motivated model that late subclones inherited the mutations existed in earlier subclones. Blue regions represent heterozygous deletion.

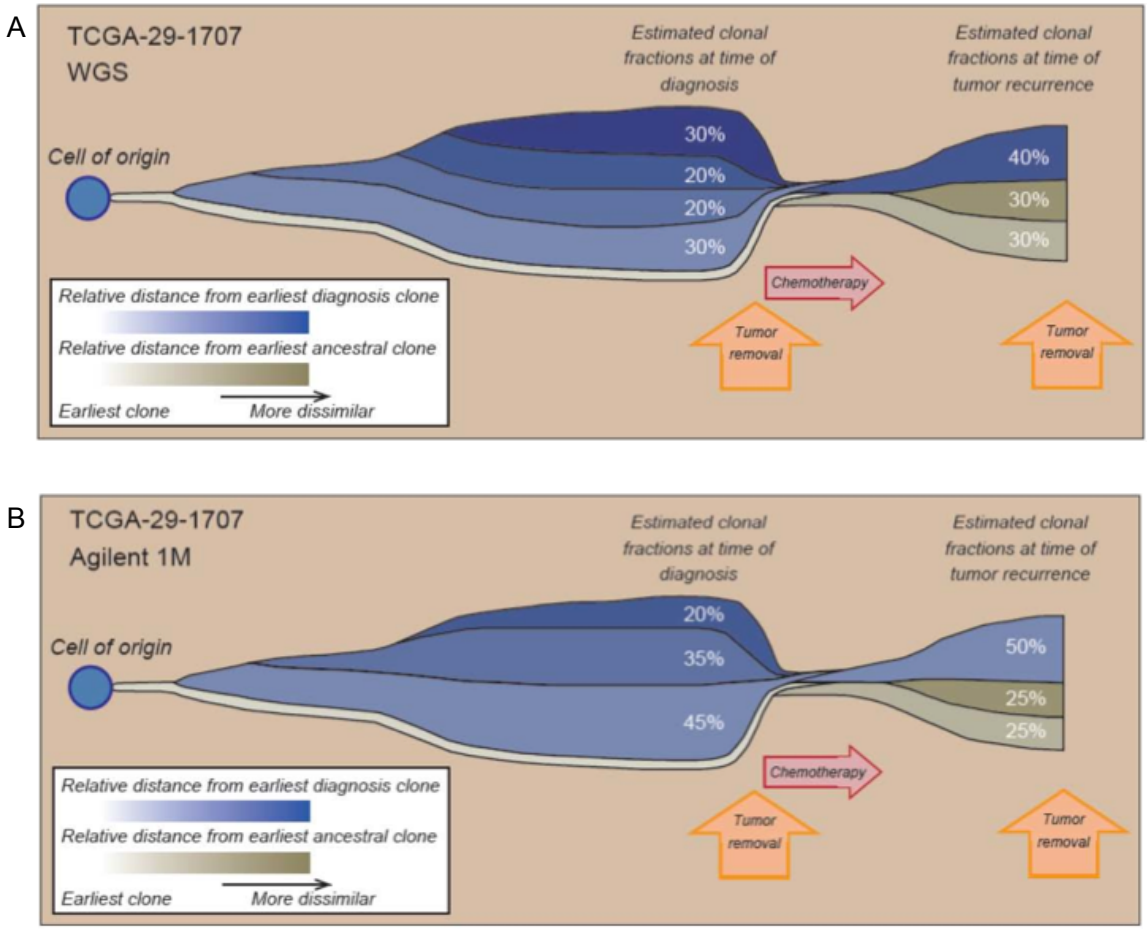


Figure 2.3.2: Schematic representation of clonal evolution in a single primary-relapse tumor pair. A) The development of primary and relapse tumor from the cell of origin is depicted over time. Using copy number levels inferred from whole genome sequencing data, four subclonal populations were predicted to be present at time of diagnosis. One of the subclones was found in the relapse tumor, combined with two new clones that were derived from an ancestral tumor cell population that was too small to be detected in the primary tumor. The percentages (in white) reflect the fraction of tumor cells for each subclone. B) Using copy number levels obtained from the Agilent[3] 1M platform, three subclones were found in the primary tumor. One subclonal population was found again in the relapse tumor, while two novel subclones developed from ancestral cells. The percentages (in white) reflect the fraction of tumor cells for each subclone.

after removal of all chromosomal regions of copy number gain: 1. Diploid / wild-type; 2. Heterozygous loss; 3. Homozygous loss. Any levels deviating from this model were likely resulted from the presence of copy number alterations in subclonal tumor cell populations. We obtained tumor purity and tumor ploidy of all samples using ASCAT [134] with microarray CNV data and default parameters. Using the whole genome sequencing derived segmented copy number data from the primary tumor sample of patient TCGA-29-1707, we inferred the presence of five copy number level clusters, which are explained by four tumor subclones and one normal tissue component. Similarly, we found three clones in the TCGA-29-1707 relapse tumor (Figure 2.3.1, Figure 2.3.2 A). Next, a distance matrix of all possible primary-relapse clone pairs was generated, using Pearson correlation as a distance metric, to establish the evolutionary trajectory from primary tumor to regrowth (data not shown). Genome correlation was assessed by comparing absolute levels of gains and losses across the entire genome. In the example TCGA 29-1707, a subset of copy number levels and alterations were found in all clones suggesting that all primary and relapse clones were derived from a common ancestor cell. Two clones in the relapse sample lacked copy number changes found in all primary clones, and must therefore originate from an ancestral tumor cell that was present in the primary tumor but at levels that we were unable to detect. The third relapse subclone harbored alterations also found in the first primary clone, but not other primary clones, and we therefore suggest that the first primary subclone is the founder population that gave rise to the third relapse subclone.

Using whole genome sequencing of primary and relapse tumor pairs, the tumor progression structure of two additional ovarian carcinomas was reconstructed (Figure 2.3.3). Of note, a substantial fraction of ovarian cancer is thought to harbor highly aneuploid[5] genomes, resulting in a modal number of chromosomes of three or higher [134, 135]. We predicted tumor ploidy for each sample using ASCAT [134] and ac-

[5]Having a ploidy that is other than 2

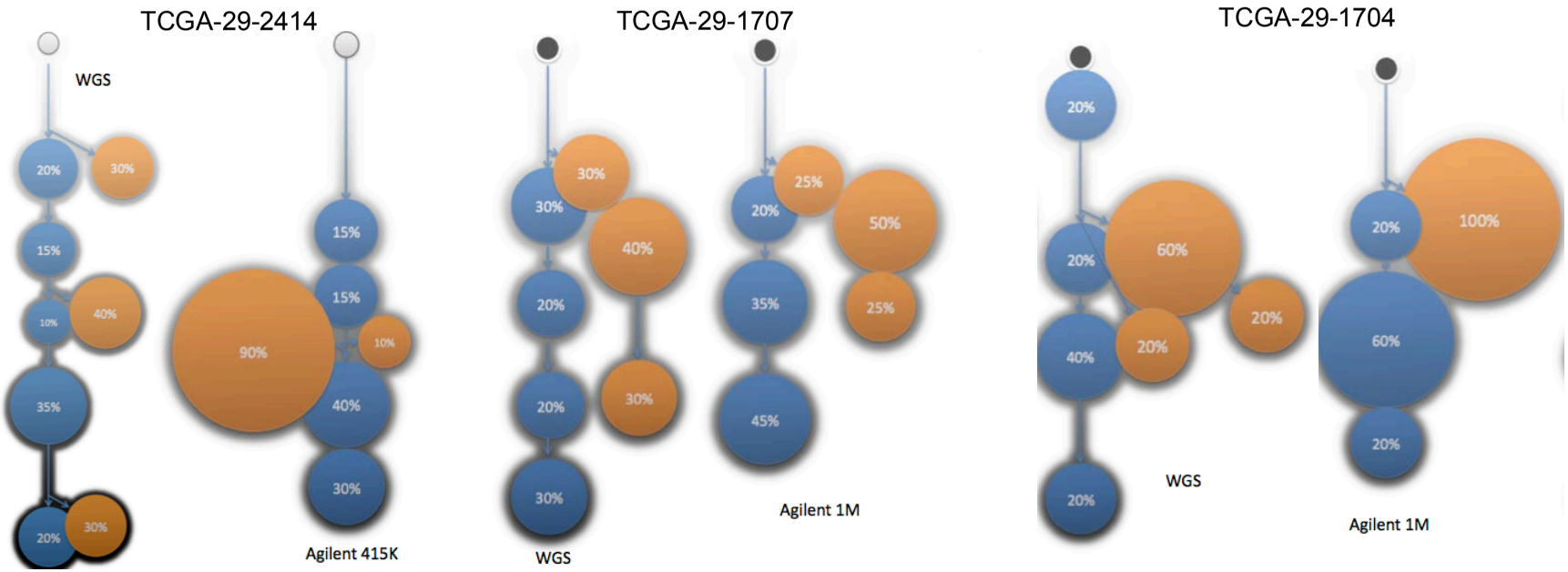


Figure 2.3.3: Clonal evolution of ovarian carcinoma, inferred using whole genome sequencing and array based copy number data. Bubble plots showing pattern of tumor progression and relapse as inferred using whole genome sequencing data and Agilent array CNV data, for three matching primary-relapse pairs. Each bubble, indicating a subclone, is drawn relative to its cellular fraction and in the relative temporal order they were derived. Primary tumor subclones are displayed in blue, relapse subclones are shown in orange. A black root node represent that the tumors are diploid, and a white root node aneuploid.

counted for triploid[6] genomes by allowing the cluster structure to contain three possible chromosome levels. Two cases for which whole genome sequencing data were available were excluded from the analysis, due to complicating factors. Sample TCGA-24-2852 was predicted by ASCAT [134] to harbor tetraploid[7] genomes in both tumors, whereas a difference in ploidy between primary and relapse tumor was predicted for TCGA-61-1916.

As whole genome sequencing data was available on five of seventeen triplets[8], but Agilent 1M array based copy number levels were generated for all triplets, we compared the results of clonal structure reconstruction of array based clonal subsets to whole genome sequencing based tumor clones. A similar pattern of primary and relapse subclones were observed, with the primary subclone 3 and 4 identified using whole genome sequencing data, being grouped into a single clone 3 when evaluating array based copy numbers (Figure 2.3.2 B, Figure 2.3.3). We thus concluded that the DNA copy number profiles obtained using arrays are able to provide reliable subclone reconstruction, but with lower granularity than whole genome sequencing.

2.3.2.2 Clonal evolution patterns associate with clinical response

Using the Agilent array data, we generated clonal structures and tumor progression models for thirteen of seventeen triplets. At least two different subclones, representing at least 10% of the tumor cell population, were detected in eleven of thirteen primary tumors. Similarly, we found that eleven of thirteen relapse tumors consisted of at least two subclones (data not shown). The number of subclonal populations in the primary tumor was not predictive of the number of clones identified in the relapse sample. Interestingly, subclones in all thirteen relapse tumors were found to have evolved from ancestral cells from which the primary tumor had also been derived.

[6] Having three copies of the complete genome.
[7] Having four copies of the complete genome.
[8] A set of samples consisting of the normal, primary tumor and relapse tumor biopsies.

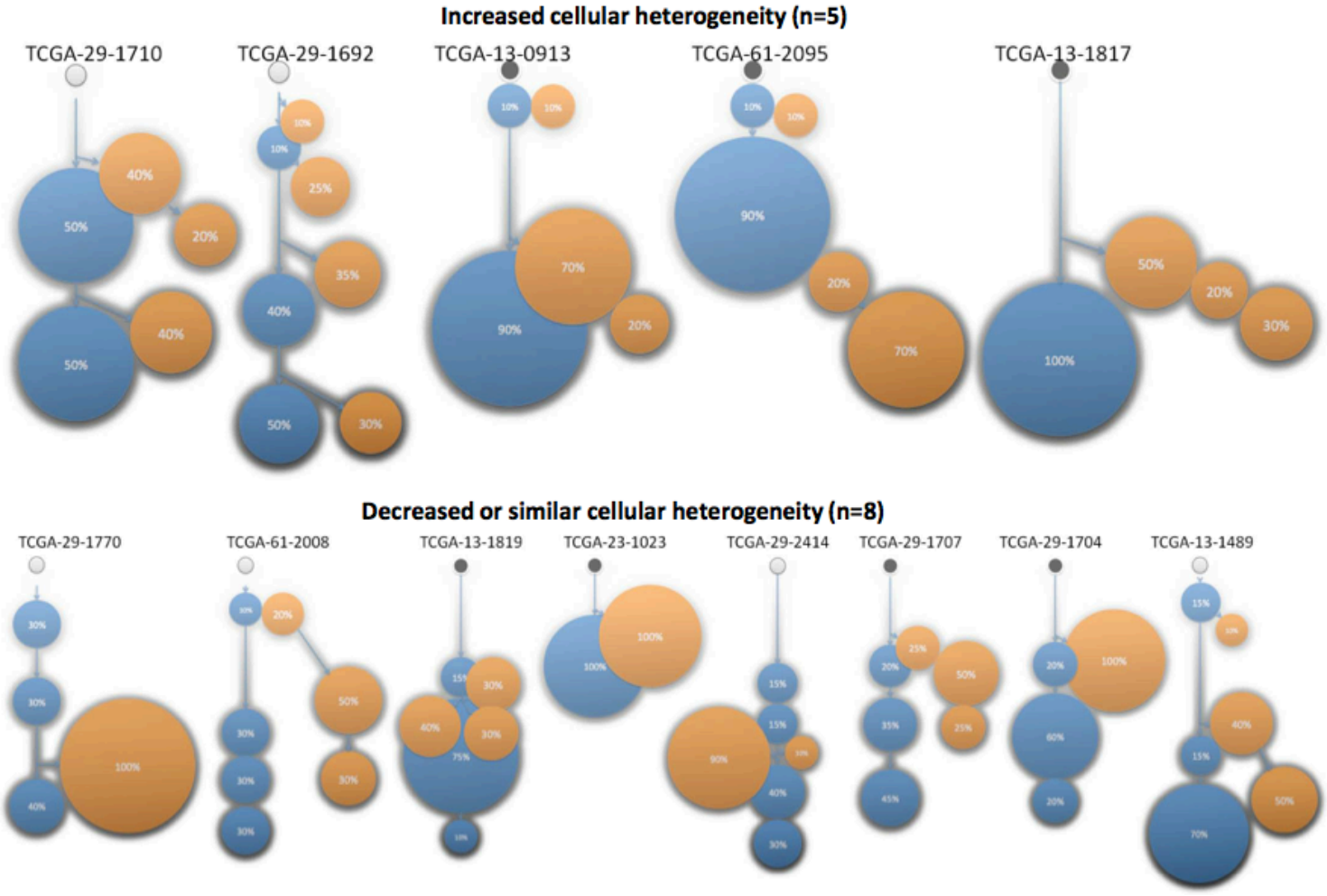


Figure 2.3.4: Comparative analysis of ovarian carcinoma tumor evolution patterns. The same bubble plots as shown in Figure 2.3.3 are included, but grouped by samples with increased tumor heterogeneity and similar/decreased tumor heterogeneity.

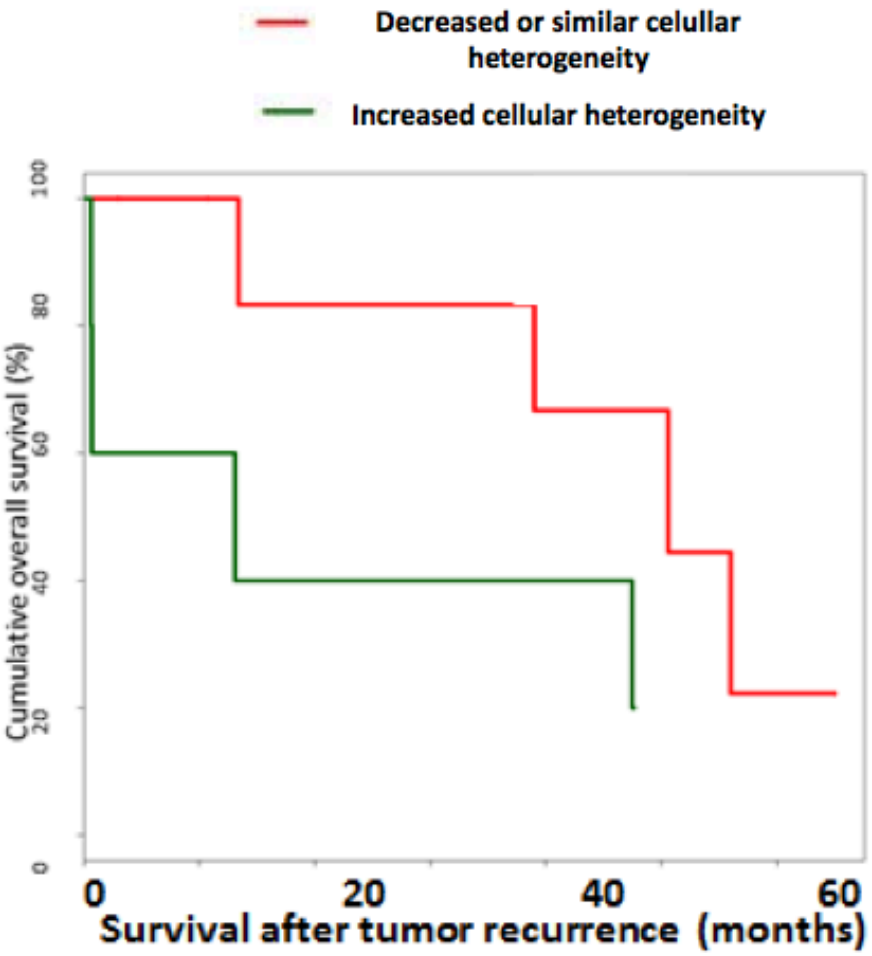


Figure 2.3.5: Patient survival after second surgery. Kaplan-Meier curve of survival from the time of second surgery.

The number of clonal populations may be indicative of the level of intratumoral heterogeneity that is present in a specific tumor sample. We did not observe a difference in the number of subclones between tumors that relapsed within 24 months after surgery ($n = 5$) and tumors that relapsed later than 24 months post surgery ($n = 8$). However, we noticed that four out of five tumor pairs displaying an increase in the number of subclonal populations resulted in a relapse tumor that was subsequently resistant to platinum therapy[9], whereas all eight tumors in which the number of subclones was similar or less than the number observed in the primary tumors responded to subsequent platinum administration (Figure 2.3.4). As a result, the overall survival after the last surgery between these two groups trended towards statistical significance, despite the small number of patients included in the analysis (p-value $= 0.09$, $n = 13$, Figure 2.3.5).

2.3.2.3 Discussion

We performed our parsimonious subclone structure reconstruction on a set of seventeen matching primary and relapse ovarian carcinomas and showed that intratumoral heterogeneity plays in important roles in this disease. Through an approach of clustering copy number segments into bins of similar magnitude, we identified multiple subclones in all primary and relapse samples. Comparative analysis of the number of relative proportion of clonal subpopulations in primary and relapse tumor samples enabled us to construct hierarchical trees of tumor progression in thirteen of seventeen cases. Interestingly, all relapse tumors that showed an increase in clonal complexity relative to the primary tumor were found to be resistant to chemotherapy, regardless of the number of subclonal populations found in the primary tumor.

The implications of this findings for treatment may be multi-fold. In an era with an increased interest in individualized therapies, our results suggest that the adequate

[9]Chemotherapy with cisplatin as the anti-neoplastic reagent

choice of a therapeutic modality should depend on the molecular profiling of each tumor sample separately, and the relative level of genomic complexity pre- and post-treatment. Importantly, precision medicine may only be able to be curative when targeting genomic abnormalities found in all tumor subclones. We showed that as a result of surgical resection and cytotoxic chemotherapy, some subclones are lost whereas others remain and result in tumor relapse. We demonstrated that each tumor follows a unique path of disease relapse and tumor progression. These observations confirm previous results obtained in pediatric acute lymphoblastic leukemia and acute myeloid leukemia (AML) [71, 139]. Our results extend previous studies of pre- and post-treatment ovarian carcinomas which showed intra- and inter-patient genomic diversity but that were limited by small sample sizes and single genomic platforms [140, 141].

2.4 Analysis of Intracranal Germ Cell Tumor loss of heterozygosity dataset

2.4.1 Introduction

Intracranal Germ Cell Tumors (IGCT) are a group of rare heterogeneous brain tumors that are clinically and histologically similar to the more common gonadal GCTs. IGCTs show great variations in their geographical and gender distribution, histological composition and treatment outcomes [142–148]. We have participated in an in-depth analysis of the genetic abnormalities of IGCTs through the collaboration with the Wheeler's group (Baylor College of Medicine, Huston, TX 77030), performed subclonal structure reconstruction based on genome-wide SNP array probe intensity data. Here we mostly focus on the aspect of subclone analysis. Please refer to Appendix Section A.2.2 for methods regarding data acquisition, and Linghua *et al.* Nature (2014) [149] for a detailed report.

This experiment procedure resulted in two data tracks, B allele frequency (BAF),

which describes the amount of heterozygosity, and probe intensity $\log_2 R$ ratio (LRR), which describes the total amount of DNA, at each probed location. For germline alleles that are present in all cells, BAF can either be 0, if the two copies of the allele are all reference, or homozygous reference; 0.5, if one of the two copies of the allele is different from the reference, or heterozygous; and 1.0, if the two copies of the alleles are all different from the reference, or homozygous alternate. The fact that BAF segments that are of other values than these three have been observed suggests that the tumor samples contain subclones. Through processes such as balanced loss-of-heterozygosity (LOH), in which a cell lose a segment of its chromosome, and then later repaired by copying the corresponding region from its homologous chromosome, such region could appear to be all either homozygous reference or alternate, effectively shifting the BAF segment away from 0.5 if such events are not present in all cells. We took advantage of this signal to construct the parsimonious subclonal structure.

Please refer to Table 4.0.1 for a summary of input data types and major conclusions.

2.4.2 METHODS

The fact that Equation 2.11 and Equation 2.12 do not care about data type specific properties, such as $e.ACN$, makes the method expandable to other data types as well. Here we give an expansion to LOH events based on microarray BAF data.

Definition 2.16. *BAF is defined as*

$$BAF = \frac{\textit{Non-Reference Allele Count}}{\textit{Total Allele Count}}$$

Lemma 2.2. *For a heterozygous region in a diploid genome, $E\{BAF\} = 0.5$; for a homozygous region in a diploid genome, $E\{BAF\} = 0$ or $E\{BAF\} = 1$.*

Proof. In a heterozygous region,

$$
\begin{aligned}
E\{\text{(Non-Reference Allele Count)}\} &= 0.5\text{(Total Allele Count)} \\
&\Longrightarrow \\
E\{\text{BAF}\} &= 0.5
\end{aligned}
$$

In a homozygous region,

$$
\begin{aligned}
E\{\text{(Non-Reference Allele Count)}\} &= 0 \\
&\text{or} \\
E\{\text{(Non-Reference Allele Count)}\} &= \text{(Total Allele Count)} \\
&\Longrightarrow \\
\text{BAF} &= 0 \\
&\text{or} \\
\text{BAF} &= 1
\end{aligned}
$$

□

Definition 2.17. *mirrored B allele frequency (mBAF) is defined as*

$$
mBAF = \begin{cases} BAF & BAF \geq 0.5 \\ 1 - BAF & BAF < 0.5 \end{cases} \tag{2.14}
$$

Corollary 2.3. *For a heterozygous region in a diploid genome,* $E\{mBAF\} = 0.5$*; for a homozygous region in a diploid genome,* $E\{mBAF\} = 1$*. (Lemma 2.2 and Definition 2.17)*

Definition 2.18. *A segmental somatic LOH event, e^{LOH}, is defined as $e^{LOH} = \{S\}$ for a segment on the genome, specified by S, whose heterozygosity has been lost (mBAF = 1).*

Definition 2.19. *An observed somatic LOH event, oe^{LOH}, is defined as $oe^{LOH} = \{e^{LOH}, CP\}$ for a LOH event e^{LOH} observed in $CP > \epsilon_d$ fraction of the total cells, with some detection sensitivity $\epsilon_d > 0$.*

Definition 2.20. *An observation, O^{LOH}, is defined as $O^{LOH} = \{oe_1^{LOH}, oe_2^{LOH}, \ldots, oe_n^{LOH}\}$ for a segmented mBAF profile with n segments of mBAF $\neq 0.5$. Observed somatic LOH events are identified with genomic segmentation algorithms, which will result in non-overlapping segments. The CP for any oe^{LOH}, with a segmental mBAF mean (u, output of segmentation algorithm), can be calculated by the following equation*

$$\begin{aligned} \because 1 \cdot CP + 0.5 \cdot (1 - CP) &= u \\ \therefore CP &= 2u - 1 \end{aligned} \tag{2.15}$$

Definition 2.8 and Definition 2.9 remain the same, thus Lemma 2.1 and Corollary 2.1 remain true.

Definition 2.21. *An actual (observed) genomic profile, A^{LOH}, is defined as $A^{LOH} = \{A_l^{LOH}\}, l = 1, 2, \ldots$ for each unique location L_l on the genome. Since at most 1 oe^{LOH} exists so that $L_l \in oe^{LOH}.e^{LOH}.S$ (Lemma 2.1), A_l^{LOH} is calculated as following*

$$A_l^{LOH} = \begin{cases} 1 \times oe^{LOH}.CP + 0.5 \times (1 - oe^{LOH}.CP) & \exists! oe^{LOH} \in O^{LOH} : L_l \in oe^{LOH}.e^{LOH}.S \\ 0.5 & otherwise \end{cases}$$

which can be simplified as

$$A_l^{LOH} = \begin{cases} 0.5(1 - oe^{LOH}.CP) & \exists! oe^{LOH} \in O^{LOH} : L_l \in oe^{LOH}.e^{LOH}.S \\ 1 & otherwise \end{cases} \tag{2.16}$$

Definition 2.22. *A model genomic profile, M^{LOH}, is defined as $M^{LOH} = \{M_l^{LOH}\}, l = 1, 2, \ldots$ for each unique location L_l on the genome. Since for any given subclone C^j, at most 1 e^{LOH} exists for $C^j.G$ so that $L_l \in e^{LOH}.S$ (Corollary 2.1), M_l^{LOH} is calculated as following*

$$M_l = \sum_{j=0}^{m} \begin{cases} 1 \times C^j.f & \exists! e^{LOH} \in C^j.G : L_l \in e^{LOH}.S \\ 0.5 \times C^j.f & otherwise \end{cases} \tag{2.17}$$

All other definitions remain the same as the case for CNV events, and theorems, lemmas, and corollaries can be proven in similar fashions. A subclone profile can thus be constructed, using the same method as described by Equation 2.11, Equation 2.12, and Listing 2.1.

First, the whole genomic BAF data of a tumor sample was filtered to exclude those locations that were identified as "homozygous" in the paired-normal sample to generate somatic LOH event profile (Figure 2.4.1 A), and from it a mBAF (Figure 2.4.1 B) [150] profile was calculated by the following rules:

The mBAF profile was then subjected to segmentation with Circular Binary Segmentation algorithm [151] (Figure 2.4.1 C, D), and the CP values are calculated with Equation 2.15 (Figure 2.4.1 E).

Next, CP values were clustered to further reduce noise by assigning each cluster a centroid value (Definition 2.15). A subclone profile was then constructed (Figure 2.4.1 F) according to the parsimonious model described in Section 2.2.

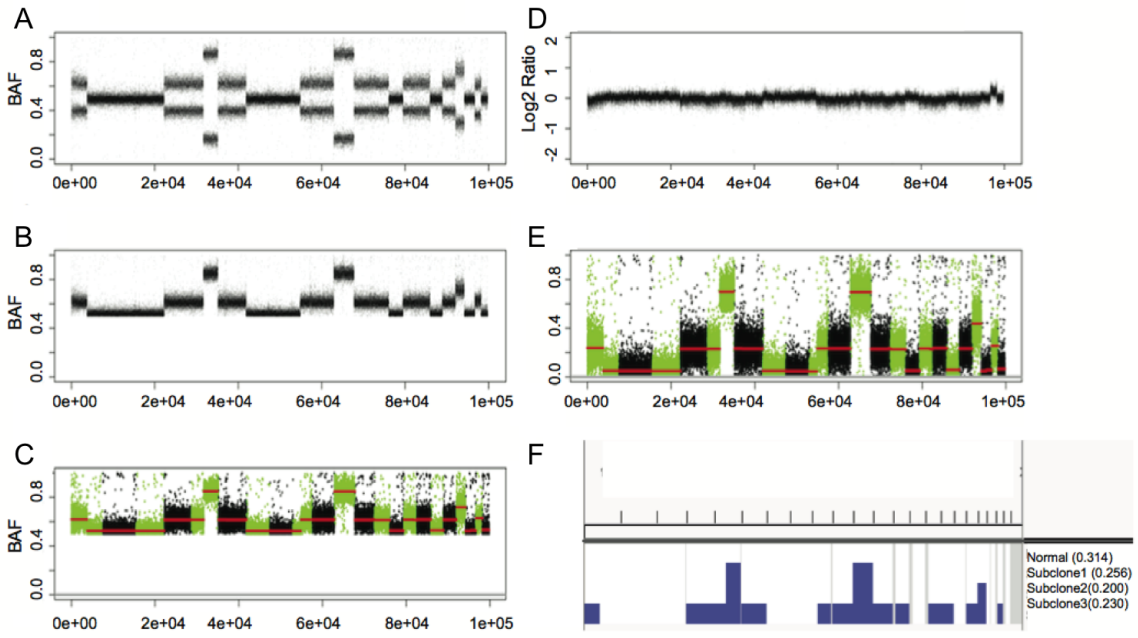


Figure 2.4.1: Clonality analysis of a representative case (N10, an immature teratoma[10]). A) The somatic BAF. The BAF data is filtered to only retain those that are heterozygous in its normal sample. B) The mBAF. The mBAF data is acquired by mapping all BAF data points smaller than 0.5 to 1 - BAF. C) The segmented mBAF frequency. The mBAF is then subjected to circular binary segmentation so that continuous segments of LOH can be identified. D) The copy number probe log 2 ratio track of the microarray is shown to illustrate that there is no observable copy number alteration that is correlating with the observed LOH pattern, indicating that the multi-level LOH is a result of multi-clonality. E) The segmented mBAF values are converted to CP. CP represents, for any given LOH event, what is the fraction of cells that are harboring the event, out of the entire cell population measured. F) Utilizing the CP, a subclone profile is constructed according to a linear heritage model, in which more prevalent events are present in earlier clones, on which less frequent events are accrued.

2.4.3 RESULTS

The analysis revealed that 71% of all the investigated IGCT genomes are subclonal. It also verified that SNP array BAF track can be a viable source for subclonal analysis. More details about this study can be found in Linghua *et al.* Nature (2014) [149].

2.5 CONCLUSION

In this chapter, I described our initial efforts to attempt the recovery of the underlying subclonal structure from genomic profiling data. The data clearly indicated that the cell populations found in a tumor sample are more complex than a simple normal + tumor segregation. Using the signals, we were able to reconstruct linear parsimonious subclonal structures from CNV measurements in the form of RDR, as well as LOH events derived from BAF. We showed examples that NGS, as well as array assay, can all be viable sources for identifying clonality, although NGS has been shown to provide higher resolution. This work was to our knowledge the first attempt to tackle this issue, and our method was presented at *TCGA 1st Annual Scientific Symposium* as an oral presentation (Nov, 2011).

However, an important aspect of the described method, which should not be dismissed easily, is that the linear parsimonious subclonal structure, although motivated by a sound biological modal that the more mutated a genome is, the more unstable it would become and the easier to produce more mutations, may not be the only possible structure that is able to explain the observation. This is largely due to the fact the the widely used whole genome profiling techniques, such as array or NGS, require tissue homogenization and DNA fragmentation as part of the sample preparation process, effectively losing the linkage information as in whether two events were originally from the same cell. In the case that the CP of two events are all greater than 50%, the linear heritage model IS in fact the only possibility, as, due to pigeon hole principle, they can-

not separately exist on different cell populations whose fractions sum up to be greater than 100%. But for low frequency events, this ambiguity would be present. Consider two events A and B, with CP values of 20% and 40% respectively, the linear parsimonious method would result in a structure that subclone 1 containing only B derived form the normal cells, taking up 20% of the entire population; and subclone 2 derived from subclone 1, inheriting the event B, but in addition also contains event A, taking up 20% of the entire population. An alternative structure would be that, irregardless of biological feasibility, subclone 1 and subclone 2 are independently derived from the normal tissue, containing event B and A respectively, and each takes up 40% and 20% of the entire population respectively. Mathematically, these two structures will result in exactly the same observation, yet they represent fundamentally different tumorigenesis mechanisms (*E.g.* in model 1, B could be the driver event; In model 2, the patient might be genetically pre-exposed to cancer, and A and B are separate, second hits to an already weakened tumor suppressing pathway). In the next chapter, I will describe an improved method that enumerate all possible structures, instead of just the linear one.

(This page is left blank intentionally . . .)

3

Exhaustive Enumeration

IN THE PREVIOUS CHAPTER, I described a method that is able to reconstruct the linear parsimonious subclone structure based on somatic Copy Number Variation (CNV) or loss-of-heterozygosity (LOH) data. Yet the difficulty still remains that the Cell Prevalence (CP) of individual events measured in a large population of tumor cells, as is the case in "bulk" tumor sequencing or microarray genotyping experiments, do not retain the underlying linkage information that exists between individual somatic events *i.e.* whether or not two or more mutation events are present within the same cell. Unfortunately, given n mutation events, there are in total $n!$ possible subclone structures, and often a large number of these can account

for the CP measurements equally well. This makes it very difficult or impossible to unambiguously reconstruct subclonal evolution from per-locus CP observations. Recently, computational methods have been developed to reconstruct clonal structures that either exploit specific biological assumptions [152] to choose between mathematically equivalent structures (most importantly, the assumption of "Shallowness", which dictates that the depth of the evolutionary tree is minimal, would be in favor of branching structures); or by using Markov Chain Monte Carlo (MCMC) sampling based Bayesian inference [153] to explore the solution space of highly possible phylogenies with a Dirichlet process prior. Both of these methods require high-precision allele frequency (AF) measurements of one specific variant type: Single Nucleotide Variation (SNV). For a comparison of performance, please refer to Appendix Section A.1. Even more recently, several new method have published that either estimate model parameters with Expectation Maximization (EM) while take advantage of physically separated samples [154]; or provide the ability to integrate multiple data types (*e.g.* CNV, LOH, SNV), and jointly estimate the subclone profile. Both of these methods do not explicitly maintain the constraint that the subclones fit within a consistent phylogeny [155]; or model the potentially multi-furcating tumor phylogeny with a bifurcating tree, without the ability to consider multiple tumors from a single patient (such as primary / relapse pairs) [156].

3.1 Introduction

Based on our experience working with two similar yet distinct data types: CNV and LOH, we came to realize that the fundamental signal for subclone structure reconstruction is not tied to any specific data type, but the CP value that describes the fraction of cells harboring a somatic event. Here we discuss a more general approach that is able to accept many types of somatic variation data (*e.g.* SNVs, CNVs from either microarray or Next Generation Sequencing (NGS), LOH, etc.) as input. Moreover, the method

enumerates all possible subclone structures that are consistent with the bulk CP measurements from the input. It is capable of reducing this solution space significantly, often to a single, unique solution when data from multiple tumor biopsies, such as primary and relapse from the same patients, are available. In the event that more than a single alternative subclone structure still remains after such trimming, it is often possible to derive high-confidence linkage information between subsets of loci based on the consensus of all remaining structures. In such cases, we focus not on efforts to disambiguate mathematically equivalent solutions, but rather on using the complete set after pruning procedure in a statistical framework to determine *e.g.* the probability that two given mutations are present within the same subclone (***mutation co-localization***), or that whether one mutation pre-dates another (***mutation ordering***). Such co-localization information may reveal *e.g.* that two distinct mutations that each sensitizes the cancer cells to specific drugs are, in fact, present in a single subclone. Given the high incidence and therapeutic challenges posed by chemoresistant tumors, knowledge of mutation co-localization may allow for more accurate and potentially more efficacious targeted therapeutics aimed at countering or preventing chemoresistance. Moreover, if such a novel mutation in a chemoresistant tumor is present in every cell of the relapse sample, it may be a top candidate in the search for driver mutation in chemoresistance (***variant prioritizing***).

3.2 METHODS

A full implementation of the described method is freely available under the MIT license at `https://github.com/yiq/SubcloneSeeker`. At the time this dissertation is written, the code repository is at commit e01e9b.

An overview of the entire workflow of subclone structure reconstruction using SubcloneSeeker is as the following:

1. Depending on the type of input data, mutation events and their associated allele frequencies are called by detection methods (Chapter 1, Section 1.3)

2. The allele frequencies of events are converted into CP values, and then subjected to clustering. If more than one sample is available, the clustering will be done in a multidimensional space, in which the number of dimensions is equal to the number of samples.

3. The resulting somatic event clusters (clustered by CP) serve as the input to the exhaustive enumeration based subclone structure reconstruction algorithm "SubcloneSeeker". This will result in a set of solutions that are biologically plausible, and mathematically consistent with the input.

4. Further trimming can be performed on the solution set, such as trying to merge multiple samples into a unified evolution tree.

5. Mutation (cluster) co-localization can be inferred from the trimmed solution set.

3.2.1 A UNIFIED FRAMEWORK FOR SUBCLONE STRUCTURE RECONSTRUCTION THAT INCORPORATES ALL TYPES OF GENOMIC VARIANTS

We define a subclone as a collection of cells in the tumor sample that harbor the same set of genomic variants, including SNVs, Structural Variations (SVs), CNVs, LOHs, etc. The only requirement for a data type to be included in the analysis is the ability to derive the fraction of the cells within the tumor sample in which the mutation is present, a quantity that has also been referred to as "cell prevalence" or CP [157]. In a simplified example, a heterozygous SNV in a copy number neutral region with an AF of 30% would correspond to a CP of 60% (Figure 3.2.1 A). It is worth mentioning that the estimation of CP is no trivial task, and should be given ample consideration. Situations such as SNVs in CNV regions will need correction techniques. The same CNV region, interpreted with different absolute copy number (ACN) states, would also results in

ambiguity. A number of tools have been developed to facilitate CP estimation, including ASCAT [134] and ABSOLUTE [135], which estimates the absolute copy number states of CNV regions, and PyClone [157], which jointly estimates the CP of SNVs. Our method requires as input CP measurements, regardless whether these measurements represent SNVs, CNVs, or some other type of genetic variation, allowing it to consider each such variant type, or any combination of them from a given sample. We note that, as a preprocessing step, our method clusters together variants with the same (or similar) CP values to minimize measurement uncertainties, and assumes *a priori* that clusters are the smallest independently inherited unit (*i.e.* all variants in each such cluster are co-localized in the same genomes). The input to our downstream methods is an ordered list of CP values, corresponding to those clusters.

3.2.2 Data preparation of various genomic variation types

Various types of raw data are processed, in data-type specific ways, into somatic events.

Whole genome copy number measurements This is done either by whole genome sequencing (WGS) or Array comparative genome hybridization (aCGH) measurement on paired tumor-normal samples from a cancer patient. In the case of WGS, read depth (RD) is measured within large genomic window (*e.g.* 10kb). For aCGH, hybridization probe intensities are measured, and often averaged across multiple probes. Relative copy number (RCN) measurement is obtained by normalizing tumor read depth or hybridization intensity first to the total amount of DNA per sample (e.g. the total number of reads), followed by normalization to the corresponding measurements in the normal sample. These normalization steps eliminate germline events shared by both the tumor and the normal genomes, and keep only the somatic events

Whole genome LOH measurement The procedure to work with LOH measurement has been described in Chapter 2, Section 2.4.2.

Segmentation The RCN derived from CNV or mirrored B allele frequency (mBAF) measurements in LOH dataset is then subjected to segmentation algorithms, such as DNAcopy [138, 151] or HMMSeg [158], to identify continuous regions with the same underlying copy number or LOH state, and to delineate event boundaries of the corresponding events.

SNV AF estimation Ultra deep sequencing SNV data does not need to be segmented, however their AF needs to be accurately estimated, *e.g.* using PyClone [157], which also performs CP estimation

3.2.2.1 Cell Prevalence Calculation

CP is defined as in what percent of all the cells being examined does one specific event exist. Different data types require different techniques to perform this calculation. **Whole Genome CNV events:** For whole genome CNV events derived from either WGS RD or aCGH probe intensity data, it is important to have a good estimation, or better yet, direct measurement, on the ploidy[1] (p). Various software packages already exist to estimate p, such as ASCAT [134], CNAnorm [159] and ABSOLUTE [135]. Moreover, an ACN needs to be called for every CNV event from RCN that is usually in the form of $\log_2$(Tumor/Normal ratio). In the examples shown in this chapter, the ACNs are called with the same Maximum Likelihood Estimation detailed in Chapter 2, Section 2.4.2. In the case of WGS, log 2 ratio of read depths are obtained as described in Section 3.2.2. For microarray, since the probe intensity (PI) is often already in log 2 ratio form to some common reference sample, the tumor to normal ratio is calculated with subtraction instead of division.

Assuming that at any genomic location where a CNV event is identified, a cell either does or does not have the event, with ACN estimated, the CP can then be calculated as

[1]The number of copies of a complete genome in a cell. Normal cells have two copies, thus diploid. Tumor cells with copy number variation could potentially contain three copies, or triploid, or more.

$$\because \text{ACN} \cdot CP + 2 \cdot (1 - CP) = \text{RCN}$$
$$\therefore CP = \frac{\text{RCN} - 2}{\text{ACN} - 2} \tag{3.1}$$

Whole genome LOH events: Please refer to Chapter 2. Section 2.4.2

SNVs: With accurate allele frequency estimation made available by ultra-deep sequencing and software advancements [157], CP can also be derived from SNVs along with allele specific copy number quantifications. For example, in diploid[2] regions, $CP = 2 \cdot AF$ for heterozygous SNVs, and $CP = AF$ for homozygous SNVs.

3.2.2.2 Clustering

Because the measurement of CP is potentially noisy, we attempt to mitigate its effect through clustering on CP to identify distribution modes. Examples shown in this chapter are clustered with the kernel density function in R, with its bandwidth calculated by the Pilot Estimation of Derivatives [160]. Users can choose to substitute with more advanced techniques, such as MCLUST [161]. When multiple samples are available, it is important to perform clustering on multi-dimensional space, in which the dimensions equal the number of samples, to identify independently inherited clusters.

3.2.3 Subclone structure reconstruction

Given n somatic events (or clusters, referred to as clusters henceforth; see Definition 3.3 and Definition 3.7), each with an associated, distinct CP value, we enumerate all possible "evolution trees" where mutation events occurring along the tree branches give raise to new subclones in a successive fashion (Figure 3.2.1 B). For n clusters, this procedure results in $n!$ distinct subclone structures (Theorem 3.5) assuming that 1) cells

[2]Having two copies of the complete genome.

in a tumor mass are derived from normal tissue cells or existing tumor cells through mitosis, in which recombination is unlikely to occur; and 2) the same mutation event does not spontaneously occur in two different subclones, nor does a mutation get lost from a subclone. Each subclone structure contains exactly n distinct subclones with associated subclone frequency (SF), plus a "null" subclone without any mutation, representing the normal tissue component within the tumor sample (and its SF the "normal tissue contamination"). SF is assigned to each subclone so that all subclones within a given structure, when put together, give rise to the same cluster CP list as the input. In order to satisfy this condition, our procedure may need to assign negative SF values to one or more subclones; such subclone structures are not biologically plausible, and are removed from further consideration. As demonstrated later (Figure 3.3.2), only a small fraction of the structures are biologically plausible (we term these "viable subclone structures").

3.2.3.1 Formal definition of subclone reconstruction problem

Definition 3.1. *A chromosomal location, L, is defined as $L = \{chromosome, position\}$, which describes a location on the genome.*

Definition 3.2. *A chromosomal segment, S, is defined as $S = \{L, length\}$, which describes a continuous region on the genome.*

Definition 3.3. *A somatic event (henceforth referred to as "event" if without specification), e, is a general symbol representing any genomic variation presented in the tumor sample, that is not found in the paired normal sample. e can be one of the following:*

- *A segmental somatic CNV event, e^{CNV}, which is defined as $e^{CNV} = \{S, ACN\}$ for a segment on the genome specified by S, with the absolute copy number state of ACN.*
- *A segmental somatic LOH event, e^{LOH}, which is defined as $e^{LOH} = \{S\}$ for a segment on the genome, specified by S, whose heterozygosity has been lost ($mBAF = 1$).*

- *A somatic SNV event, e^{SNV}, which is defined as $e^{SNV} = \{L, GT\}$ at a position on the genome, specified by L, with a genotype, specified by GT, either being "heterozygous" or "homozygous alternate" in the case of diploid genome, or other more complex genotypes in the case of aneuploid[3] genome.*

Definition 3.4. *An observed somatic event (henceforth referred to as "observed event" if without specification), oe, is defined as $oe = \{e, CP\}$ for an event e observed in $CP > \epsilon_d$ fraction of the total cells, with some detection sensitivity $\epsilon_d > 0$.*

Definition 3.5. *An observation, O, is defined as $O = \{oe_1, oe_2, \ldots, oe_n\}$ for a tumor genome with n detected events. The detection process if data-type specific. In any case, $oe.CP$ should be, or can be derived, from the output of the detection process (Section 3.2.2.1).*

Definition 3.6. *The complete events set, E, is defined as $E = \{oe.e | oe \in O\}$.*

Definition 3.7. *A set of event clusters, P, is defined as a partition over an observation O, that, for a given error margin $\epsilon_P \geq 0$, satisfies*

$$\forall p \in P : (\forall oe \in p, oe' \in p : |oe.CP - oe'.CP| \leq \epsilon_P)$$

$$and$$

$$\forall p \in P, p' \in P, p \neq p' : (\forall oe \in p, oe' \in p' : |oe.CP - oe'.CP| > \epsilon_P)$$

Each element $p \in P$ is called an event cluster. We denote $p.CP = \frac{\sum_{oe \in p} oe.CP}{|p|}$ as the cluster centroid.

We further impose, without loss of generality that, P is a sorted set, with respect to the cluster centroids, in descending order.

[3] Having a ploidy that is other than 2

$$\forall i \in [1, n'], j \in [1, n'], i < j : p_i.CP > p_j.CP$$

Definition 3.8. *A subclone profile, C, is defined as* $C^j = \{B, f\}^j, j = 0..m$, *in which* $C^j.B = \{b_1, b_2, \ldots, b_{n'}\}^j$ *is a row vector whose value* b_i^j *indicates whether the j-th subclone contains the somatic events belonging to the event cluster* p_i. *The 0-th subclone is a special one representing the normal tissue component, thus* $C^0.B = \{0, 0, \ldots, 0\}$. $\forall j \leq m, j' \leq m, j \neq j' : C^j.B \neq C^{j'}.B$. $C^j.f$ *represents the fraction the j-th subclone occupies over the entire cell population, or SF, and that* $\sum_{j=0}^{m} C^j.f = 1$.

Definition 3.9. *For any given subclone profile C, it is said that*

- *C is biologically plausible, if* $\forall j \leq m : C^j.f \geq 0$.
- *C is not biologically plausible, if* $\exists j \leq m : C^j.f < 0$.

Definition 3.10. *A subclone profile C with* $m+1$ *subclones is a solution to an observation O with a clustering partition P having* n' *clusters when the following is satisfied*

$$[C^1.f, C^2.f, \ldots, C^m.f] \times \begin{bmatrix} b_1^1, & b_2^1, & \cdots & b_{n'}^1 \\ b_1^2, & b_2^2, & \cdots & b_{n'}^2 \\ \vdots & \vdots & \ddots & \vdots \\ b_1^m, & b_2^m, & \cdots & b_{n'}^m \end{bmatrix} = [p_1.CP, p_2.CP, \ldots, p_{n'}.CP]$$

Definition 3.11. *A subclone structure, SS, over a given subclone profile C, is defined as a multi-furcating tree whose nodes are individual subclones in C. It is apparent that all SS over a given C with* $m+1$ *subclones have exactly* $m+1$ *nodes. If* $C^j \in SS$ *is the parent node to another node* $C^{j'} \in SS$, *such relationship is denoted as* $C^j \Rightarrow C^{j'}$ *or* $C^{j'} \Leftarrow C^j$; *If* C^j *is an ancestral (not necessarily direct parent) node to another node* $C^{j'}$, *such relationship*

is denoted as $C^j \rightarrow C^{j'}$ or $C^{j'} \leftarrow C^j$. If neither C^j nor $C^{j'}$ is an ancestral node to the other, such relationship is said as "C^j and $C^{j'}$ are parallel", and denoted as $C^j || C^{j'}$.

Due to the unique biology of tumorigenesis, we make the following assumptions

- Cells in a tumor mass are derived from germline cells or parental, existing tumor cells through mitosis, in which recombination is unlikely to occur.
- The same event (with respect to a boundary resolution) would not spontaneously occur in two subclones without a descendant relationship, nor would pre-existing events revert back to the normal state in a descendant subclone

Definition 3.12. *A subclone structure SS over a given subclone profile C is said to be evolutionary, if the following conditions are satisfied*

- $\forall C^j, C^{j'}, C^j \rightarrow C^{j'} : \forall b_i^j = 1 : b_i^{j'} = 1$
- $\forall C^j, C^{j'}, C^j \Rightarrow C^{j'} : \forall b_i^j = 0, b_i^{j'} = 1 : \forall C^{j''} || C^{j'} : b_i^{j''} = 0$
- $\forall C^j, C^{j'}, C^j \Rightarrow C^{j'} : \forall b_i^j = 0, b_i^{j'} = 1 : \forall C^{j''} \rightarrow C^j : b_i^{j''} = 0$

We term $\exists C^j, C^{j'}, C^j \Rightarrow C^{j'} : \exists b_i^j = 0, b_i^{j'} = 1$ as "Event cluster p_i first appeared in subclone $C^{j'}$"

We term $\exists C^j, C^{j'}, C^j \rightarrow C^{j'} : \exists b_i^j = 0, b_i^{j'} = 1$ as "Event cluster p_i appeared after subclone C^j"

Theorem 3.1. *For any given C over which evolutionary subclone structures exist, only one evolutionary subclone structure exists.*

Proof. Assume that there are two subclone structures, SS_1 and SS_2, over the same subclone profile C, that are both evolutionary. Due to Definition 3.12

$$\exists C^{p1}, C^{p2}, C^{j} : \begin{cases} C^{p1} \Rightarrow C^{j} \text{ in } SS_1 \implies \begin{cases} \forall b_i^{p1} = 1 : b_i^{j} = 1 \\ \forall b_i^{j} = 0 : b_i^{p1} = 0 \end{cases} \\ C^{p2} \Rightarrow C^{j} \text{ in } SS_2 \implies \begin{cases} \forall b_i^{p2} = 1 : b_i^{j} = 1 \\ \forall b_i^{j} = 0 : b_i^{p2} = 0 \end{cases} \\ C^{p1}.B \neq C^{p2}.B \neq C^{j}.B \end{cases}$$

Lemma 3.1. $\forall b_i^{p1} = 0, b_i^{j} = 1 : b_i^{p2} = 0$

Proof. Assume that

$$\exists b_i^{p1} = 0, b_i^{j} = 1 : b_i^{p2} = 1$$

Due to Definition 3.12, the only possible relationship between C^j and C^{p2} in SS_1 is that $C^j \rightarrow C^{p2} \implies \exists b_i^{p2} = 1 : b_i^{j} = 0$. However, this contradicts with the condition that $C^{p2} \Rightarrow C^j$ in $SS_2 \implies \forall b_i^{p2} = 1 : b_i^{j} = 1$. □

Lemma 3.2. $\forall b_i^{p2} = 0, b_i^{j} = 1 : b_i^{p1} = 0$

Proof is similar to Lemma 3.1

Combine Equation 3.1, Lemma 3.1, and Lemma 3.2, we have

$$\forall b_i^{j} = 0 : b_i^{p1} = 0, b_i^{p2} = 0$$

$$\forall b_i^{j} = 1 : (b_i^{p1} = 0, b_i^{p2} = 0) \vee (b_i^{p1} = 1, b_i^{p2} = 1)$$

Thus we have

$$\forall i : b_i^{p1} = b_i^{p2}$$

This contradicts with the condition that $C^{p1}.B \neq C^{p2}.B$ □

Definition 3.13. *The problem of subclone structure reconstruction, is that given an observation O, along with a clustering partition P with n' event clusters, find all evolutionary subclone structures SS whose corresponding, biologically plausible, subclone profiles C are solutions to O with clustering partition P.*

Theorem 3.2. *For a given evolutionary subclone structure SS whose corresponding subclone profile C is biologically plausible and a solution to a given observation O with a clustering partition P having n' clusters, let C^{p_i} denote the subclone in which p_i first appeared, the following condition is true*

$$\forall p_i \in P : p_i.CP = C^{p_i}.f + \sum_{C^j}^{C^{p_i} \to C^j} C^j.f$$

Proof. Because p_i first appeared in C^{p_i}, from Definition 3.12 we have

$$\begin{aligned} b_i^{p_i} &= 1 \\ \forall C^j \leftarrow C^{p_i} : b_i^j &= 1 \\ \forall C^{j'} \to C^{p_i} \vee C^{j'} \| C^{p_i} : b_i^{j'} &= 0 \end{aligned}$$

Because C is a solution to O with clustering partition P, from Definition 3.10, we have

$$\begin{aligned} p_i.CP &= \sum_{C^j} C^j.f \times b_i^j \\ &= \sum_{C^j}^{C^j \to C^{p_i} \vee C^j \| C^{p_i}} C^j.f \times b_i^j + C^{p_i}.f \times b_i^{p_i} + \sum_{C^j}^{C^j \leftarrow C^{p_i}} C^j.f \times b_i^j \\ &= C^{p_i}.f + \sum_{C^j}^{C^{p_i} \to C^j} C^j.f \end{aligned}$$

□

We define $\widehat{C^j}$ *as*

$$\widehat{C^j} = C^j.f + \sum_{C^{j'}}^{C^j \to C^{j'}} C^{j'}.f$$

which represent the sum of the subclone frequencies of all nodes in the subtree with C^j *being the root.*

Corollary 3.1. *For a given evolutionary subclone structure* SS *whose corresponding subclone profile* C *is biologically plausible and a solution to a given observation* O *with a clustering partition* P *having* n' *clusters, if* $p_i \in P$ *first appeared in subclone* C^{p_i}*,* $p_{i'} \in P$ *first appeared in subclone* $C^{p_{i'}}$*,* $C^{p_i} \to C^{p_{i'}}$*, then we have* $p_i.CP \geq p_{i'}.CP$

Proof. From Theorem 3.2, we have

$$\begin{aligned} p_i.CP &= \widehat{C^{p_i}} \\ &= C^{p_i}.f + \sum_{C^j}^{C^{p_i} \to C^j} C^j.f \\ p_{i'}.CP &= \widehat{C^{p_{i'}}} \\ &= C^{p_{i'}}.f + \sum_{C^j}^{C^{p_{i'}} \to C^j} C^j.f \end{aligned}$$

Because $\forall C^j \leftarrow C^{p_{i'}} : C^j \leftarrow C^{p_i}$

$$p_i.CP - p_{i'}.CP \geq C^{p_i}.f$$

$$p_i.CP \geq p_{i'}.CP$$

□

Corollary 3.2. *For a given evolutionary subclone structure SS whose corresponding subclone profile C is biologically plausible and a solution to a given observation O with a clustering partition P having n' clusters, $\forall p_i \in P, p_{i'} \in P, p_i.CP < p_{i'}.CP$, The subclone C^j that contains p_i but not $p_{i'}$ cannot be an ancestral node to a subclone that contains $p_{i'}$*

Proof. $C^j \rightarrow C^{j'}$ implies that p_i first appeared either in C^j or some subclone $C^k \rightarrow C^j$, and $p_{i'}$ first appeared in $C^{j'}$, according to Corollary 3.1, $p_i.CP \geq p_{i'}.CP$, which contradicts with the condition that $p_i.CP < p_{i'}.CP$ □

Theorem 3.3. *For a given evolutionary subclone structure SS whose corresponding subclone profile C, having $m+1$ subclones, is biologically plausible and a solution to a given observation O with clustering partition P, having n' clusters, it is true that $m \geq n'$.*

Proof. Assume that $m < n'$. Due to pigeonhole principle,

$$\exists C^j : \exists p_i \in P, p_{i'} \in P : p_i \text{ and } p_{i'} \text{ both first appeared in } C^j$$

Thus

$$p_i.CF = p_{i'}.CF = \widehat{C^j}$$

This contradicts with Definition 3.7 □

Theorem 3.4. *For a given evolutionary subclone structure SS whose corresponding subclone profile C, having $m+1$ subclones, is biologically plausible and a solution to a given observation O with clustering partition P, having n' clusters, it is true that $m \leq n'$.*

Proof. Assume that $m > n'$. Due to pigeonhole principle,

$$\exists C^j \neq C^0 : \forall p_i \in P : p_i \text{ first appeared in a subclone } C^{j'} \neq C^j$$

Thus, $\exists C^k \Rightarrow C^j : C^k.B = C^j.B$, which contradicts Definition 3.8 □

Corollary 3.3. *For a given evolutionary subclone structure SS whose corresponding subclone profile C, having $m+1$ subclones, is biologically plausible and a solution to a given observation O with clustering partition P having n' clusters, from Theorem 3.3 and Theorem 3.4, we have $m = n'$.*

Corollary 3.3 states that, for a given evolutionary subclone structure SS whose subclone profile C is biologically plausible and a solution to a given observation O with clustering partition P having n' clusters, there are exactly $n'+1$ subclones in C.

Theorem 3.5. *For a given observation O with clustering partition P having n' clusters, there are at most $n'!$ different subclone structures that are evolutionary and their corresponding subclone profiles are biologically plausible and solutions to O.*

Proof. First, we denote an evolutionary subclone structure SS with m nodes SS_m.

Base case: when $|P| = 1$, there is only one SS_1, $C^0 \Rightarrow C^1$, that p_1 first appeared in C^1.

Induction: assume that when $|P| = k$, the total number of SS_k is $k!$. We introduce p_{k+1} into $|P|$, and assume without loss of generality that $\forall p_i \in P : p_i.CP > p_{k+1}.CP$. SS_{k+1} can be derived by attaching a new node C^{k+1}, in which p_{k+1} first appeared, onto existing SS_k. From Theorem 3.2, in any given SS_{k+1}, C^{k+1} cannot be the parent of any subclones in SS_k, leaving the only possible placement for C^{k+1} to be a child of the existing $k+1$ subclones in SS_k. Therefore, the number of possible SS_{k+1} is $k! \times (k+1) = (k+1)!$ □

3.2.3.2 Exhaustive enumeration method

An exhaustive enumeration algorithm is designed to derive all possible structures in the similar fashion as described in the induction step of the proof to Theorem 3.5, and outlined as Listing 3.1.

The function "treeEnum(T, P)" enumerate all SS_{k+1} from all SS_k, exactly as described in the proof of Theorem 3.5, through a recursive call to itself when P is not exhausted,

```
Initialize a tree T with a root that contains no event
treeEnum(T, P);

function treeEnum(T, P):
  p = first_elem(P)  ; the event cluster with highest CP
  for n in all existing nodes of T:
    create a new node n'
    n'.first_event = p
    add n' to T as a child of n
    if P.size == 1:
      Evaluate(T)
    else:
      treeEnum(T, P-{p})
    end-if
    remove n' from T
  end-for
end-function

function Evaluate(T):
  for n in post-order-traverse(T):
    if n is leaf:
      n.SF = n.first_event.CP
    else:
      n.SF = n.first_event.CP - sum(n.children.SF)
    end-if
    if(n.SF < 0) abort
  end-for
  output T
end-function
```

Listing 3.1: Pseudo code of the exhaustive enumeration algorithm.

and evaluate whether the subclone profile, to which the resulting structure T corresponds, is biologically plausible and a solution to the observation O with clustering partition P.

The function "Evaluate(T)" will, through a post-order tree traverse, try to assign a SF (f as mentioned in Definition 3.8) value to each of the tree nodes so that at the end the subclone structure is a solution to the observation(O, P). If the function visits a leaf node, it will assign the CP of the event clusters first appeared in the node; If the function visits an internal node, it will assign the CP of the event clusters first appeared in the node, minus the sum of the SF of all its descendant nodes (because they all inherit those events). It it can do so without assigning any node a less-than-zero SF, the subclone profile C the specific tree structure corresponds to is biologically plausible, and the tree structure is recorded as a feasible solution.

This method will result in a tree-set, which contains all the possible subclone structures whose subclone profile is a solution to the observation, and the phylogeny between the subclones. One can choose to further trim the set by external or internal linkage information, or perform coexistence prediction. An example of all the resulting structures, along with the assigned CP values, when three event clusters are considered is given in Figure 3.2.1B.

3.2.3.3 Trimming the space of viable subclone structures.

Often there are more than one viable subclone structures in the resulting solution set, corresponding to multiple alternative subclone evolutions. However, if additional "linkage" data is available, further trimming is usually possible. Such linkage information may be either directly observed, such as in the case of spectral karyotype images [46, 162, 163], single cell colony assays, or single cell sequencing; or indirectly inferred from *e.g.* primary and relapse tumor from the same patient. Because typically, the relapse tumor is derived from the primary tumor, they share mutations originating from

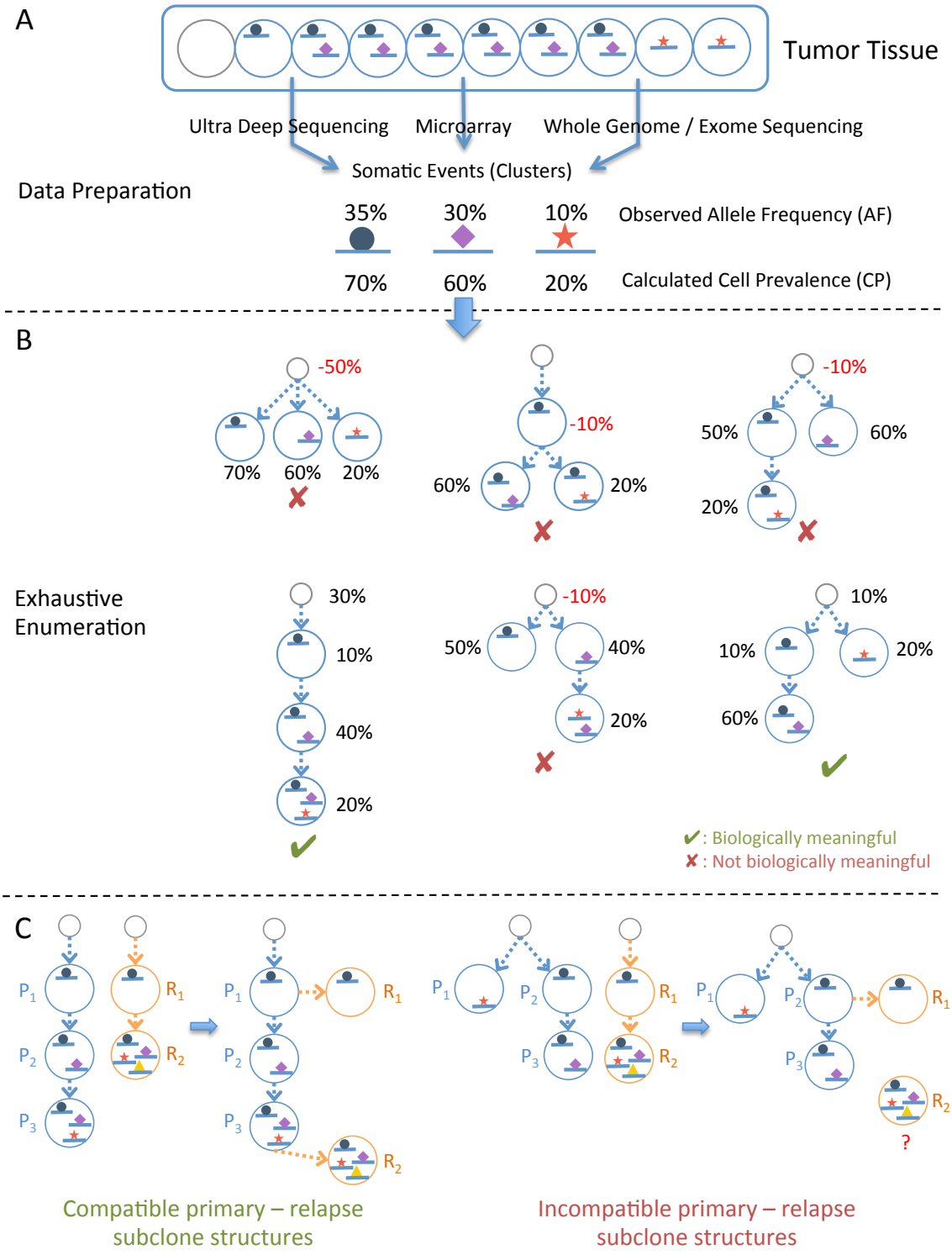
common ancestor subclones, and through such shared evolutionary history the primary and relapse subclones can be merged into one unified subclone structure, while satisfying the following two conditions:

1. After merging, for any given non-leaf node, its children node must have all the mutations presented in the node itself (extra relapse specific mutations are allowed).

2. No two branches shall have the same mutation simultaneously without sharing a common parent node who has that mutation.

These two conditions assure the fundamental assumptions concerning tumorigenesis aforementioned are met. Through this process, if a specific primary (or relapse) tree cannot be merged with any relapse (or primary) tree, that specific tree is then an invalid solution, and can be discarded. Figure 3.2.1 C shows examples of two compatible primary / relapse structures (left) as well as two incompatible ones (right). In the latter example, the relapse subclone R2 contains two mutations that are found in different branches on the primary tree (P1 and P3), violating the assumptions above. Any structure in the primary that has no compatible structure in the relapse, or vice versa, is discarded from consideration, reducing the solution space.

Figure 3.2.1 ***(following page)*****:** SubcloneSeeker Method Overview. A) Data Preparation: Genomic variation data (SNVs, CNVs, etc.) is converted into the corresponding CP values, and clustered into distinct groups. B) Structure Enumeration: Based on the identified CP clusters, all possible subclone structures, represented as branching tree structures where one subclone is derived from its "predecessor" by the addition of a mutation (or cluster of mutations), are visited. During the visit, each subclone on the tree structure is assigned a SF value so that the implied total CP values for mutations are in agreement with the input CP values. Those structures with negative SF values are removed from the solution set. C) Solution Trimming: The aim of this procedure is to merge the subclone structures from the relapse tumor (orange circles) those from the primary tumor (blue circles) from the same patient. Left Panel: Example showing a compatible pair of relapse/primary structures. Right Panel: Example showing a pair of incompatible relapse/primary subclone structures. A subclone in the relapse, R_2, cannot be positioned anywhere within the primary subclone structure because it contains mutations found in separate primary subclones (P_1 and P_3.), and therefore cannot be derived from either one or the other.


A
Tumor Tissue
Ultra Deep Sequencing
Microarray
Whole Genome / Exome Sequencing
Somatic Events (Clusters)
Data Preparation
35%
30%
10%
Observed Allele Frequency (AF)
70%
60%
20%
Calculated Cell Prevalence (CP)
B
-50%
70%
60%
20%
-10%
60%
20%
-10%
50%
60%
20%
Exhaustive
Enumeration
30%
10%
40%
20%
-10%
50%
40%
20%
10%
10%
20%
60%
✔: Biologically meaningful
✘: Not biologically meaningful
C
P1
P2
P3
R1
R2
Compatible primary – relapse
subclone structures
Incompatible primary – relapse
subclone structures
?

3.2.4 MUTATION CO-LOCALIZATION PREDICTION

Useful knowledge can be derived even in cases where there are multiple alternative subclone structures. Although one cannot determine the precise subclone evolution with certainty in such cases, the collection of all possible solutions can be used to predict whether or not two mutations are present in the same cell, *i.e.* whether or not they are co-localized within the same subclone. This prediction is based on the fraction of all viable subclone structures in which two mutations (or more generally, a given set of mutations) are present in at least one subclone. Such information could potentially be important in *e.g.* designing personalized chemotherapy treatment plans. Given n clusters, there are in total C_n^2 (n choose 2) unique, unordered cluster pairs, each of which is assigned a status of either "co-localized", "not co-localized", or "ambiguous" (Figure 3.2.2). Furthermore, for two events that are localized in the same subclone, the timing of the mutations can be easily determined: the event with the higher CP value appeared earlier, and the event with the lower CP value emerged later.

For any given pairs of somatic event clusters, a co-localization frequency matrix (CLF) can be calculated as

$$CLF = \sum_{i=1}^{\#\text{ of solutions}} PS_i \cdot CL \tag{3.2}$$

in which PS_i is the probability that solution i is the correct solution, which in case no prior knowledge is available, can be estimated as

$$PS_i = \frac{1}{\#\text{ of solutions}} \tag{3.3}$$

CL is a binary variable that describes whether a given pair of event clusters co-localize in solution i, which can either be 1, if in at least one subclone the event clusters co-localize, or 0 if in none of the subclones the event clusters co-localize. This framework allows us to estimate co-localization giving all structures equal probability to be true,

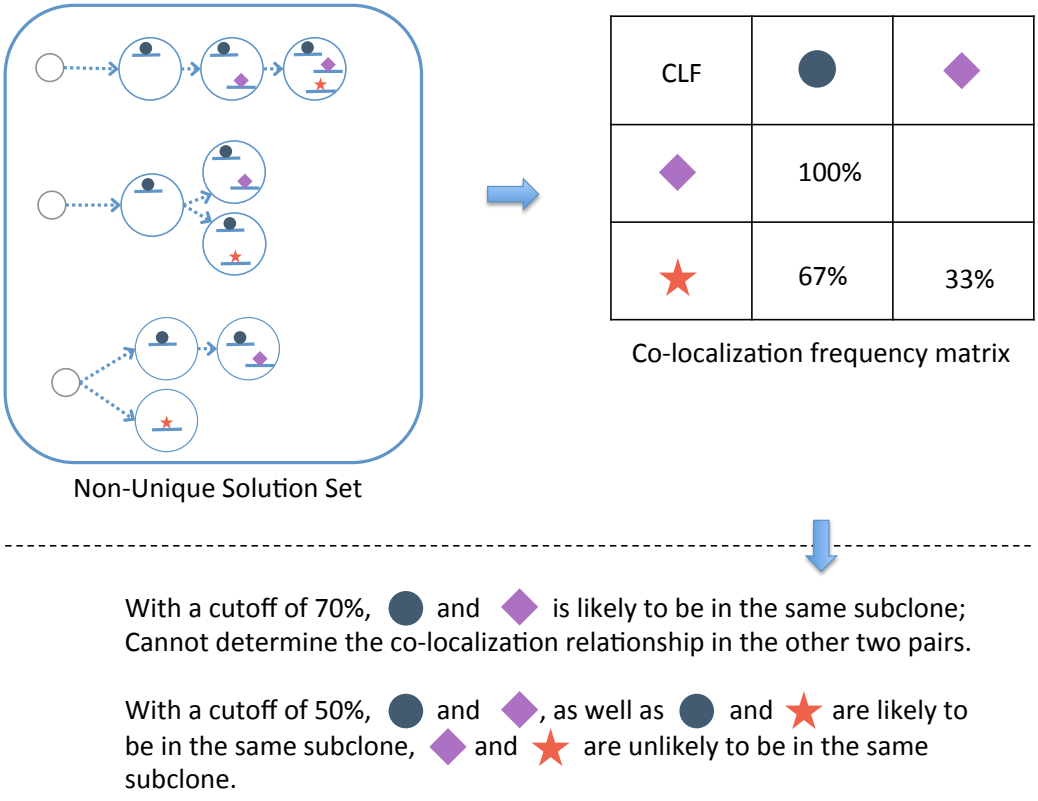


Figure 3.2.2: Predicting mutation co-localization. In cases where there are multiple viable subclone structures, we count the fraction of all structures within which two mutation events are co-localized. This fraction is the probability that the two events are present in the same subclone. One can also make a "co-localization call" by declaring that two events are co-localized, if this probability is above a pre-defined threshold.

or weight towards, or against, specific structures. (*e.g.* one can reasonably argue that it is generally unlikely for a patient to develop two, separate tumor subclones without related by a common ancestor which contains the initial driver mutations, thus placing a lower prior on those structures in which multiple subclones are derived directly from the normal tissue.)

3.2.5 Subclone structure simulation process

In order to understand the behavior of our subclone reconstruction algorithm, we designed a tumor subclone structure simulator. The simulator initialize in a state that it only contains one subclone with no somatic event. This "null" subclone logically represents the normal tissue before tumor expansion, and mathematically represents the normal tissue contamination usually found in tumor sample. We also assign a "viability" value of 100 to this null subclone. The viability value represent the ability for a certain subclone to grow, and will ultimately determine the SF of each subclone. The simulator will then repeat the following steps exactly m times to simulate one subclone structure with $m+1$ subclones

1. From the existing subclones, a "parent" subclone is selected randomly by sampling from a roulette wheel. The proportion of each subclone on the roulette wheel is determined by the viability value of the subclone.

2. A new subclone is created, with one additional mutation, and attached as a children node to the parent subclone. The mutation is only symbolic, so that allele frequency can be calculated at the end

3. The viability value of the new subclone is determined by randomly sampling from a uniform distribution with a range of $(0.5 \times \text{Viability}_{\text{Parent}}, 2 \times \text{Viability}_{\text{Parent}})$, signifying that a mutation can be beneficial, detrimental, or neutral to the growth advantage.

The process is not meant to accurately model the actual tumor micro-evolution, but to create a large number of subclone structures with varying topology and CP values. After the structure is created, each subclone is assigned a SF proportional to its viability value

$$C^j.f = \frac{\text{Viability}_j}{\sum \text{Viability}} \tag{3.4}$$

The CP value for each of the introduced mutations is then calculated, which will serve as the input to the subclone reconstruction algorithm, as

$$CP_i = \sum_{j=0}^{j \leq m} C^j.f \cdot b_i^j;\ b_i^j = \begin{cases} 1 & \text{subclone } j \text{ contains mutation } j \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

The output of the simulation procedure will be a subclone structure, along with the CP value of all the mutations. The CP values will be used as input to the reconstruction algorithm, and the subclone structure will be used to check if, among the results produced by reconstructing, the correct structure has been found.

3.3 Results

3.3.1 The method always capture the correct structure

We generated simulated tumor samples (Section 3.2.5) comprising $3, 4, \ldots, 8$ mutation events with distinct CP values (from our experience, we usually see less than 6 subclones in a clinical tumor sample). For each of these "tumor samples", we produced a random subclone structure serving as a "true" structure. We repeated this procedure 1,000 times. In every case, SubcloneSeeker was able to reproduce the "true" subclone structure as one of the solutions in the complete solution set of viable subclone structures. This "sanity check" was necessary to ensure that our software worked appropriately for

simulated datasets.

3.3.2 The number of biologically plausible subclone structures is low

We also found that the number of viable subclone structures is very low compared to the number of all possible structures. As Figure 3.3.1 illustrates, the expected number of viable subclone structures is far less than the theoretical upper-limit ($n!$ for n distinct CP values, Theorem 3.5).

3.3.3 Normal cell component estimation procedure is accurate

As described above, our subclone structure reconstruction method provides in each its resulting structures a null subclone with no mutations. This is the normal cell component of the tumor biopsy, and its fraction the normal cell fraction. We investigated the accuracy with which our method estimates the normal cell fraction in experimental data. We applied our method to a dataset created by mixing $10\%, 20\%, \ldots, 90\%, 95\%$ and 100% sequencing reads from a SNUC cell line sample [164], with reads sequenced from paired normal tissue (Figure 3.3.2). In this dataset, the non-branching, stepwise mutation accumulation model (red-cross), a parsimonious solution that always exists (Chapter 2), produced very accurate estimate for normal cell content among all alternative structures ($R^2 = 0.9705395$ to the line $y = x$).

3.3.4 Our algorithmic procedure improves on interpretation in previously published data

In a recent study, Ding, *et al.* [113] investigated clonal evolution in eight acute myeloid leukemia (AML) patients. To ensure comparability with the published results, we started with the somatic mutation clusters and AF values provided in the study (Table 3.3.1), rather than re-computing them ourselves. With two exceptions, SubcloneSeeker produced the same subclone structures, and with one exception, came to the same conclu-

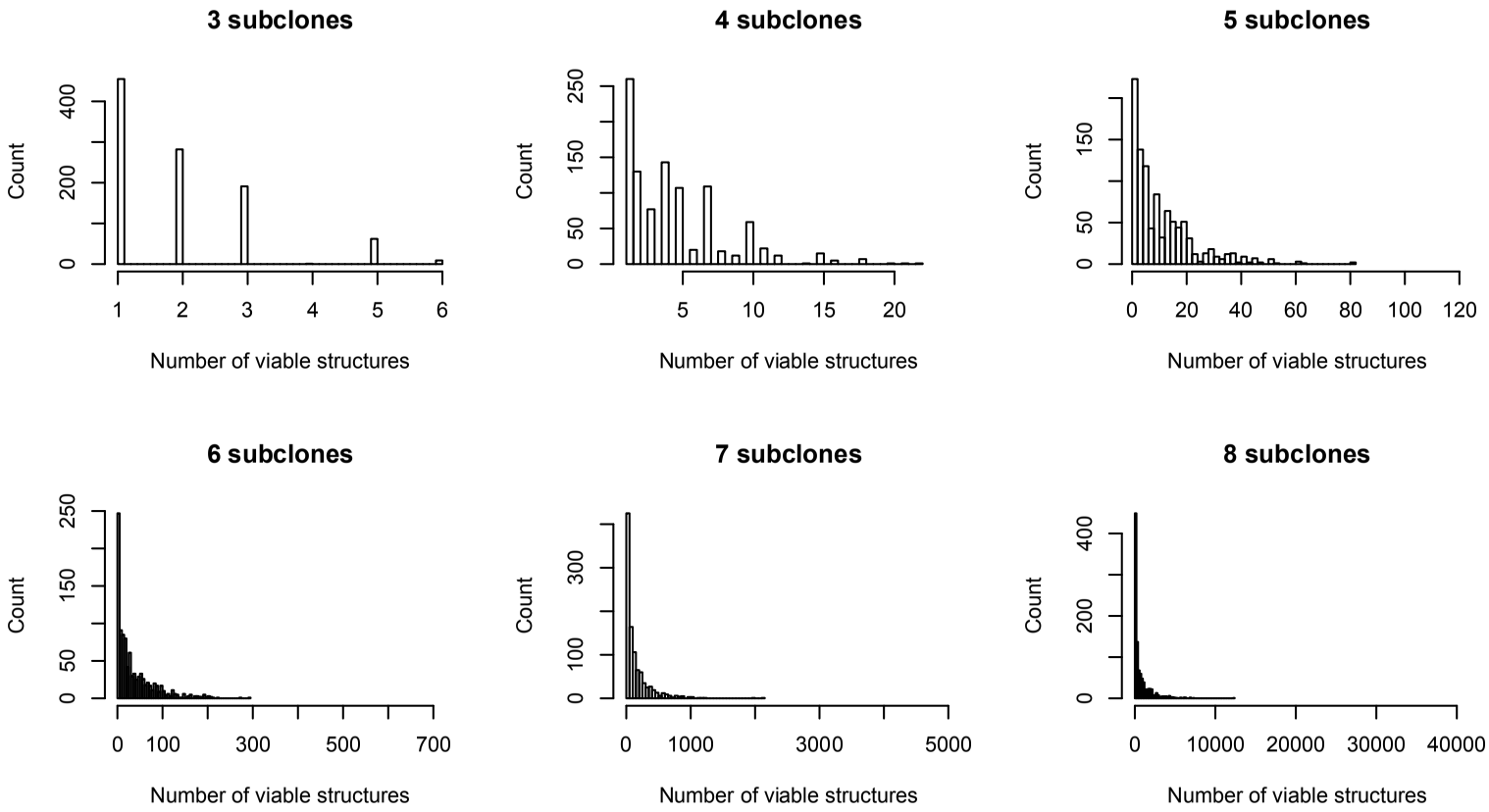


Figure 3.3.1: Number of biologically plausible structures histogram based on simulation. Each plot is based on a set of 1000 randomly generated subclone structures (procedure described in Section 3.2.5). The maximum value on the x axis of each plot represent the theoretical upper limit on the number of solutions, $n!$, given n CP values. The distributions are heavily compressed towards the left, suggesting that the actual number of biologically plausible structures is usually small.

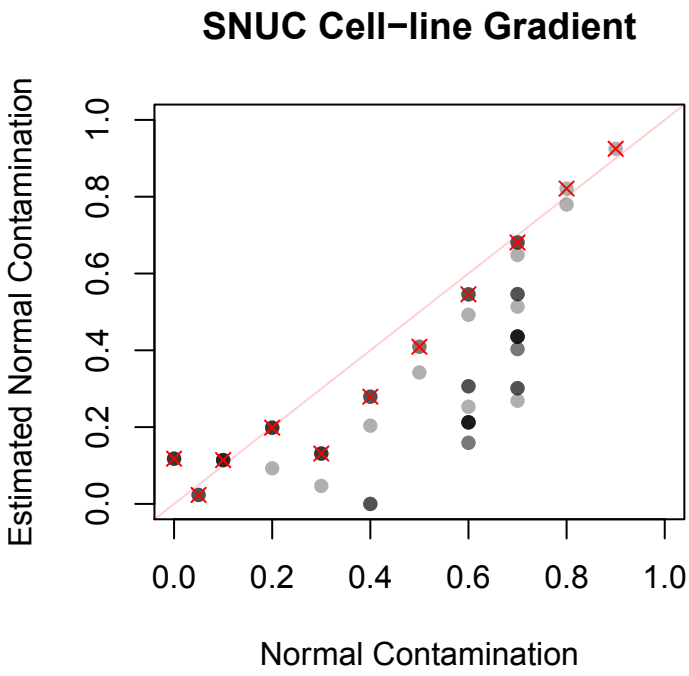


Figure 3.3.2: Normal cell content estimated by subclone reconstruction in a controlled mixing experiment. Dataset is generated by mixing sequencing reads from a Sinonasal Undifferenciated Carcinoma (SNUC) cell-line and matched normal tissue. Data points corresponding to the subclone structure representing linear mutation accumulation are shown with a red cross.

sions (Table 3.3.2). Please refer to Table 4.0.1 for a summary of input data types and major conclusions.

In the case of patient UPN933124, the primary sample contained two low frequency clusters, which resulted in a total of 6 different viable subclonal structures, including the one reported in the original study. However, only one of these was compatible with the sole viable subclone structure in the relapse, and the resulting single primary / relapse subclone structure was in agreement with the model presented in the original paper (Figure 3.3.3 A). In the case of patient UPN758168, the relapse sample yielded two possible structures, both of which were compatible with the primary structure. However, the tumor expansion model suggested by either of these structures disagrees with the expansion model described in the original paper as "a minor clone carrying the vast majority of the primary tumor mutations survived and expanded at relapse". Our subclone structures (Figure 3.3.3 B) suggest, in contrast, that both primary subclones survived in the relapse. The difference between the two relapse models is which primary subclone expanded with extra mutations.

3.3.5 Analysis of whole-exome sequencing data from chemoresistant *vs.* primary ovarian tumors demonstrates that our method can be used to prioritize somatic mutations for further follow-ups

We are investigating how high-grade serous ovarian cancers become chemoresistant by applying SubcloneSeeker to whole exome sequencing datasets on normal, primary tumor and chemoresistant relapse tumor tissue samples from the same patient. Please refer to Table 4.0.1 for a summary of input data types and major conclusions. Figure 3.3.4 shows our analysis workflow for prioritizing mutations observed in patients "S15" and "S17". Somatic mutations were first clustered in the "Primary AF — Relapse AF" space to identify discrete modal values, corresponding to distinct subclones (Figure 3.3.4 A, B, D, E). The allele frequencies of these clusters were then converted to CP

UPN	Cluster	Primary AF (%)	Relapse AF (%)
933124	Cluster1	46.86	42.23
933124	Cluster2	24.89	0.24
933124	Cluster3	17	40.04
933124	Cluster4	2.39	38.53
933124	Cluster5	0.04	39.65
758168	Cluster1	45.5	44.8
758168	Cluster2	41.8	26
758168	Cluster3	0	17
400220	Cluster1	44.6	36.6
400220	Cluster2	0	13.3
426980	Cluster1	45.4	41.3
426980	Cluster2	45.4	11.5
426980	Cluster3	45.4	0
426980	Cluster4	18.2	0
426980	Cluster5	0	30.1
452198	Cluster1	45.4	18
452198	Cluster2	36	0
452198	Cluster3	11	0
452198	Cluster4	0	18
573988	Cluster1	41.7	14.3
573988	Cluster2	0	21.7
869586	Cluster1	45.4	20
869586	Cluster2	23.3	0
869586	Cluster3	16.4	20
869586	Cluster4	0	20

Table 3.3.1: Summary of data published in Ding *et al.* [113] that were used in the analysis of the same AML dataset.

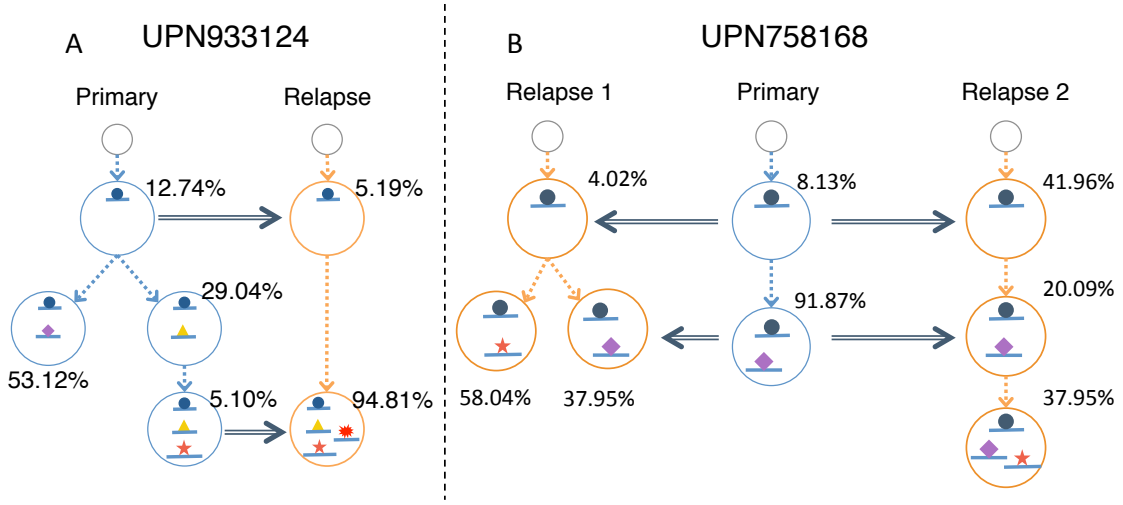


Figure 3.3.3: Our re-analysis of published primary/relapse AML dataset in Ding *et al.* Primary, relapse, and merged subclone structures for two patients, reconstructed with SubcloneSeeker. A) SubcloneSeeker analysis found 6 alternative primary subclone structures for patient UPN933124. Only one is compatible with the relapse subclone structure, and the pair is in agreement with the original study. B) Each of the two viable merged primary/relapse subclone structures for patient UPN75816 suggests that the two primary subclones made it to the relapse tumor, and further expanded.

Patient ID	Primary tumor structures	Relapse tumors structures	Compatible structure pairs	Same conclusion as Ding, *et al.*
933124	6	1	1	Yes
758168	1	2	2	No
400220	1	1	1	Yes
426980	1	1	1	Yes
452198	1	1	1	Yes
573988	1	1	1	Yes
804168	1	1	1	Yes
869586	2	1	1	Yes

Table 3.3.2: Summary of the re-analysis results of AML patient samples reported in Ding, *et al.*

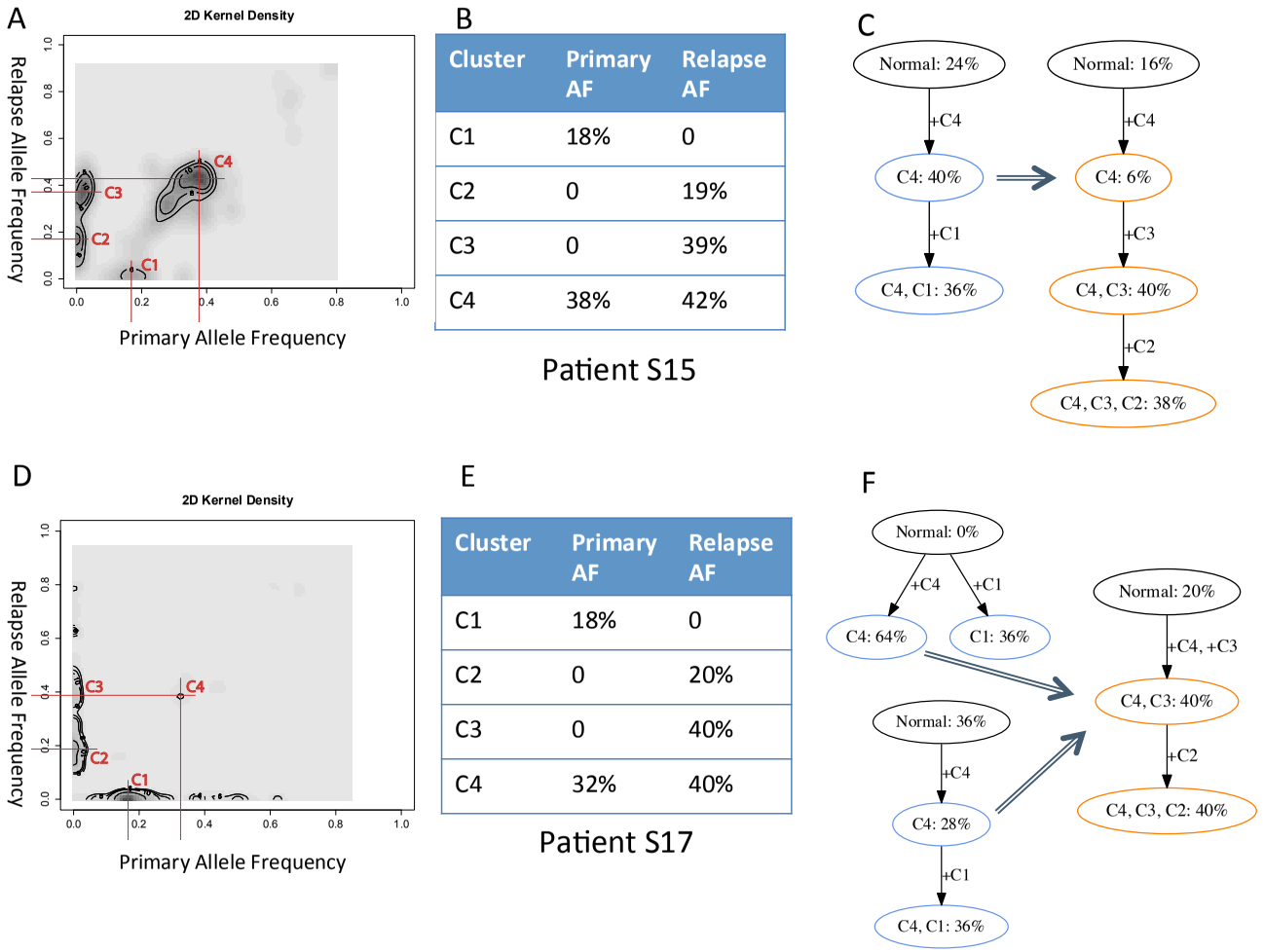


Figure 3.3.4: Analysis of whole-exome sequencing data on patient S15 and S17 from chemoresistant relapse *vs.* primary ovarian cancer dataset. A) Clustering of somatic mutations in patient S15. B) Mutation clusters and CP values in S15 primary and relapse. C) Uniquely identified, compatible S15 primary and relapse subclone evolution tree. D) Clustering of S17 somatic mutations in patient S17. E) Mutation clusters and CP values in S17 primary and relapse. F) Two viable structures for S17 primary, and a sole structure for S17 relapse.

values, and subjected to subclone structure reconstruction. In the case of "S15", both the primary and the relapse sample yielded a unique structure; these are compatible with each other (Figure 3.3.4 C). The mutations in cluster "C4" are early events in the primary, present in every cell of the relapse, and likely contain the driver mutation responsible for initial tumor expansion. On the other hand, in the relapse sample, the vast majority (93%) of tumor cells contain the mutations that make up cluster "C3". This makes it likely that the mutation(s) conferring the tumor phenotype are part of this cluster.

In the case of sample "S17", the primary sample yielded two viable subclone structures, both compatible with the sole structure in the relapse (Figure 3.3.4 F). Similarly to sample "S15", cluster "C4" is likely to contain the initial driver mutation(s), and cluster "C3", which is present in all relapse subclones, is likely to contain the mutation leading to chemoresistance. In both samples, the use of subclone analysis resulted in information that one can use for variant prioritization, in order to narrow down the set of somatic events in the search for the causative mutation, both for initial tumor expansion, and for chemoresistance.

3.3.6 Simulation studies demonstrates that our statistical framework is able to accurately predict whether two somatic mutations, or clusters, are localized in a subclone together

To understand the behavior of our methods predicting co-localization of mutations within subclones, we simulated tumors with 5, 6, and 7 subclones (in each case, 1000 replicates), performed our subclone reconstruction procedure, and carried out mutation co-localization analysis (Section 3.2.4). We used threshold values of 0.7 and 0.5 to call whether two mutations are co-localized, not co-localized, or that the results are ambiguous (see Figure 3.3.5 for 6 subclones, and Figure 3.3.6 for the complete set). Importantly, at a call threshold of 0.7, our method calls co-localized and not co-localized

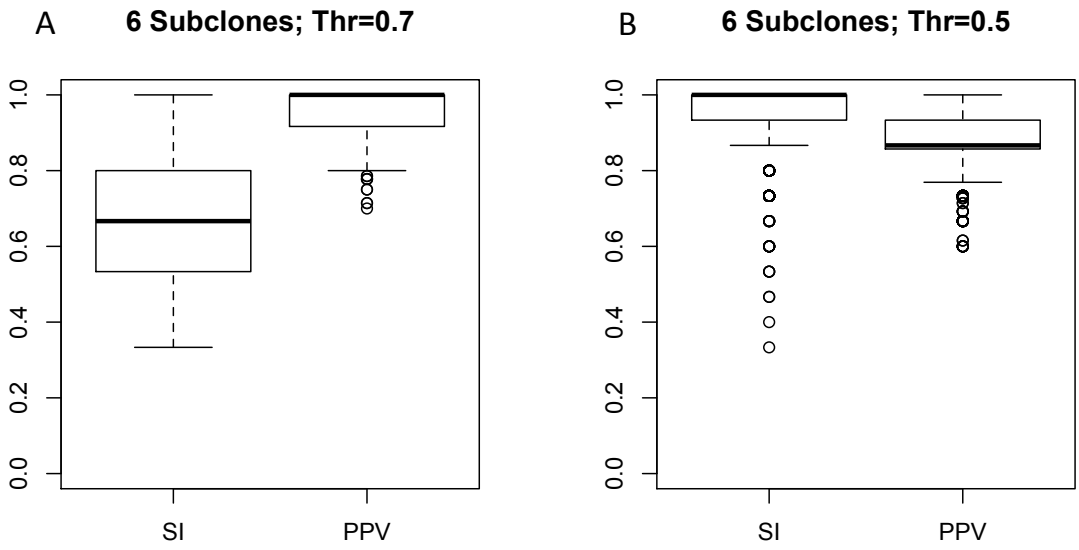


Figure 3.3.5: Performance of mutation co-localization prediction on simulated data. A) Co-localization prediction statistics on simulated dataset with 6 subclones in each tumor sample, and a threshold of 0.7. SI — Combined Sensitivity; PPV — Combined positive predictive value; B) Co-localization prediction statistics on simulated dataset with 6 subclones in each tumor sample, and a CLF threshold at 0.5.

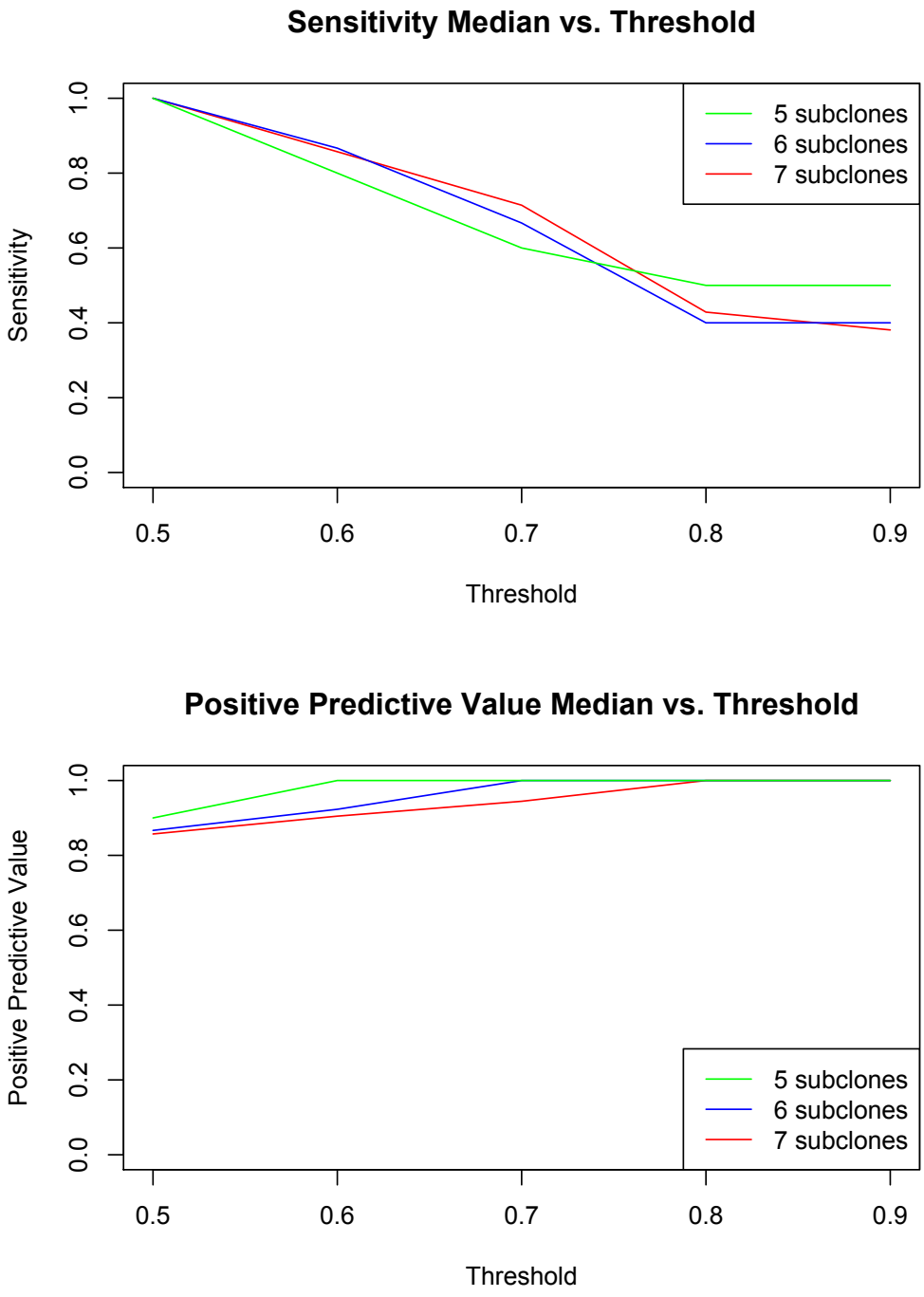


Figure 3.3.6: Performance statistics over the complete set of mutation co-localization prediction performance on simulated data. Values plotted are median over 1000 simulations in each case.

pairs with 70% sensitivity and nearly 100% positive predictive value (PPV, the fraction of correct calls in all the calls made). At a threshold of 0.5, sensitivity goes up to nearly 100%, while PPV drops to 80%.

3.3.7 Re-analysis of bulk *vs.* single cell colony assay data demonstrates that we are able to accurately identify mutations that are present in the same subclone

In a recent study by Jan *et al.* [114], hematopoietic stem cell (HSC) from several AML patients were sequenced to >20,000 depth to measure somatic mutation allele frequencies at several targeted loci. In addition, colonies grown from single cells separated from the sample were subjected to allele-specific SNV TaqMan assay[4] at the same SNV sites, resulting in direct observations of subclones within the tissue. We used the bulk AF values obtained from the sequencing data as input to our subclone reconstruction method, followed by our mutation co-localization prediction procedure. We then compared our co-localization predictions to the colony assay results. Please refer to Table 4.0.1 for a summary of input data types and major conclusions. Among four patient samples for which colony assay data was available, SU030 and SU008 did not yield conclusive results because the AFs at the tested sites were so low (well below 1%) that they were indistinguishable from measurement noise (Table 3.3.3). SU070 yielded a unique subclone structure that is in agreement with the structure identified by colony assay (Figure 3.3.8). SU048 (Figure 3.3.7) produced a result set of 48 viable subclone structures. Every structure supports that *TET2-E1375STOP* is the earliest event, followed by *SMC1A* and *ACSM1* (Figure 3.3.7 A, Table 3.3.4). With a co-localization calling threshold of 0.5, *TET2-D1384V*, *OLFM2* and *ZMYM3* co-localize with *TET2-E1375STOP* and *SMC1A*, which is in agreement with the conclusion in the original analysis by Jan *et al.* that AML precursor HSC cells contain double mutations (presumably forming a com-

[4] A quantitative PCR technique, using the TaqMan probe, for genotyping.

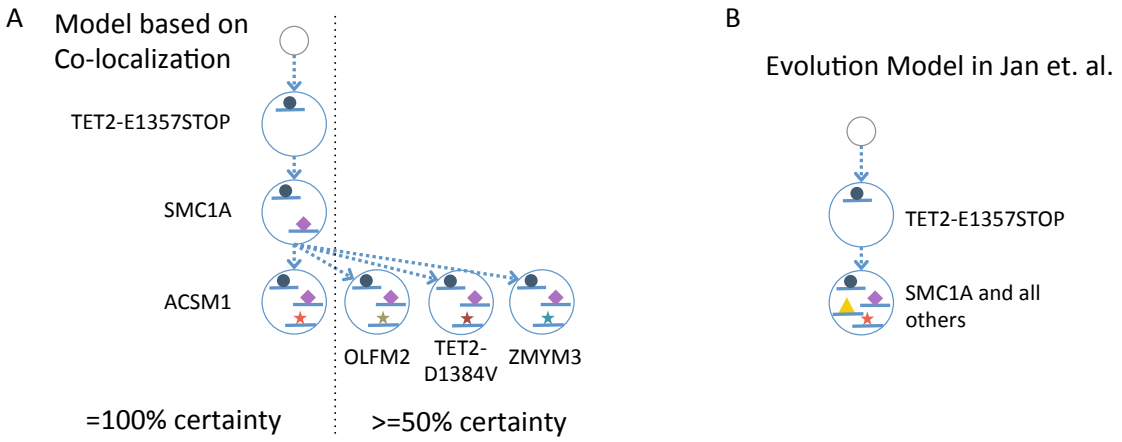


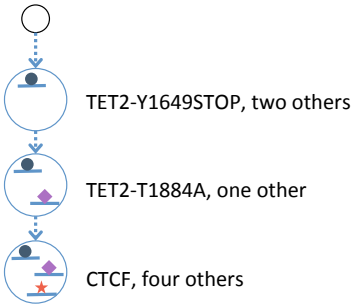
Figure 3.3.7: Analysis results on patient SU048 HSC sample in Jan *et al.* A) Our model of subclone evolution constructed based on co-localization probabilities. Left: Consensus structure supported by all subclone structures. Right: Consensus structure supported by at least 50% of subclone structures. B) Model of subclone evolution reported in Jan *et al.* constructed based on colony assay results.

A

SU070, 189 colonies assayed (Jan et. al.)	
# of colonies	Mutations
0	None
2	TET2-Y1649STOP, two others
35	Above plus TET2-T1884A and one other
152	Above plus CTCF and four others

B

Evolution Model in Jan et. al.



C

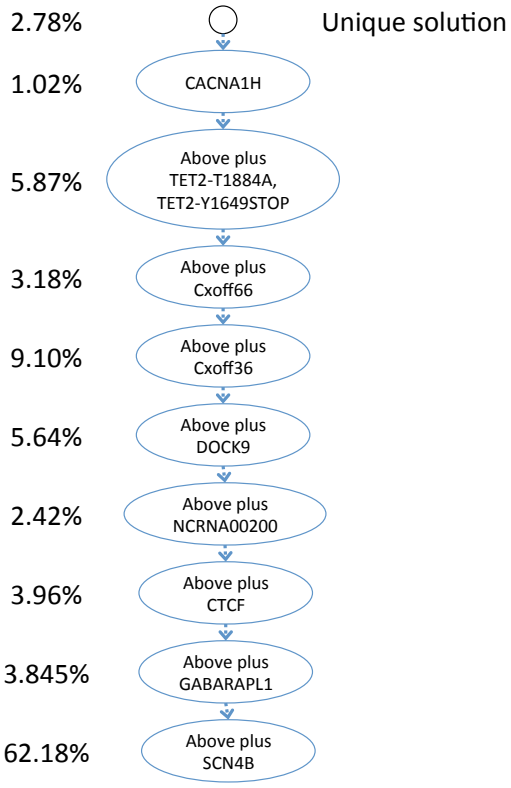


Figure 3.3.8: Reported and Analysis results on patient SU070 HSC sample in Jan *et al.* A) Colony assay results reported in Jan *et al.* B) Evolution Model reported in Jan *et al.* based on the colony assay results. C) The unique evolution tree constructed from the deep sequencing results on heterogeneous HSC sample

pound heterozygote) in the *TET2* gene. According to our analysis, *TET2-E1375STOP* and *SMC1A* are the two early events, and the two *TET2* mutations are already present in the same, early subclone. This is biologically sensible given that *TET2* is involved in DNA demethylation [165] and *SMC1A* in chromosome structure maintenance [166]. In addition, the depletion of *TET2* in mouse model leads to HSC expansion [167, 168], and the lack of *SMC1A* protein predicts poor survival in AML [169]. On the other hand, the relatively low co-localization probabilities among *ACSM1*, *TET2-D1384V*, *OLFM2* and *ZMYM3* suggest a branching structure for these mutations (Figure 3.3.7 A), rather than linear mutation accumulation consistent with the colony assay for this patient (the colony assay found one cell in which all these mutations are present). This points out the relatively weak power of our method to resolve co-localization among mutations with very low allele frequencies, as such low frequency mutations can be placed with relative freedom on multiple branches of the evolutionary tree.

3.4 Discussion

In this chapter I present a novel algorithm to elucidate tumor subclonal structure using as input CP values of individual, unlinked somatic mutations. This method is able to analyze many different types of genomic variant data, as long as AF measurements can be converted into CP values. Because bulk mutation frequency measurements from fragmentary sequence data or per-site microarray measurements do not retain "linkage" across such somatic variant sites, often there are many alternative subclone structures that can account for the input measurements. This method exhaustively enumerates all such viable subclone structures, tackling the short-comings of the parsimonious method described in Chapter 2. We were able to show that the number of solutions is usually much smaller than the theoretical upper limit. Often tumor tissues from multiple phases of tumor development (*e.g.* primary and relapse biopsies) are available. In such cases, the number of subclone structures that are not only consistent with the respective input

Patient	Mutation	Variant Allele Read count	Reference Allele Read count	Variant AF
SU008	SKP2	45937	624754	0.068492048
SU008	ELP2	1915	504335	0.003782716
SU008	PDZD3	161	100433	0.001600493
SU008	CNDP1	2238	475621	0.00468339
SU030	KCTD4	116061	2090267	0.052603693
SU030	SLC12A1	7754	1163598	0.006619701
SU048	ACSM1	16819	110087	0.132531165
SU048	NPM1	30	11079	0.002700513
SU048	OLFM2	13717	108695	0.112056008
SU048	PYHIN1	16	12952	0.001233806
SU048	SMC1A	181167	477095	0.275220201
SU048	TET2-D1384V	1797	15854	0.101807263
SU048	TET2-E1357STOP	7416	12117	0.379665182
SU048	ZMYM3	18518	288810	0.060254842
SU070	TET2-Y1649STOP	7732	8419	0.478731967
SU070	CXOFF36	3503	4537	0.435696517
SU070	CACNA1H	12083	12775	0.48608094
SU070	TET2-T1884A	4218	4552	0.480957811
SU070	CXOFF66	3678	4466	0.451620825
SU070	SCN4B	5086	11273	0.310899199
SU070	NCRNA00200	9199	16212	0.362008579
SU070	GABARAPL1	1648	3344	0.330128205
SU070	DOCK9	3382	5285	0.390215761
SU070	CTCF	10529	19561	0.349916916
SU070	PXDN	78	4712	0.016283925
SU070	TMEM20	157	14986	0.010367827
SU070	TMEM8B	69	7791	0.008778626

Table 3.3.3: Somatic Variations used in the re-analysis of the HSC targeted deep sequencing dataset in Jan *et al.*

	TET2-E1357STOP	SMC1A	ACSM1	OLFM2	TET2-D1384V
SMC1A	1				
ACSM1	1	1			
OLFM2	0.67	0.67	0.33		
TET2-D1384V	0.75	0.5	0.25	0.25	
ZMYM3	0.75	0.5	0.25	0.25	0.25

Table 3.3.4: Mutation co-localization frequency matrix for patient SU048 HSC targeted deep sequencing data from Jan *et al.* Mutations are sorted in descending order by AF.

frequency data but also across *e.g.* the primary and the relapse is lower, further trimming the "solution space", often to a single, unique structure. Using both simulations and experimental data, we have extensively characterized and validated our methods. We have illustrated with a number of datasets that this approach is often able to identify key patterns underlying tumor progression and relapse, including information to guide mutation prioritization.

In the case that the solution space cannot be further trimmed, we provide methods to derive useful knowledge, in terms of mutation cluster co-localization and timing. Our subclone structure enumeration procedure is exhaustive, and is free from the biases introduced by the choice of parameters or prior distributions often required for statistical sampling of the subclone structure solution space. We demonstrated that the co-localization and timing of mutations predicted from the HSC bulk targeted sequencing (Jan *et al.*) correlate well with their function, and can be used in a similar fashion to prioritize functional study.

The analysis of previously published datasets and our own datasets suggests that SubcloneSeeker will be applicable for a number of clinical / biological problems. Using serous ovarian cancer as an illustrative example, we have demonstrated that chemoresistance and relapse in this disease is a clonally driven process, and that such clones can be either present in the primary tumor or "arise" during progression or relapse. The patterns of temporal mutational order and cellular co-localization provide clinically relevant insight into the genomic basis for chemoresistance. In ovarian cancer, 80% of tumors are classified as chemosensitive while 20% of cancers progress during or recur shortly after platinum-based adjuvant chemotherapy. Unfortunately, there are no known genetic markers at present that can reliably predict inherent or acquired chemoresistance. This is likely the result of the complex and multi-factorial biological basis for this phenotype. However, whereas one or a small number of them may not be informative, analysis of many resistant clones and identification of the correspond-

ing mutational order and cellular co-localization may lead to a better understanding of chemoresistance, and form a rational basis for targeting the chemoresistant clones.

We envision similar utility for this type of analysis in advancing the current understanding of genomic alterations involved in the pre-malignant phases of cancer. Once again using ovarian cancer as a prototypical case, it has been established that *TP53* mutations are ubiquitous and early events in serous ovarian carcinogenesis [170]. However, the prevalence of other relapse somatic mutations is about 10% or less [170] suggesting that the additional requirements for transformation may be met through a combination of more diverse co-localized or temporally related somatic mutations (plus possible contributions from epigenetics and other molecular alterations, etc.). Thus genomic investigation of putative precursor lesion for serous carcinoma using approaches presented here is likely to identify subclonal hierarchies whose constituent mutations define cooperative classes on oncogenic event whose sum total results in malignant transformation.

(This page is left blank intentionally . . .)

4

Summary & Future Prospect

DESPITE DECADES OF EFFORT, cancer still remain as one of the deadliest diseases mankind struggles with. With the onset of high throughput genomic profiling technologies, we were granted, for the first time, the power to glimpse into the inner working of cancer genomics. However, it gradually became clear that a cancer biopsy, much in contrast to normal tissues, exhibits high intra-tumoral genomic heterogeneity, as in the cells in an entire tumor sample are divided into groups of genetically different subpopulations, or subclones. As it was indicated by several recent studies (Section 1.4), the interrogation of the degree of the heterogeneity, the genomic profiles of each subclone, and the evolution dynamics between samples (e.g. primary

and relapse) often holds the key to the further understanding of tumorigenesis, drug resistance, or metastasis.

We realized this challenge early on (Section 2.1), and designed method to reconstruct tumor subclone structure with a simplifying model that was biologically motivated (Section 2.2). The method always returns a parsimonious solution that represents a linear subclonal heritage. Although certain ambiguity exists, the method will result in accurate structure for somatic events that are existing in more than 50% of the tumor cells. Two dataset representing different cancer types, ovarian serious carcinoma (OV) and Intracranal Germ Cell Tumors (IGCT), were analyzed with the method, and the results are presented in Section 2.3 and Section 2.4. In the case of the OV dataset, based on whether the relapse sample contained more subclones than the primary, the patients were classified into two groups with trending significant difference in survival after the second surgery. For the IGCT dataset, the analysis helped elucidating the complexity of this rare disease [149]. This work was to our knowledge the first attempt to tackle the subclonality issue with whole genome microarray and Next Generation Sequencing (NGS) dataset, and our method was presented at *The Cancer Genome Atlas (TCGA) 1st Annual Scientific Symposium* as an oral presentation (Nov, 2011).

One apparent weakness of the method mentioned above is that it only return one structure, when in fact multiple structures may result in the same input data that is the observed somatic events. We developed an extended version of the method to enumerate all possible structures followed by trimming (Section 3.2). Given n somatic mutations (or mutation clusters), there are in total $n!$ potential structures. Not all of them are biologically plausible, as in order to achieve the same cell prevalence values found in the observation, some subclones in some structures need to be assigned a negative subclone frequency (SF). These structures are discarded from any further consideration. Simulation study showed that the size of the solution space is tightly restricted by this property (Section 3.3.2).

Realizing that related samples, such as the primary and relapse tumor samples from the same patient, often represent different time points of the same evolution process, we take advantage of these extra samples to further trim the solution space (Section 3.2.3.3). The reanalysis of published dataset showed a successful example in which the primary data alone resulted in 6 equivalent structures, yet only 1 of them was compatible to the unique relapse structure (Section 3.3.4). Utilizing a similar approach, we showed that how subclone structure can be of potential help in identifying the "top" candidates, among many somatic events, in the search for driver mutations which result in chemoresistance (Section 3.3.5)

In cases when the solution space cannot be reduced to a single, unique solution, we have developed a statistical framework to treat all the remaining structures as a distribution, and identify the probability any two mutations (or mutation clusters) co-localizing in the same subclone. We showed an example in which our prediction of co-localization correlated well with the biological functions of the mutated genes (Section 3.3.7), and provided discussion in how it could be of value in understanding chemoresistance, and in designing personalized treatments (Section 3.4).

That concludes all the constituents that are part of this dissertation work. The rest of this chapter will be devoted to my thoughts regarding the future of cancer subclonality research.

4.1 More accurate subclone structure reconstruction

As it was pointed out in Chapter 3, often there exist multiple mathematically equivalent and biologically plausible subclone structures for the same observation data. In some situations, further trimming with extra samples (Section 3.2.3.3), or deriving mutation co-localization knowledge (Section 3.2.4) is possible, but ultimately it would be desirable to reduce this ambiguity to the minimal. Due to the fact that the linkage information is lost during large scale genomic profiling (Section 2.5), this is impossible

Referenced in	Cancer Type	Sample Type	Data Type	Method	Major Conclusion
Section 2.3	OV	triplets	CNV events based on data from Agilent 415K / 1M microarray as well as WGS (x19 median coverage, Illumina GA-IIx, Appendix Section A.2.1)	Linear evolution model (Chapter 2)	Each patient followed a unique history of tumor relapse; Patients with higher heterogeneity in relapse exhibited less survival rate.
Section 2.4	IGCT	Primary tumor only	LOH events based on Illumina HumanOmni2.5-8 BeadChip Kit (microarray)	Linear evolution model (Chapter 2)	IGCT samples investigated were largely polyclonal (∼71%)
Section 3.3.4	AML	triplets	SNV clusters in Primary AF — Relapse AF space, based on deep sequencing (>30x) of validated SNVs	Exhaustive enumeration (Chapter 3)	The exhaustive enumeration method was capable of reproduce, and in one case improve upon, results based on manual reasoning from published datasets
Section 3.3.5	OV	triplets	SNV clusters in Primary AF — Relapse AF space, based on SNVs called from whole exome sequencing dataset	Exhaustive enumeration (Chapter 3)	In several patients, there were apparent clusters of SNVs that are likely to be responsible for chemoresistance, exhibiting the value of the method in variant prioritization
Section 3.3.7	AML	Tumor precursor HSCs	SNVs from targeted deep sequencing (>20,000x), treated as clusters of their own; Single cell colony assay results serving as "truth" set	Exhaustive enumeration (Chapter 3)	Co-localization prediction correlates well with the "truth" set, as well as their biological functions, suggesting that co-localization prediction can be of value in functional studies when the number of solutions are high, and cannot be further trimmed

Table 4.0.1: Summary of the clinical datasets analyzed in this work

with a pure mathematically approach. However, new experiments should be designed with intra-tumoral heterogeneity in mind, and gather as much information as possible regarding cellular linkage.

4.1.1 Collect additional data types from the same sample

In Chapter 3, the concept of Cell Prevalence (CP) was introduced as in what percent of all cells does a specific somatic variation exist. Since the method heavily relies on the clusters identified with the CP values, it is thus crucial to have accurate CP estimations, something that is not easily achieved. Imagine that a SNV with 40% reads supporting the alternate allele would corresponds to 80% CP, yet it is only true in copy number neutral region in which two copies of the locus exist in all cells. The same observation would correspond to 40% CP should the SNV fall within a heterozygous deletion region where only one copy of the locus exists. With this instance, it is made clear that the copy number information at the locus of the SNV in question, or allele specific copy number state, will be of tremendous value in correcting the CP estimation. Consequently, data gathered without copy number estimation in mind, such as exome sequencing, would suffer from difficulties in CP correction. It is therefore important, for future cancer genomic profiling experiments, to incorporate as many types of data as possible. The same sample, being investigated simultaneously by whole genome sequencing, whole exome sequencing, and Array comparative genome hybridization (aCGH), would yield much better and conclusive results in subclone structure reconstruction, as it allows better correction of CP values, as well as more data points be observed given the unified framework (Section 3.2.1).

4.1.2 Design experiments that specifically consider tumor subclonality

Often the amount of cells used for DNA sample preparation is minuscule compared to the entire tumor biopsy. It is therefore difficult to guarantee that the DNA sample pro-

vides an accurate representation of the distribution of all (if not only partial) subclones. There have been studies based on multiple spatially separated samples [118]. Other studies took serial samples at multiple time points throughout the course of the disease [106, 109, 115]. These spatially and / or temporally segregated samples, when pooled together, or analyzed separately before merging, would provide much representative view on the tumor subclone dynamics with much greater resolution.

Single cell colony assay is another fruitful path when it comes to mutation timing. In Section 3.3.7, we utilized a dataset by Jan *et al.* [114], which contains targeted deep sequencing on single cell colonies of HSCs, to demonstrate how our method was able to correctly predict the co-localization of mutations. Genomic profiling on colonies derived from single cells would allow direct observation on the linkage between mutations, and would consequently provide valuable inputs to the solution trimming (Section 3.2.3.3) step.

4.2 The impact of new technologies on the problem of subclone reconstruction

Before the dawn of large scale genomic profiling techniques, such as aCGH and NGS, it was impossible to interrogate cancer genomes into finer details, let alone the realization and attempts to reconstruct the subclone structure. A wave of new technologies are now on the horizon that promises further advancement. Two specific techniques, single molecule sequencing and single cell sequencing, are of particular interest.

4.2.1 Single molecule sequencing

One of the fundamental sources of ambiguity during subclone structure reconstruction comes from the loss of linkage information as in whether two independently identified variation events were from the same cell. A potential remedy, albeit with limited power,

is to check if a single sequencing read spans the loci of both events. If the reads spanning the loci either contain both events or none of them, the events were from the same cells. Otherwise, if the events were mutually exclusive on the same read, chances are that different subclones independently harbor one of these events. However, limited by the length of the current generation of NGS, which is no longer than 250 bp [171], and the somatic mutation rate of ~1 to ~100 per Mb in the whole exome [9], the number of reads spanning two somatic events is too low to be of any practical use.

However, with the emergence of single molecule sequencing [172–177] methods that promise much longer reads (> 1000 bp), the significance of the aforementioned signal would be much more applicable. In addition, longer reads would also enable modified versions of *de novo* genome assembly algorithms to assemble subclone genomes directly from the sequencing data. New methods will need to be developed to handle mapping, assembling, variant calling, and subclone structure reconstructing problems with the unique properties of the new sequencing data types, but the result will be much more accurate and less ambiguous.

4.2.2 Single cell sequencing

Further up the culprit of uncertainty in structure reconstruction is the fact that the cells, which are the ultimate unit of asexual inheritance, are broken down before genomic profiling. It is akin to the problem of phylogenetics, only the individual "species" are mixed together, and from which a single observation is obtained. Should we be able to obtain genomic profiles of individual cells, or "species", we would then be able to tap into the vast knowledge in the field of molecular phylogenetics [178, 179] and phylogenomics [180, 181]. Single cell sequencing (SCS) [182–187], an emerging technology, provides just the mean. Several studies that utilizing SCS to investigate tumor heterogeneity were mentioned in Section 1.4, and there have already been reports on methods that reconstruct evolution history of tumors using SCS data [188–190].

Although success stories are many, SCS could pose a different type of inaccuracy. In most of the studies, the number of cells investigated is between 20 to 100, which cannot guarantee an unbiased sampling on the underlying tumor population that often contain cells orders of magnitude more. Each individually determined genome, though, would serve as good source of confirming or denying whether events co-localize in the same cell, and help in the process of trimming equivalent structures resulting from the analysis of bulk sequencing data.

4.3 Concluding remarks

Cancer is a formidable foe the humanity faces together. 40 years after the declaration of "War on Cancer" by then U.S. President Richard Nixon [191–194], it still remains as a terminal disease. According to the 2010 United States Cancer Statistics [195], the combined incidence rate among the top 10 cancer types is 458.2 per 100,000, or roughly 1 in every 200 people, and the death rate is 1 in every 1,000. With the explosion of high throughput genomic profiling technologies, we are starting to peek behind the curtain and for the first time realizing the complexity of cancer genome. It is to my most sincere wish that the work this dissertation presented, along with many others, would be ultimately of help in shedding lights on the mystery of tumorigenesis, metastasis, drug resistance, and other cancer related mechanisms, as well as facilitate the push in advancing cancer treatments, such as personalized therapeutic strategy design, so that mankind could ultimately be rid of the dire fate cancer brings.

A

Supplemental Materials

Materials that are relevant to this work, yet don't really fit anywhere else, are organized here.

A.1 Comparison of performance among TrAp, PhyloSub and SubcloneSeeker, and example of SubcloneSeeker utilizing CNV data based on microarray

As mentioned in Chapter 3, while all three methods (TrAp, PhyloSub, and SubcloneSeeker) attempt subclone reconstructions, TrAp and PhyloSub require as input raw al-

lele counts at individual Single Nucleotide Variation (SNV) sites, whereas SubcloneSeeker expects Cell Prevalence (CP) estimates, ideally of clusters of variants with a shared CP value. Using each method as prescribed by their authors, TrAp and SubcloneSeeker are both able to refine the results originally published by Ding *et al.* [113]; PhyloSub and SubcloneSeeker were both used to analyze the hematopoietic stem cell (HSC) bulk sequencing + single cell colony assay dataset from Jan *et al.* [114] and produced comparable results.

We further tested and compared the performance of these packages on a dataset consisting of ultra-deep sequencing based read count data at a set of 21 validated SNVs from primary / relapse ovarian tumor samples with matched normal tissues. As TrAp and PhyloSub are designed to work on the "raw" allele count measurements, we first provided this input to each of these methods. The TrAp method ran out of memory, and provided no output, which we assume is because the method is not able to handle such high number of individual SNVs. PhyloSub did produce output that, we fear, was minimally informative to a user wishing to understand the resulting subclone structures (Figure A.1.1).

A.1.1 Subclone Reconstruction by TrAp and PhyloSub, using raw 454 sequencing read counts for each SNVs

We first attempted to perform subclone reconstruction using the raw read counts of 21 validated somatic SNVs with 738x median and 1080x mean coverage, as this is the format these packages are designed to take as their input. However, TrAp[152] (v0.3) issued an OutOfMemory error with 4G memory allocated to the JVM, and PhyloSub [153] (commit 540fdfb003, as of Jun 17, 2014) produced a partial order plot that made little sense due to the high number of nodes and edges. The data used for the analysis is shown in Table A.1.1

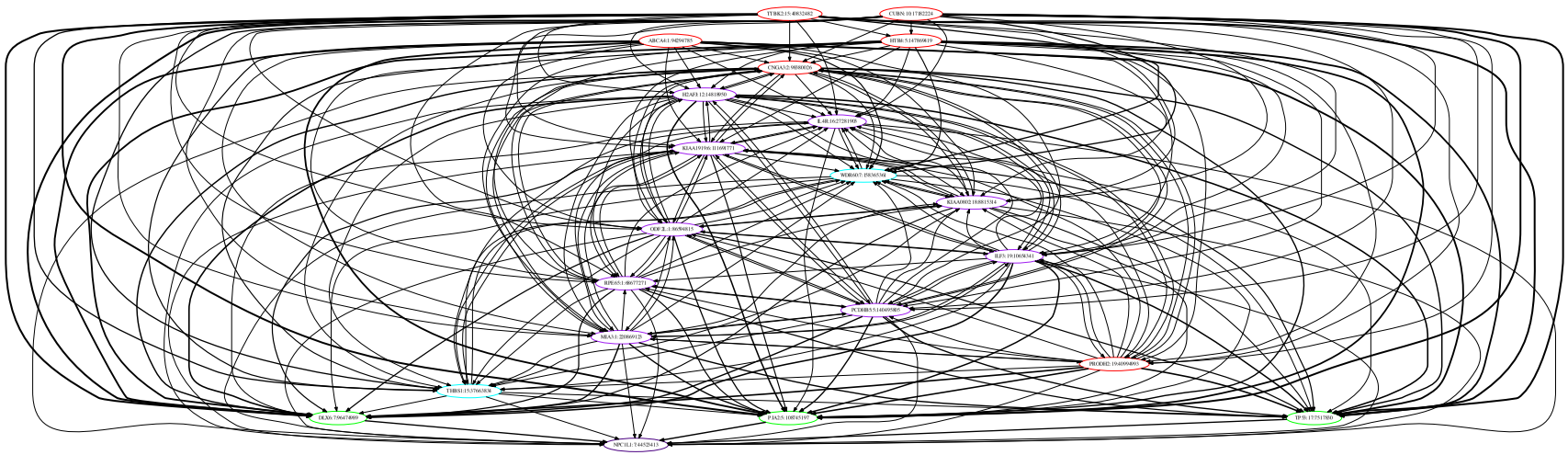


Figure A.1.1: PhyloSub partial order plot on the raw SNV read count data of TCGA-13-0913

Mutation ID	Primary AF	Relapse AF	Primary TotCov[a]	Relapse VarCov[b]	Relapse TotCov	Relapse VarCov
ABCA4:1:94294785	0.480536913	0.336842105	745	358	475	160
ADHFE1:8:67535201	0	0.366812227	N/A	N/A	458	168
ASNA1:19:12717512	0	0.428093645	N/A	N/A	1196	512
CNGA3:2:98380026	0.444897959	0.392380952	490	218	525	206
CSRNP2:12:49744690	0	0.425101215	N/A	N/A	1235	525
CUBN:10:17182224	0.49580574	0.390004097	2265	1123	2441	952
DLX6:7:96474989	0.434621493	0.370212766	1889	821	470	174
H2AFJ:12:14818950	0.472072072	0.422423556	555	262	883	373
HTR4:5:147869419	0.473563218	0.399074074	3045	1442	1080	431
IGFBP3:7:45923532	0	0.302281369	N/A	N/A	526	159
IL4R:16:27281903	0.438202247	0	534	234	N/A	N/A
ILF3:19:10654341	0.456818182	0.467445743	440	201	599	280
KIAA0802:18:8815314	0.443645084	0.374193548	417	185	620	232
KIAA1919:6:111691771	0.412048193	0.344155844	415	171	462	159
KIRREL3:11:125799964	0	0.397341211	N/A	N/A	677	269
MIA3:1:220869123	0.455463728	0.378734622	1089	496	1138	431
NPC1L1:7:44523413	0.065945946	0.357142857	925	61	924	330
ODF2L:1:86594815	0.459533608	0.418088737	729	335	586	245
PCDHB5:5:140495805	0.440585009	0.416526138	547	241	593	247
PJA2:5:108745197	0.356521739	0	460	164	N/A	N/A
PRODH2:19:40994993	0.474254743	0.35915493	738	350	852	306
RPE65:1:68677271	0.44295302	0.377440347	894	396	461	174
S1PR2:19:10196436	0	0.375714286	N/A	N/A	700	263
THBS1:15:37663834	0.447455387	0.346749226	1513	677	646	224
TP53:17:7517830	0.443045564	0.445190157	1668	739	894	398
TP63:3:191073389	0	0.391791045	N/A	N/A	536	210
TTBK2:15:40832482	0.515151515	0.35391924	627	323	421	149
WDR60:7:158365361	0.446931408	0.419091967	1385	619	859	360

Table A.1.1: SNVs used for assessing the performance of TrAp and Phylosub

[a]Total Coverage
[b]Variant Coverage

A.1.2 Subclone Reconstruction by SubcloneSeeker, using SNV clusters

We clustered the same 21 SNVs in Primary allele frequency (AF) — Relapse AF space, and identified 4 clusters (Figure A.1.2). SubcloneSeeker produced two structures with the primary clusters and one solution with the relapse clusters. One of the primary structures was trimmed away during the primary / relapse tree merging, resulting in a unique subclone structure for this patient.

A.1.3 SubcloneSeeker's unique ability to perform structure reconstruction on additional data types

We obtained Copy Number Variation (CNV) segments from TCGA-13-0913 microarray level 2 probe intensity data, and clustered them in Primary CP — Relapse CP space. The reconstruction result (Figure A.1.3) suggests the same conclusion as the SNV data does (Figure A.1.2, ancestral, as well as more recent, subclones in the primary are present in the relapse.), although the exact structure for the primary tumor sample differs. This is potentially due to that, although these two datasets were from the same patient, the DNA samples are different preparations, resulting in different sampling on the underlying tumor cell population, and consequently would not necessarily correspond to the same subclone structure / fraction distribution, or that each could be providing a partial view on the overall subclone structure.

A.2 Additional Materials and Methods

A.2.1 Sequencing procedure for the TCGA Ovarian Serious Carcinoma dataset

A.2.1.1 Illumina library construction

DNA samples were constructed into Illumina Paired-end libraries according to a modified version of the manufacturer's protocol (Paired-End Sample Preparation Guide, Part

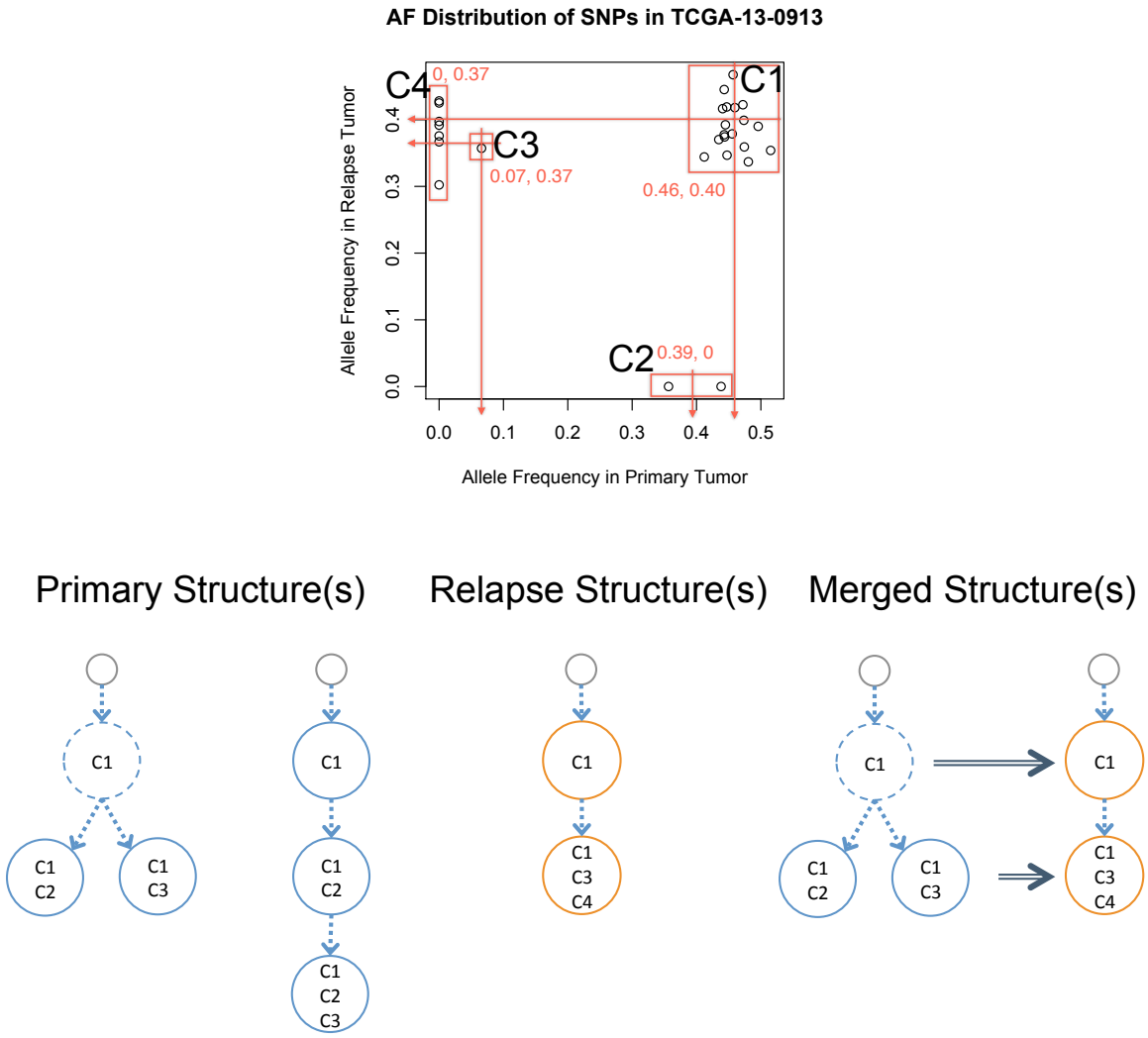


Figure A.1.2: Subclone structure reconstruction results based on SNV clusters of TCGA-13-0913. Top) The clusters, as well as their centroid allele frequency values. Bottom) The primary, relapse, and merged primary / relapse pair structures identified by SubcloneSeeker.

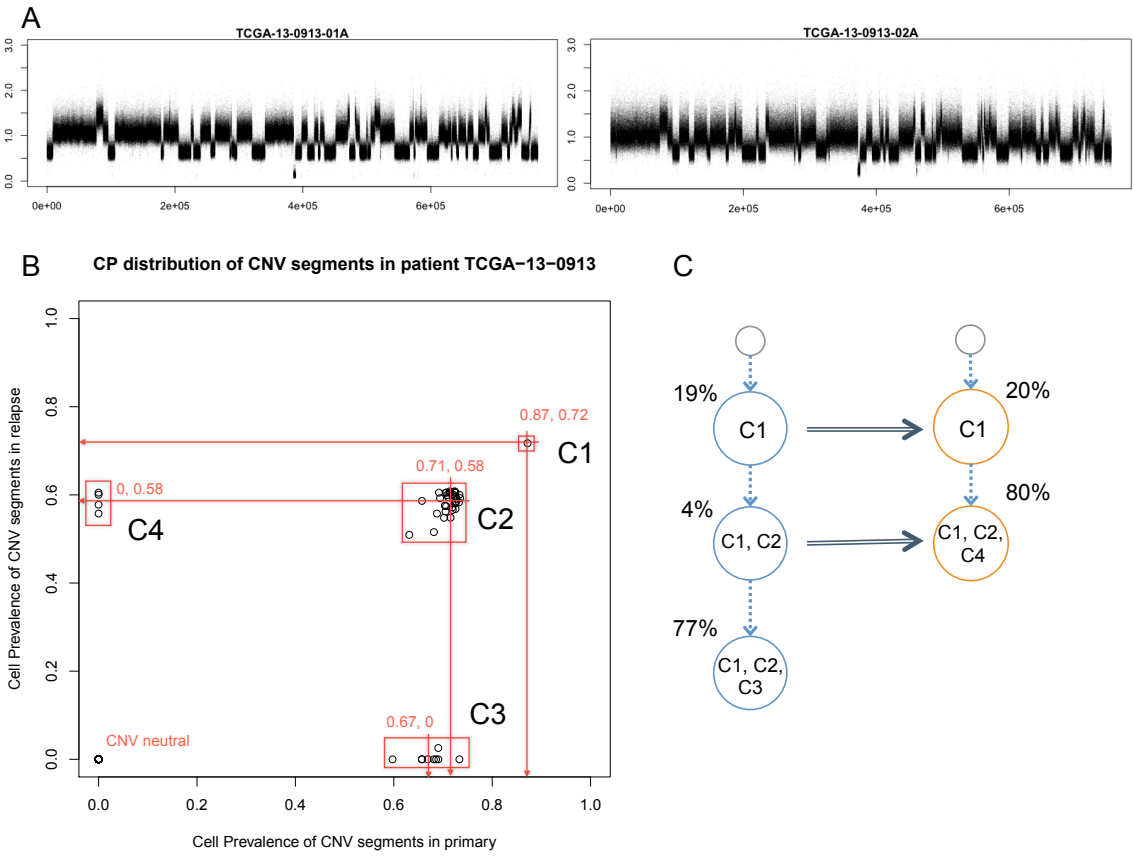


Figure A.1.3: Subclone structure reconstruction results based on microarray CNV clusters of TCGA-13-0913. A) Probe Intensity plot of both the primary and relapse tumor samples. B) CNV segments clustered in Primary CP — Relapse CP space. C) Subclone structure and relapse pattern from the identified clusters.

no. 1005063). Briefly, 500ng of native DNA was sheared into 200-500 bp fragments by nebulization followed by end-repair, 3'-end adenylation and ligation of the Illumina PE adapters using the Illumina Paired-End DNA Sample Prep Kit (Part no. PE-102-1001). Fragments with sizes between 290 and 350 bps were selected using 2% agarose gel electrophoresis. Ligation Mediated-PCR was performed for 18 cycles of amplification using primers and enzyme mix supplied in the sample preparation kit. Purification was performed with the QIAquick PCR purification kit (Qiagen, Part no. 28106) after enzymatic reactions. Following the final PCR purification, quantification and size distribution of the PCR products were determined using the Agilent Bioanalyzer 2100 DNA 7500 chip.

A.2.1.2 ILLUMINA DNA SEQUENCING

Library templates were prepared for sequencing using Illumina's cBot cluster generation system with TruSeq PE Cluster Generation Kits (Part no. PE-401-1001). Briefly, these libraries were denatured with sodium hydroxide and diluted to 3-6 pM in hybridization buffer in order to achieve a load density of ~800K clusters / mm2. Each library was loaded in 3 lanes of a flow cell, and each lane was spiked with 2% phiX control library for run quality control. The sample libraries then underwent bridge amplification to form clonal clusters, followed by hybridization with sequencing primer. Sequencing runs were performed in paired-end mode using the Illumina HiSeq 2000 platform. Using the TruSeq SBS Kits (Part no. FC-401-1001), sequencing-by-synthesis reactions were extended for 101 cycles from each end. Real Time Analysis software was used to process the image analysis and base calling. Sequencing runs generated approximately 80-120 million successful reads (2x100bp) on each lane of a flow cell, yielding ~60Gb per sample.

A.2.1.3 ILLUMINA READ MAPPING

Illumina reads were aligned to Human NCBI Build 36 using BWA (bwa-0.5.9rc1). Default parameters are used for alignment except for a 40 bp seed sequence, 2 mismatches in the seed, and a total of 3 mismatches allowed. BAM files generated from alignment of Illumina sequencing reads were preprocessed using GATK to recalibrate and locally realign reads.

A.2.2 SUPPLEMENTAL METHODS REGARDING DATA ACQUISITION FOR THE IGCT SNP ARRAY DATASET

DNA copy number analysis were performed using the high resolution Illumina HumanOmni2.5-8 (Omni2.5) BeadChip Kit (Illumina). In brief, 200ng genomic DNA was first denatured by NaOH. After nebulization of the sample, isothermal whole genome amplification was conducted to uniformly increase the DNA amount. The amplified DNA was enzymatically fragmented and hybridized to BeadChip for 16–24 h at 48 °C. After washing off unhybridized and non-specifically hybridized DNA fragments, allele-specific single-base extension reaction was performed to incorporate labeled nucleotides into the bead-bound primers. Following multi-layer staining to amplified signals from the labeled extended primers and final washing and coating, beadchips were imaged using the Illumina iScan system. SNV calls were collected using the Illumina GenomeStudio Version 2011.1 Genotyping Module 1.9.4. For improved CNV analysis, B allele frequencys (BAFs) were calculated and probe intensity $\log_2 R$ ratios (LRRs) were extracted after re-clustering the raw data by applying the GenomeStudio clustering algorithms.

(This page is left blank intentionally . . .)

Glossary

Agilent The company that produces Human Genome CGH Microarrays. Details regarding the Agilent 1M and 415K platform can be found at http://www.genomics.agilent.com/. 47, 49, 50

aneuploid Having a ploidy that is other than 2. 48, 71

angiogenesis The development of new blood vessels. 2

BAM A binary file format widely used for storing sequencing reads alignments.. 20

chromothripsis a catastrophic phenomenon that the chromosomes appear to be shattered and then stitched back together. 3

diploid Having two copies of the complete genome.. 45, 55, 56, 69, 71

emPCR emulsion-based PCR. 10, 12

ER-positive Endocrine receptor (estrogen or progesterone receptor) positive. 3

kataegis a hypermutation region characterized by multiple base substitutions. 3

non-Hodgkin lymphoma Any of a large group of cancers of lymphocytes (http://www.cancer.gov/cancertopics/types/non-hodgkin). 7

platinum therapy Chemotherapy with cisplatin as the anti-neoplastic reagent. 53

ploidy The number of copies of a complete genome in a cell. Normal cells have two copies, thus diploid. Tumor cells with copy number variation could potentially contain three copies, or triploid, or more.. 17, 26, 48, 50, 68

qPCR quantitative real time PCR. 14

TaqMan assay A quantitative PCR technique, using the TaqMan probe, for genotyping.. 97

teratoma a tumor composed of tissues not normally presented at the site (http://en.wikipedia.org/wiki/Teratoma). 59

tetraploid Having four copies of the complete genome.. 50

triplet A set of samples consisting of the normal, primary tumor and relapse tumor biopsies.. 50, 108

triploid Having three copies of the complete genome.. 50

trisomy having three instances of a particular chromosome, instead of the normal two. 16

List of Abbreviations

ACC Adrenocortical carcinomas. 6

aCGH Array comparative genome hybridization. 9, 67, 68, 109, 110

ACN absolute copy number. 25, 26, 45, 66, 68

AF allele frequency. 13, 64, 66, 68, 87, 90, 97, 100, 108, 117

AID activation-induced deaminase. 3

AML acute myeloid leukemia. 15, 54, 87, 97, 100, 108

APOBEC apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like. 3

BAF B allele frequency. 54–56, 58–60, 121

BI Broad Institute. 2

CCR Complex Chromosomal Rearrangement. 7

CGH Comparative Genomic Hybridization. 8, 9

CLF co-localization frequency matrix. 83

CLL Chronic lymphocytic leukemia. 8, 15, 16

CNV Copy Number Variation. 2, 6, 13, 20, 25, 27, 48, 49, 58, 60, 63, 64, 66–68, 70, 81, 108, 117, 121

CP Cell Prevalence. 25, 26, 42, 57–61, 63–70, 80, 81, 83, 86–88, 90, 93, 100, 109, 114, 117

CRC colorectal carcinoma. 4

ddNTP dideoxynucleotide. 9, 10

dNTP deoxynucleotide. 9–11

EM Expectation Maximization. 64

ESCC Esophageal squamous cell carcinoma. 6

FFT Fast Fourier Transform. 20

FISH fluorescence *in situ* hybridization. 8

HGP Human Genome Project. 2

HSC hematopoietic stem cell. 15, 97, 100, 102, 108, 110, 114

ICGC International Cancer Genome Consortium. 2

IGCT Intracranal Germ Cell Tumors. 54, 60, 106, 108

INDEL Insertion or Deletion. 2

LOH loss-of-heterozygosity. 55, 57–60, 63, 64, 66–70, 108

LRR probe intensity $\log_2 R$ ratio. 55, 121

M-FISH multiplex-FISH. 8

mBAF mirrored B allele frequency. 56–59, 68

MCL Mantle cell lymphoma. 7, 16

MCMC Markov Chain Monte Carlo. 64

MM multiple myeloma. 14

MPN myeloproliferative neoplasm. 16

NCI National Cancer Institute. 2

NGS Next Generation Sequencing. 2, 5, 9, 10, 12, 13, 16, 18, 20, 60, 64, 106, 110, 111

NSCLC non-small-cell lung cancer. 4, 5

OV ovarian serious carcinoma. 106, 108

PI probe intensity. 68

PPi pyrophosphate. 11

RCN relative copy number. 67, 68

RD read depth. 20, 23, 46, 67, 68

RDR read depth ratio. 20, 23, 25, 60

SCC squamous cell carcinoma. 4, 5

SCLC small-cell lung cancer. 4, 5

SCS single cell sequencing. 13, 111, 112

SF subclone frequency. 28, 42, 70, 72, 80, 81, 85, 86, 106

SKY spectral karyotyping. 8

SNP Single Nucleotide Polymorphism. 6, 20, 60

SNUC Sinonasal Undifferenciated Carcinoma. 87, 89

SNV Single Nucleotide Variation. 2, 12–14, 64, 66–69, 71, 81, 97, 108, 109, 114, 117, 121

sPLC secondary plasma cell leukemia. 14

SV Structural Variation. 66

TCGA The Cancer Genome Atlas. 4, 6, 7, 20, 60, 106

TNBC triple-negative breast cancer. 15

WGS whole genome sequencing. 67, 68, 108

Bibliography

[1] Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

[2] Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173 (1976).

[3] Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300**, 143–149 (1982).

[4] Becker, W. M., Kleinsmith, L. J., Hardin, J. & Raasch, J. *The world of the cell* (San Francisco, CA, 2009), seventh edn.

[5] Weinberg, R. *The biology of cancer* (Garland Science, 2014).

[6] Macconaill, L. E. & Garraway, L. A. Clinical implications of the cancer genome. *J. Clin. Oncol.* **28**, 5219–5228 (2010).

[7] Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

[8] Wang, E. Understanding genomic alterations in cancer genomes using an integrative network approach. *Cancer Lett.* **340**, 261–269 (2013).

[9] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

[10] Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

[11] Crasta, K. *et al.* DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482**, 53–58 (2012).

[12] Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).

[13] Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).

[14] Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* **2**, e00534 (2013).

[15] Meyerson, M. & Pellman, D. Cancer genomes evolve by pulverizing single chromosomes. *Cell* **144**, 9–10 (2011).

[16] Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589–593 (2012).

[17] Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59–71 (2012).

[18] Cazier, J. B. & Tomlinson, I. General lessons from large-scale studies to identify human cancer predisposition genes. *J. Pathol.* **220**, 255–262 (2010).

[19] Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).

[20] Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14**, 703–718 (2013).

[21] Boehm, J. S. & Hahn, W. C. Towards systematic functional characterization of cancer genomes. *Nat. Rev. Genet.* **12**, 487–498 (2011).

[22] Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

[23] Goto, T., Marusawa, H. & Chiba, T. Landscape of genetic aberrations detected in human colorectal cancers. *Gastroenterology* **145**, 686–688 (2013).

[24] Wang, K. *et al.* Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat. Genet.* **43**, 1219–1223 (2011).

[25] Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* **46**, 573–582 (2014).

[26] Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).

[27] Seo, J. S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.* **22**, 2109–2119 (2012).

[28] Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

[29] Vignot, S. *et al.* Next-generation sequencing reveals high concordance of recurrent somatic alterations between primary tumor and metastases from patients with non-small-cell lung cancer. *J. Clin. Oncol.* **31**, 2167–2172 (2013).

[30] Rudin, C. M. *et al.* Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.* **44**, 1111–1116 (2012).

[31] Peifer, M. *et al.* Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.* **44**, 1104–1110 (2012).

[32] Pietanza, M. C. & Ladanyi, M. Bringing the genomic landscape of small-cell lung cancer into focus. *Nat. Genet.* **44**, 1074–1075 (2012).

[33] Assie, G. *et al.* Integrated genomic characterization of adrenocortical carcinoma. *Nat. Genet.* **46**, 607–612 (2014).

[34] Wandoloski, M., Bussey, K. J. & Demeure, M. J. Adrenocortical cancer. *Surg. Clin. North Am.* **89**, 1255–1267 (2009).

[35] Soon, P. S. & Sidhu, S. B. Molecular basis of adrenocortical carcinomas. *Minerva Endocrinol.* **34**, 137–147 (2009).

[36] Lin, D. C. *et al.* Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat. Genet.* **46**, 467–473 (2014).

[37] Weinstein, J. N. *et al.* Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).

[38] Zhang, J. *et al.* The genomic landscape of mantle cell lymphoma is related to the epigenetically determined chromatin state of normal B cells. *Blood* **123**, 2988–2996 (2014).

[39] Forment, J. V., Kaidi, A. & Jackson, S. P. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat. Rev. Cancer* **12**, 663–670 (2012).

[40] Kridel, R. *et al.* Whole transcriptome sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma. *Blood* **119**, 1963–1971 (2012).

[41] Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).

[42] Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52 (2012).

[43] Wang, L. *et al.* SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* **365**, 2497–2506 (2011).

[44] Ferreira, P. G. *et al.* Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res.* **24**, 212–226 (2014).

[45] Speicher, M. R., Gwyn Ballard, S. & Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat. Genet.* **12**, 368–375 (1996).

[46] Schrock, E. *et al.* Multicolor spectral karyotyping of human chromosomes. *Science* **273**, 494–497 (1996).

[47] Shayesteh, L. *et al.* PIK3CA is implicated as an oncogene in ovarian cancer. *Nat. Genet.* **21**, 99–102 (1999).

[48] Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).

[49] Shaw-Smith, C. *et al.* Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.* **41**, 241–248 (2004).

[50] Shaffer, L. G. *et al.* The identification of microdeletion syndromes and other chromosome abnormalities: cytogenetic methods of the past, new technologies for the future. *Am J Med Genet C Semin Med Genet* **145C**, 335–345 (2007).

[51] Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).

[52] Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).

[53] Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).

[54] Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13**, 2399–2412 (1985).

[55] Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).

[56] Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).

[57] McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).

[58] Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363, 365 (1998).

[59] Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).

[60] Nyren, P. The history of pyrosequencing. *Methods Mol. Biol.* **373**, 1–14 (2007).

[61] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

[62] Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinformatics* **11**, 473–483 (2010).

[63] Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**, 2711–2718 (2012).

[64] Abel, H. J. & Duncavage, E. J. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet* **206**, 432–440 (2013).

[65] Koboldt, D. C., Larson, D. E., Chen, K., Ding, L. & Wilson, R. K. Massively parallel sequencing approaches for characterization of structural variation. *Methods Mol. Biol.* **838**, 369–384 (2012).

[66] Hall, I. M. & Quinlan, A. R. Detection and interpretation of genomic structural variation in mammals. *Methods Mol. Biol.* **838**, 225–248 (2012).

[67] Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6**, 13–20 (2009).

[68] WU, X. & XIAO, H. Progress in the detection of human genome structural variations. *Sci. China, C, Life Sci.* **52**, 560–567 (2009).

[69] Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).

[70] Altmann, A. *et al.* A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.* **131**, 1541–1554 (2012).

[71] Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).

[72] Green, M. R. *et al.* Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood* **121**, 1604–1611 (2013).

[73] Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).

[74] Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

[75] Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

[76] Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).

[77] Roth, A. *et al.* JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics* **28**, 907–913 (2012).

[78] Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

[79] Hansen, N. F., Gartner, J. J., Mei, L., Samuels, Y. & Mullikin, J. C. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics* **29**, 1498–1503 (2013).

[80] Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).

[81] Kassahn, K. S. *et al.* Somatic point mutation calling in low cellularity tumors. *PLoS ONE* **8**, e74380 (2013).

[82] Xu, H., DiCarlo, J., Satya, R. V., Peng, Q. & Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **15**, 244 (2014).

[83] Kim, S. Y., Jacob, L. & Speed, T. P. Combining calls from multiple somatic mutation-callers. *BMC Bioinformatics* **15**, 154 (2014).

[84] Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 13081–13086 (2008).

[85] Shipitsin, M. *et al.* Molecular definition of breast tumor heterogeneity. *Cancer Cell* **11**, 259–273 (2007).

[86] Benetkiewicz, M. *et al.* Chromosome 22 array-CGH profiling of breast cancer delimited minimal common regions of genomic imbalances and revealed frequent intra-tumoral genetic heterogeneity. *Int. J. Oncol.* **29**, 935–945 (2006).

[87] Fujii, H., Marsh, C., Cairns, P., Sidransky, D. & Gabrielson, E. Genetic divergence in the clonal evolution of breast cancer. *Cancer Res.* **56**, 1493–1497 (1996).

[88] Glockner, S., Buurman, H., Kleeberger, W., Lehmann, U. & Kreipe, H. Marked intratumoral heterogeneity of c-myc and cyclinD1 but not of c-erbB2 amplification in breast cancer. *Lab. Invest.* **82**, 1419–1426 (2002).

[89] Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).

[90] Teixeira, M. R., Pandis, N., Bardi, G., Andersen, J. A. & Heim, S. Karyotypic comparisons of multiple tumorous and macroscopically normal surrounding tissue samples from patients with breast cancer. *Cancer Res.* **56**, 855–859 (1996).

[91] Torres, L. *et al.* Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res. Treat.* **102**, 143–155 (2007).

[92] Macintosh, C. A., Stower, M., Reid, N. & Maitland, N. J. Precise microdissection of human prostate cancers reveals genotypic heterogeneity. *Cancer Res.* **58**, 23–28 (1998).

[93] Alvarado, C. *et al.* Somatic mosaicism and cancer: a micro-genetic examination into the role of the androgen receptor gene in prostate cancer. *Cancer Res.* **65**, 8514–8518 (2005).

[94] Konishi, N. *et al.* Intratumor cellular heterogeneity and alterations in ras oncogene and p53 tumor suppressor gene in human prostate carcinoma. *Am. J. Pathol.* **147**, 1112–1122 (1995).

[95] Gonzalez-Garcia, I., Sole, R. V. & Costa, J. Metapopulation dynamics and spatial heterogeneity in cancer. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13085–13089 (2002).

[96] Samowitz, W. S. & Slattery, M. L. Regional reproducibility of microsatellite instability in sporadic colorectal cancer. *Genes Chromosomes Cancer* **26**, 106–114 (1999).

[97] Giaretti, W. *et al.* Intratumor heterogeneity of K-ras2 mutations in colorectal adenocarcinomas: association with degree of DNA aneuploidy. *Am. J. Pathol.* **149**, 237–245 (1996).

[98] Coons, S. W., Johnson, P. C. & Shapiro, J. R. Cytogenetic and flow cytometry DNA analysis of regional heterogeneity in a low grade human glioma. *Cancer Res.* **55**, 1569–1577 (1995).

[99] Mora, J., Cheung, N. K. & Gerald, W. L. Genetic heterogeneity and clonal evolution in neuroblastoma. *Br. J. Cancer* **85**, 182–189 (2001).

[100] Califano, J. *et al.* Genetic progression model for head and neck cancer: implications for field cancerization. *Cancer Res.* **56**, 2488–2492 (1996).

[101] Sauter, G., Moch, H., Gasser, T. C., Mihatsch, M. J. & Waldman, F. M. Heterogeneity of chromosome 17 and erbB-2 gene copy number in primary and metastatic bladder cancer. *Cytometry* **21**, 40–46 (1995).

[102] Fujii, H. *et al.* Frequent genetic heterogeneity in the clonal evolution of gynecological carcinosarcoma and its influence on phenotypic diversity. *Cancer Res.* **60**, 114–120 (2000).

[103] Horvai, A. E., DeVries, S., Roy, R., O'Donnell, R. J. & Waldman, F. Similarity in genetic alterations between paired well-differentiated and dedifferentiated components of dedifferentiated liposarcoma. *Mod. Pathol.* **22**, 1477–1488 (2009).

[104] Pantou, D. *et al.* Cytogenetic manifestations of multiple myeloma heterogeneity. *Genes Chromosomes Cancer* **42**, 44–57 (2005).

[105] Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473 (2006).

[106] Keats, J. J. *et al.* Clonal competition with alternating dominance in multiple myeloma. *Blood* **120**, 1067–1076 (2012).

[107] Egan, J. B. *et al.* Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. *Blood* **120**, 1060–1066 (2012).

[108] Walker, B. A. *et al.* Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4;14) and t(11;14) myeloma. *Blood* **120**, 1077–1086 (2012).

[109] Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997 (2014).

[110] Melchor, L. *et al.* Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. *Leukemia* (2014).

[111] Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

[112] Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).

[113] Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).

[114] Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci Transl Med* **4**, 149ra118 (2012).

[115] Schuh, A. *et al.* Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120**, 4191–4196 (2012).

[116] Lundberg, P. *et al.* Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. *Blood* **123**, 2220–2228 (2014).

[117] Bea, S. *et al.* Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18250–18255 (2013).

[118] Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

[119] Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).

[120] Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).

[121] Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

[122] Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).

[123] Assie, G. *et al.* SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. *Am. J. Hum. Genet.* **82**, 903–915 (2008).

[124] Bengtsson, H., Irizarry, R., Carvalho, B. & Speed, T. P. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759–767 (2008).

[125] Goransson, H. *et al.* Quantification of normal cell fraction and copy number neutral LOH in clinical lung cancer samples using SNP array data. *PLoS ONE* **4**, e6057 (2009).

[126] Lamy, P., Andersen, C. L., Dyrskjot, L., Torring, N. & Wiuf, C. A Hidden Markov Model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. *BMC Bioinformatics* **8**, 434 (2007).

[127] Nancarrow, D. J., Handoko, H. Y., Stark, M. S., Whiteman, D. C. & Hayward, N. K. SiDCoN: a tool to aid scoring of DNA copy number changes in SNP chip data. *PLoS ONE* **2**, e1093 (2007).

[128] Peiffer, D. A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16**, 1136–1148 (2006).

[129] Yamamoto, G. *et al.* Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am. J. Hum. Genet.* **81**, 114–126 (2007).

[130] Yu, G. *et al.* BACOM: in silico detection of genomic deletion types and correction of normal cell contamination in copy number data. *Bioinformatics* **27**, 1473–1480 (2011).

[131] Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**, 1586–1592 (2009).

[132] Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).

[133] Ivakhno, S. *et al.* CNAseg–a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* **26**, 3051–3058 (2010).

[134] Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16910–16915 (2010).

[135] Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).

[136] Bowtell, D. D. The genesis and evolution of high-grade serous ovarian cancer. *Nat. Rev. Cancer* **10**, 803–808 (2010).

[137] Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12**, 323–334 (2012).

[138] Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

[139] Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377–1380 (2008).

[140] Castellarin, M. *et al.* Clonal evolution of high-grade serous ovarian carcinoma from primary to recurrent disease. *J. Pathol.* **229**, 515–524 (2013).

[141] Cooke, S. L. *et al.* Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma. *Oncogene* **29**, 4905–4913 (2010).

[142] Packer, R. J., Cohen, B. H., Cooney, K. & Coney, K. Intracranial germ cell tumors. *Oncologist* **5**, 312–320 (2000).

[143] Matsutani, M. *et al.* Primary intracranial germ cell tumors: a clinical analysis of 153 histologically verified cases. *J. Neurosurg.* **86**, 446–455 (1997).

[144] Kamakura, Y., Hasegawa, M., Minamoto, T., Yamashita, J. & Fujisawa, H. C-kit gene mutation: common and widely distributed in intracranial germinomas. *J. Neurosurg.* **104**, 173–180 (2006).

[145] Sakuma, Y. *et al.* c-kit gene mutations in intracranial germinomas. *Cancer Sci.* **95**, 716–720 (2004).

[146] Rickert, C. H., Simon, R., Bergmann, M., Dockhorn-Dworniczak, B. & Paulus, W. Comparative genomic hybridization in pineal germ cell tumors. *J. Neuropathol. Exp. Neurol.* **59**, 815–821 (2000).

[147] Schneider, D. T. *et al.* Molecular genetic analysis of central nervous system germ cell tumors with comparative genomic hybridization. *Mod. Pathol.* **19**, 864–873 (2006).

[148] Terashima, K. *et al.* Genome-wide analysis of DNA copy number alterations and loss of heterozygosity in intracranial germ cell tumors. *Pediatr Blood Cancer* **61**, 593–600 (2014).

[149] Wang, L. *et al.* Novel somatic and germline mutations in intracranial germ cell tumours. *Nature* (2014).

[150] Staaf, J. *et al.* Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.* **9**, R136 (2008).

[151] Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).

[152] Strino, F., Parisi, F., Micsinai, M. & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.* **41**, e165 (2013).

[153] Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).

[154] Zare, H. *et al.* Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10**, e1003703 (2014).

[155] Fischer, A., Vazquez-Garcia, I., Illingworth, C. J. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep* **7**, 1740–1752 (2014).

[156] Hajirasouliha, I., Mahmoody, A. & Raphael, B. J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics* **30**, 78–86 (2014).

[157] Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).

[158] Day, N., Hemmaplardh, A., Thurman, R. E., Stamatoyannopoulos, J. A. & Noble, W. S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**, 1424–1426 (2007).

[159] Gusnanto, A., Wood, H. M., Pawitan, Y., Rabbitts, P. & Berri, S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28**, 40–47 (2012).

[160] Thomas, M., Brabanter, K. D. & Moor, B. D. New bandwidth selection criterion for Kernel PCA: Approach to dimensionality reduction and classification problems. *BMC Bioinformatics* **15**, 137 (2014).

[161] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. & Ruzzo, W. L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987 (2001).

[162] Liyanage, M. *et al.* Multicolour spectral karyotyping of mouse chromosomes. *Nat. Genet.* **14**, 312–315 (1996).

[163] Purdue, P. E., Zhang, J. W., Skoneczny, M. & Lazarow, P. B. Rhizomelic chondrodysplasia punctata is caused by deficiency of human PEX7, a homologue of the yeast PTS2 receptor. *Nat. Genet.* **15**, 381–384 (1997).

[164] Takahashi, Y. *et al.* Genomic Characterization of Sinonasal Undifferentiated Carcinoma. *Journal of Neurological Surgery Part B: Skull Base* **75**, A084 (2014).

[165] Hu, L. *et al.* Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545–1555 (2013).

[166] Schmiesing, J. A. *et al.* Identification of two distinct human SMC protein complexes involved in mitotic chromosome dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12906–12911 (1998).

[167] Abdel-Wahab, O. *et al.* Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. *Blood* **114**, 144–147 (2009).

[168] Quivoron, C. *et al.* TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25–38 (2011).

[169] Homme, C. *et al.* Low SMC1A protein expression predicts poor survival in acute myeloid leukemia. *Oncol. Rep.* **24**, 47–56 (2010).

[170] Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).

[171] MacConaill, L. E. Existing and emerging technologies for tumor genomic profiling. *J. Clin. Oncol.* **31**, 1815–1824 (2013).

[172] Ohshiro, T. *et al.* Single-molecule electrical random resequencing of DNA and RNA. *Sci Rep* **2**, 501 (2012).

[173] Zwolak, M. & Di Ventra, M. Electronic signature of DNA nucleotides via transverse transport. *Nano Lett.* **5**, 421–424 (2005).

[174] Lagerqvist, J., Zwolak, M. & Di Ventra, M. Fast DNA sequencing via transverse electronic transport. *Nano Lett.* **6**, 779–782 (2006).

[175] Lagerqvist, J., Zwolak, M. & Di Ventra, M. Influence of the environment and probes on rapid DNA sequencing via transverse electronic transport. *Biophys. J.* **93**, 2384–2390 (2007).

[176] Rajavelu, A., Jurkowska, R. Z., Fritz, J. & Jeltsch, A. Function and disruption of DNA methyltransferase 3a cooperative DNA binding and nucleoprotein filament formation. *Nucleic Acids Res.* **40**, 569–580 (2012).

[177] Scholz, S., Liebler, E. K., Eickmann, B., Fritz, H. J. & Diederichsen, U. Variation of the intercalating proline in artificial peptides mimicking the DNA binding and bending IHF protein. *Amino Acids* **43**, 289–298 (2012).

[178] Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13**, 303–314 (2012).

[179] Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L. & Tamura, K. Statistics and truth in phylogenomics. *Mol. Biol. Evol.* **29**, 457–472 (2012).

[180] Forterre, P. & Gadelle, D. Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. *Nucleic Acids Res.* **37**, 679–692 (2009).

[181] Dagan, T. Phylogenomic networks. *Trends Microbiol.* **19**, 483–491 (2011).

[182] Eberwine, J., Sul, J. Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat. Methods* **11**, 25–27 (2014).

[183] Nawy, T. Single-cell sequencing. *Nat. Methods* **11**, 18 (2014).

[184] Korfhage, C., Fisch, E., Fricke, E., Baedker, S. & Loeffert, D. Whole-genome amplification of single-cell genomes for next-generation sequencing. *Curr Protoc Mol Biol* **104**, Unit 7.14 (2013).

[185] Kohn, A. B., Moroz, T. P., Barnes, J. P., Netherton, M. & Moroz, L. L. Single-cell semiconductor sequencing. *Methods Mol. Biol.* **1048**, 247–284 (2013).

[186] Lasken, R. S. Single-cell sequencing in its prime. *Nat. Biotechnol.* **31**, 211–212 (2013).

[187] Yilmaz, S. & Singh, A. K. Single cell genome sequencing. *Curr. Opin. Biotechnol.* **23**, 437–443 (2012).

[188] Kim, K. I. & Simon, R. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics* **15**, 27 (2014).

[189] Ren, S. C., Qu, M. & Sun, Y. H. Investigating intratumour heterogeneity by single-cell sequencing. *Asian J. Androl.* **15**, 729–734 (2013).

[190] Potter, N. E. *et al.* Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Res.* **23**, 2115–2125 (2013).

[191] authors listed, N. The national cancer act of 1971. *J. Natl. Cancer Inst.* **48**, 577–584 (1972).

[192] Weinhouse, S. National Cancer Act of 1971–an editorial. *Cancer Res.* **32**, i–ii (1972).

[193] authors listed, N. National Cancer Act of 1971. Conference report. *Cancer* **29**, 917–923 (1972).

[194] Rauscher, F. J. Proceedings: The National Cancer Program and the National Cancer Act of 1971. *Natl Cancer Inst Monogr* **40**, 3–6 (1974).

[195] Group, U. C. S. W. United States cancer statistics: 1999–2010 incidence and mortality web-based report. *Atlanta, GA* (2013).