Stony Brook University

Inferring Tumor Progression from Genomic Heterogeneity

A Dissertation Presented

by

Nicholas E. Navin

to

The Graduate School

in Partial Fulfillment of the
Requirements
for the Degree of

Doctor of Philosophy

in

Molecular Genetics and Microbiology

Stony Brook University
August 2010

Stony Brook University

The Graduate School

Nicholas E. Navin

We, the dissertation committee for the above candidate for the

Doctor of Philosophy, hereby recommend
acceptance of this dissertation

Dr. Michael Wigler (Advisor)
Professor, Cold Spring Harbor Laboratory

Dr. Patrick Hearing (Chair)
Department of Molecular Genetics & Microbiology
Professor, Stony Brook University

Dr. Bruce Futcher
Department of Molecular Genetics & Microbiology
Professor, Stony Brook University

Dr. Scott Lowe
Professor, Cold Spring Harbor Laboratory

Dr. Scott Powers (Outside Member)
Associate Professor, Cold Spring Harbor Laboratory
This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Inferring Tumor Progression from Genomic Heterogeneity
by
Nicholas E. Navin

Doctor of Philosophy
in
Molecular Genetics and Microbiology

Stony Brook University
2010

Cancer progression in humans is difficult to infer because we do not routinely sample patients at multiple stages of their disease. However, heterogeneous breast tumors provide a unique opportunity to study human tumor progression because they still contain evidence of early and intermediate subpopulations in the form of the phylogenetic relationships. We developed a method we call Sector-Ploidy-Profiling to study the clonal composition of breast tumors. SPP involves macro-dissecting tumors, flow-sorting genomic subpopulations by DNA content, and profiling genomes using comparative genomic hybridization. Breast carcinomas display two classes of genomic structural variation: (1) monogenomic and (2) polygenomic. Monogenomic tumors appear to contain a single major clonal subpopulation with a highly stable chromosome structure. Polygenomic tumors contain multiple clonal tumor subpopulations, which may occupy the same sectors, or separate anatomic locations. In polygenomic tumors, we show that heterogeneity can be ascribed to a few clonal subpopulations, rather than a series of gradual intermediates.

While very informative, the SPP method yields only approximate results when applied to mixed populations of rapidly evolving cells. In such cases our understanding would be improved by dissecting genetic events at the single cell level. We therefore developed a method to quantify genomic copy number in single cells using next-generation sequencing. This method, single nucleus sequencing (SNS), involves flow-sorting single nuclei, whole genome

amplification and sequencing random DNA fragments. We validated our method in a normal fibroblast cell line that has been deep-sequenced along with a genetically complex breast cancer cell line. We then used SNS to analyze 100 single cells isolated from a heterogeneous basal-like breast carcinoma. From this data, we constructed a detailed phylogenetic lineage, showing that the majority of cells belong to one of five major clonal subpopulations. Additionally, we observed a subpopulation of pseudodiploid cells with random amplifications and deletions that are not present in the major aneuploid subpopulations and may represent an unstable precursor. Our data support a model of tumor progression by sequential clonal expansions to form the mass of the tumor.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

Michael Wigler and Jim Hicks

Wigler Laboratory
Jude Kendall, Jennifer Troge, Linda Rodgers, Yvonne Eberling, Kerry Cook

McCombie Laboratory
Richard McCombie, Laura Gelley, Elena Ghiban, Melissa Kramer

Zetterberg Laboratory
Anders Zetterberg, Susanne Maner, Par Lundin

CSHL FACS Facility
Pamela Moody, Tara Spencer

Thesis Committee
Scott Lowe, Scott Powers, Bruce Futcher, Patrick Hearing

Additional Acknowledgements
Michael Ronemus, Diane Esposito, Anthony Leotta, Yamrom Boris,
Patrick Blake, Nancy Navin

Funding Support:
NCI T32 Ruth L. Kirschstein Fellowship
Swedish Cancer Society
Department of the Army
Breast Cancer Research Foundation
Simons Foundation

INTRODUCTION

Defining the pathways through which tumors progress is critical to our understanding and treatment of cancer**.** We do not routinely sample patients at multiple time points during the progression of their disease, and thus our research on humans is limited to inferring progression *a posteriori* from the examination of a single tumor sample. Despite this limitation, inferring progression is possible because the tumor genome contains a natural history of the mutations that occur during the formation of the tumor mass.

In theory, there are two ways to infer progression from primary tumor genomes: (1) comparing different tumors, and (2) comparing clones within single tumors. Until recently most studies have used the former approach, which involves surveying single samples from archived tumor collections and cataloguing the order and frequency of genetic events (Figure 1.A). This approach has been widely applied to reconstruct progression in many different cancer types using large collection of tumors (Bilke et al., 2005; Hicks et al., 2006a; Hicks et al., 2005; Hoglund et al., 2005; Hoglund et al., 2002; Liu et al., 2009a; Pathare et al., 2009; Selvarajah et al., 2008). In these studies the underlying assumption is that mutations accumulate as the tumor progresses and only rarely are lost. Specific genetic lesions can thus be classified as early or late, relative to the total complexity of the tumor genome. The limitation to this approach is that while a few structural aberrations can be clearly classified as early, placement of high frequency events into ordered pathways has been problematic. Surgically resected tumors from archived collections represent relatively advanced cases with large numbers of genomic aberrations. With a few exceptions, the vast majority of mutational events occur at low frequency across tumor collections, indicating that each tumor travels down a unique mutational pathway.

Here, we present an alternative approach to studying tumor progression, by comparing multiple samples *within* an individual tumor (Figure 1.B). Many studies have suggested that breast cancers show significant heterogeneity in their

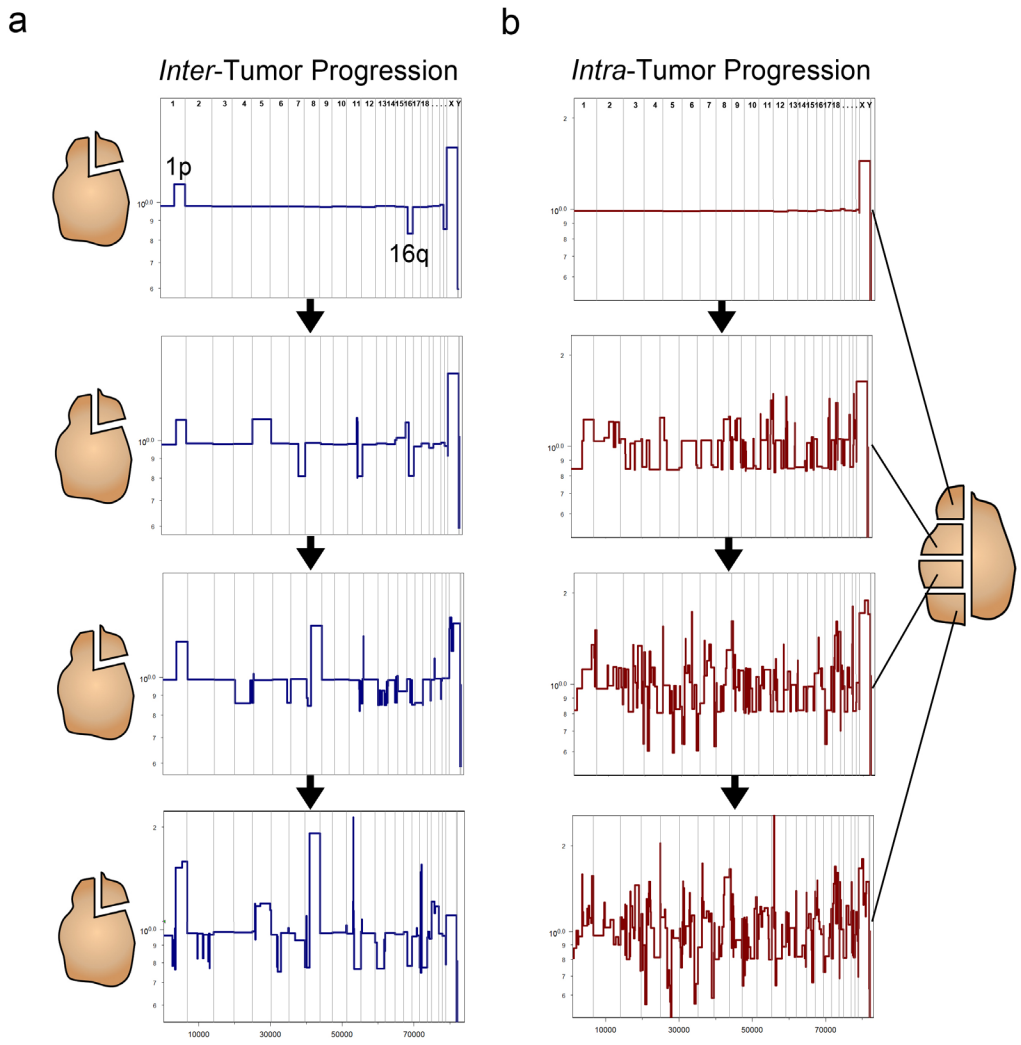


Figure 1.1 Inter and Intra-tumor Comparisons of Copy Number Profiles
(a) Inter-tumor comparisons. A single sample was resected from four different luminal A breast tumors and CGH profiles were measured and segmented. The profiles shown were ordered based on increasing genomic complexity. (b) Intra-tumor comparisons. Four samples were taken from a single heterogeneous basal-like breast carcinoma. Nuclei were isolated from each quadrant and samples were flow-sorted by ploidy, followed by microarray CGH profiling. The profiles are ordered based on increasing numbers of chromosome breakpoints.

genomic profiles, making it possible to identify clones that represent various time points in the progression of the tumor. By isolating and comparing tumor clones from a single tumor, we can reconstruct a detailed lineage of how the tumor developed, assuming that mutations are persistent and inherited between clones. This approach has the advantage of identifying mutations that belong to a unique mutational pathway within each tumor, leading to malignancy.

To study the clonal composition of solid tumors we developed a method called Sector-Ploidy-Profiling (SPP). SPP involves macro-dissecting tumors, flow-sorting genomic subpopulations by DNA content, and profiling genomes using comparative genomic hybridization. We applied this method to twenty ductal carcinomas and found that they display two classes of genomic structural variation: (1) monogenomic and (2) polygenomic. Monogenomic tumors appear to contain a single major clonal subpopulation with a highly stable chromosome structure. Polygenomic tumors contain multiple clonal tumor subpopulations, which may occupy the same sectors, or separate anatomic locations. In a single basal-like tumor, we investigated the topography of tumor clones at the single cell level using a cytological approach called fluorescence in situ hybridization (FISH). This method showed that divergent tumor clones are intermixed in tissues, raising interesting questions about cooperativity. By comparing multiple subpopulations from different anatomic locations, we have inferred pathways of cancer progression and the organization of tumor growth. In polygenomic tumors, we show that heterogeneity can be ascribed to a few clonal subpopulations, rather than a series of gradual intermediates, suggesting that the tumor grows by a series of clonal expansions.

While SPP could isolate tumor subpopulations by region and ploidy, the tumor sectors we analyzed consisted of mixed populations of millions of cells. It is possible that monogenomic and polygenomic tumors consist of mixtures of many different tumor clones rather than clonal subpopulations. In order to address this problem, we developed a method to quantify genomic copy number in single cells, called Single Nucleus Sequencing (SNS). This method involves flow-sorting single nuclei, whole genome amplification and sequencing random DNA fragments using next-generation sequencing to measure copy number by read depth. We validated our method in a normal fibroblast cell line that has been deep-sequenced along with a genetically complex breast cancer cell line. We then used SNS to analyze 100 single cells isolated from a heterogeneous, basal-

like, breast carcinoma. From this data, we constructed a detailed phylogenetic lineage, showing that the majority of cells belong to one of four major clonal subpopulations, that progress in a sequential pattern. Additionally, we observed a rare subpopulation of pseudo-diploid cells that contain random amplifications and deletions that are not present in the major aneuploid subpopulations and may represent an unstable precursor.

Together, our results show that breast tumors can grow by sequential clonal expansions (SCE). We present this as a general model for tumor progression, which relates to the clonal evolution model (Nowell, 1976). However, our model differs by assuming that gradual intermediates are rare, and that one or more major subpopulation clonally expands to form the tumor mass. Our model assumes that evolution occurs in bursts and that the genome is remarkably stable during the growth of the tumor mass. This model has important clinical implications, by suggesting that targeting the genomes of one or more major subpopulations will eradicate the majority of the tumor mass. Moreover, the methods that we developed open up new avenues for studying genomic mutations in single cells in cancer, and other human diseases.

CHAPTER 1

History of Tumor Heterogeneity

In breast cancer the malignant cells often arise from ductal tissue and are constrained by the duct structure until they begin to invade surrounding stromal tissue. They exhibit regions of growth, regions of hypoxia and necrosis and regions of interaction with blood vessels and lymph ducts. It would be surprising if all cells in a tumor were identical. As early as the 1800's, Rudolf Virchow and other early pathologists observed the morphological heterogeneity of tumor cells using the first compound microscopes (Brown and Fee, 2006). The subsequent development of sophisticated staining methods allowed pathologists to visualize and categorize the morphology of tumor cells in detail, and to score various characteristics including nuclear size, mitotic index and differentiated structures. These characteristics are used to score the grade of a tumor, which aids clinicians in determining how aggressively to treat a patient. However, many pathologists have noted that cells from different regions of a tumor differ in their morphological characteristics. (Fitzgerald, 1986; Hirsch et al., 1983; Kruger et al., 2003; van der Poel et al., 1997). Taking into account this heterogeneity, pathologists will examine many tissue sections from several regions of the tumor, but generally report only the highest grade for clinical treatment (Ignatiadis and Sotiriou, 2008; Komaki et al., 2006).

In the early 1980s, a new arsenal of tools was developed by cytogeneticists to investigate tumor heterogeneity at the genome level: chromosome G-banding, spectral karyotyping (SKY) and fluorescence in situ hybridization (FISH). A particularly large body of data concerning genetic heterogeneity comes from interphase FISH studies. Using specific DNA probes, FISH can reveal the copy number of a limited number of chromosomal loci across a large number of cells. By comparing the copy numbers of representative genomic loci using specific DNA probes across multiple tumor samples, various studies reported tumors as either 'homogeneous' (monoclonal) or 'heterogeneous' (polyclonal) (Farabegoli et al., 2001; Maley et al., 2006; Mora et al., 2001; Pantou et al., 2005; Roka et al., 1998; Sauter et al., 1995; Shipitsin et al., 2007; Teixeira et al., 1996; Zojer et al., 1998).

A more complete characterization of the tumor genome was obtained by visualizing metaphase chromosomes by Giemsa staining. The resulting G-banding karyotypes provided chromosome-specific landmarks and made it possible to accurately identify chromosome abnormalities in tumor genomes (Mitelman et al., 1997; Trent, 1985). As with FISH, it was observed that subpopulations of cells from the same tumor showed distinct sets of chromosomal rearrangements, indicating the presence of multiple clones (Coons et al., 1995; Pandis et al., 1995; Teixeira et al., 1996; Teixeira et al., 1995). Using this technique, recurrent chromosome events began to be catalogued, providing the first notion that such events might be ordered in tumor development.

The heterogeneity of tumors has since been repeatedly validated using various molecular markers, including mRNA expression (Bachtiary et al., 2006; Cole et al., 1999); protein expression (Allred et al., 2008; Johann et al., 2009); and DNA sequencing (Khalique et al., 2007; Lips et al., 2008). The question then becomes one of understanding the role of heterogeneity in tumor progression. A number of studies have shown that despite the genetic diversity in heterogeneous tumor, neighboring clones often share many common mutations (Maley et al., 2006; Navin et al., 2010; Pantou et al., 2005; Shipitsin et al., 2007; Teixeira et al., 1996; Torres et al., 2007). Thus it seems unlikely that genetic heterogeneity is simply the result of random unselected variation. Instead, heterogeneous clones may represent discrete time points in the progression of the disease.

With the advent of genomic techniques, such as microarrays and next-generation sequencing, it has become possible to survey the entire genome at much higher resolution than previously possible. Deep sequencing of heterogeneous tumors using next-generation sequencing has shown that some tumors contain more alleles than would be expected in single clones (Campbell et al., 2008; Shah et al., 2009). But it is difficult, if not impossible, to determine from sequence alone the number of clones present (and to which genomes the reads belong). As an alternative strategy, comparative genomic hybridization (CGH) microarrays can measure the precise location of chromosome breakpoints and the amplitude of copy number events that differ between divergent tumor subpopulations (Benetkiewicz et al., 2006; Navin et al., 2010; Shah et al., 2009; Shipitsin et al., 2007; Torres et al., 2007). This information can be used to track chromosome breakpoint markers as they are inherited and persist through successive subpopulations of clones that progress to form the tumor.

CHAPTER 2

General Models for Tumor Progression

Several general models have been proposed to explain tumor progression. These models make different assumptions concerning the proliferative capacity of the major populations of tumor cells and thus lead to testable predictions concerning their lineage. The first model for tumor progression to gain widespread acceptance appeared in a landmark theoretical paper by Peter Nowell in 1976, where he combined two seemingly unrelated fields: evolutionary biology and tumor biology (Nowell, 1976). Nowell proposed that tumor cells obey the laws of natural selection, undergoing positive selection when advantageous mutations occur and negative selection when deleterious mutations arise. The two major schemes based on this fundamental tenet are collectively referred to as clonal evolution. They share the common assumption that the majority of tumor cells have the potential to undergo unlimited proliferation, but differ in the number of clonal subpopulations that they predict will form the mass of tumor.

2.1 Monoclonal Evolution

The monoclonal evolution model states that solid tumors undergo a brief period of heterogeneity in the early stages of tumor progression, followed by a clonal expansion of a single population of cells, which forms the mass of the tumor (Figure 2.1A). It is assumed that a single clone undergoes positive selection and outcompetes all other subpopulations by the time the tumor is large enough to be detected. Evidence supporting the monoclonal evolution model originally came from methods that followed only a small number of traits, such as X-inactivation in tumors, RFLP analysis of carcinomas, plasma cell immunoglobulin sequences and microsatellite markers (Endoh et al., 2001; Fialkow, 1974; Linder and Gartler, 1965; Matsumoto et al., 2004; Noguchi et al., 1992, 1994; Sawada et al., 1994). Genomic data also supports this model in a subset of breast cancers by showing that multiple samples within the same tumors contain highly similar copy number profiles by CGH microarrays (Navin et al., 2010).

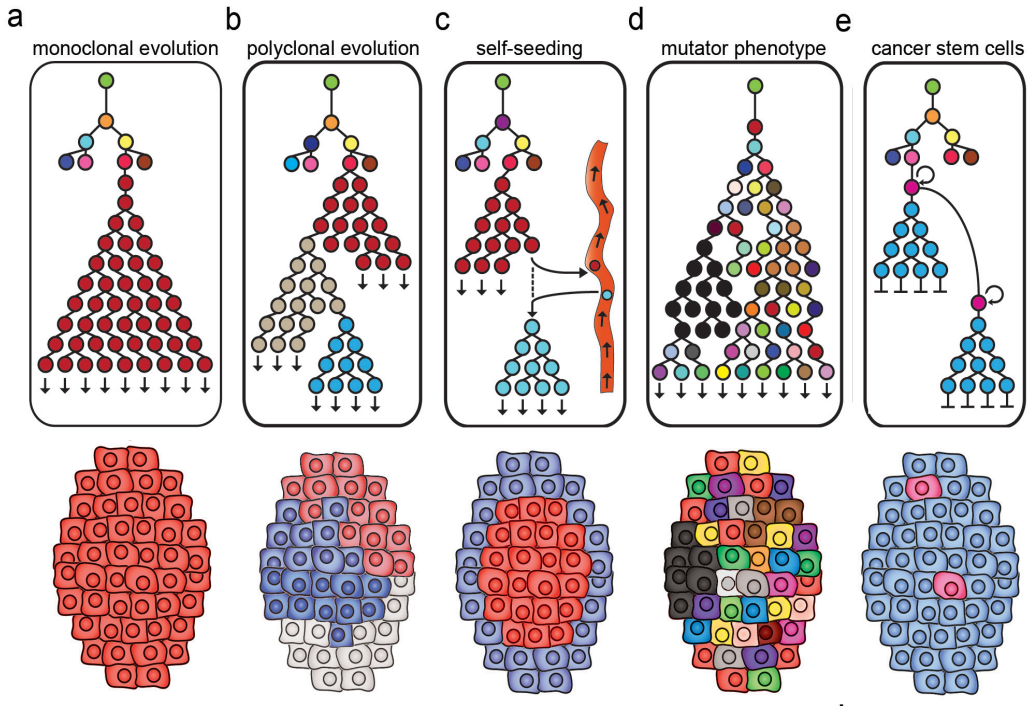


Figure 2.1 General Models for Tumor Progression
Root nodes are colored in green and represent the normal diploid cells from which the tumor arose. (a) In monoclonal evolution a burst of genomic instability results in the expansion of a single dominant clone, forming a monogenomic tumor (b) In polyclonal evolution a brief period of genomic diversity results in the clonal expansion of a few stable genomes to form a polygenomic tumor (c) In self-seeding the expansion of the primary tumor mass results in clones intravasating the circulatory system, diverging and returning to expand the peripheral mass of the primary (d) In the mutator phenotype evolution is driven by the accumulation of random mutations generating many diverse tumor clones, with few clonal expansions, resulting in a very heterogeneous tumor (e) In the cancer stem cell model, mutations lead to rare progenitor cell that continuously regenerate the tumor mass.

2.2 Polyclonal Evolution

In contrast, the polyclonal evolution model posits that solid tumors undergo an early period of heterogeneity followed by the expansion of multiple, divergent clones to form the mass of the tumor (Figure 2.1B). Empirical evidence supporting this model comes from a variety of studies including interphase FISH experiments, immunohistochemistry of tumor sections, gene expression studies and array CGH experiments (Aubele et al., 1999; Bachtiary et al., 2006). Recent experiments have supported the polyclonal evolution model by showing that genetic related clones with divergent genomes may cohabit the same tumor

(Navin et al., 2010; Shipitsin et al., 2007). This raises an interesting question about polyclonal evolution: do cohabiting clones within a single tumor suggest a cooperative relationship? In contrast to monoclonal evolution, in which a single dominant clone outcompetes all others, the polyclonal model implies an evolutionary advantage to cohabitation. In the tumor microenvironment resources – including oxygen, vasculature, stroma and growth factors - are scarce, so a selective advantage for having two or more clones seems highly plausible. The nature of their interaction may be mutualistic, commensal or perhaps even parasitic (Reviewed in (Marusyk and Polyak, 2009)). Clone interactions merit further study, as they may imply that targeting a single clone with therapy could lead to the rise or the demise of neighboring subpopulations.

2.3 Self-Seeding

In recent years, several variations of the polyclonal evolution model have been proposed, including the self-seeding hypothesis and the mutator phenotype. The self-seeding hypothesis posits that tumor clones leave the primary site, intravasate into the circulatory system, develop or subsist at a distant site for a period of time, then return to the primary tumor where they establish new subpopulations (Norton, 2008; Norton and Massague, 2006). This variation on the venerable 'seed and soil' theory (Paget, 1889) implies that circulating tumor cells have a homing mechanism that attracts them back to their site of origin. It also predicts that new clones will aggregate at the periphery of the tumor surface, or where vasculature leads into the tumor (Figure 2.1C) and that tumors are 'built' out of the sequential accretion of clones. Recently, homing behavior and self-seeding was demonstrated in a mouse model using both human tumor cell lines and pleural effusion cells (Kim et al., 2009). These investigators showed that specific cytokine attractants (IL-6 and IL-8) and mediators of infiltration (MMP1 and fascin-1) were integral factors in this self-seeding process. Among the various implications of these results, the authors raise the counter-intuitive notion that the presence of a primary tumor mass might act as a 'sponge' for circulating tumor cells, and by 'soaking them up' actually reduce the potential for distant metastases, the major cause of breast cancer mortality.

2.4 Mutator Phenotype

The mutator phenotype model, originally set forth by Lawrence Loeb (Loeb et

al., 1974), is related to polyclonal evolution but differs by proposing that tumors consist of a large diversity of small clones rather than a few dominant clonal subpopulations (Figure 2.1D). In this model the rate of random mutations in tumor cells is thought to increase drastically perhaps by the introduction of mutations into DNA polymerase itself (Bielas and Loeb, 2005; Loeb et al., 1974). Clonal expansions may occur, but a large diversity of tumor genomes are generated by random, non-expanded mutations. Evidence for this model comes largely from a DNA capture sequencing approach, from which it was estimated that the mutation rate increased to more than 200 fold in neoplastic tissues (Bielas et al., 2006). The mutator phenotype has also been extended to copy number changes, suggesting that tumor progression is driven by random, non-expanded amplifications and deletions that generate genomic instability (Heng et al., 2006a; Heng et al., 2006b). While this model differs in predicting a larger diversity of genetic clones, it shares the primary assumption of clonal evolution: that the majority of tumor cells have the potential to proliferate indefinitely.

2.5 Cancer Stem Cells

In the late 1990's an alternative model emerged that challenged the primary assumption of the previous models by assuming that only a minority of tumor cells could proliferate indefinitely. The cancer stem cell (CSC) hypothesis became widely accepted as the leading model for tumor progression. The CSC hypothesis posits that a rare population of stem cells within the solid tumor is the only subpopulation with the ability for unlimited proliferation (Figure 2.1E). The model assumes: (1) a rare population of cancer stem cells proliferate indefinitely, (2) the majority of tumor cells have limited proliferation, and (3) the rare cells continuously give rise to the major population. Cancer stem cells were originally believed to arise from normal stem cells, but it is now thought that any somatic cell may become a cancer stem cell (Clarke et al., 2006).

Evidence for the CSC hypothesis originally came from studying normal hematopoietic stem cells and the malignant stem cells that arise during leukemogenesis. The first empirical evidence came with the invention of fluorescence-activated cell sorting (FACS) which allowed the isolation of human leukemic stem cells using surface markers (Lapidot et al., 1994). These human cancer stem cells were reimplantated into immunocompromised mice, in which they were fully capable of initiating leukemia, while other reimplanted cancer

cells could not (Bonnet and Dick, 1997). The isolation and reimplantation assay has become the gold standard for identifying cancer stem cells and has been used to identify cancer stem cells in breast carcinomas (Al-Hajj et al., 2003) , brain tumors (Shen and Singh, 2004), colon cancers (O'Brien et al., 2007) and pancreatic tumors (Li et al., 2007). The CSC model is also attractive to clinicians, because it suggests that the entire tumor can be eradicated by targeting only the cancer stem cell population (Campbell and Polyak, 2007)

In summary, several general models have been proposed for tumor progression. We do not aim to resolve which of these models apply to breast cancer, but instead take an agnostic view. Using genomic heterogeneity to infer tumor progression we seek to identify pattern of tumor growth. This chapter serves as background information for the reader to understand the current general models for tumor progression. In the final chapters, we will discuss how these general models relate to the genomic patterns of tumor progression that we have inferred.

CHAPTER 3

Inferring Progression From Tumor Genomes

3.1 *Inter*-tumor Comparisons

Early studies of tumor genome progression involved longitudinal comparisons of the karyotypes of tumors from large collections (Heim and Mittleman 2009). The general theory was that tumor genomes with the fewest chromosomal aberrations contained the earliest mutations in tumor progression. An extension to this approach involved separating non-invasive precursors to breast cancer (DCIS) or low grade tumors and comparing them to high grade samples (Tsarouha et al., 1999). Most of these tumors were near-diploid and therefore easy to karyotype by G-banding. In tumors with few chromosomal aberrations, the most frequent event involved the gain of the entire 1q chromosome arm and the loss of the 16q arm. (Hoglund et al., 2002; Tsarouha et al., 1999). This combination of gain and loss seems to be the earliest event in some breast cancer and often occurs through pericentric recombination and the generation of an either 1q:16p translocation (followed by loss of the reciprocal product) or a 1q:1q isochromosome. These tumors were mostly hormone receptor- positive, and had the best prognoses. However, in these studies of tumor progression, the collections often consisted of a diverse mixture of subtypes, with each evolving down a different mutational pathway.

A milestone in understanding the diversity of breast cancers came with advances in gene expression microarrays. In 2000, Sørlie and Perou et al proposed that breast tumors could be classified into five different subtypes based on the expression of a few hundred mRNA transcripts. This stratification of breast cancer had important implications for studying tumor progression in that each subtype could be studied as an independent disease. The original six subtypes have now been refined to five: Luminal A, dominated by the ER+ tumors with the best prognosis; Luminal B, characterized as more advanced and often more genomically complex; Erbb2-like, often amplified at the *ERBB2* growth factor receptor locus; basal-like, most often negative for ER, PR and *ERBB2* (‘triple-negative’); and normal-like, with expression patterns most closely related

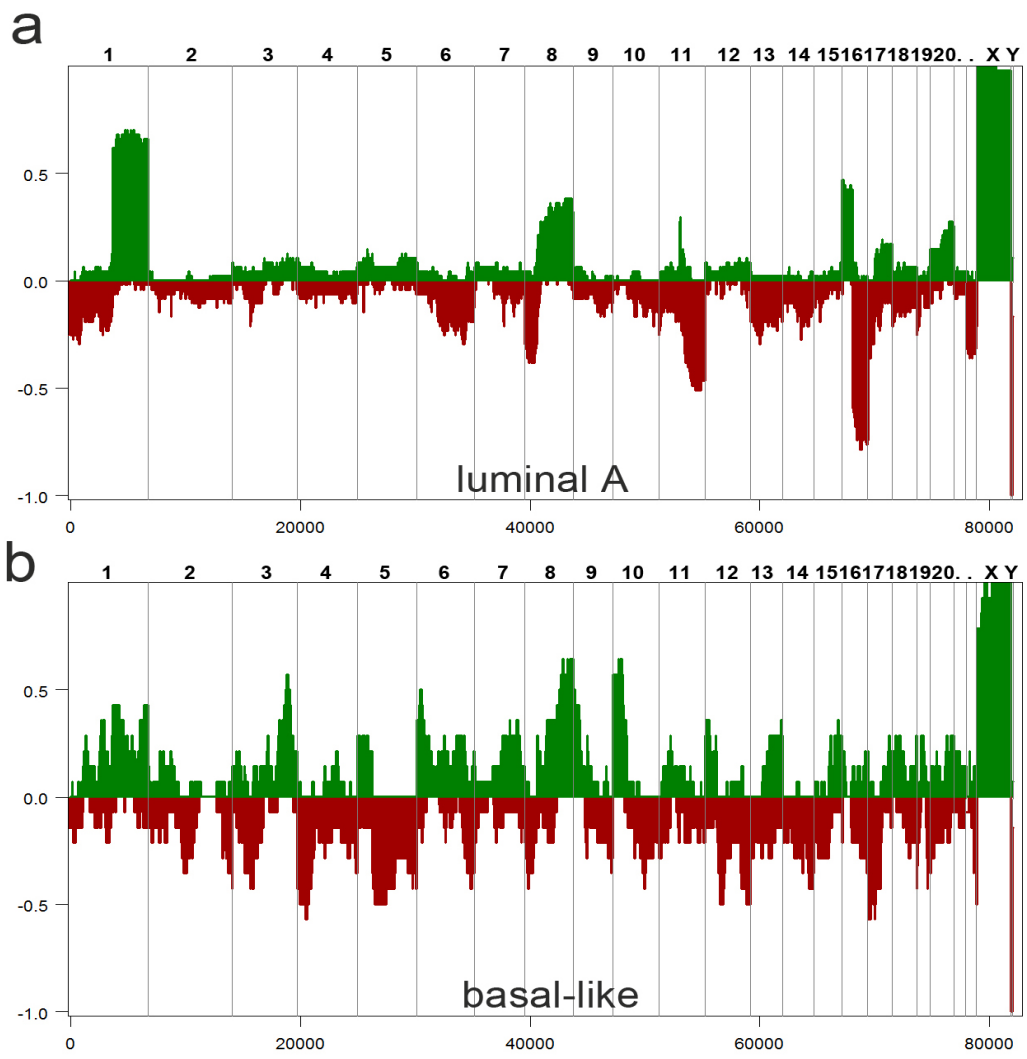


Figure 3.1 Frequency Plots of Luminal A and Basal-like Breast Tumors
Microarray CGH was used to generate profiles from collections of luminal A and basal-like breast tumors. The frequency plots were calculated from segmented copy number profiles. (a) Frequency plot of Luminal A breast tumors calculated from 45 copy number profiles (b) Frequency plot of Basal-like breast tumors was calculated from 23 copy number profiles.

to normal breast tissue (Perou et al., 2000; Sorlie et al., 2001). This classification has been shown to be extremely robust and has further been refined using more advanced technology on larger cohorts (Calza et al., 2006; Carey et al., 2006; Hu et al., 2006; Sorlie et al., 2003).

More recently, high-resolution CGH microarrays have been used to study the genome structure of these subtypes and shown that they progress by

different genomic rearrangement patterns (Bergamaschi et al., 2006; Chin et al., 2006). In the basal-like tumors, Bargamaschi identified a higher numbers of gains and losses than Luminal A, and the Luminal B and erbB2+ had more frequent high-level amplifications (Bergamaschi et al., 2006). Work in our lab has also shown that breast cancers can be classified into at least four distinct patterns of genomic rearrangements (Simplex, Complex I, Complex II or Flat), suggesting different progression patterns (Hicks et al., 2006b). The 'Simplex' pattern had broad segments of duplications and deletions. 'Complex I' had a "sawtooth" appearance with narrow segments of deletions and duplications affecting more or less all chromosomes. 'Complex II' resembled the 'Simplex' but had at least one localized region of clustered peaks of amplifications called "firestorms." The fourth pattern was called "Flat" defined profiles with no clear gains or losses. A calculated index reflecting the complex rearrangements called firestorms, was found to be significantly associated with survival independent of other clinical parameters. Another aCGH study identified three subtypes that varied with respect to level of genomic instability and shared characteristics with the Hicks et al. classes (Chin et al., 2006). In summary, the expression subtypes of breast cancer were highly correlated to different genomic rearrangement patterns, suggesting that inter-tumor comparison studies should be restricted to individual subtypes.

These microarray CGH studies corroborated many of the previously identified chromosome arm imbalances and translocations that were reported by Teixeria using G-banding in metaphase cells (Teixeira et al., 1996; Teixeira et al., 1994; Teixeira et al., 1995). However, CGH microarrays also identified many additional focal aberrations (<1mb) that could not be detected by G-banding and allowed uncultured tumors to be analyzed. CGH data can also be mathematically segmented to detect numerous chromosome 'breakpoints' that characterize tumor genomes. When such methods are applied to tumor profiles it is possible to distinguish imbalances ranging from whole chromosomes and chromosome arms to events of ~30kb. For example in the segmented copy number pattern shown in Figure 1.1A (upper panel), and in the frequency plot of Figure 3.1A, we observe the gain of at least one copy of the q arm of chromosome 1 and the loss of 1 copy of chromosome 16q. These two changes in copy number are the most frequent events observed in breast cancer and are also the most highly correlated with each other. That they are highly correlated is not surprising, because these changes are

likely the result of a single event - a pericentromeric and apparently reciprocal translocation between chromosomes 1 and 16, followed by the loss of the hybrid containing 16q and 1p. A similar event often occurs between chromosome 16 and chromosome 8 leading to profiles such as that in Figure 1.1B; the 16q arm is lost along with the 8p arm, followed by the doubling of the 16p-8q hybrid. Interestingly, the breakpoints of these translocation and rearrangements do not pinpoint a specific location or gene important for the cancer process.

In luminal A tumors, the earliest event that can be detected by inter-tumor comparisons is the translocation of chromosomes 1p and 16q. By ordering tumor profiles based on increasing numbers of chromosome rearrangements, we can readily identify 'early' profiles that contain only a gain of 1q and loss of 16p (Figure 1.1A, upper panels). These profiles are simple in that they contain no other copy number changes. We can also detect 'late' profiles that often contain the 1q gain and/or 16p loss but have also acquired a numerous additional amplifications and deletions (Figure 1.1B, lower panels). This early 1p-16q event can also be seen in frequency plots of hundreds of luminal A tumors, which clearly show the gain of 1q and 16p in the progression of this subtype (Figure 3.1A). With the exception of the concurrent gain and loss of the 8q and 8p arms (often appearing simultaneously), loss of 11q and 22q is also apparently accomplished through arm swapping with multiple other chromosome partners (personal communication with Dr. Anders Zetterberg). The rest of these events appear to be distributed more or less evenly across multiple tumors. The genome profiles of luminal B tumors are in general more complex than luminal A profiles, usually characterized by the appearance of multiple amplified regions or 'firestorms' (Hicks et al., 2006b). Their frequency plots, however, do not differ a great deal from Luminal A, indicating that the additional events are distributed throughout the genome.

Conversely, the inter-tumor comparisons of Basal-like tumors show a very different pattern of genome progression. As exemplified by the tumor profile in Figure 1.1B (which was measured from intra-tumor comparisons), the basal-like subtype most often presents a 'sawtooth' pattern characterized by multiple broad deletions rather than reciprocal gains and losses seen in the luminals. Also, the deletion breakpoints are not necessarily pericentromeric. Although the sawtooth pattern varies greatly from tumor to tumor even in the early stages of its development, these tumors ultimately share a series of common markers

distinct from the luminals. By calculating frequency plots from the segmented profiles, we can show that these genomes are characterized by frequent gains at the ends of 3q, 6q and 10p and losses at 4p, 5q and 17p (Figure 3.1B). In fact, the progression patterns of the luminal A and basal-like subtypes are so drastically different, that the only events they share are the loss of 8p and gain of 8q (Figure 3.1). Thus, the basal-like and luminal A breast tumors show markedly different patterns of genome progression.

In summary, only a limited number of conclusions can be drawn from longitudinal surveys of breast tumor genomes. Although certain events are frequently observed in certain subtypes, it is difficult to draw a roadmap in which even a few of the observed events are precisely ordered. Furthermore, the roles that these genomic events play in the initiation or proliferation of cancer is still open to speculation. The broad distribution of breakpoints makes it unlikely that they act through gene fusion or disruption and these events represent at most twofold changes in gene dosage. It is also difficult to discern the biological impact of gene dosage effects during progression, when whole chromosome arms are deleted or amplified. In the case of 16q deletion, the copy number of six cadherin genes is decreased, perhaps decreasing cell-cell interaction, but the copy number of hundreds of other genes is also reduced. In the case of 8q arm amplification, there is a drastic increase in the gene dosage of *CMYC*, but also many other potential oncogenes. Inter-tumor analysis is also confounded because most samples represent a single time point in the later stages of tumor progression, often containing numerous genetic aberrations. Thus it is difficult to understand the importance of any single amplification or deletion event during a specific stage of tumor progression.

3.2 *Intra*-tumor Comparisons

Here we present an alternative approach to studying large sets of tumors, by inferring progression from multiple samples within heterogeneous tumors (Figure 1.1B). Our philosophy differs from longitudinal studies in suggesting that much can be learned from intensely studying individual tumors. Generally, inferring tumor progression in humans is difficult, because we cannot sample the patient at multiple time points during the progression of the disease. Often, we are left with a single tumor sample that has been surgically excised at a specific time point. However, it has been shown that many solid tumors are genetically

heterogeneous. We hypothesized that these heterogeneous tumors contain a natural history of the mutations that occurred in different tumor genomes, and that this 'permanent record' could be used to reconstruct progression. A good analogy is an archeologist inferring the evolution of a species from the fossil record, however genomic data is far more quantitative.

Assuming that mutational complexity increases with time, we can temporally order a set of genomes based on increasing numbers of mutations. This is possible because the tumor genome acquires a myriad of stable mutations during tumor progression with only a low probability of reversion. A major advantage of this approach, compared to longitudinal studies, is that cells within individual tumors will share many common mutations, because the tumor as a whole shares a common genetic lineage. In contrast, tumors sampled from many different patients in longitudinal studies will share only high frequency events, containing many mutations that are unique to each patient.

Inter-tumor comparisons are, however, confounded by mixed populations of cells. Unlike many cytological techniques where individual tumor cells can be observed, genomic techniques measure signal from complex mixture of cell types, including various tumor clones and an amalgamation of normal cells collectively referred to as stroma (Hanahan and Weinberg, 2000). To more accurately compare the genomes of heterogeneous clones, we need to first isolate the individual subpopulations and remove any normal cells in order to 'purify' the measured signal.

One method for isolating subpopulations from within a single tumor involves using surface receptors that are displayed on different clones. For example, by isolating tumor subpopulations via FACS using $CD44^{+}$ $CD24^{-}$ and $CD44^{-}$ $CD24^{+}$ receptors and measuring copy number profiles, it was shown that these subpopulations were highly similar, but did differ by a few genomic aberrations (Shipitsin et al., 2007). This approach requires *a priori* knowledge of which receptors can distinguish clonal subpopulations. Another method for isolating tumor subpopulations involves sampling multiple distinct regions of a tumor by macro-dissection or laser capture micro-dissection. Using macro-dissection it has been shown that different quadrants of single breast tumors show divergent copy number profiles, suggesting the presence of multiple, genetically related clones in the tumor (Teixeira et al., 1996; Teixeira et al., 1995; Torres et

al., 2007). Similar results have been found, using laser-capture micro-dissection and copy number quantification to identify divergent clonal subpopulations (Aubele and Werner, 1999; Glockner et al., 2002). A caveat of this method is that tumor clones must be regionally segregated in the tumor in order to be detected. Furthermore, the mixing of normal cells may severely decrease the overall signal of the tumor subpopulations in different sectors. Coupling macro-dissection with flow-sorting nuclei by DNA content provides an alternative method for isolating heterogeneous tumor subpopulations. Cytometrists have long been aware that many solid tumors contain multiple aneuploid distributions of cells with different mean DNA indices (Coons et al., 1995; Giaretti et al., 1996; Kallioniemi, 1988). However, until recently a genomic analysis of subpopulations that differ in ploidy had not been investigated (Corver et al., 2008).

Initially, our approach involved isolating regionally segregated tumor subpopulations by macro-dissecting tumors for analysis by ROMA. We then coupled this approach with flow-sorting to isolate tumor subpopulations by region and differences in ploidy. We call this method Sector-Ploidy-Profiling (SPP) and applied it to twenty breast tumors to study progression(Navin et al., 2010). However, despite our efforts to isolate tumor subpopulations by region and ploidy, we were still analyzing mixed populations of millions of cells. To address this problem, we developed a method to measure genome-wide copy number in single cells called Single Nucleus Sequencing (SNS). Analyzing tumor cells at single cell resolution will effectively eliminate mixing problems that confound the analysis of progression. We applied SNS to analyze 100 single cells from a heterogeneous breast carcinoma, which confirmed our SPP results and further showed strong evidence that tumor growth is driven by a series of sequential clonal expansions.

CHAPTER 4

Regional Genomic Heterogeneity

We hypothesized that some tumor subpopulations occupy discrete regions within the tumor. To test this hypothesis we macro-dissected tumors into sectors and measured genomic copy number using Representational Oligonucleotide Microarray Analysis (ROMA). ROMA is an array comparative genomic hybridization (aCGH) technique that was developed in our laboratory (Lucito et al., 2003). In this technique, two genomes (a reference and an experimental) are digested with a restriction enzyme, ligated with specific adapters and PCR amplified to generate libraries (Figure 4.1). These libraries are labeled with different fluorophores, generally Cy3 and Cy5, and co-hybridized to a custom DNA microarray with 50mer probes that are designed to target the restriction fragments. In this study we used both 85K and 390K custom DNA microarrays (manufactured by Nimblegen) with a genomic resolution of approximately 50 kilobases. The raw intensity values were Lowess normalized and segmented using the Kolmogorov-Smirnov segmentation algorithm (as described in (Grubor et al., 2009)) to identify contiguous regions of the human genome with significant differences in copy number. When ROMA is applied to normal vs. normal genomes, the profiles show diploid copy number across all autosomes. In contrast ROMA profiles from cancer vs. normal genomes show numerous amplifications and deletions that are often within regions of cancer genes.

To investigate regional genomic heterogeneity we macro-dissected breast tumors into 8 sectors, isolated DNA and applied ROMA to four sectors (Figure 4.2). We also stained corresponding tissues from each sector with Hematoxylin & Eosin to observe any changes in grade in the four sectors. We applied this approach to four high grade (III) primary ductal carcinomas (T1-T4) that were randomly selected from a large collection of frozen ductal carcinomas in the Wigler laboratory. Two tumors analyzed by this method (T1, T2) contained minimal variation in their genomic copy number profiles in all four sectors. Our data indicated that T1 contained 39 chromosomal breakpoints that were common to all tumor sectors, and that multiple amplifications and deletions were present at similar copy number in every sector (Figure 4.3A). Similarly, T2 contained

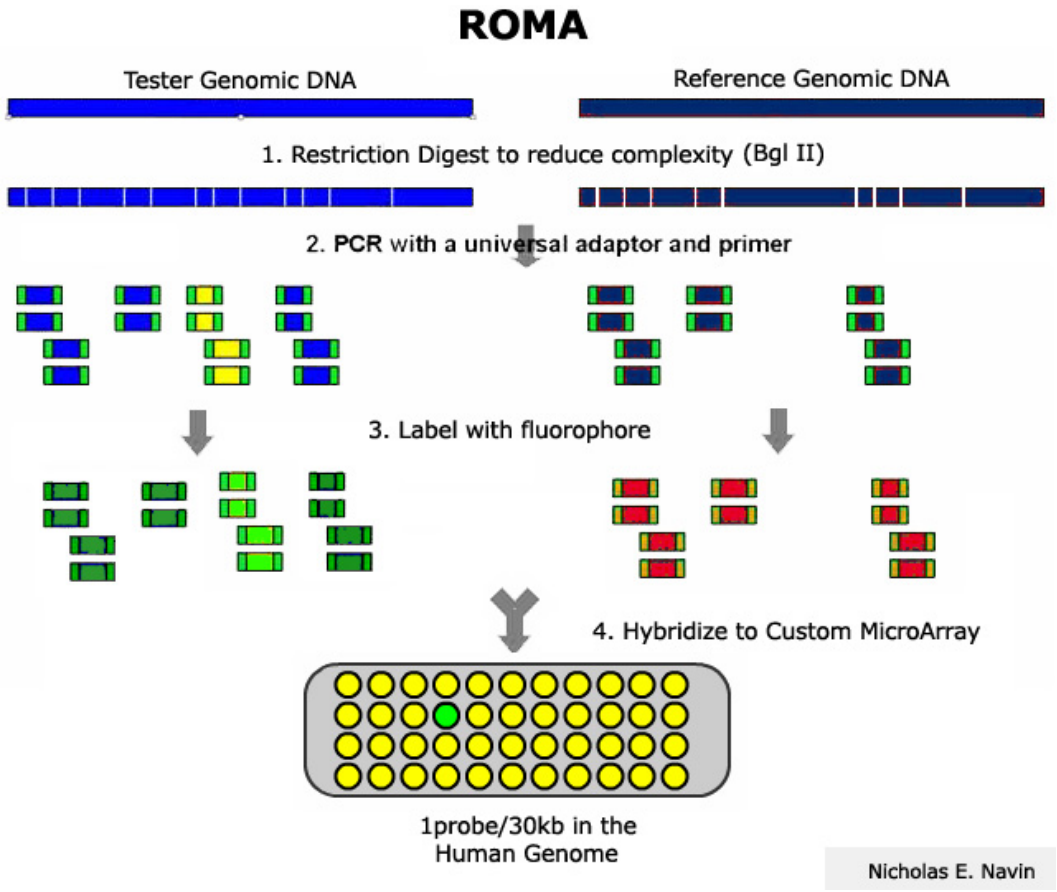


Figure 4.1 Representational Oligonucleotide Microarray Analysis (ROMA)
ROMA involves cutting two genomes (Tester and Reference) with a restriction enzyme, ligating univeral adapters and PCR amplifying both pools. In this diagram there is an amplification shown in yellow that is present only in the tester genome. The tester fragments are labeled with cy3 (green) and the reference fragments with cy5 (red) and co-hybridized to a custom DNA microarray. The majority of probes are shown in yellow, due to equal contributions of the fluorophores from the tester and reference genome, representing diploid copy number in both genomes. On this microarray a single probe in the tester genome has a greater intensity of green signal representing an amplification.

44 amplification and deletion breakpoints that were common in position and magnitude in all four tumor sectors. This analysis indicates that these tumors contain highly similar profiles in every sector, suggesting that T1 and T2 are each composed of a single major tumor subpopulation, or a homogeneous mixture of subpopulations that are not resolvable by dissection alone.

In contrast, when we analyzed tumors T3 and T4, we noticed a large degree of variation in the genome patterns of distinct sectors. T3 contains 21 chromosomal breakpoints common to all four sectors, but S3 of T3 also contains

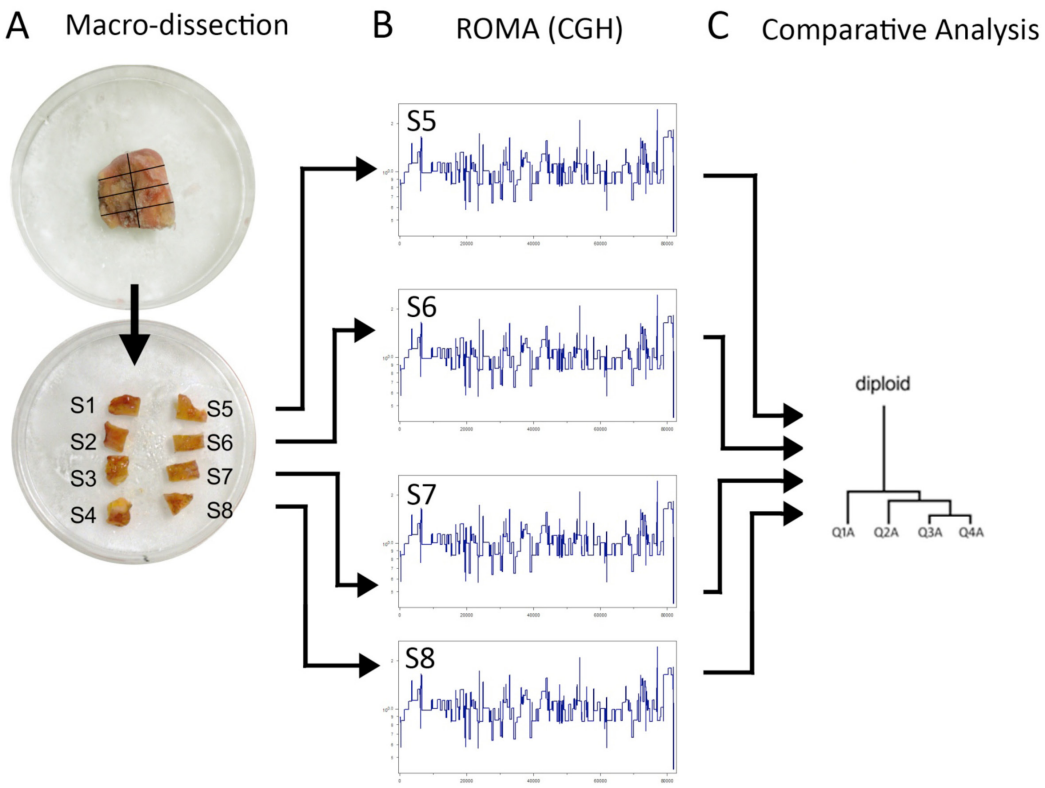


Figure 4.2 Sector-ROMA Approach
(a) Our approach involved macro-dissecting frozen breast tumors into 8 sectors, four of which were used for histological H&E staining. (b) Genomic DNA was isolated from the remaining four sectors and analyzed by ROMA. (c) Copy number profiles from each sector were segmented and compared by calculating a Pearson correlation tree, which included a simulated diploid profile.

16 new divergent chromosome breakpoints not present in the other tumor sectors. These chromosome breakpoints encompass three genomic amplifications (6p22.1, 6p21.1, 17q21.32) and a deletion (21q11), none of which are detectable in S1, S2, or S4. Thus at least two subpopulations are evident in this polygenomic tumor. T4 displays yet another pattern (Figure 4.3B). Two sectors (S1 and S2) that contain high proportions of tumor cells as assessed by histopathology from the H&E sections (71% and 69%, respectively) do not display prominent genomic rearrangements. In these tumor profiles, normal copy number variation is also observed (Sebat et al., 2004). Sampling from this part of the tumor (S1 and S2), and using previous genomic measures (Hicks et al., 2006), we would not judge the tumor to be highly malignant. However, had we sampled from sectors 3 and 4 (which display many prominent rearrangements, including 98 breakpoints not present in sectors S1 and S2), we would judge the tumor to be highly malignant.

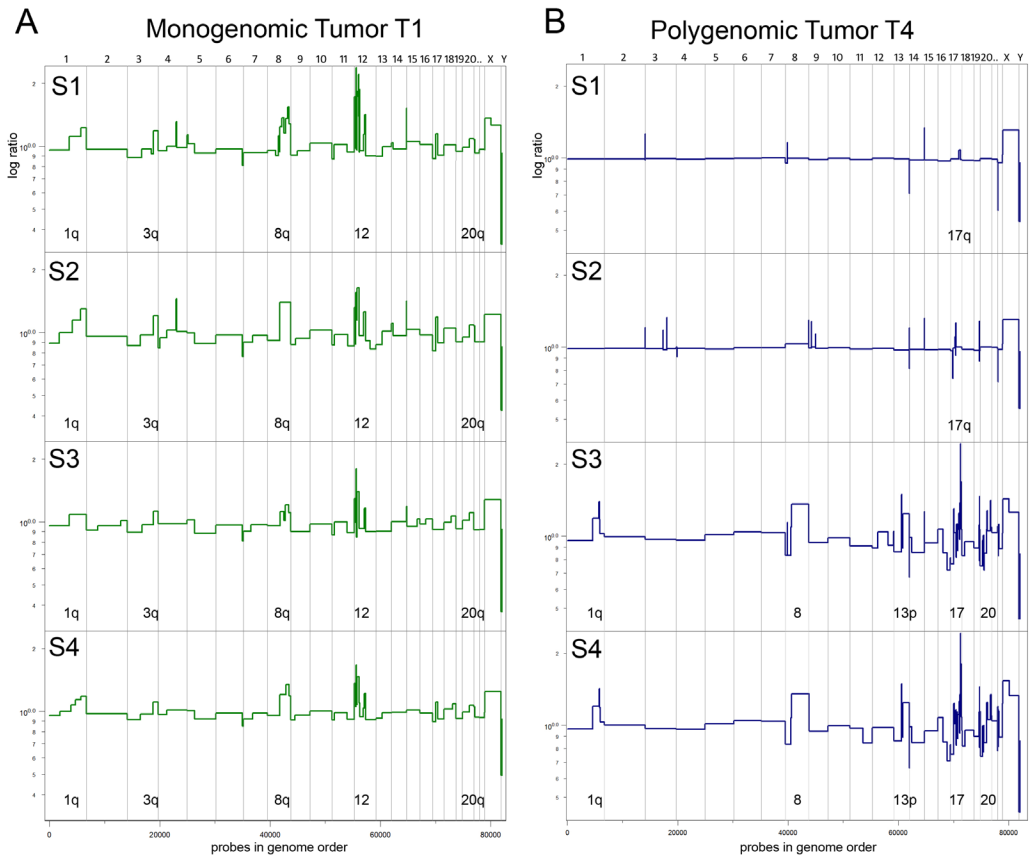


Figure 4.3 ROMA Analysis of Tumor Quadrants
Sector-ROMA was applied to four ductal carcinomas (T1-T4), two of which are shown. (A) T1 is a monogenomic tumor displaying a highly similar copy number profile in all four sectors (S1-S4), suggesting that it consists of a single dominant subpopulation (B) T4 is a polygenomic tumor, displaying a near diploid copy number profile in sectors S1-S2, but progressing to a highly aneuploid copy number profile in sectors (S3-S4), suggesting that it consist of at least two major tumor subpopulations.

Our initial study showed that regional tumor heterogeneity is quite common. Two of the tumors that we analyzed (T1 and T2) consisted of a highly stable genome profile that was found throughout the tumor sectors. We refer to this class of tumor as 'monogenomic' to indicate that it appears to be composed of a single stable genome profile that dominates the tumor mass. In contrast, the two other tumors analyzed (T3 and T4) show that multiple genomic profiles were present in different anatomical sectors. We refer to this class as 'polygenomic' to indicate the presence of multiple tumor subpopulations, each with distinct genomic rearrangements.

Our initial results were very encouraging, however, a major caveat of this approach is that each tumor sector was a mixture composed of both normal stroma and tumor cells. The proportion of these cells can vary greatly between sectors, which confounds the comparison of copy number profiles, because genomic profile are limited to measuring mixed population of cells (Figure 4.4). A monogenomic tumor with a high percentage of diploid cells (90%) and low proportion of aneuploid cells (10%) in one sector, and the opposite proportions in an adjacent sector would falsely appear to be progressing by acquiring chromosome aberrations and would be misclassified as polygenomic. Thus, better methods were needed to separate out tumor subpopulations prior to measuring genome-wide copy number.

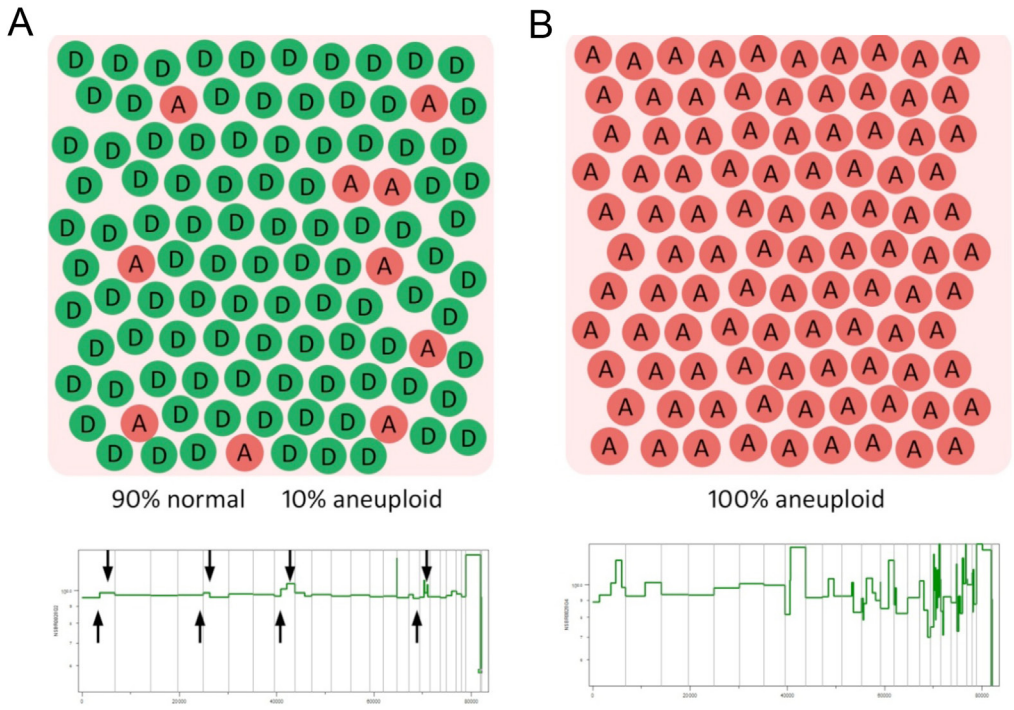


Figure 4.4 Mixing Caveat
The mixing of diploid and tumor genomes may confound the interpretation of copy number profiles based on millions of cells. (A) Tissues with a high proportion of diploid cells (green) and low number of aneuploid tumor cells (red) will show a copy number profile that is averaged down towards a diploid signal. (B) Tissues with a high proportion of aneuploid tumor cells will show profiles with strong copy number amplitudes.

CHAPTER 5

Sector-Ploidy-Profiling Method

To gain a clearer picture of the number of subpopulations and their clonal relationship, and to mitigate the mixing effects of normal and tumor cells, we added a further tool for separating subpopulations, fluorescence-activated cell sorting (FACS). Previous studies have shown that FACS can be used to separate tumor cells by ploidy for genomic analysis (Corver et al., 2008). We use FACS to separate subpopulations of tumor cells, and tumor cells from normal cells, by differences in their total genomic DNA content, or ploidy. Because tumor cells are often aneuploid and normal cells have 2N diploid copy number, FACS can sort these cells into different wells for analysis by ROMA. We call the combined approach Sector-Ploidy-Profiling (SPP), and illustrate our method for a single example (T10) in Figure 5.1

T10 was cut in half along one axis, and six cuts were made along an orthogonal axis, resulting in 12 pieces (Figure 5.1A). Nuclei were prepared from six of these pieces using a DAPI-NST buffer ((800 mL of NST [146 mM NaCl, 10 mM Tris base at pH 7.8, 1 mM CaCl2, 21 mM MgCl2, 0.05% BSA, 0.2% Nonidet P-40]), 200 mL of 106 mM MgCl2, 10 mg of DAPI, and 0.1% DNase-free RNase A) by finely mincing sections using two no. 11 surgical scalpels. The isolated nuclei from each sector were then separated by flow-sorting subpopulations distinguishable by total DNA content using the BD Diva FACS Machine (Figure 5.1B). In all FACS analysis, a small amount of prepared nuclei from each tumor sample was mixed with a diploid control sample (derived from a lymphoblastoid cell line of an apparently normal person) to accurately determine the diploid peak position within the tumor DNA content distribution for FACS collection gates and to calculate ploidy.

DNA was isolated from the gated FACS peaks using the QIAGEN Genomic DNA Isolation Kit. A total of 200ng of DNA was used to make complexity-reducing representations of genomic DNA for whole-genome copy number analysis by ROMA (as described by (Grubor et al., 2009)) using 390K custom Nimblegen microarrays. Hybridizations of the 390K experiments were performed in color reversal to prevent color bias and ensure data quality. All

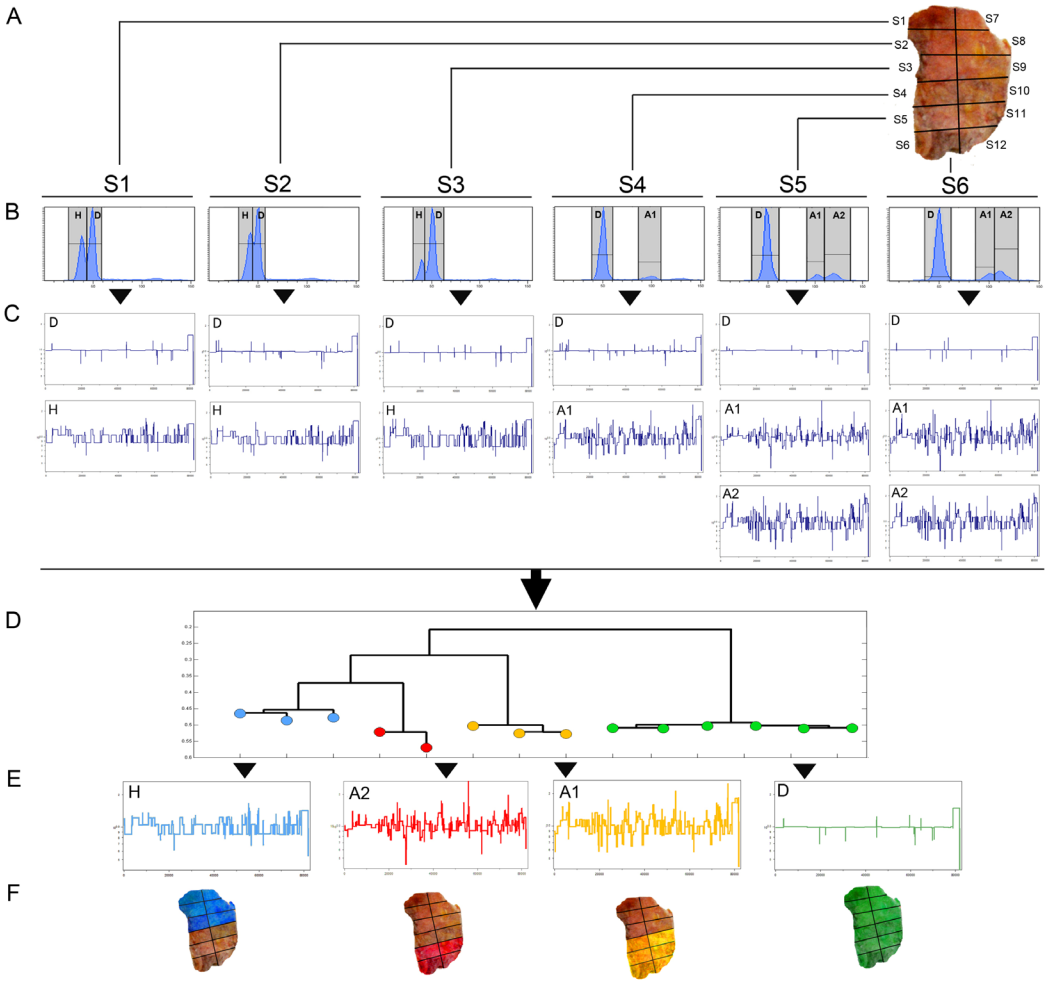


Figure 5.1 – Sector-Ploidy-Profiling Approach
The SPP approach separates tumor subpopulations by region and ploidy. (A) The tumor is macro-dissection into sectors (B) Nuclei are isolated and stained with DAPI for flow-sorting, then gated by FACS according to differences in total genomic DNA content (C) The gated FACS peaks are individually analyzed for genome-wide copy number by ROMA (D) A neighbor-joining tree is calculated using all of the profiles from the tumor (E) Highly similar groups of copy number profiles are coalesced into consensus profiles (F) The topography of the subpopulations in the tumor are colored. In this diagram of T10, the tumor sectors S7-S12 are colored according to the adjacent subpopulations in S1-S6.

tumor samples were cohybridized with a reference genome from fibroblast DNA. The ROMA experiments were scanned, gridded, and normalized with a Lowess curve-fitting algorithm followed by a local normalization as described by Hicks et al. (2006). The data were imported and analyzed using Splus (Insightful) and Matlab (Mathworks), and the geometric mean ratio was computed from each color channel. In color-reversal experiments, the geometric mean of two log ratios was calculated. The data were then segmented using both the Kolmogorov-Smirnov algorithm (Grubor et al. 2009) and the circular binary segmenter (Olshen et al., 2004; Venkatraman and Olshen, 2007). The segmented profiles within each tumor were always clearly related but often indistinguishable by their chromosome breakpoint pattern (Figure 5.1C).

For each tumor we calculated 1-Pearson correlations and used neighbor-joining algorithms to form distance trees, clustering profiles into similar or distinguishable subgroups (Figure 5.1D). In each case where we claim that a genomic breakpoint distinguished two subgroups, we examined the raw data to rule out the possibility of segmentation artifacts, namely, that the differences were not merely of degree. To facilitate further comparisons between subgroups, we coalesced profiles within subgroups by calculating the means of the segmented values from subgroups of individual CGH profiles (Figure 5.1E). In all cases, we found that genome-wide copy number patterns corresponded to specific ploidy distributions, which will be discussed in detail in the next chapter. We also show the topography of genetically distinct subpopulations by coloring the sectors of the tumor (Figure 5.1F).

In summary, the SPP method allows us to identify copy number differences in tumor subpopulations that differ in ploidy or region within a tumor. This method is limited when two or more tumor subpopulations with genetically distinct copy number patterns share the same ploidy distribution, for they will be represented as a single profile. However, this is not a problem when two tumor subpopulations with similar ploidy are anatomically segregated in the tumor, because they can be separated by macro-dissection. This method is also limited by requiring millions of cells, which is very time consuming since each nucleus from the tumor must be flow-sorted. Moreover, since each genome profile represents a population of millions of cells, rare subpopulations or intermediate cells may be entirely missed. Nevertheless, SPP is a powerful method for isolating and studying the genomes of tumor subpopulations.

CHAPTER 6

A Study of 16 Breast Tumors by SPP

In order to understand the substructure of breast tumors, we applied SPP to a collection of 16 breast tumors. We hypothesized that some of these tumors would contain multiple subpopulations from which we could infer tumor progression. For this study we randomly selected 16 ductal carcinomas from a collection of hundreds. These breast tumors were high grade (III) and consisted of diverse receptor statuses (Estrogen, Progesterone and Her2) as assessed by histopathology. They ranged in both morphology and size (from 0.5 x 0.5 x 0.3cm to 6.0 x 6.0 x 5.0cm) (Figure 6.1). Each tumor was randomly selected from hundreds collected by Dr. Michael Wigler from various sources including the Cooperative Human Tissue Network (T1–T7), North Shore University Hospital (T7–T8), Asterand Corporation (T16–T17), Memorial Sloan-Kettering Cancer Center (T12–T14), and Columbia University (T19–T20).

As we flow-sorted the tumors, two different classes of ploidy distributions became evident: (1) tumors with single consistent aneuploid peaks in all sectors, and (2) tumors with multiple aneuploid peaks that often shifted ploidy between sectors. Subsequent analysis of these peaks by ROMA showed that the former class contained a highly similar copy number profile in all sectors, corresponding to the 'monogenomic' tumor class. We found that monogenomic tumors are very common (9/20) and consisted of a single dominant subpopulation throughout the tumor. Tumors with multiple aneuploid peaks were also analyzed by ROMA and revealed that each peak consisted of a genetically related, but divergent clonal subpopulation. We refer to this class as 'polygenomic' to indicate they are composed of multiple clonal subpopulations that often occupy different regions of the tumor. Thus, the ploidy distributions of breast tumors often corresponded to the genomic classification of monogenomic and polygenomic tumors.

Monogenomic tumors typically contained a single 2N (1.0 DNA index) peak of normal cells, presumably composed of stroma and immune cells, and a single aneuploid peak with a consistent ploidy in all sectors. T11, for example, contained a 2N distribution of normal cells and a 3.2N distribution of aneuploid cells that is present in all six sectors (Figure 6.2A). Another monogenomic tumor,

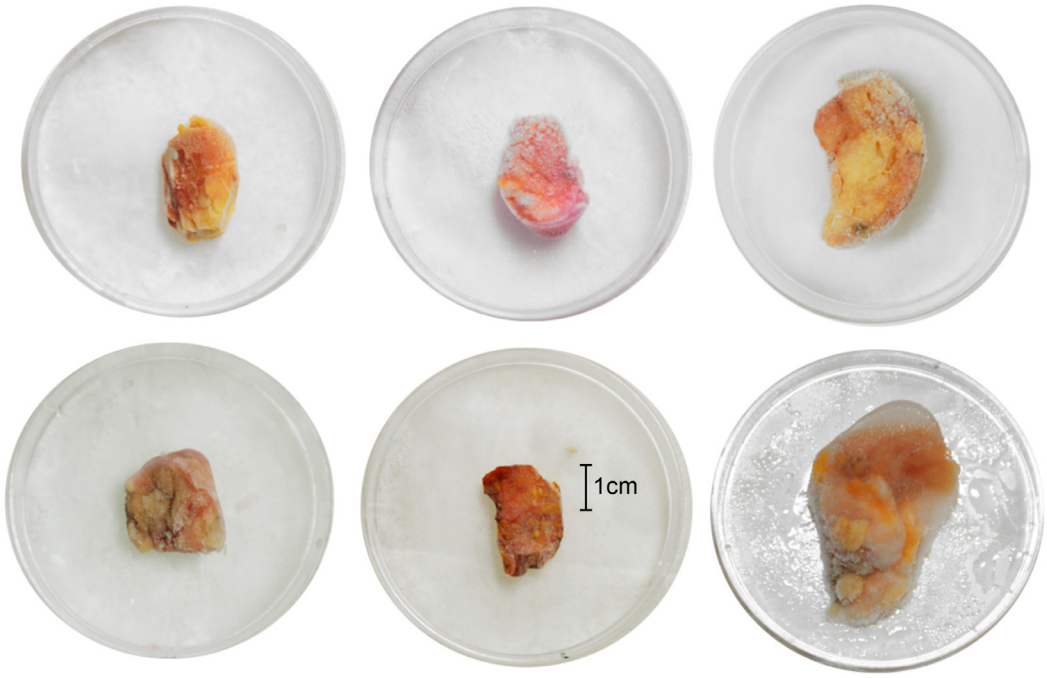


Figure 6.1 – Breast Tumor Morphology
Twenty frozen ductal carcinomas were randomly selected from a collection of hundreds belonging to Dr. Michael Wigler. These tumors ranged in morphology and size, averaging approximately 2cm x 1cm x 1cm. They represent whole tumors removed by lumpectomy, not fragmented samples.

T9, contains a 2N distribution of normal cells and a 4.3N tetraploid distribution of aneuploid cells in all six sectors (Figure 6.2B). The height of the peaks does vary between sectors, signifying different proportions of normal and tumor cells, however there is small change in the mean ploidy index in monogenomic tumors. This information can be summarized in a FACS matrix (Figure 6.5), which shows the ploidy index in each tumor sector (the relative proportions of cells are not shown). These graphs are useful for comparing overall ploidy patterns between tumors. In monogenomic tumors (T6, T7, T9, T11, T15, and T20) all sectors contain a single distribution of aneuploid cells with plodies of 2.4N to 6.0N along with the expected diploid fraction of 2N. The aneuploid fractions all showed abnormal CGH profiles, but within each tumor this profile was highly similar in every sector. One tumor (T16) had a single FACS peak (with a DNA content of 2N), but this peak contained a highly rearranged tumor subpopulation in every sector, as revealed by array CGH.

In contrast, polygenomic tumors contained single or multiple aneuploid

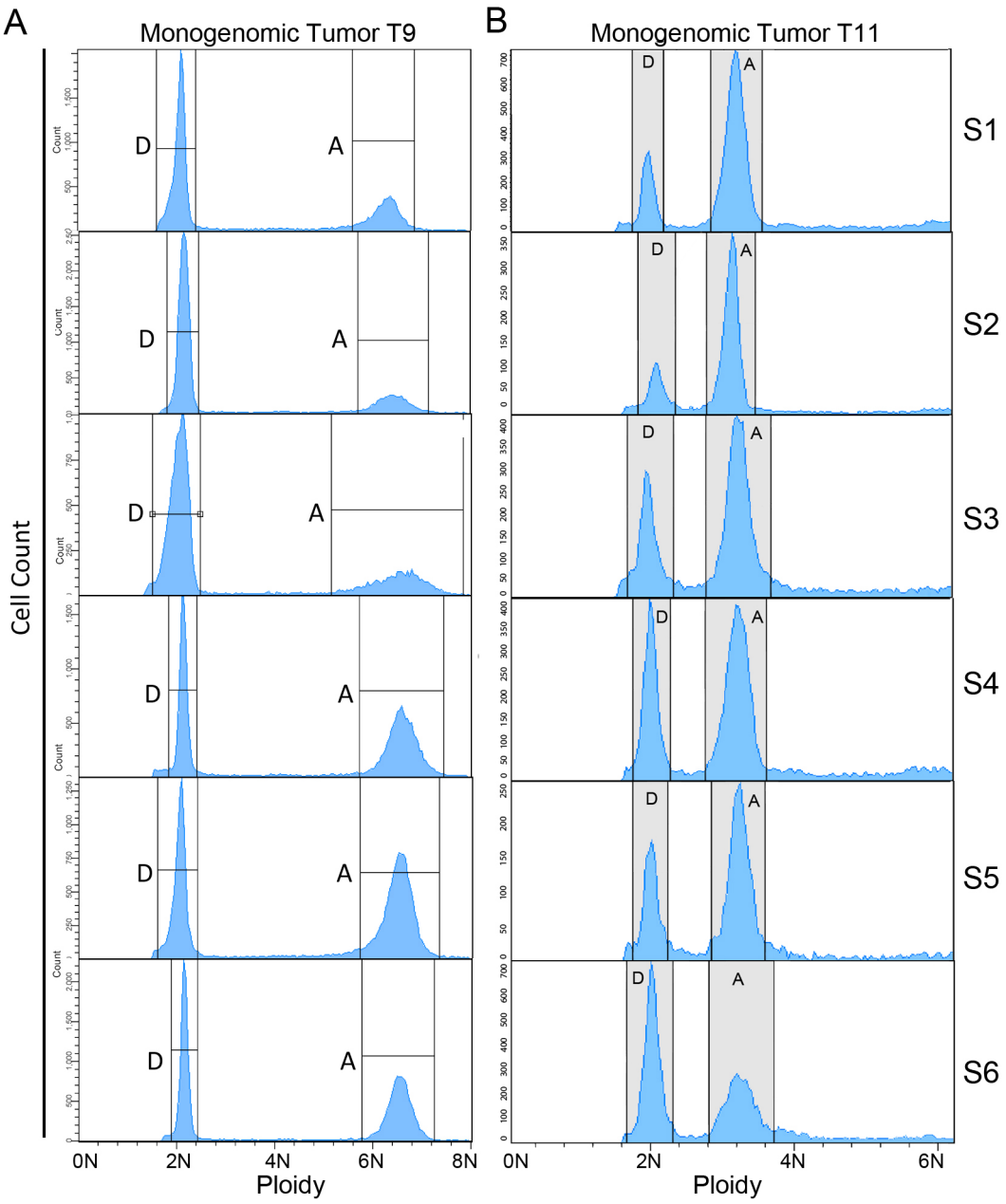


Figure 6.2 – FACS Histograms of Monogenomic Tumors
Nuclei were isolated from each tumor sector (S1-S6) and subpopulations were sorted by differences in ploidy. (A) The monogenomic tumor T9 shows a single 2N distribution of normal cells and a single aneuploid distribution at 6.4N that was highly similar in all six sectors. (B) The monogenomic tumor T11 shows a single 2N distribution of normal cells and an aneuploid distribution at 3.3N that was nearly identical in all six sectors. In both tumors, some variation is seen in the aneuploid cell counts between the sectors.

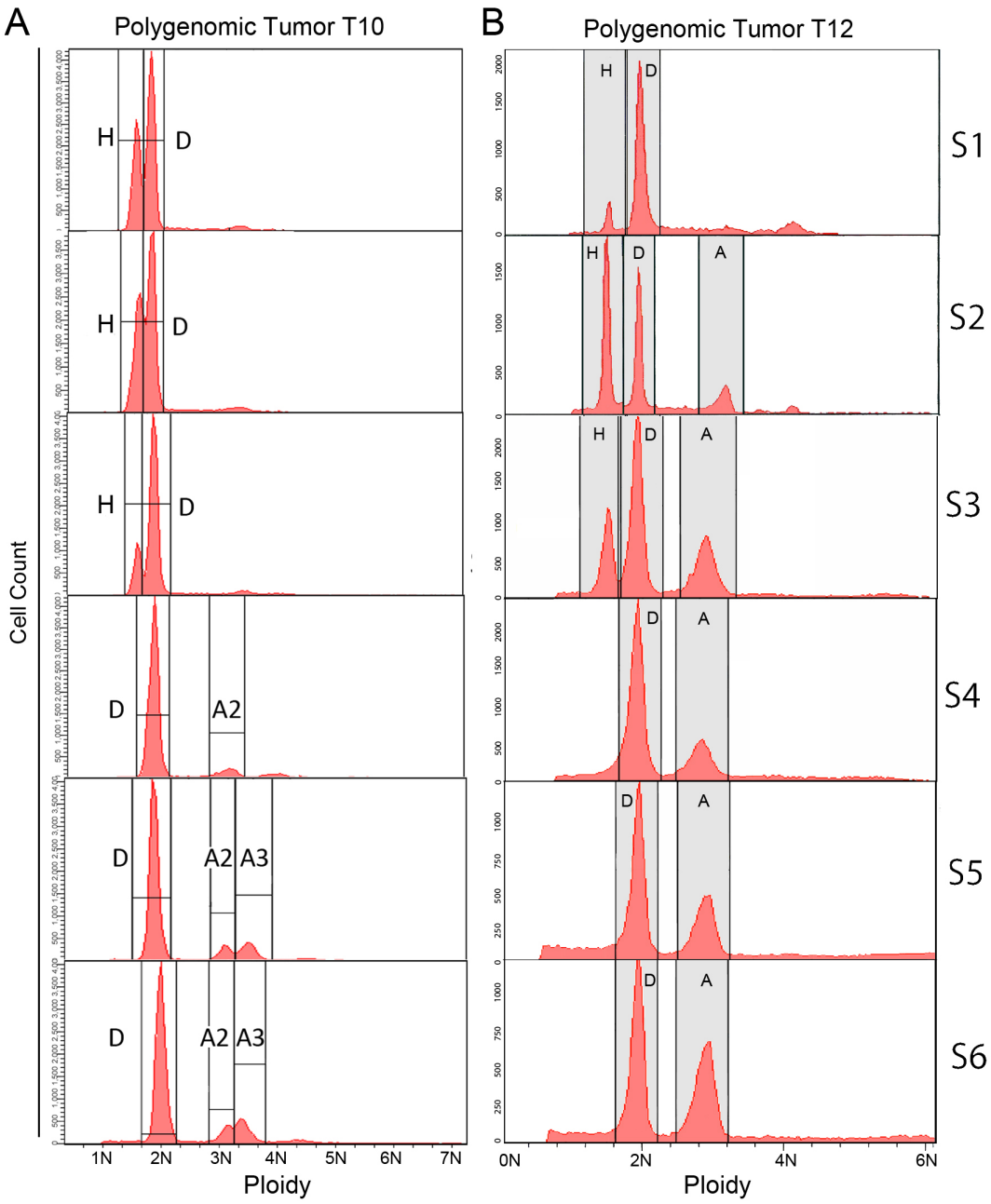


Figure 6.3 – FACS Histograms of Polygenomic Tumors
Nuclei were isolated from tumor sectors, stained with DAPI and sorted by total genomic DNA content (ploidy). (A) T10 contains one diploid 2N distribution in all six sectors (S1-S6) and three aneuploid distributions, H, A1 and A2 that occupy discrete regions. The H distribution at 1.7N is exclusive to the upper sectors (S1-S3), while the A1 and A2 distributions are intermixed in the lower sectors (S4-S6) (B) T12 contained three cellular distributions: hypodiploid (H), diploid (D) and aneuploid (A). The diploid 2N distribution was present in all six sectors. The H distribution at 0.8N was present in only three sectors (S4-S6), while the A distribution AT 3.0N was present in five sectors (S1-S5).

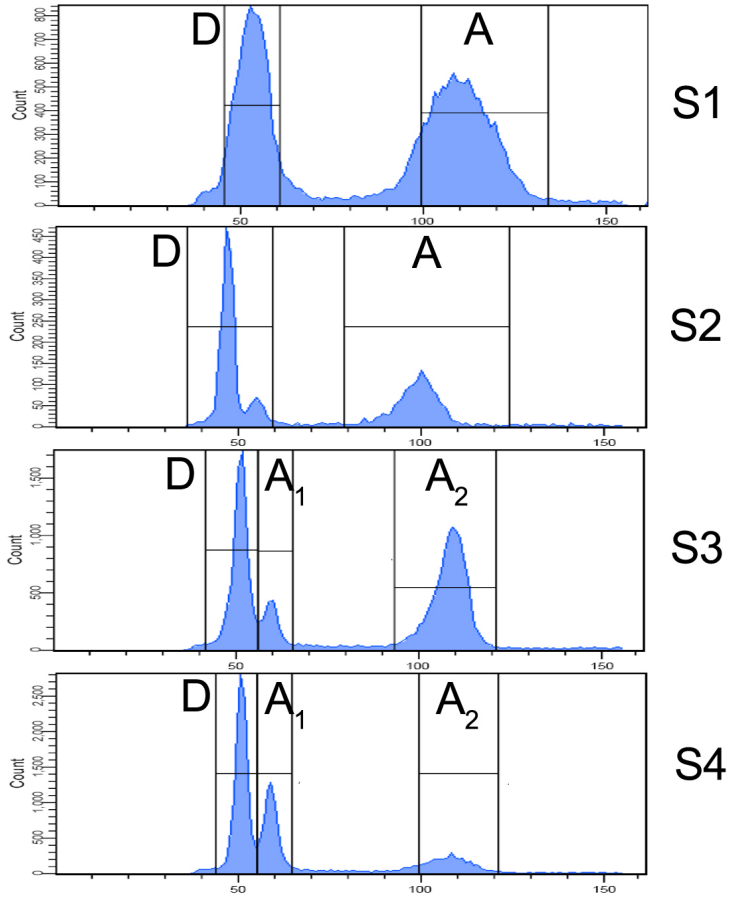


Figure 6.4 – FACS Profile of Polygenomic Tumor T5
Nuclei were isolated from four sectors of T5, stained with DAPI and sorted by total genomic DNA content (ploidy). In all four sectors, a 2N diploid (D) distribution and a 4.2N aneuploid (A2) distribution are present. In sectors S2-S4 second aneuploid distribution (A1) emerges at 2.4N. Cell counts vary in the aneuploid distributions between tumor sectors.

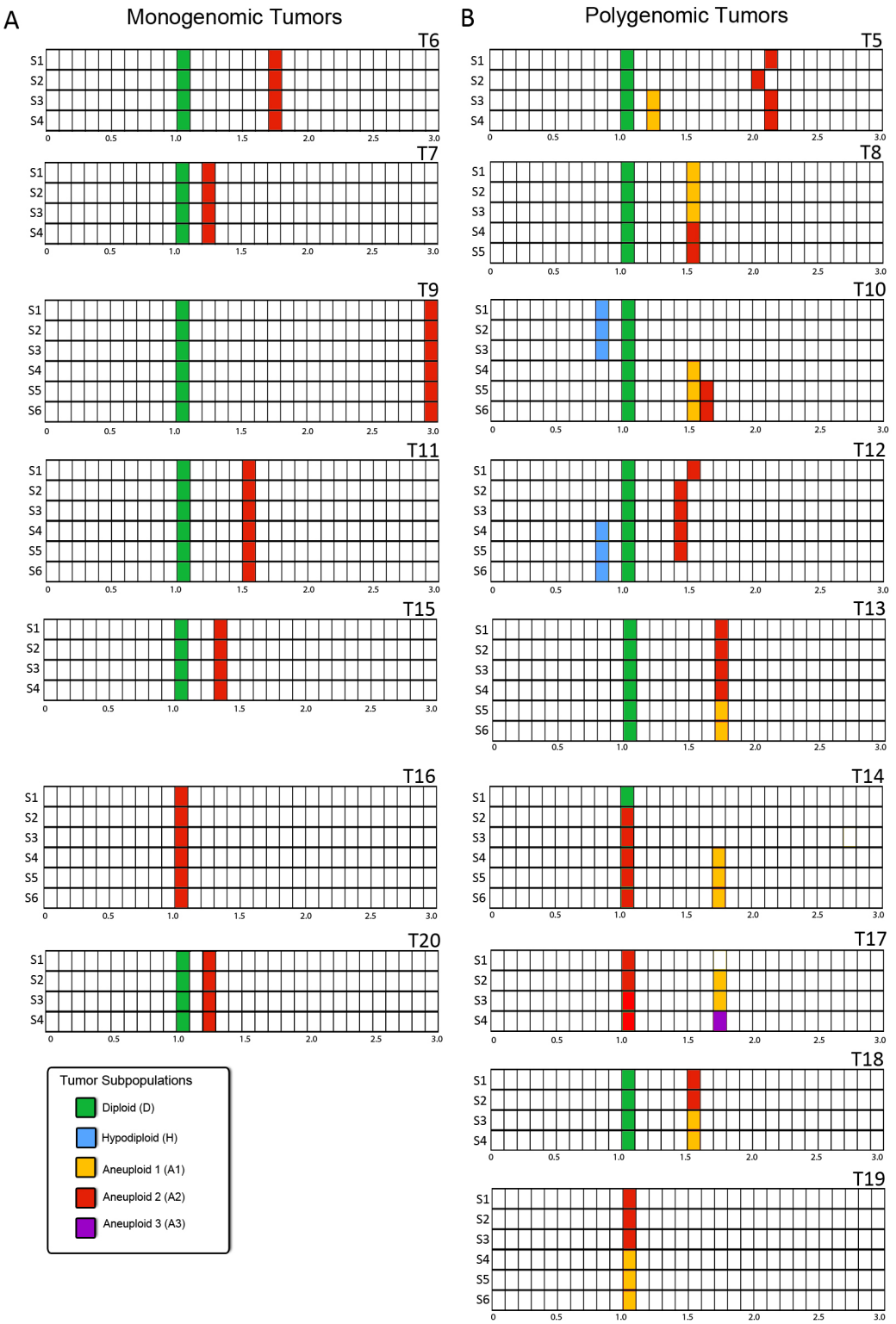
peaks that often shifted in ploidy between adjacent sectors. All of these tumors also contained a large proportion of normal diploid cells in every sector. In the polygenomic tumor T12 we found a 2N distribution of cells in all six sectors and two aneuploid distributions: hypodiploid (H), with a lower than 2N ploidy in sectors S1-S3, and an aneuploid (A), with a ploidy of 2.8N that was present only in sectors S2-S6 (Figure 6.3B). Another polygenomic tumor, T10 contained three different aneuploid distributions, H, AA, AB (1.7N, 3.1N and 3.3N) that were anatomically segregated to different regions within the tumor (Figure 6.3A). T6 showed another interesting pattern of ploidy changes between sectors: A1 was present exclusively in S2-S4, while A2 was present in all sectors (Figure 6.4). In

summary, we found that polygenomic tumors often contained ploidy distributions, representing genomic subpopulations, which occupied exclusive regions within the tumor.

Altogether, nine tumors were classified as polygenomic and displayed considerable complexity. Their FACS histograms are summarized in Figure 6.5B. Eight had multiple peaks of ploidy. In every case, subpopulations distinguishable by total DNA content were also clearly distinguishable by variation in their CGH profiles. Three tumors had more than one aneuploid subpopulation distinguishable by FACS (T5, T10, T12). Three tumors had subpopulations of near-diploid cells exhibiting aberrant CGH profiles (T14, T17, and T19). Five tumors had subpopulations with genomic transitions that were not evident from ploidy, but were distinguishable by sector when analyzed by CGH (T8, T13, T17, T18, T19). Two tumors had hypodiploid subpopulations (T10 and T12) with total DNA contents lower than the diploid distribution.

In most cases, monogenomic tumors can be distinguished from polygenomic tumors by analyzing the ploidy distributions of DAPI stained nuclei, however there are exceptions. T19 and T13 appear to be monogenomic tumors by FACS alone, containing a single aneuploid distribution with a constant ploidy in all sectors, however ROMA analysis clearly shows that the lower sectors (S5-S6 in T13 and S4-S6 in T19) contain copy number profiles that have acquired additional amplifications and deletions that are not present in the upper sectors (Figure 6.5A). In most cases, however, FACS analysis of ploidy shows that monogenomic tumors contain single aneuploid and diploid distributions with consistent ploidies in all sectors, while polygenomic tumors contain multiple aneuploid peaks with genetically distinct tumor subpopulations. Analyzing ploidy distributions in tumor sectors alone may have clinical utility in identifying polygenomic tumors, but as we show from this study, monogenomic tumors will require further analysis by genome-wide copy number methods to distinguish divergent subpopulations that share similar ploidies.

Figure 6.5 – FACS Tumor Sector Matrices
For each tumor (T6-T20) the FACS ploidy data from sectors (S1-S6) are displayed in bins showing the mean DNA index of each cellular distribution. (A) The monogenomic tumors may contain a single diploid subpopulation (green) and a single aneuploid subpopulation (red); (B) The polygenomic tumors may contain a single diploid subpopulation (green), a single hypodiploid subpopulation (blue) and/or multiple aneuploid subpopulations (red, yellow, purple) which have been distinguished by differences in their copy number profiles


A
Monogenomic Tumors
B
Polygenomic Tumors
T6
T7
T9
T11
T15
T16
T20
T5
T8
T10
T12
T13
T14
T17
T18
T19
S1
S2
S3
S4
S5
S6
0
0.5
1.0
1.5
2.0
2.5
3.0
Tumor Subpopulations
Diploid (D)
Hypodiploid (H)
Aneuploid 1 (A1)
Aneuploid 2 (A2)
Aneuploid 3 (A3)

CHAPTER 7

Phylogenetic Analysis of Tumor Subpopulations

We initially identified groups of highly similar profiles within tumors, by applying hierarchical clustering to the segmented copy number profiles. Clusters were calculated using average-linkage and a Euclidean distance metric, which measures the density of amplification and deletions between profiles. Large chromosome aberrations (for example loss of a whole arm) that were shared between profiles carry more weight than focal events. A density metric is justified biologically, since loss or gain of a whole chromosome arm will affect gene dosage at numerous loci in comparison to focal events.

Clustering revealed that monogenomic tumors always formed two highly similar groups: one cluster of diploid profiles, and one cluster of aneuploid profiles, as shown with T9 (Figure 7.1A). In contrast, polygenomic tumors formed multiple (2-3) clusters of aneuploid profiles in addition to the diploid cluster, as shown in T10 (Figure 7.1B). Thus, although we measured copy number profiles from the tumor cells 8 times, we did not observe 8 different copy number profiles. Instead, we observed that the 8 profiles belonged to three highly similar groups, representing clonal subpopulations within the tumor. This was common to all polygenomic tumors, in which the 4-10 tumor profiles typically clustered into 2-3 homogeneous groups.

To more rigorously discern the variation between tumor profiles, we used mathematical methods that scale with large numbers of profiles. For each tumor, we computed a matrix of 1-Pearson correlations from the segmented profiles and used neighbor-joining (Saitou and Nei, 1987) to construct distance trees. Neighbor-joining has an advantage over ultrametric methods, in that it does not assume an equal distance of each node from the root node, and can thus display single profiles that have diverged significantly from a population. Moreover, ultrametric methods assume a constant rate of mutation, which is not justified biologically in tumor cells. We omitted sex chromosomes to diminish extraneous correlation, and computed distance using segmented profiles to avoid the noise inherent in raw copy number data. The trees were rooted using flow-sorted

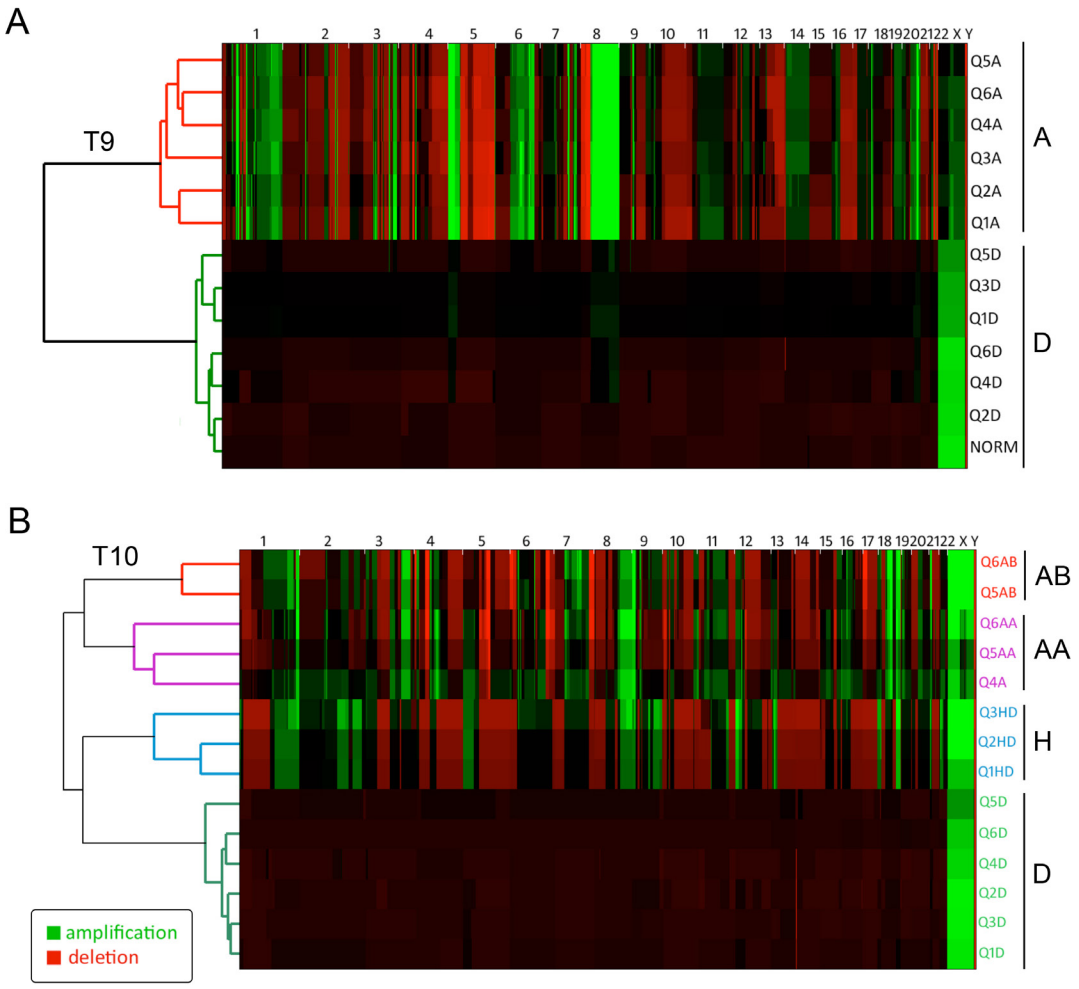


Figure 7.1 – Hierarchical Clustering of Tumor Profiles
Segmented copy number profiles from each tumor were hierarchically clustered and displayed as a heatmaps. Amplifications are shown in green, deletions are red and diploid copy number is shown as black (A) T9 is a monogenomic tumor showing a single cluster of diploid profiles (D) and a single cluster of aneuploid profiles (A) (B) T10 is a polygenomic tumor with a single cluster of diploid cells (D) and three aneuploid clusters (H, AA and AB).

diploid copy number profiles, represented by a green node. The resulting trees for each profile are shown in Figures 7.2 and 7.3. The trees divide into two groups: those with a high correlation, >0.9 between all subpopulations (Figure 7.2), and others that were less correlated (Figure 7.3). The former group corresponds to the monogenomic tumors and the latter to polygenomic tumors, with one exception (T8). In this case, the number of events that distinguishes subpopulations is very small: three focal amplifications on chromosome 12q21.1 (Figure 8.2A). These differences are readily apparent by examining graphs of the segmented profiles, but less so by the mathematical measures.

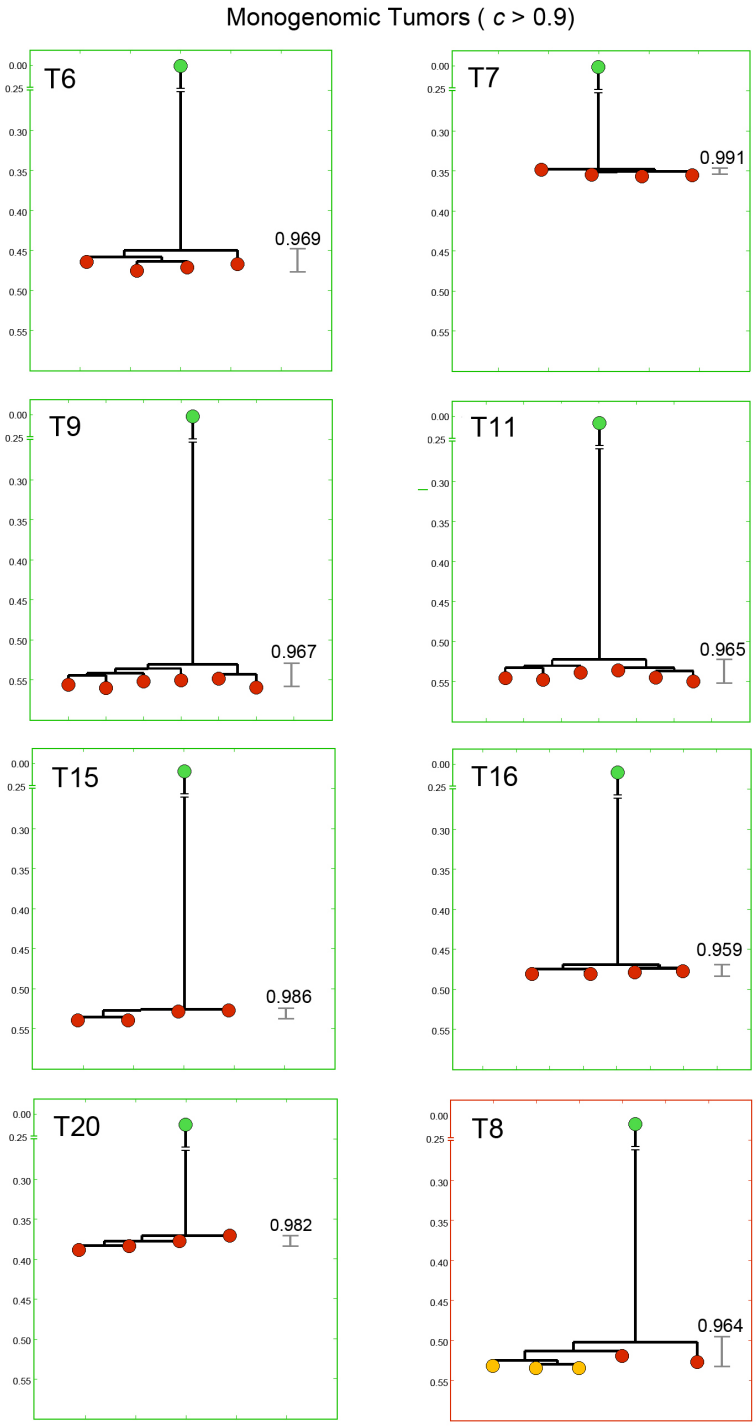


Figure 7.2– NJ Distance Trees of Monogenomic Tumors
Neighbor-joining trees of tumors with a minimum correlation of profiles greater than 0.9. All Monogenomic tumors are outlined in green and polygenomic tumors are outlined in red. Most tumors with highly correlated profiles are monogenomic, with one exception T8. Trees are rooted with a single consensus diploid profile colored in green.

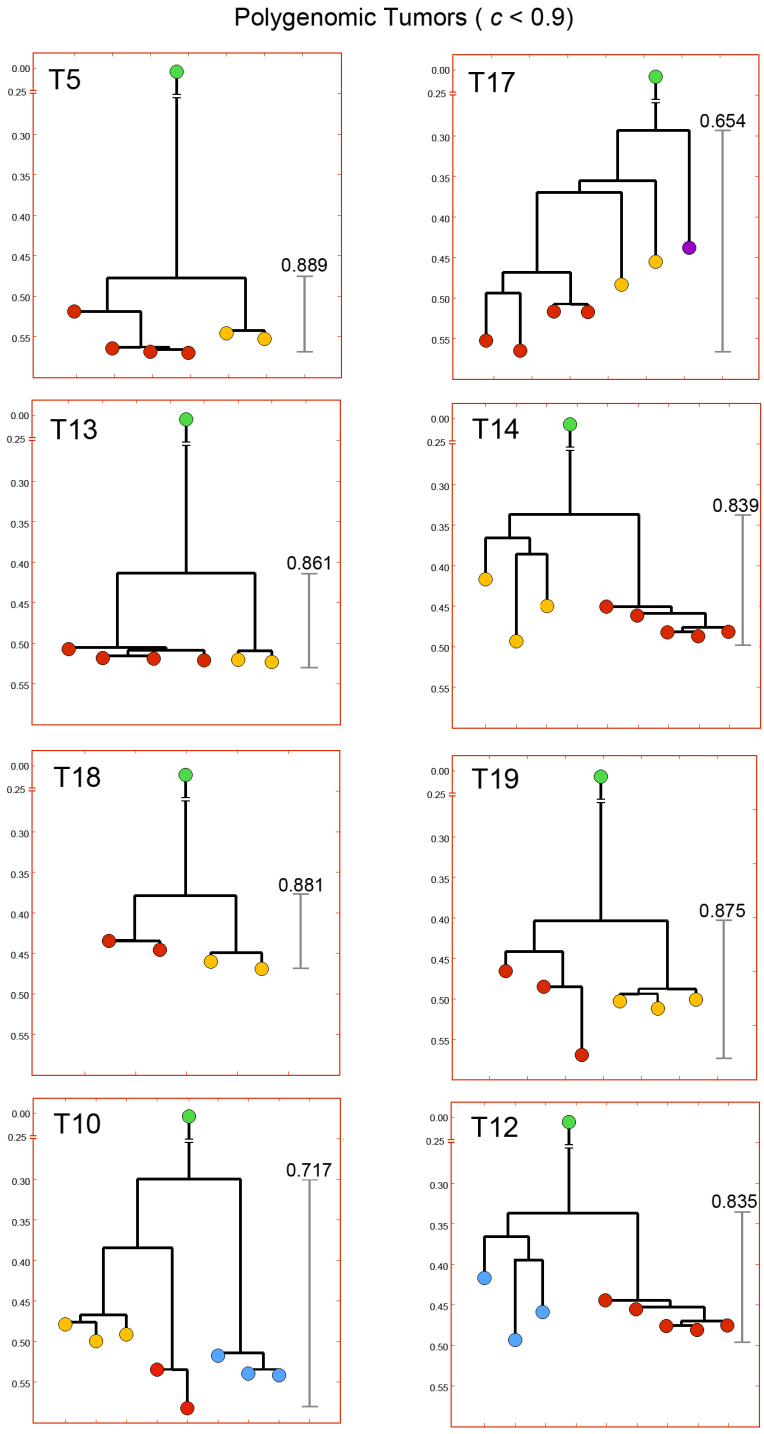


Figure 7.3 – NJ Distance Trees of Polygenomic Tumors
Neighbor-joining trees of tumors with a minimum correlation of profiles less than 0.9. All tumors with a correlation less than 0.9 are polygenomic and outlined in red.The leaves are colored in red, yellow and blue to show different subpopulations as determined by comparing ROMA copy number profiles. Trees are rooted with a single consensus diploid profile colored in green.

In general, subpopulations within a tumor are very similar and share many or most chromosome breakpoints. On the other hand, we see very few common breakpoints between different tumors. This strongly implies that all subpopulations within a tumor have a common clonal origin. Given the potential importance of this conclusion, we validated it by purely computational analysis. The result of distance clustering of all tumor subpopulations clearly confirms that the subpopulations within a tumor are vastly more related to each other than the subpopulations between tumors (Figure 7.4). We cannot rule out that some tumors are mixtures of totally distinct clones, but we have never seen evidence for this alternate hypothesis (e.g., by observing two completely unrelated subpopulations within the same tumor).

The lineage trees that we constructed from copy number profiles within tumors further support the classification scheme of monogenomic and polygenomic tumors. Moreover, these neighbor-joining trees show the relative genetic distance between divergent subpopulations, as well as an estimate of genetic variation within clonal subpopulations. In the monogenomic tumors we often observed a flat tree structure in which all nodes were highly correlated and diverged an equal distance from the root node (Figure 7.2). In these homogenous tumors, a more detailed genetic lineage is difficult to infer, because no other intermediate subpopulations, representing time points in evolution can be measured. In contrast, polygenomic tumors allow us to infer a detailed genetic lineage. For example in Figure 7.3, the T10 tree shows three major clonal subpopulations (red, yellow and blue nodes). This tree shows that the blue subpopulation is closer to the normal diploid cells, while the red and yellow subpopulations are more related to each other and have diverged the greatest distance from normal. In many polygenomic breast tumors, the inferred trees showed that copy number profiles were clearly related, and shared the majority of chromosome breakpoints, suggesting a common genetic lineage from a single precursor cell. Moreover, the copy number profiles in these tumors were always organized into highly similar groups, representing clonal subpopulations. Thus, genomic heterogeneity can be ascribed to a few major subpopulations rather than a series of gradual intermediates. From these data we conclude that the majority of chromosome breakpoints are inherited from previous subpopulations and persist through the evolution of more advanced subpopulations, as clones expand to form the mass of the tumor.

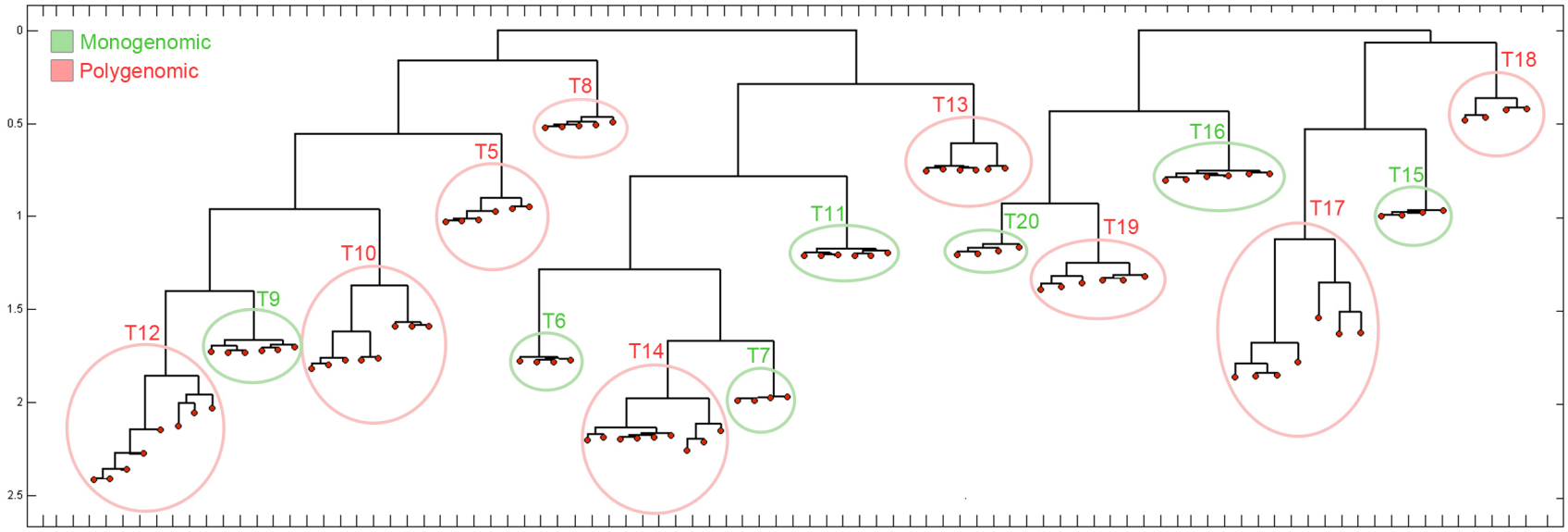


Figure 7.4 – NJ Tree of All Breast Tumors
Neighbor-joining trees was calculated from all tumors using segmented copy number profiles. Two trees were calculated independently: one from 85K experiments (T4-T14) and one from the 390K experiments (T15-T20). Each green circle represents one monogenomic tumor, while each red circle represents one polygenomic tumor.

CHAPTER 8

Inferring Tumor Progression from Polygenomic Tumors

The order of progression can be inferred from genomic markers in tumor subpopulations if we make two assumptions. The first assumption is that the tumor subpopulations have arisen from a common progenitor tumor cell. The second assumption is that there is no "reversion to normal" in a lineage once a change occurs. In other words, observable mutations only accumulate. There can be violations of this assumption, for example, if a chromosome with changes is subsequently lost. Also, violations of this assumption can arise due to observing mixtures of subpopulations.

As we have shown in the previous chapter, in almost all cases the subpopulations within a tumor have many similar copy number changes (Figure 7.2 and 7.3), but have few in common with other tumors (Figure 7.4B), justifying the assumption of a common origin for subpopulations in each individual tumor. However, tumor T4 had sectors with essentially no discernible copy number changes ("flat" profiles), and other sectors with many chromosomal breakpoints (Figure 4.3). The sectors with flat profiles nevertheless were full of malignant cells as judged by histopathology. Thus a common origin for tumor cells with flat profiles and for those with copy number changes cannot strictly be inferred.

In the general case, we assume that mutational complexity increases in time and make inferences about the order of progression. To compare clonal subpopulations we coalesce segmented profiles within a cluster into a consensus profile by taking the segmented value that was most frequent in all profiles (majority rules, rounding up). The pair-wise difference between coalesced profiles was then calculated to identify subpopulation-specific amplifications and deletions. The profiles were then ordered based on increasing numbers of chromosome breakpoints.

8.1 Progression in Basal-like Breast Tumors

Two of the most extreme examples of progression are seen in two basal-like breast tumors T10 and T12, a particularly aggressive subtype of breast cancer

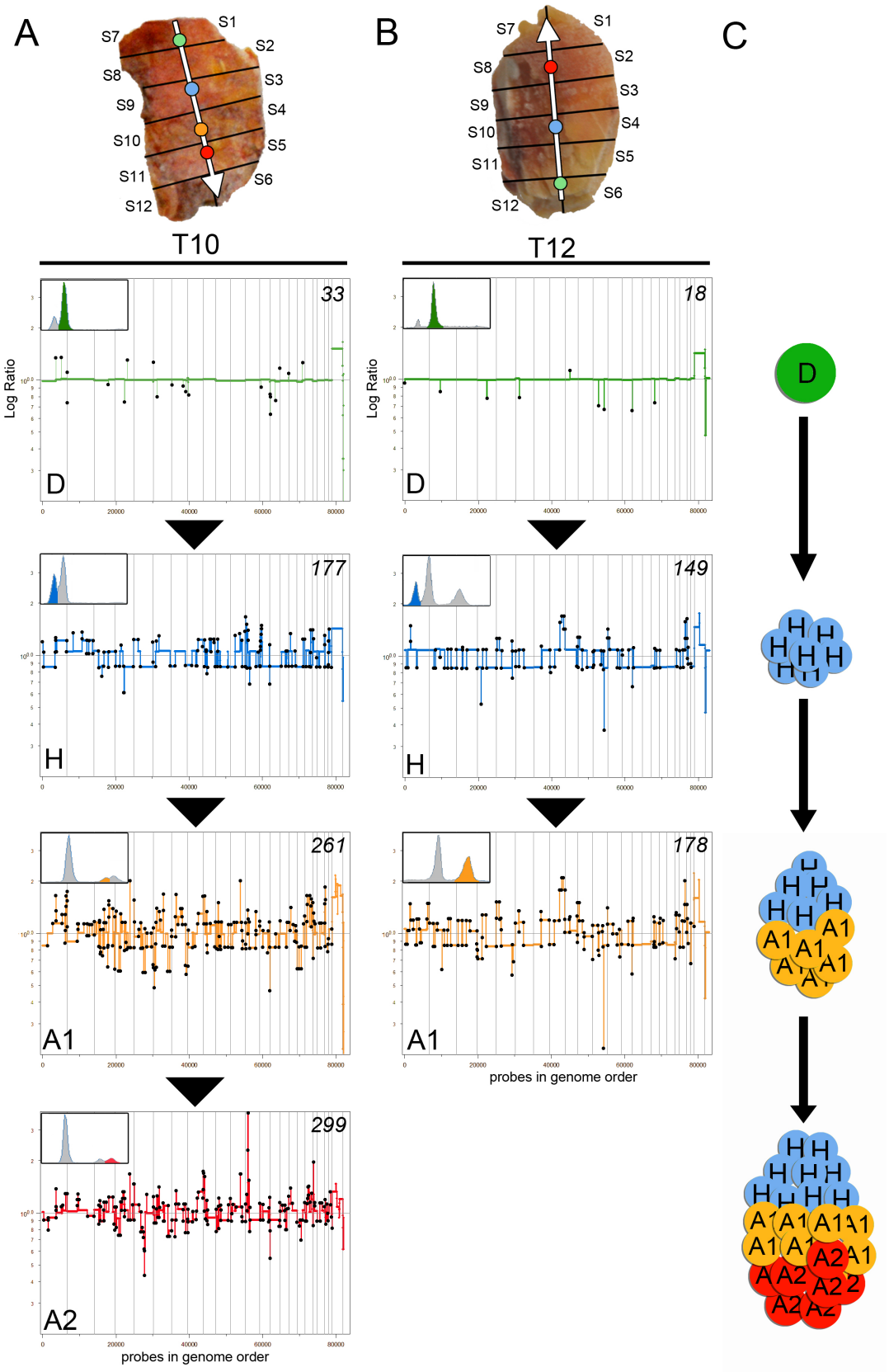

A
B
C
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
T10
T12
Log Ratio
33
18
177
149
261
178
299
D
H
A1
A2
probes in genome order
0
20000
40000
60000
80000

Figure 8.1 – Progression in Basal-like Breast Tumors
Consensus copy number profiles from each basal-like breast tumor (T10 and T12) were ordered by increasing numbers of chromosome breakpoints. (A-B) White arrows represent the direction of growth as subpopulations evolve from H (blue) to A1 (orange) to A2 (red). FACS histograms are shown with the gated subpopulation highlighted in color. (A) Tumor T10 progresses from diploid (D, green) to hypodiploid (H, blue), to hyperaneuploid (A1, yellow), to hyperaneuploid (A2, red), as the number of chromosome breakpoints increases. (B) Tumor T12 progresses from diploid (D, green) to hypodiploid (H, blue) to hyperaneuploid (A1, yellow). (C) Illustration of the clonal expansion of subpopulations as the tumor grows.

with triple-negative receptor status (ER-, PR-, Her2) and poor survival outcome. Recently, the genome structure of basal-like breast tumors has been explored by array CGH, showing a 'sawtooth' copy number pattern caused by the loss of many broad genomic regions (Bergamaschi et al., 2006; Chin et al., 2007; Hicks et al., 2006a). This genome pattern correlates to one of the subpopulations we identified in these tumors, which we refer to as Hypodiploid (H) that was present in both T10 and T12. Additionally, we found one (T12) or two (T10) more advanced aneuploid subpopulations (A1 and A2) in these tumors by SPP, as well as diploid cells in every sector. The hypodiploid subpopulations were isolated from a ploidy distribution of 1.7N, just below the diploid distribution. The aneuploid subpopulations contained much higher ploidies by FACS, and the CGH profiles showed that they had acquired many focal amplifications and deletions not seen in the hypodiploid subpopulation.

Assuming that mutational complexity increases with time, we ordered the genomic profiles by increasing numbers of chromosome breakpoints (Figure 8.1). We found that the basal-like tumors progressed from diploid to hypodiploid, which correlated with a downward shift in total DNA content (as indicated by the FACS histogram) and loss of many broad chromosomal regions in the genome profiles. The hypodiploid subpopulation then diverged to form the aneuploid subpopulations, correlating with a drastic increase in total DNA content and multiple genome-wide focal amplifications and deletions (Figure 8.1, lower panels). In T10, the A1 subpopulation continued to evolve into the A2 subpopulation, acquiring a massive amplification of the *KRAS* locus on chromosome 12p12.1 and a homozygous deletion of the *EFNA5* tumor suppressor to form the A2 subpopulation, correlating which another upward shift in total DNA content by FACS (Figure 8.1A, lowest panel).

From this, we infer a sequential pattern of progression in the basal-

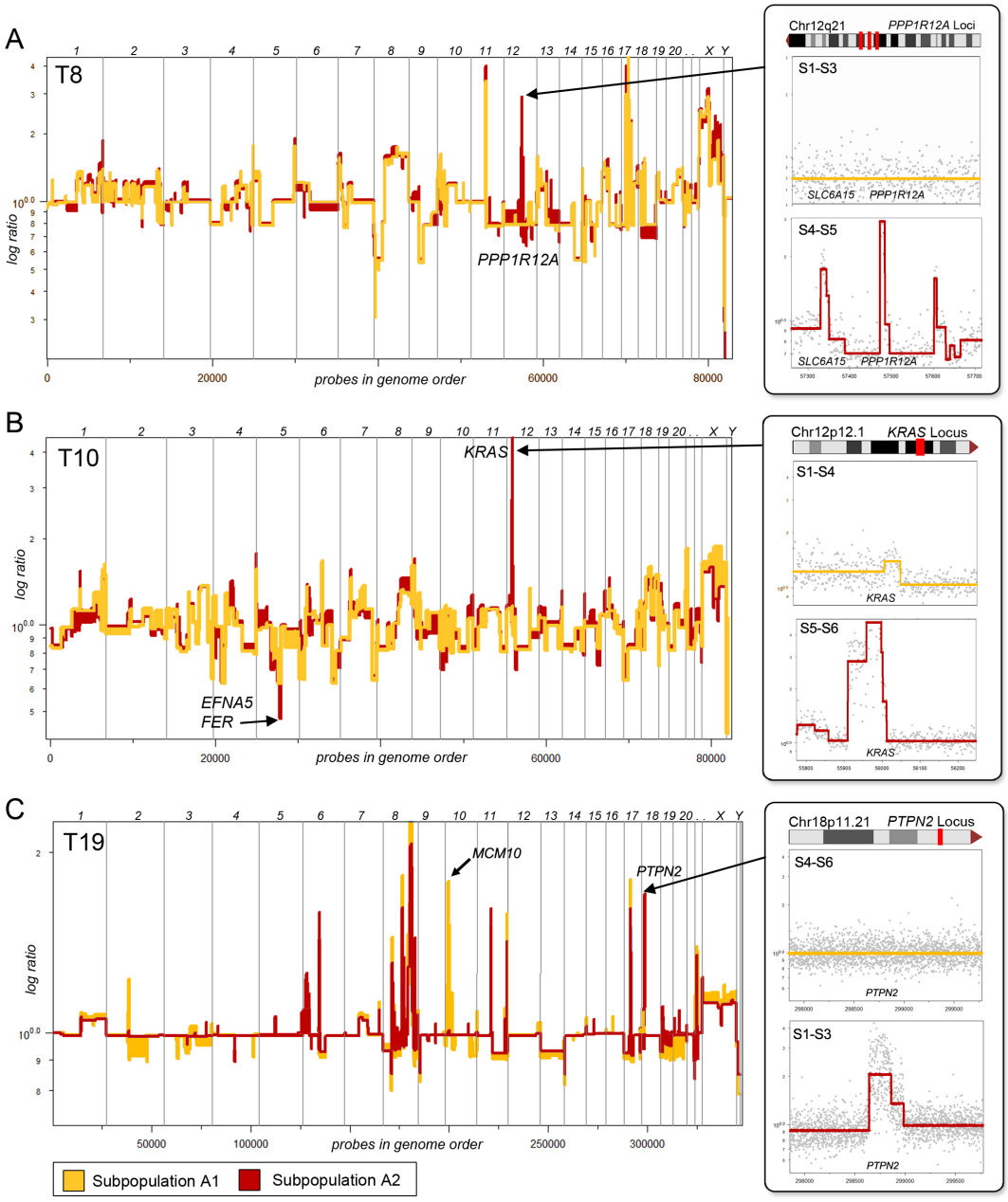


Figure 8.2 – Focal Lesions that Differ Between Subpopulations
Segmented copy number data from coalesced tumor profiles are plotted in genome order. (A) Tumor T8 contains three focal amplifications, including the amplification of the *PPP1R12A* locus on Chr12q21, which is present in the A2 tumor subpopulation (red), but absent in A1 (yellow). (B) Tumor T10 contains a focal amplification of the *KRAS* locus on Chr12p12.1, which is present in the A2 tumor subpopulation (red), but absent in A1 (yellow). T8 also contains a homozygous deletion of the *EFNA5* and *FER* locus on chrom 5q21.3 in the A2 subpopulations (red) which is a hemizygously deleted in A1 (yellow) (C) Tumor T19 contains a focal amplification of the *PTPN2* locus on chrom18p11.21, which is present in the A2 subpopulation (red), but absent in A1 (yellow). T19 also contains a focal amplification of the *MCM10* locus on chromosome 10p13 in the A1 tumor subpopulation that is absent in A2.

like tumors in which much of the genome is first deleted, followed by endoreduplication or cell fusion to generate a highly aneuploid genome that continues to acquire many focal amplifications and deletions of cancer genes. The large degree of genomic heterogeneity in this subtype of breast cancer make them an ideal tool for studying tumor progression and may result in clinically useful diagnostic markers to determine how far along these tumors have progressed in breast cancer patients.

8.2 Focal Differences Between Tumor Subpopulations

The most prominent differences between subpopulations within a tumor were changes in the copy number of broad chromosomal regions. However, many polygenomic tumor subpopulations diverged by a small number of focal (narrow) genetic events, and we may infer that these focal changes occurred "late," after tumor initiation and considerable expansion. Overall, we identified 24 focal lesions that differed between tumor subpopulations: 12 amplifications and 12 deletions (Table 8.3). Each genomic lesion was annotated to identify UCSC genes (Hsu et al., 2006) and cancer genes. Cancer genes were identified using a compiled database from the cancer gene consensus (Futreal et al., 2004) and the NCI cancer gene index (Sophic Systems Alliance Inc., Biomax Informatics A.G). As we expected, many focal amplifications encompassed known oncogenes, including *KRAS*, *PPP1R12A, HRASLS, MYC, RAD52,* and *RARA*; while the deletions eliminated known tumor suppressors, such as *CDKN2A*, *CASK, EFNA5, FER, PAX8*, and *ERCC3* (Futreal et al. 2004). Furthermore, we identified many focal deletions and amplifications containing single genes not previously implicated in cancer, including *CACNA1C, HYDIN, SLC6A15, DCLK2, DNER,* and *C11ORF87*. The latter group are ideal candidate for *in vitro* overexpression (oncogenes) or RNAi (tumor suppressor) experiments to determine if they play a functional role in breast cancer progression.

We illustrate focal differences with three polygenomic tumors (T8, T10, and T19). The T8 tumor subpopulations diverged by only three tandem genomic amplifications on chromosome 12q21.1 present in the A1 tumor subpopulations in sectors 4 and 5, but not sectors 1 to 3 (Figure 8.2A). These focal regional amplifications encompassed three single genes—*BC061638*, *SLC6A15*, and *PPP1R12A*—the former of which have not previously been implicated in cancer. The T10 tumor subpopulations diverged a massive (10 fold) amplification and a

#	Tumor	Present	Absent	Loc	Exc	Event	Ratio	Size (kb)	Chr	Cytoband	Start HG18	Stop HG18	Cancer Genes	Known Genes
1	T5	A2	A1	S3-S4	S1-S2	del	1:2	1,148	1	q32.2	207859891	209007930	-	*LAMB3, G0S2, HSD11B1, IRF6, SYT14, HHAT, KCNH1*
2	T5	A1	A2	S1-S2	S3-S4	del	1:2	97	12	p13.33	2060341	2157396	-	*CACNA1C*
3	T5	A1	A2	S1-S2	S3-S4	del	1:2	37	16	q12.1	47582130	47619657	-	*NT_010498.59*
4	T5	A1	A2	S1-S2	S3-S4	amp	3:2	287	16	q22.2	69470004	69757528	-	*HYDIN*
5	T8	A2	A1	S4-S5	S1-S3	amp	4:2	864	12	q21.1	72269679	73134226	-	*BC061638, BC094833*
6	T8	A2	A1	S4-S5	S1-S3	amp	6:2	373	12	q21.2-21.31	78696935	79069950	*PPP1R12A*	-
7	T8	A2	A1	S4-S5	S1-S3	amp	4:2	206	12	q21.31	83688476	83895005	-	*SLC6A15*
8	T10	A1,A2	H	S5,S6	S1-S4	del	1:2	149	3	q21.3	127728953	127878837	-	*CHST13, TR2IT1*
9	T10	A1,A2	H	S5-S6	S1-S4	amp	4:2	5	4	q31.3	151282090	151287122	-	*DCLK2*
10	T10	A2	H,A1	S5,S6	S1-S4	del	1:2	7978	5	q21.1-22.1	101814799	109793050	*EFNA5, FER*	*PAM, FBXL17, SLC06A1, PJA2, MAN2A1*
11	T10	A2	H,A1	S5,S6	S1-S4	amp	10:2	3652	12	p12.1	22083693	25736050	*KRAS*	*SOX5, ETNK1, CMAS, BCAT1, LRMP, CASC1*
12	T12	A1	H	S1-S4	S5-S6	del	1:2	128	5	q33.2	153282447	153410942	-	*MFAP3, FAM114A2*
13	T12	A1	H	S1-S4	S5-S6	del	0:2	153	11	q22.3	108696368	108849416	-	*c11orf87*
14	T12	A1	H	S1-S4	S5-S6	amp	3:2	215	17	q21.1-q21.2	35505295	35720207	*CDC6, RARA*	*NR1D1, CASC3, RAPGEFL1, WIRE, WIPF2*
15	T12	A1	H	S1-S4	S5-S6	amp	4:2	419	20	q13.13	48157873	48577190	*PTPN1*	*UBE2V1, CEBPB. TMEM189*
16	T14	A2	A1	S2-S4	S1,S5-6	del	1:2	371	2	q36.3	229951523	230322758	-	*DNER*
17	T14	A2	A1	S2-S4	S1,S5-6	del	0:2	220	11	q12.1	58007425	58227622	*LPXN*	*ZFP91, CNTF*
18	T14	A2	A1	S2-S4	S1,S5-6	del	0:2	639	22	q13.31	46146803	46786015	-	*FLJ46257*
19	T17	A1	A2	S1-S3	S4	amp	3:2	1247	1	q44	242836931	244084235	*SMYD3*	*FAM36A, HNRNPU, EFCAB2, KIF26B*
20	T17	A1	A2	S1-S3	S4	amp	3:2	671	22	q11.21	17671011	18342500	*SEPT5, CDC45L*	*HIRA, UFD1L, CDC45L, CLDN5, TBX1, TXNRD2, COMT*
21	T18	A1	A2	S1-S2	S3-S4	del	1:2	97	7	q21.13	89450127	89547319	*CREB3L2*	-
22	T18	A1	A2	S1-S2	S3-S4	del	1:2	422	X	p11.4	41494040	41916836	*CASK*	-
23	T19	A1	A2	S1-S3	S4-S6	amp	3:2	6652	10	p14-p12.33	11137382	17789776	*MCM10*	*32 known genes*
24	T19	A2	A1	S4-S6	S1-S3	amp	3:2	1790	18	p11.21	12150130	13940735	*PTPN2*	*CIDEA, TUBB6, SPIRE1, SEH1L, CEP192, RNMT, MC5R*

Table 8.3 – Summary Table of Subpopulation-specific Focal Lesions
Focal lesions that differ between tumor subpopulations were annotated for cancer genes and known genes. Twelve amplifications and twelve deletions were mapped to the UCSC human genome 18 (March, 2006). Cancer genes were annotated using the NCI Cancer gene index by Sophic Alliance (www.sophicalliance.com) and the Sanger Cancer Gene Census (www.sanger.ac.uk/genetics/CGP/Census). Known genes were annotated using the UCSC known gene index (genome.ucsc.edu). The columns are:

identification number of the focal lesion
Tumor tumor identification number
Present indicates the tumor subpopulation that contains the lesion
Absent indicates the tumor subpopulation that does not contain the lesion
Loc the anatomical sector(s) that contains the lesion
Exc the anatomical sector(s) from which the lesion is excluded
Event describes if the focal lesion is an amplification (amp) or deletion (del)
Ratio log ratio of the focal lesion from the segmented coalesced copy number profile
Size genomic interval of the focal lesion in kilobases (kb)
Chr chromosome to which the lesion has been mapped
Cytoband cytogenetic band in which the lesion has been mapped
Start HG18 start coordinate of the focal lesion
Stop HG18 stop coordinate of the focal lesion

single homozygous deletion (Figure 8.2B). The region of chromosome 12p12.1 contains the *KRAS* oncogene and was present at greater than 10 copies in the A2 subpopulation in sectors 5 and 6, but only present at three copies in the A1 subpopulation. The T19 tumor subpopulations diverged by two amplifications on chromosome 10p14-p12.33 and 18p11.21 containing the *MCM10* and *PTPN2* oncogenes, respectively.

In this chapter, we have shown that genomic deletions and amplifications are stable markers that are useful for studying tumor progression because they represent time points in the evolution of the tumor. Assuming that mutational complexity increases with time, we ordered the copy number profiles and identified genetic events that occurred during tumor progression. In the triple-negative basal-like tumors (T10, T12) we observed large chromosomal rearrangements that occured early in the evolution of the tumor followed by focal amplifications and deletions in the later stages of progression. However, most polygenomic tumors (T3, T5, T13, T14, T17, T18, T19) did not show such drastic rearrangements between subpopulations, but rather focal amplifications and deletions that often encompassed cancer genes. Such mutations are likely to affect gene dosage and provide a selective advantage for the clones to expand in the tumor microenvironment, forming the major subpopulations. In the next chapter, we use these focal changes as markers to analyze the spatial relationship of clones within tissue sections.

CHAPTER 9

Cytological Analysis of the Spatial Organization of Tumor Clones

It is evident even from our crude dissections and flow-sorting that some tumor subpopulations are regionally segregated, while others cohabit the same sector. Using the SPP genomic approach we were able to identify many subpopulation-specific chromosome markers (Table 8.3) that can be used with locus-specific cytological techniques to qualitatively distinguish subpopulations at single cell resolution. We focused our study on one tumor (T10), because it contained both subpopulations that were anatomically segregated (H is present only in the upper sectors) and subpopulations that were intermixed (A1 and A2 were intermixed in the lower sectors) by SPP analysis. To explore the spatial organizations of these subpopulations in tissue sections at single cell resolution, we applied fluorescence in situ hybridization (FISH) to observe subpopulation-specific chromosome markers that were identified by CGH. FISH is an orthogonal approach to SPP, but further allows us to explore the anatomic segregation and intermixing of tumor clones *in situ* in tissue sections from different sectors.

9.1 – PROBER

Traditional FISH methods use bacterial artificial chromosomes (BACs) as fluorescent DNA probes to target relatively large regions of the genome (> 200-500kb). However, this method was inadequate for our purposes for three reasons: availability, resolution and an inability to avoid repetitive genetic elements. Most of the loci that we were interested in analyzing were less than 100kb and often contained repetitive elements such as simple repeats, microsatellites, LINES and LTRs. To address these issues, we developed a new method for designing FISH probes which combines a computational algorithm with an experimental approach, called 'PROBER' (Navin et al., 2006) available at (http://prober.cshl.edu).

PROBER is an oligonucleotide primer design software application that designs multiple primer pairs for generating PCR probes useful for fluorescence in situ hybridization (FISH). PROBER generates Tiling Oligonucleotide Probes (TOPs) by masking repetitive genomic sequences and delineating essentially

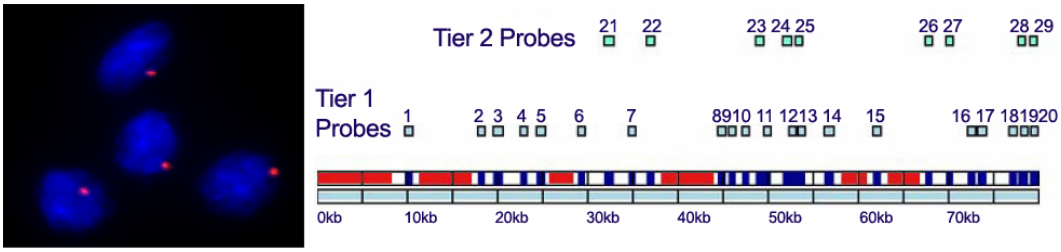


Figure 9.1 – Tiling Oligonucleotide FISH Probes
A cocktail of tiling FISH probes 1–29 were designed across an 80kb region. Highly repetitive areas (red) were avoided. Blue areas denote the genomic regions covered by Tier1 or Tier2 FISH probes. White areas did not contain sequence that was suitable for probe primers. The tiling probe cocktail was hybridized to a breast cancer cell line that was known to contain a hemizygous deletion by CGH. FISH experiments validated the hemizygous deletion on chromosome 16q1.

unique regions that can be amplified to yield small (100–2000bp) DNA probes that in aggregate will generate a single, strong fluorescent signal for regions as small as a single gene. In theory, PROBER can be applied to any genomic locus, with the limitation that the locus must contain at least 10 kilobases of essentially unique blocks. To design probes, genomic DNA sequences are retrieved from a server, masked for repetitive exact string matches in the human genome using a wheeler-Burrows transformation of the human genome into a suffix array (Healy et al., 2003), and analyzed for contiguously amplifiable, nearly repeat free regions of sufficient aggregate length. These regions are searched for optimized PCR forward and reverse primers by the following criteria: size range from 500-2000bp, matching melting temperatures, nucleotide repeats <4, and must end in G/C at the 3' end to control mispriming. The result is a collection of oligonucleotide probes within a specific locus that avoid repetitive elements (Figure 9.1). Individual tiling probes are then PCR amplified and combined into a cocktail for FISH analysis that can be fluorescently labeled by nick-translation and hybridized to cells following standard FISH protocols to detect copy number signals.

Chapter 9.2 - Regional Amplification of *KRAS*

In our study of T10 by SPP we identified a massive 10-fold amplification that was found exclusively in sectors 5 and 6 of this tumor. This 3.6mb locus amplified the *KRAS* oncogene in addition to several other genes that have not previously been implicated in cancer (Figure 9.2). To validate this finding with an orthogonal

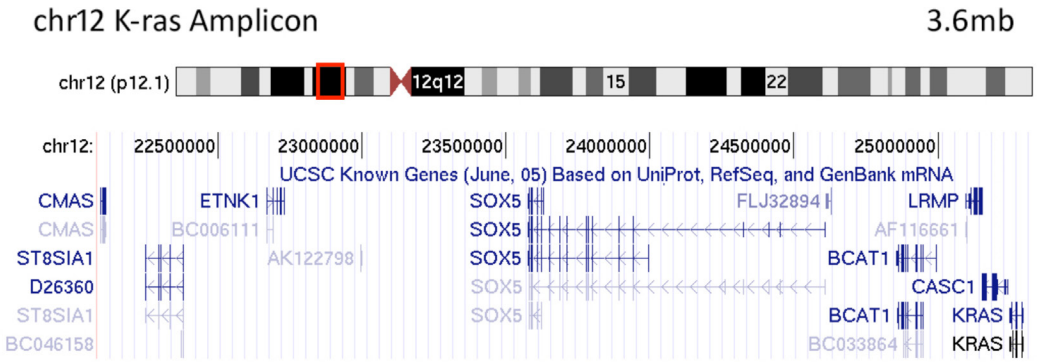
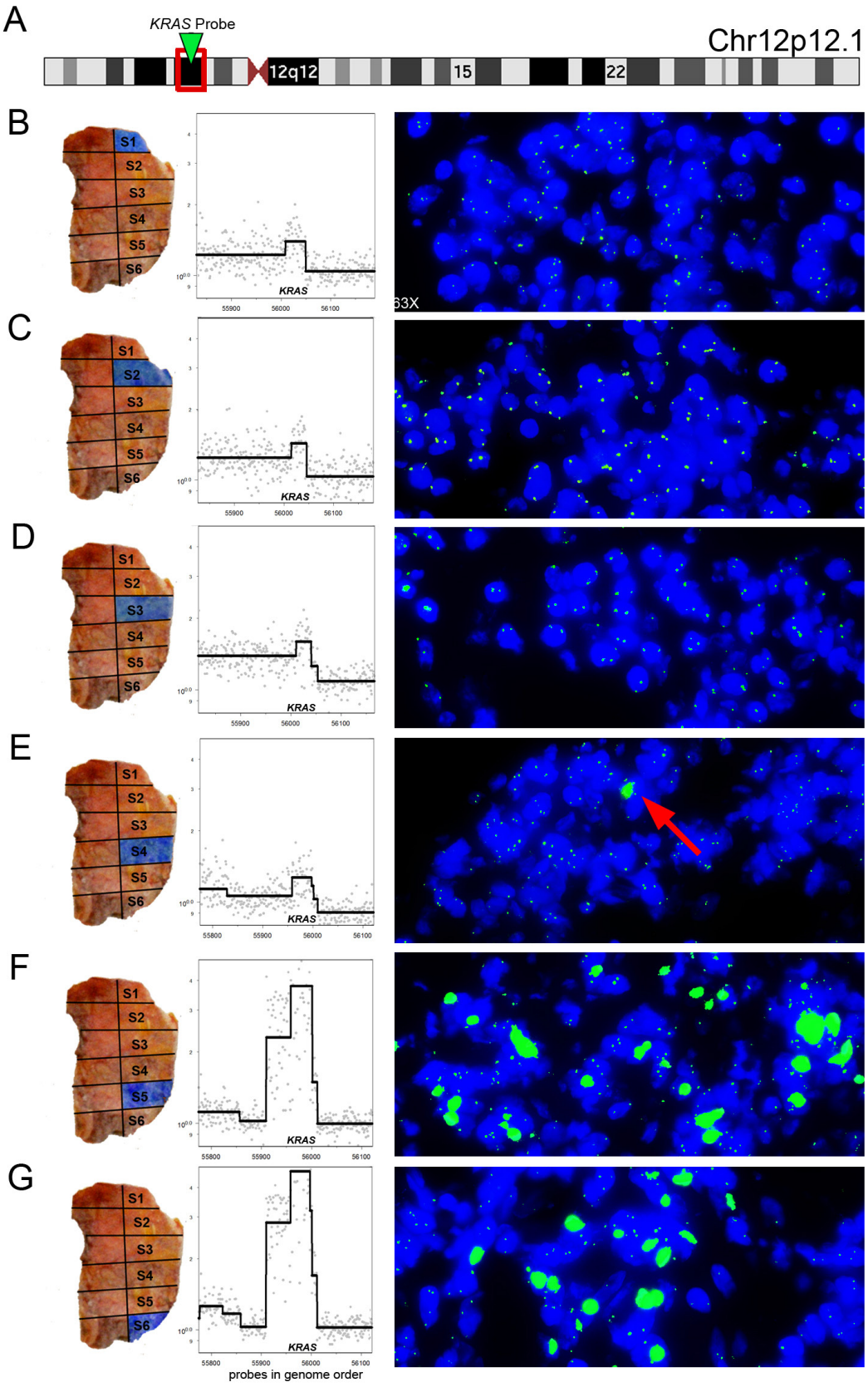


Figure 9.2 – Gene Annotations on Chromosome 12p12.1
Tumor T10 has a massive >10 fold amplification of a 3.6 megabase region on chromosome 12p12.1 in sectors 5-6. This region contains several genes, including the *KRAS* oncogene.

approach and to observe the amplification in single cells, we designed TOP probe to target the *KRAS* locus using PROBER (Navin et al., 2006). We hybridized this probe to six frozen tissue sections corresponding to the sectors analyzed by ROMA. As expected, our interphase FISH experiments validated the massive amplification of the *KRAS* locus in the lower sectors (S5-S6) of this tumor in a small subpopulation of cells (Figure 9.3). Within the other sectors (S1–S4), the stroma and tumor cells exhibited just two or three copies of the *KRAS* locus expected from the CGH profiles. Additionally, in two microscopic fields of about 500 tumor cells in sector 4, we observe one isolated cell that was highly amplified for *KRAS* (Figure 9.3E), and could not be detected by ROMA using samples of millions of cells. These results clearly show that tumor subpopulations can be anatomically segregated to different regions within the tumor mass. This finding has important clinical implications, since diagnostic molecular assays are often based on samples taken from a single location within a tumor.

Figure 9.3 – Regional Amplification of the *KRAS* Locus
FISH experiments were performed using tissue sections from sectors 1-6 from tumor T10 using a single tiling probe specific to the *KRAS* locus. (A) Ideogram showing the cytobands and location of the *KRAS* FISH probe on chromosome 12p12.1 (B-G) Left panels show the tumor sector from which the tissues sections are cut. Middle panels show the ratio and segmented CGH data for the *KRAS* locus in each tumor sector. Right panels show the resulting FISH experiments with 2 or 3 copies of the *KRAS* probe in S1-S4 and numerous copies in S5-S6 shows as a homologous staining region. (E) Red arrow shows a single *KRAS* cell in Sector 4.


A
KRAS Probe
Chr12p12.1
12q12
15
22
B
C
D
E
F
G
S1
S2
S3
S4
S5
S6
KRAS
55900
56000
56100
55800
63X
probes in genome order

Chapter 9.3 – Intermixing of Tumor Clones

The presence of multiple tumor subpopulations in sectors is most obvious in polygenomic tumors where the FACS histograms show multiple aneuploid peaks. It is not clear from FACS, however, whether these co-occupied sectors result from our gross dissection crossing a boundary between segregated neighborhoods, or, alternatively, from an organization in which the subpopulations physically intermix. In theory the tumor clones can have several organizations within tissues, such as internal clusters, at the peripheral organization or statistic intermixing (Figure 9.4). We sought to explore this organization of clones in tissues of T10 at single cell resolution using FISH.

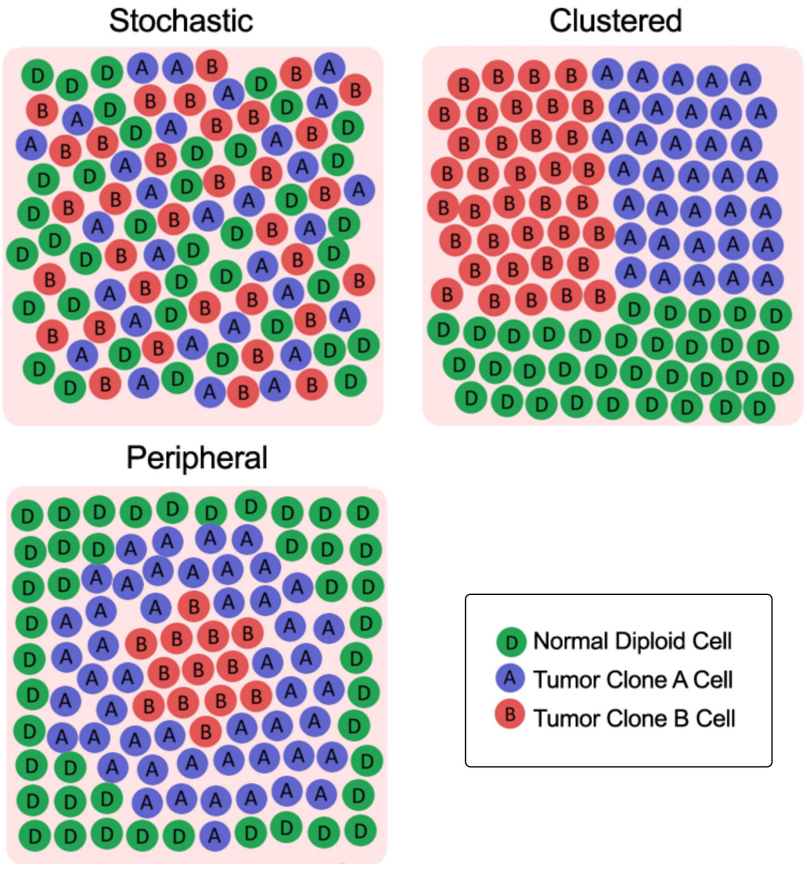


Figure 9.4 – Theoretical Organization of Clones
In theory distinct tumor subpopulations in tissues could have several organizations: stochastic intermixing, clustering in distinct domains, or a peripheral organization.

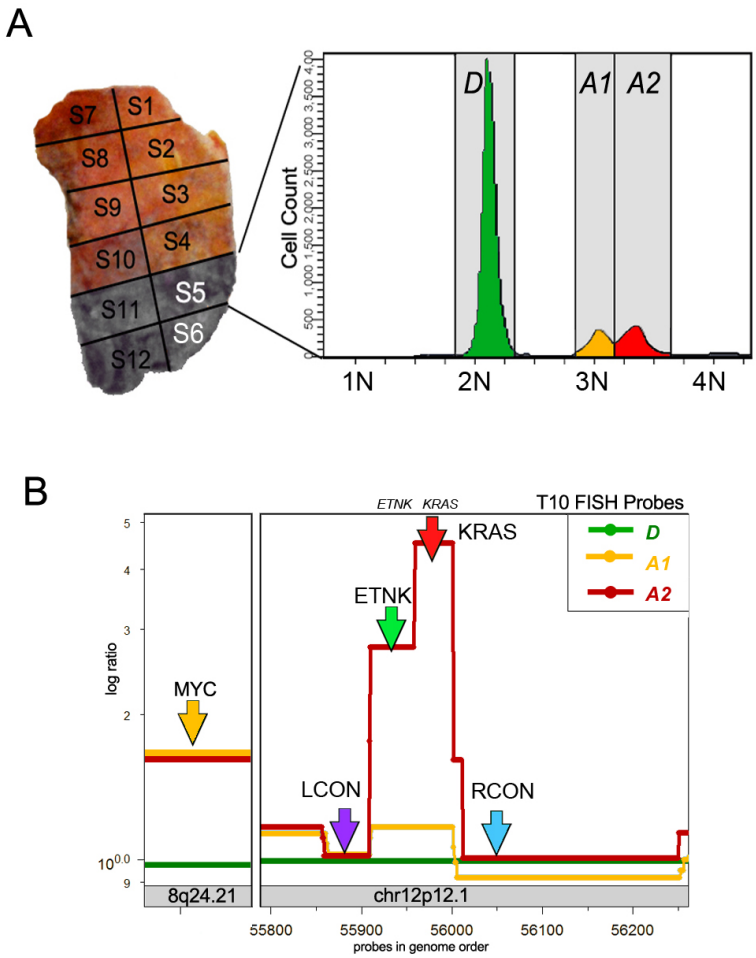


Figure 9.5 – FISH Probe Strategy
(A) FACS histograms from the lower sectors of T10 show the presence of three subpopulations (D, A1 and A2) with distinct ploidies (2.0N, 3.1N and 3.3N). (B) Segmented copy number profiles of the three subpopulations on chromosome 8q24.31 and 12p12.1. Alltogether five FISH probes will be used to distinguish subpopulations. The MYC probe discriminates between the diploid and aneuploid subpopulations. The LCON and RCON controls will show the same diploid copy number in all subpopulations. The ETNK and KRAS probes will discriminate between the A1 and A2 tumor subpopulations, showing over 10 copies in A2.

T10 contains two distributions of aneuploid cells (A1 and A2) in the lower sectors (S5-S6) of the tumor mass (Figure 9.5A). The FACS histograms shows the ploidy of A1 is 3.1N, while the ploidy of A2 is 3.3N, with an additional subpopulation of normal diploid cells at 2.0N. From the CGH propfiles, we identified two major distinguishing features between the A1 and A2 subpopulations: a homozygous deletion on chromosome 5q21.1-22.1 of *EFNA5* and an amplification of the *KRAS* locus at 12p12.1 to more than 10 copies. To explore how these subpopulations are organized in tissues, we used a complex of

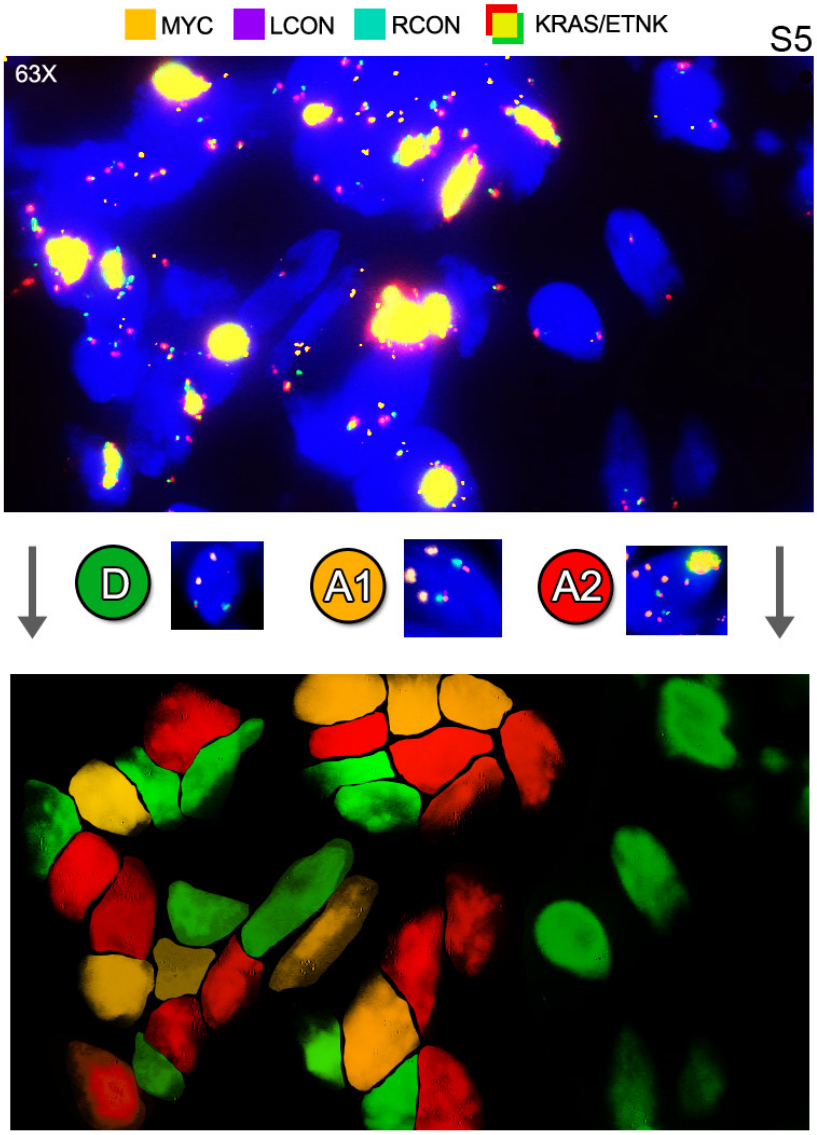


Figure 9.6 – Intermixing of Subpopulations in Tissues from Sector 5
Interphase FISH experiments were performed on frozen tissue section from sector 5 of T10. Five FISH probes (LCON, RCON, *MYC*, *KRAS*, *ETNK*) were hybridized to distinguishing the subpopulations. Upper panel shows a 63X field showing the normal diploid cells with 2 copies of all probes, the A1 subpopulation with 3 copies of *MYC* and the A2 subpopulation that shows a bright yellow signal due to the colocalization of the *KRAS* and *ETNK* probes. Lower Panel A false-colored DAPI channel shows the location of the clones from each subpopulation, revealing a stochastically intermixed organization.

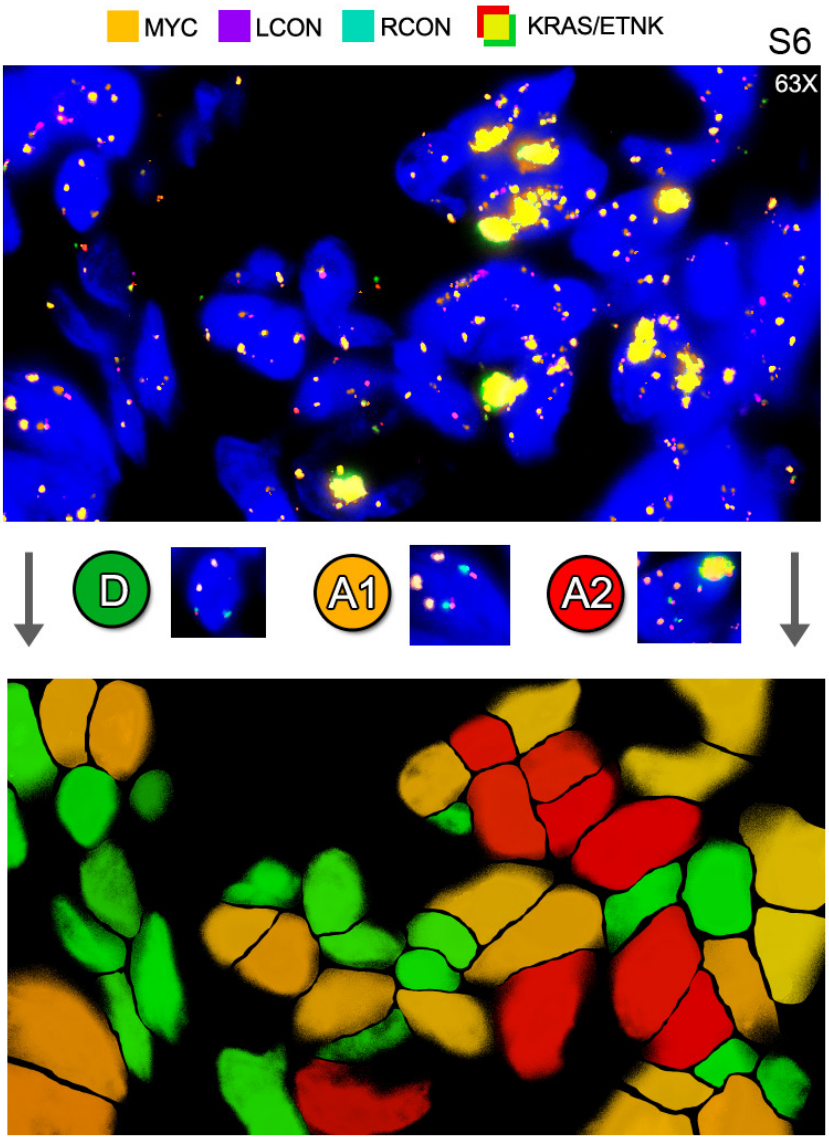


Figure 9.7 – Intermixing of Subpopulations in Tissues from Sector 6
Interphase FISH experiments were performed on frozen tissue section from sector 6 of T10. Upper panel shows a 63X field showing the normal diploid cells with 2 copies of all probes, the A1 subpopulation with 3 copies of *MYC* and the A2 subpopulation that shows a bright yellow signal due to the colocalization of the *KRAS* and *ETNK* probes. Lower Panel A false-colored DAPI channel shows the location of the clones from each subpopulation, revealing a stochastically intermixed organization.

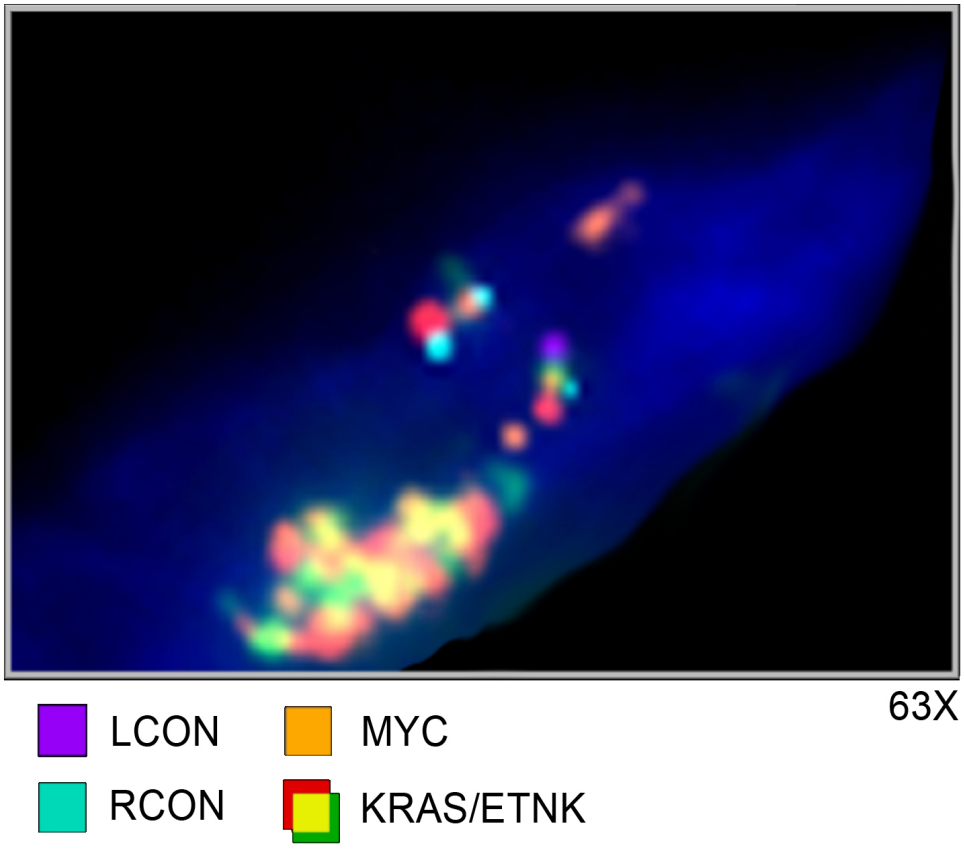


Figure 9.8 – Homologous Staining Region in a *KRAS* Cell
Interphase FISH experiment showing a single *KRAS* cell at 63X objective. The control probes (LCON and RCON) are present at diploid copy number, while *MYC* shows three copies. The two arms of chromosome 12p12.1 are evident by the linear organization of the FISH probes. The *KRAS* and *ETNK* probes colocalize showing a bright yellow signal and are located near one arm of chromosome 12p. Their signal reveals a massive increase in copy number that is localized on 12p in the form of a homologous staining region.

FISH probes capable of distinguishing A1 and A2 from normal stroma and from each other (Figure 9.5B). To distinguish A1 and A2 from normal stroma, we used a *MYC* probe present in both the A1 and A2 subpopulations at a copy number of three. To distinguish A2 from A1, we used two probes (*ETNK* and *KRAS*) that colocalize to the region with a highly amplified KRAS locus in A2. We visualized all cells, tumor and diploid, using two probes, *LCON* and *RCON*, that map just outside the amplified region on A2. The probe scheme and location of the mixed sector 5 of T10 are shown in Figure 9.5B.

The multi-color complex of FISH probes was hybridized to tissue sections from sectors 5 (Figure 9.6) and sector 6 (Figure 9.7) in T10. These experiments allowed us to clearly identify the diploid cells, A1 cells and A2 cells in the tissues, showing that the organization of these single cells are intermixed, rather than occupying separate domains. The A2 cells are easily identified in the FISH images because they contain many copies of both the *ETNK1* and *KRAS* probes as a bright yellow signal. To aid in identifying the other cells, we used false-colored DAPI channels to show the organization of the diploid cells (green), A1 cells (yellow) and A2 cells (red) in sector 5 (Figure 9.6) and sector 6 (Figure 9.7). At high magnification (63X) the *KRAS* amplification appears as a homogeneous staining region (HSR) by cytological classification, and is evident by numerous copies of the *KRAS* and *ETNK* probes colocalizing, while the other probes appear like ‘beads along a string’ showing diploid copy number on chrom 12p12.1 (Figure 9.8). These results show that genetically divergent tumor subpopulations and normal diploid cells can be intermixed within tissues, raising interesting questions about the cooperativity of tumor cells. We now turn to a detailed discussion of these results.

CHAPTER 10

Conclusions from the SPP Study

Dissecting the clonal composition of tumors at the genetic level is key to understanding the nature and progression of cancer and assessing prognosis and treatment. Genomic heterogeneity has long been reported in breast tumors, but with conflicting results, some suggesting that breast tumors are homogeneous (Endoh et al., 2001; Noguchi et al., 1992) and some heterogeneous (Farabegoli et al., 2001; Shipitsin et al., 2007; Teixeira et al., 1995). These reports were based on analysis of single samples from whole tumors, in which the subpopulations were not separated by differences in topography or ploidy. Only one study examined genomic variation in regionally separated tumor quadrants using CGH and concluded that some breast tumors had genetically distinct quadrants (Torres et al., 2007). Our preliminary analysis of T1–T4 in which we used sectoring and CGH is consistent with this earlier study. In our full study, we analyze a larger number of samples, and more sectors per tumor, and use separation of subpopulations by ploidy as well as FISH to study the clonal composition of tumors. As a result, we describe heterogeneity in both greater breadth and detail, enabling us to infer the progression of subpopulations.

In summary, we find that clonal genomic heterogeneity in breast cancers is very common. We identified 11 polygenomic tumors in our sample of 20 (Table 10.1). In heterogeneous tumors, we observed that the subpopulations may be anatomically separate or intermixed. We also find that these tumors consist of only a few major subpopulations. As we showed for one case, differences in the genome of subpopulations can be exploited to visualize the population substructure of a solid tumor by FISH, enabling us to unravel the developmental organization of tumor growth and the migratory pattern of cells within the tumor. From the shared chromosomal breakpoints, we infer that tumor subpopulations have a common genetic lineage. By comparing subpopulations, we can infer the order of certain genomic events.

In some tumors (T4, T5, T10, T12, and T14) the subpopulations differ by many genomic events. In the case of T4, we observe one subpopulation without discernible genomic copy number changes and another subpopulation

ID	Sectors	FACS	n	c	Sub	Co-oc	Class	Grade	Size (cm)	ER	PR	Her2
T1	4	no	8	-	-	-	mono	2	2.0 x 1.0 x 0.5	na	na	-
T2	4	no	8	-	-	-	mono	3	0.5 x 0.4 x 0.3	na	na	na
T3	4	no	4	-	-	-	poly	3	0.2 x 3.0 x 1.2	+	+	na
T4	4	no	4	-	-	-	poly	3	1.0 x 1.0 x 1.0	+	+	na
T5	4	yes	10	0.81	2	yes	poly	3	2.8 x 0.5 x 0.5	na	na	na
T6	4	yes	8	0.95	1	no	mono	3	2.0 x 0.8 x 0.4	na	na	na
T7	4	yes	8	0.99	1	no	mono	2	1.5 x 1.5 x 1.5	+	+	-
T8	5	yes	10	0.94	2	no	poly	3	2.8 x 2.8 x 2.8	na	na	na
T9	6	yes	12	0.92	1	no	mono	3	2.0 x 1.3 x 0.4	+	-	-
T10	6	yes	14	0.47	3	yes	poly	3	2.7 x 1.4 x 1.1	-	-	na
T11	6	yes	12	0.90	1	no	mono	3	2.0 x 1.0 x 1.0	na	na	na
T12	6	yes	16	0.64	3	yes	poly	3	6.0 x 6.0 x 5.0	na	na	na
T13	6	yes	12	0.76	2	yes	poly	3	2.0 x 2.0 x 1.0	-	-	na
T14	6	yes	15	0.68	1	no	poly	3	2.0 x 0.8 x 0.5	na	na	-
T15	4	yes	8	0.92	1	no	mono	3	0.5 x 0.5 x 0.3	na	na	na
T16	4	yes	8	0.99	1	no	mono	3	1.5 x 1.0 x 0.5	-	-	-
T17	4	yes	8	0.53	3	yes	poly	3	2.6 x 1.0 x 1.0	na	na	na
T18	4	yes	8	0.84	3	no	poly	3	2.2 x 1.0 x 0.8	-	-	-
T19	6	yes	12	0.77	1	no	poly	3	2.0 x 1.3 x 0.8	+	+	+
T20	5	yes	10	0.94	1	no	mono	3	5.0 x 3.0 x 2.0	-	-	-

Table 10.1 – Summary Table of 20 Breast Tumors Analyzed
Twenty primary ductal carcinomas were analyzed by SPP to iden¬tify tumor subpopulations. Nine tumors were classified as monogenomic and eleven tumors as polygenomic. T1-T4 were macro-dissected and analyzed by ROMA. T5-T20 were analyzed by SPP. The column descriptions are:

ID Tumor identification number
Sectors Number of tumor sectors that were macro-dissected
FACS Samples from which tumor nuclei were stained with DAPI and flow-sorted by ploidy
n Total number of copy number profiles analyzed from a single tumor
cc is the the minimum Pearson's correlation of all aneuploid copy number profiles
Sub Number of subpopulations identified
Co-oc Two or more tumor subpopulations co-occupied a single sector in the FACS histogram
Class Tumor was classified as monogenomic (mono) or polygenomic (poly).
Grade Histological tumor grade scored using the modified Bloom-Richardson system
Size Dimension of the frozen solid tumor in centimeters
ER Estrogen receptor status of the tumor determined by immunohistochemistry
PR Progesterone receptor status of the tumor determined by immunohistochemistry
Her2 Herceptin receptor status of the tumor determined by FISH or Immunohistochemistry

with many events. In a previous study (Hicks et al., 2006b), we reported that ~10% of breast cancers had profiles with no discernible events. Perhaps those profiles arose from analysis of breast cancers in very early stages or from sampling only one subpopulation in the tumor. In all the other cases reported here, the subpopulations share many chromosomal events, but the total number of events is substantially greater in certain subpopulations. In T10 and T12 the subpopulations with lower numbers of events are hypodiploid, and the subpopulations with higher numbers are clearly aneuploid, strongly suggesting

that a hypodiploid state preceded the aneuploid state. These two were the only tumors displaying the "sawtooth" pattern of genomic breaks (Hicks et al., 2006b). Recent experiments have shown evidence that the basal-like expression subtype of breast cancer and *BRCA1* tumors display the sawtoothed genome profile, with extensive low-level chromosomal loss and gains (Bergamaschi et al., 2006; Chin et al., 2007). Our results suggest that the extensive chromosomal loss may represent a common early stage in the evolution of basal-like subtypes, which is then followed by increased ploidy.

In contrast, in some tumors the subpopulations differ by only a few focal events. Events common to two profiles are "early" (prior to their divergence), while events unique to the profiles are "late" (after their divergence). In Table 8.3 we list those focal changes that we classify as "late" and are therefore implicated in progression as opposed to initiation. These loci contain many well-known cancer genes, such as *KRAS*, which were first discovered on the basis of being able to initiate malignancy; however, many loci contain single genes that have not previously been implicated in cancer and are worthy of more study.

Many of the focal amplifications and deletions that we identified are regionally segregated in the tumor (Table 8.3). Regional amplifications have previously been reported in glioblastomas, where the amplification of *EGFR* was shown to occur only in specific anatomical locations (Nafe et al., 2004). Our data show that regional amplifications and deletions occur frequently in the polygenomic breast tumors. Such events have important clinical implications, because current molecular assays are performed from samples taken from a single region of a solid tumor. If for example, a clinical test for *KRAS* amplification was performed on the upper sectors of T10, it would have been negative, however if the test was performed on the lower sectors, it would have been positive. In current practice, oncologists use FISH or IHC to evalulate the levels of the *ERBB2* receptor in breast cancer patients to determine if they should receive adjuvant treatment with a monoclonal antibody, Herceptin. However, it has been shown that many Her2 negative breast cancer patients respond well to Herceptin (Paik et al., 2008), and thus oncologists will often prescribe Herceptin regardless of the outcome of the Her2 diagnostic FISH test. One possible explanation for such response is the anatomic segregation of tumor clones. Perhaps, these patients contain Her2+ tumor clones that occupy different regions of the tumor

from which the test was sampled. As we have shown in great detail in T10, multiple regions of the tumor will need to be analyzed to determine if a patient contains subpopulations that may respond to a drug.

Several, but not all, polygenomic tumors showed evidence of two tumor subpopulations co-occupying a tumor sector. SPP is insufficient to determine if the co-occupying subpopulations are intermixed at the cellular level. However, once subpopulations are identified, molecular markers can be used to examine the spatial organization of the subpopulations at the cellular level. For example, tumor T10 had three tumor subpopulations: H, A1, and A2, with the latter two intermixed. A1 and A2 were very similar, differing by a massive amplification of the *KRAS* locus. This amplification, and the amplification of nearby genes, provided us with FISH markers to distinguish A2 from A1 in tissue sections. Based on the discrete breakpoints of the amplicon in ROMA profiles of both S5 and S6, we believe that this amplification occurred in a single cell similar to the A1 subpopulation that subsequently underwent clonal expansion and finally diverged to become the A2 subpopulation present throughout these sectors. We observed a pattern of extensive intermixing of A2 and A1 in sectors 5 and 6, and very limited penetration of A2 in sector 4. We can think of three reasonable and nonexclusive explanations for intermixing subpopulations. First, the subpopulations A1 and A2 cooperate, and their mutual presence has a selective advantage. Second, A1 provides a hospitable environment into which A2 can invade, whereas normal stroma mixed with H does not. Last, A2 originated in sector 6 and has only begun invading its way back into the remainder of the tumor. The last explanation is consistent with experiments suggesting that the overexpression of *KRAS* leads to increased cell migration (Fotiadou et al., 2007).

In our study, we analyzed high grade ductal carcinomas (18/20 grade III, and 2/20 grade II) (Table 10.1). Thus we could not correlate different tumor grades with the monogenomic or polygenomic classes. However, the fact that we observe both classes in grade III tumors suggests that they do not represent exclusive stages of progression. We also tested for correlation of clinical parameters including ER, PR, and Her2 status (when available) for each tumor with the monogenomic and polygenomic classes using the Fischer's exact test, but did not find any significant correlations. Some triple negative tumors, for example, were classified as monogenomic and some as polygenomic tumors. While our samples were limited to only 20 tumors, our current data suggest that

the ER, PR, and Her2 clinical parameters show no specific correlation with either class of genomic heterogeneity. Furthermore, we scored the tumor grade in H&E-stained tissue sections from the four to six sectors of T1–T10 to see if a change in tumor grade correlated with the polygenomic tumors. We found no significant correlations; the polygenomic tumors often contained the same high grade (III) in all four to six tumor sectors. We do not have expression data for the tumors we studied, so we cannot determine if the expression subtype correlates with genomic heterogeneity, or if heterogeneity accounts for the failure of some breast cancer expression profiles to classify neatly into subtypes.

Much can be learned by discerning the subpopulations in a tumor and their spatial organization. Such analysis can be used to explore theories of cancer progression, patterns of growth (Norton and Massague, 2006), migration, and metastasis (Liu et al., 2009b) and may be of use in clinical settings. For example, clinical pathologists have long been aware of tumor heterogeneity and report the highest tumor grade observed after a fairly exhaustive survey of the tumor mass. However, as we have shown here, histological heterogeneity does not by itself imply genomic heterogeneity or vice versa. Genome-wide measures derived by sampling a single region may not be representative of the entire tumor when subpopulations are anatomically segregated. The degree of genomic heterogeneity itself might be a useful clinical parameter and could be missed entirely if not deliberately sought.

We observe a significant proportion of tumors that are apparently monogenomic, and even in the polygenomic tumors we never distinguish more than three major tumor subpopulations. However, our assessment of tumor heterogeneity with SPP is likely to be an underestimate. Minor and very heterogeneous subpopulations will be averaged into main subpopulations if they share a common ploidy. Moreover, the tumor dissection will not in general follow the natural boundaries of subpopulations, further blurring our assessments. We are limited in our method of separating subpopulations by sector and ploidy. One way to escape this limitation is to analyze the genomes of single cells. Although not without its own limitations, single-cell analysis has the potential to further clarify the extent and origins of tumor heterogeneity, and more importantly, the genetic pathways of tumor progression. This was our impetus to develop a single cell method to quantify genomic copy number - the focus of the remaining chapters.

CHAPTER 11

Introduction to Single Cell Genomics

Genomic analysis provides insights into the role of copy number variation in cancer, but current methods are not designed to resolve mixed populations of cells. This problem is particularly acute in heterogeneous tumors, which contain genetically diverse genomes. In longitudinal comparisons, single samples from heterogeneous tumors may reflect a mixture of tumor clones at various stages of progression and thus dilute the detection of high frequency chromosome mutations. In intra-tumor experiments, such as our SPP study of 20 breast tumors, the copy number profiles represent mixed populations of millions of cells (despite our efforts to stratify by region and ploidy). However, mixing problems are effectively eliminated by single cell analysis. By analyzing individual tumor genomes we can address questions such as: Are the major subpopulation composite mixtures of diverse clones or single dominant subpopulations? Do monogenomic tumors really contain highly similar genomes in every tumor cell? Are rare or intermediate cells present that could not be detected by our crude analysis of millions of cells? Using SPP, minor subpopulations would almost certainly be masked by the overwhelming signal from the major tumor subpopulations in a mixture. Furthermore, single cell copy number profiles are very useful for reconstructing detailed phylogenetic lineages to understand the pathways of tumor progression.

To study tumor progression and heterogeneity in single cells, we developed a method called Single Nucleus Sequencing (SNS). SNS combines flow-sorting, whole genome amplification (WGA) and massively parallel DNA sequencing to achieve robust single cell copy number profiles with a resolution of nearly 50 kilobases in the human genome. We applied SNS to a number of single cells in culture to validate the method, which also showed that there is only minor genomic variation in the cell cultures we analyzed. We then applied this technique to profile 100 single cells in a basal-like breast tumor to study tumor progression. At single cell resolution, our results show strong evidence that this tumor evolved by a series of sequential clonal expansions to form the tumor mass.

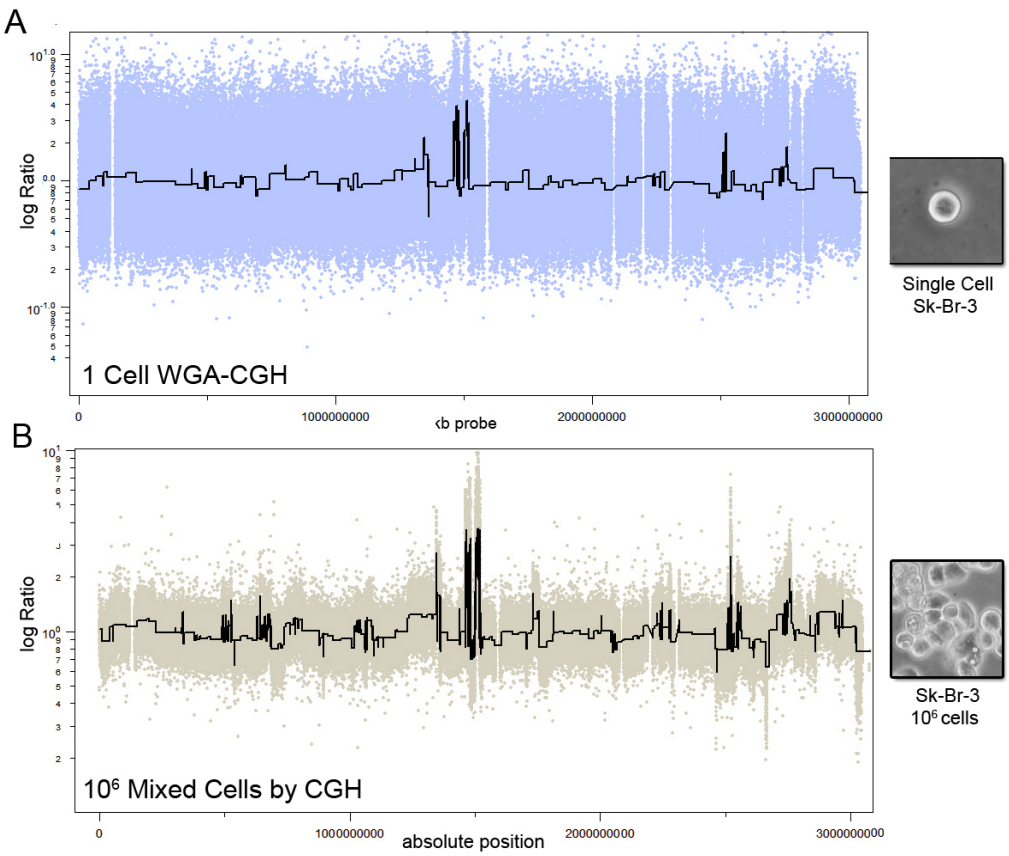


Figure 11.1 – Single Cell ROMA
ROMA CGH microarray profiles of a single SK-BR-3 cell compared to a million. (A) A single SK-BR-3 cell was isolated from culture by micromanipulations, whole genome amplified and hybridized to a 390K ROMA microarray to measure copy number as shown in blue. The data was segmented by KS statistic and is shown in black. (B) A million cells were isolated from the same SK-BR-3 culture and analyzed for copy number by ROMA and segmented by KS, as shown in black.

11.1 Background on Single Cell Methods

Recent advances in whole genome amplification (WGA) methods now allow DNA from a single cell to be amplified to microgram quantities (Sigma GenomePlex©, Rubicon PicoPlex© Kits). However, the amplified DNA is not a perfect copy of the genome, but rather a representative library of random fragments covering less than 10% of the human genome. Efforts to quantify whole genome copy number from WGA DNA by CGH have shown that it is possible, albeit at low resolution (Le Caignec et al., 2006). However, major issues exist with the overall signal:noise ratio, standard deviation and dynamic

range which permits only large (>10 megabases) chromosome aberrations to be detected in single tumor cells (Fuhrmann et al., 2008; Imle et al., 2009; Klein et al., 1999). One study did achieve a higher resolution (>3mb) by applying tiling oligonucleotide microarrays to single cell WGA fragments (Geigl et al., 2009). However, such resolution is not a big improvement over traditional cytological techniques to analyze single cells, such as G-banding, that have been available since the 1980's. Initially we attempted a similar approach by combing WGA with ROMA to measure genome-wide copy number in single cells that were isolated by micromanipulation. Similarly, these copy number profiles had a high standard deviation and low signal:noise ratio allowing only large chromosome aberrations to be resolved (Figure 11.1). Such data are not very useful for studying tumor heterogeneity.

11.2 Single Cell Microarray Analysis vs. Sequencing

In principle, CGH methods are problematic for measuring copy number from single cell WGA samples, since microarray probes target predefined sequences and only a fraction of the genome is amplified (< 10%). Thus, the probability of a microarray probes hybridizing to randomly amplified WGA fragment from a single cell is very low (P_m 0.01 x P_s 0.1 = 0.001, approximately 0.1%). To explore this idea we collaborated with Dr. Richard McCombie and used next-generation sequencing to investigate how the WGA fragments from single cells are distributed in the human genome. In a preliminary experiment, we sequenced a single fibroblast cell on a flow-cell lane, which resulted in ~4 million sequence reads that were mapped uniquely to the human genome. We compared the position of the sequence reads to the coordinates of the microarray probes, which confirmed that many probes had entirely missed hybridizing to single cell WGA fragments (Figure 11.2). Thus, we concluded that targeted approaches such as microarrays are inadequate for measuring sparse, random sequences.

This experiment led to a new idea: measuring copy number directly from sequence read depth. Using this method we would not 'miss' the randomly amplified WGA fragments from single cells, when counted at a sufficiently large genomic intervals. Recent studies have shown that read depth from next-generation sequencing can be used to accurately measure genomic copy number in DNA from millions of cells (Alkan et al., 2009; Chiang et al., 2009). We estimated that ~4 million sequence reads would be sufficient to measure copy

number in 50kb intervals, allowing us to measure at least 50 reads in each interval. Thus, our approach would involve sequencing WGA amplified single cells to quantify genome-wide copy number.

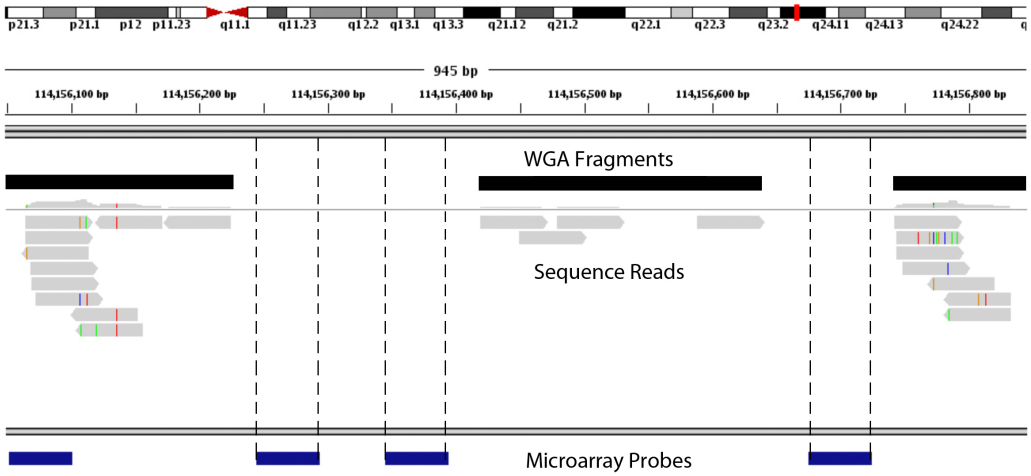


Figure 11.2 – Microarray Probes vs Sequence Reads
A single SK-BR-3 cell was WGA amplified and sequenced. The resulting sequence reads (grey) are compared to the location of ROMA microarray probes (blue) within a 945bp region. The estimated location of WGA fragments are shown (black) as estimated from sequence read density. Dotted lines show microarray probes that failed to hybridize to their respective targets

CHAPTER 12

Single Nucleus Sequencing (SNS)

12.1 SNS Method

We combined FACS, WGA and next-generation sequencing in a method we call Single Nucleus Sequencing (SNS). FACS allows us to efficiently isolate single nuclei from populations of cells, from which we amplify random fragments of the genome by WGA for next-generation sequencing, to estimate genome-wide copy number. To perform SNS, nuclei are isolated from cells in culture, or from frozen tumor sections, using a DAPI-NP40 buffer and filtered through 37-μm plastic mesh as described in chapter 5 (Figure 12.1A-B). The nuclei are sorted by FACS using the BD Biosystems Aria IIu flow cytometer by gating cellular distributions with differences in their total genomic DNA content (ploidy) according to DAPI intensity. Initially, we used micromanipulation to isolate single cells, but found that this method often led to reactions with multiple or no cells. We found a more efficient approach to be FACS, which is often used in cell culture studies to 'subclone' single cells and establish new clonal cultures.

During FACS, we first determine 2N copy number by sorting a small amount of prepared nuclei derived from a control lymphoblastoid cell line of a normal person to establish FACS collection gates. Before sorting single nuclei, a few thousand cells were sorted to determine the DNA content distributions for gating (Figure 12.1C). A 96-well plate was then prepared with 10ul of lysis solution in each well from the Sigma-Aldrich GenomePlex© WGA4 kit. Single nuclei were deposited into individual wells in a 96-well plate (Figure 12.1D) along with several negative controls in which no nuclei were deposited. To initially estimate the error rate of the Aria IIu in sorting more than 1 nuclei, we sorted single DAPI-stained nuclei into flat-bottom 96-well plates and examined the wells by fluorescent microscopy. We found that the Aria IIu had a very low error rate, sorting a single nucleus in 94/96 wells.

Whole genome amplification was performed on single flow-sorted nuclei as described in the Sigma-Aldrich GenomePlex WGA4 kit kit (cat #

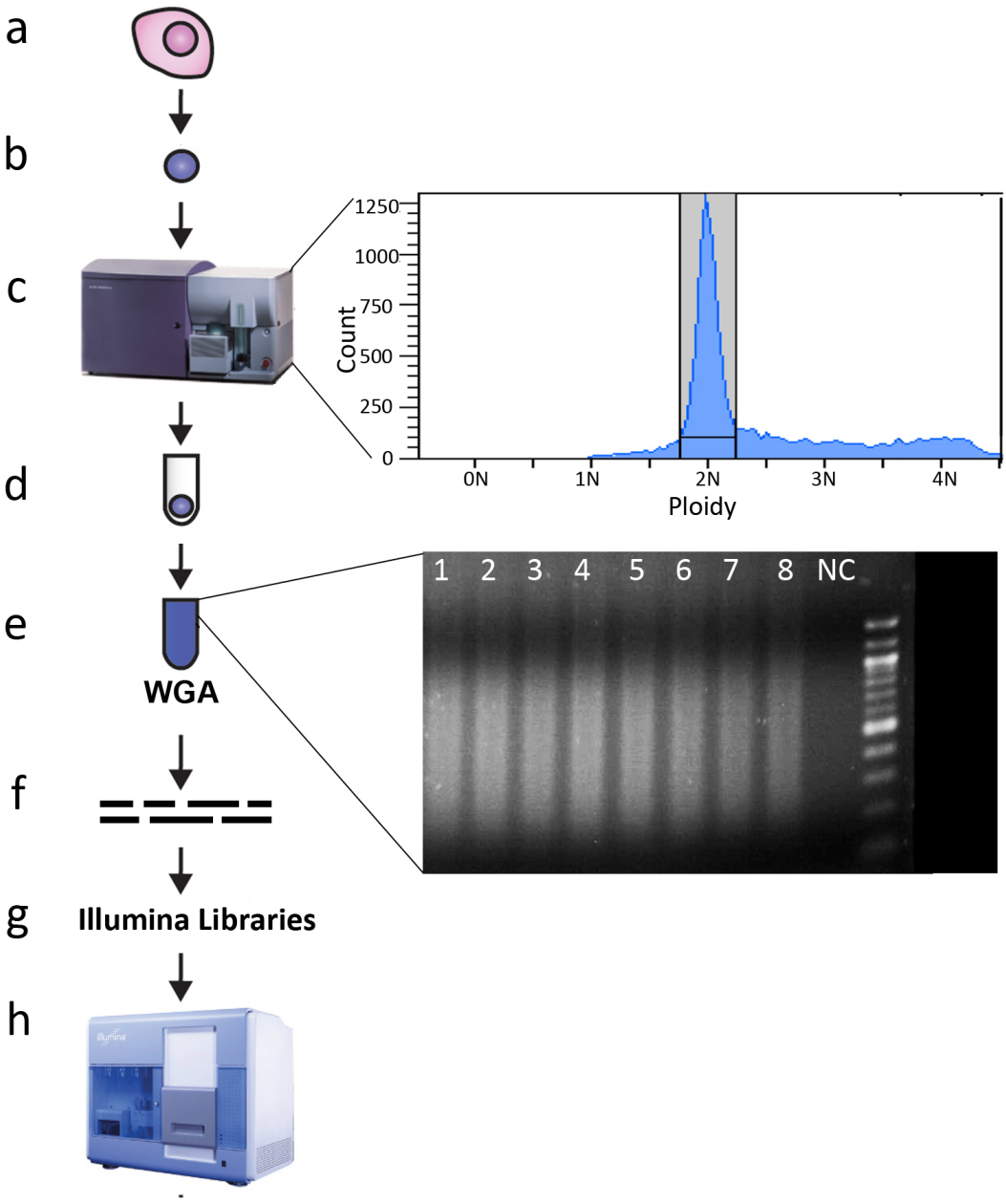


Figure 12.1 – Single Nucleus Sequencing (SNS) Method
(A) A suspension of nuclei are isolated from a population of cells and filtered (B) Nuclei are stained with DAPI (C) A number of nuclei are flow-sorted to generate a FACS histogram of the ploidy distributions, and a population is gated (D) A single nucleus is deposited directly into well containing lysis solution (E) The DNA is fragmented and amplified by WGA. The DNA is separated by electrophoresis showing a distribution of fragments from 100 to 1000bp, and an empty negative control (F) DNA fragments are sonicated (optional) (G) Single-read Illumina libraries are constructed (H) DNA is sequenced on a single lane of a flow cell by an Illumina Analyzer.

WGA4-50RXN) protocol (Figure 12.1E). Initially, before commercial kits for WGA were available, we developed our own WGA method involving random hexamer primers and the Φ29 polymerase to amplify DNA from single cells by multiple strand displacement (MDS). However, our method yielded only nanogram quantities of DNA from the initial 6 picograms in a single cell. As commercial kits became available, we explored their use for amplifying DNA in single cells. These kits were not intended for single cells and were generally designed to amplify DNA from small numbers of cells (>100). They included the GenomePlex WGA kit (Sigma-Genosys), REPLI-G kit (Qiagen) and PicoPlex kit (Rubicon Genomics), which we evaluated for amplifying DNA from single cells. Eventually, Rubicon Genomics developed the first single cell WGA kit which was purchased by Sigma-Genosys called the GenomePlex Kit (WGA4), which clearly worked better than the others, amplifying DNA to microgram quantities, randomly and uniformly across the genome, while other kits showed strong biases and overrepresentation. Moreover, this kit had the great advantage of not amplifying DNA in the negative control reactions, when no template DNA was added, while other kits would always amplify DNA through self-priming.

The molecular details of the Sigma-Genosys WGA kit are described in the patent (U.S. Patent #7718403). In summary, a single cell is added to a well and the 6 picograms of DNA is heat fragmented in an alkali solution (Figure 12.2A). Special adapters are added to the solution, containing both a specific primer sequence and a stretch of random nucleotides (Figure 12.2B). The random portion of the adapters anneal to the fragmented genomic DNA and Φ29 polymerase extends these regions by MDS (Figure 12.2C). After Φ29 polymerase extends the nascent strands, a specific adapter sequence is added to the 5' end of the molecule. By chance, a second priming event occurs when another random-adapter primer anneals within the new molecule and Φ29 polymerizes by MDS (Figure 12.2D-E). The final molecules from these reactions have specific adapter sequences at both ends, allowing them to be amplified by standard PCR reaction protocol, using specific primers and a DNA polymerase (Figure 12.2F). This PCR reaction generates microgram quantities of DNA fragments with a distribution of 100-1000bp.

The resulting WGA fragments can be used directly for single-read library construction using the Illumina Genomic DNA Sample Prep Kit (cat # FC-102-1001), following standard protocol with a gel purification size range of 250-300bp

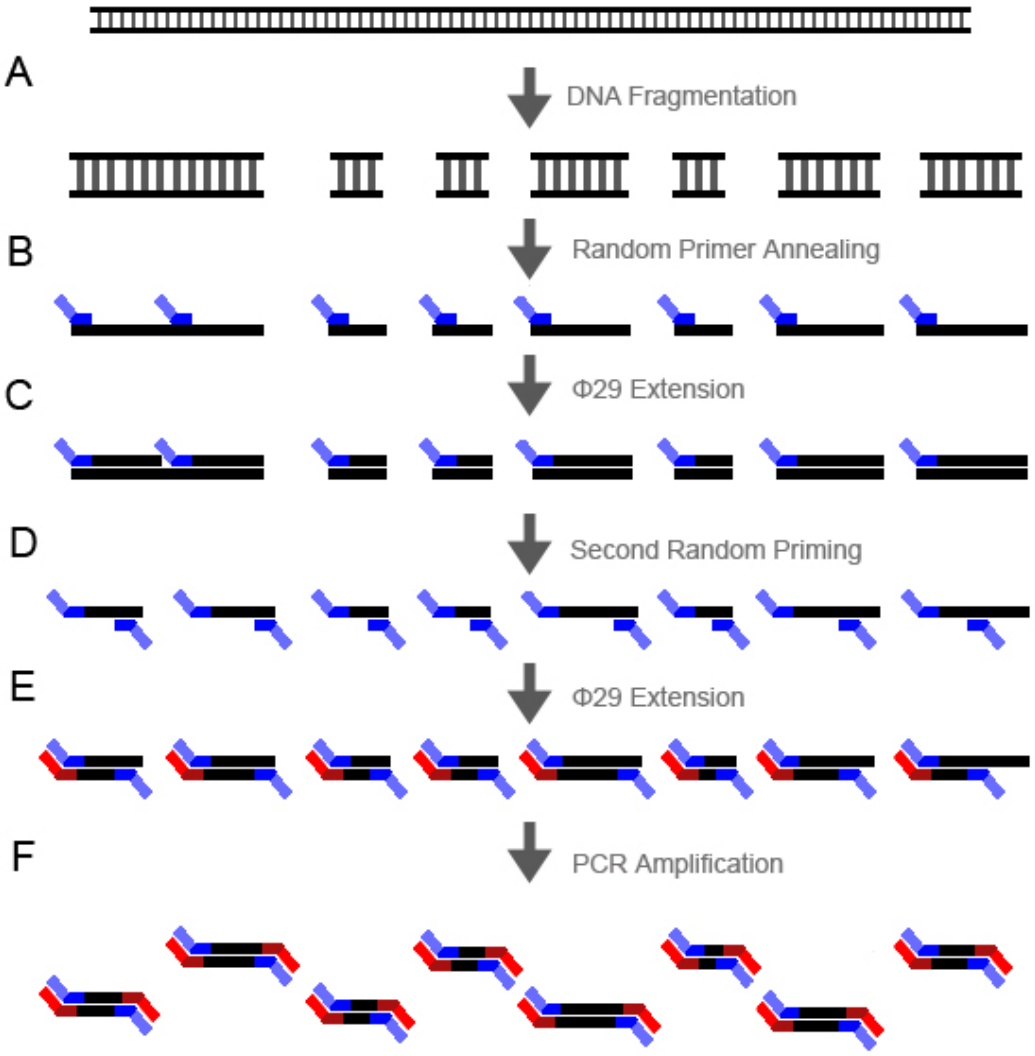


Figure 12.2 – Molecular Mechanism of WGA
Molecular approach to amplifying genomic DNA as described in the Sigma-Genosys patent (A) A single genome is heat fragmented in alkali solution (B) Specific adapters are added with random oligonucleotides that anneal to the genomic fragments (C) The Φ29 polymerase locates the primed DNA and extends by multiple strand displacement (MSD) (D) By chance another specific-random adapter sequence anneals to the nascent strand and initiates a second priming reaction by the Φ29 polymerase (E) After extension, the resulting library of molecules contains specific adapter sequences at both ends (F) The library is then amplified by standard PCR using specific primers sequences.

(Figure 12.1G). Alternatively, the WGA fragments can be sonicated to remove the 28bp adapter sequence (which was added on during the WGA reaction) using the Diagenode Bioruptor© with the following program: 2 times, 7 minutes with 30 seconds high on/off mode in ice cold water (Figure 12.1F). Sonication will greatly improve the sequencing cluster amplification reaction, and the total number of sequencing reads per lane.

Single-read libraries from single nuclei are then sequenced on individual flow-cell lanes using the Illumina GA2 analyzer for 76 cycles (Figure12.1H). Data was processed using the Illumina GAPipeline-1.3.2 to 1.6.0. Sequence reads were aligned to the human genome (HG18/NCBI36) using the Bowtie alignment software (Langmead et al., 2009), with the following parameters: 'bowtie –S –t –m 1 –best –strata –p16' to report only top scoring unique mappings for each sequence read. To eliminate PCR duplicates, we remove sequences with identical start and stop coordinates.

On average, running SNS on a singe cell generated 12.3 million sequence

Statistic	Single Cells	Million Cells
Filtered Reads	12,321,629	25,265,342
Mapped Reads	6,884,789	20,013,676
% Mapped	55.87%	79.21%
% Genome Coverage	4.39%	17.70%
Reads/50kb Bin	86.72	368.22

Table 12.3 – Sequence Run Statistics
Sequencing statistics for single cell compared to million cell samples run on single Illumina flowcell lanes. Values represent means calculated from many single cells (N=100) or million cell runs (N=10). Samples were run at 76 cycles on an Illumina GA2 analyzer and mapped with Bowtie. Reads/50kb bin represent the average number of reads within variable bin intervals of approximately 50kb in the human genome.

reads, of which 55.87% mapped uniquely to the human genome (Table 12.3). In comparison, bulk DNA from millions of cells generated about twice as many sequence reads (25.2 million) of which a larger proportion (79.21%) mapped uniquely to the human genome. This difference may be explained the large numbers of adapter sequences from the WGA reactions that are sequenced, but cannot be mapped back to the human genome. It should be noted that due to

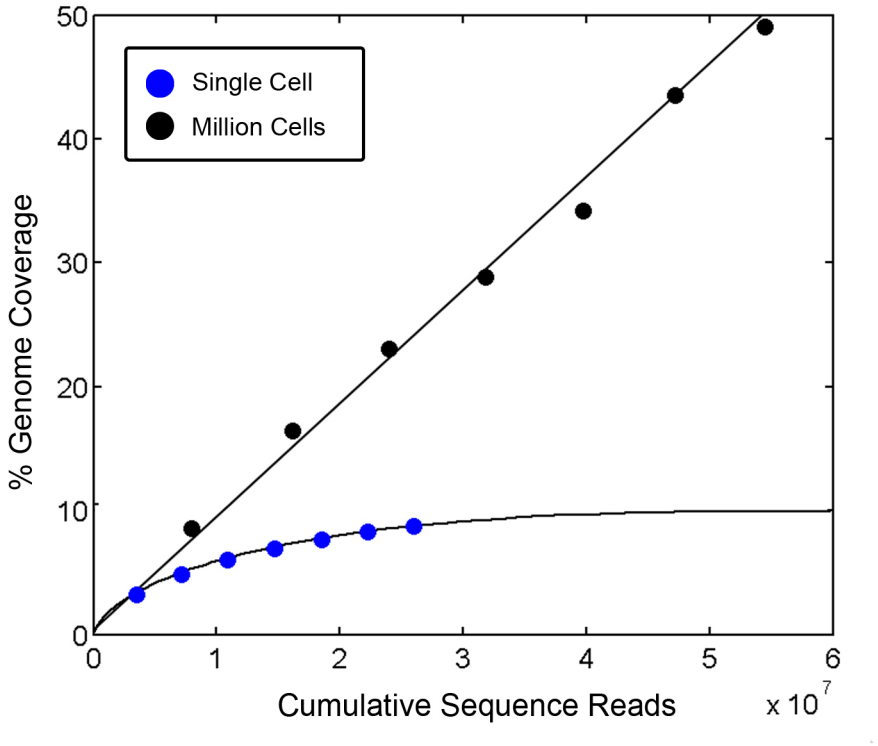


Figure 12.4 – Limitations to Coverage in Single Cells
A single SK-BR-3 cell and a million cell sample were 'deep' sequenced on seven Illumina flowcell lanes. This graph shows the increase in genomic coverage as more sequence reads are added. The million cell sample increases linearly, while the single cell sample increases less as more reads are added, never exceeding 10%

technological advances in Illumina sequencing the throughput quadrupled from 2009-2010, and thus the mean values reflect a large range that corresponds temporally to when the sequence run was performed.

12.2 Limitations

We calculated the total number of bases that uniquely covered the human genome and found the mean to be 4.39% in single cells, compared to 17.70% coverage in million cell samples. This was a significant difference, and thus we wanted to determine if the relationship between coverage and sequence read numbers was linear in single cells. In other words, could we increase coverage by sequencing single cell WGA libraries more. To investigate this question, we 'deep-sequenced' a million cell library prepared from a fibroblast culture on seven lanes, and a single cell WGA library prepared from a single fibroblast cell

on seven lanes. We calculated the cumulative coverage as each lane was added and plotted the number of sequence reads against total coverage (Figure 12.4). In the million cell sample, coverage increased linearly as more reads were added, starting at 8% and increasing to 48%. In contrast, the single cell WGA library started at 5% coverage, but did not exceeded 10%, each additional lane added no more than 1% coverage, following a unipolar convex curve. In single cells, the relationship between the number of sequence reads and coverage followed the 'law of diminishing return' in which less and less unique coverage is gained by additional sequencing.

This data suggests that the initial WGA reaction amplifies less than 10% of the genome in a single cell, and imposes a theoretical limitation on the resolution of the SNS method, at approximately 10 kilobases. For our purposes, a single flowcell sequencing lane generates around 4 million reads, allowing copy number to be detected at 50 kilobases with a mean of 86.72 reads per bin (Table 12.2). At this resolution the bins follow a normal Gaussian distribution of read counts. In contrast, million cell samples have no theoretical limit to which copy number can be detected, but to simplify our comparative analysis of genome profiles, we use the same resolution.

In summary we have shown that SNS can isolate single nuclei and randomly amplify genomic DNA to sufficient quantities for massively parallel sequence, allowing us to quantify genome-wide copy number at an approximate resolution of 50kb. Higher resolutions in single cells will require developing a better WGA technique, capable of amplifying more than 10% of the genome in the initial Φ29 strand-displacement reaction. We use this method to measure read counts in intervals across the human genome for estimating genomic copy number.

CHAPTER 13

Absolute Copy Number Quantification

Previous studies using WGA have shown that a common problem is the oversampling of regions of the genome (Pugh et al., 2008; Talseth-Palmer et al., 2008; Huang et al., 2009). However, as long as the oversampled regions are distributed largely at random across the genome, then at a sufficient genomic scale and read depth the sequence read density of a WGA product should be proportional to gene copy number. We demonstrate this by generating copy number profiles of WGA DNA from single cells, and comparing these to profiles of bulk DNA directly prepared from $>10^6$ cells. Our results show that absolute copy number can be detected in single cells at high resolution (60 kilobases), and that these profiles are highly similar to copy number profiles measured from millions of cells.

13.1 WGA Stacking

A problem with some WGA methods is oversampled regions of the genome, or 'stacking', in which some WGA fragments are grossly over-amplified during WGA (Pugh et al., 2008; Talseth-Palmer et al., 2008; Huang et al., 2009). To investigate this phenomenon in our single cell sequencing experiments, we calculated pileup plots. To construct these plots we first remove PCR duplicates with identical start and stop positions and calculate a vector of zeroes for each nucleotide in the human genome. To this vector we add 1 to every position where a nucleotide from a sequence read maps, thus at a read length of 76bp the vector cannot exceed a maximum value of 76. We show the data for seven single cells, with stacking regions (> 20) marked by an asterisk across a one megabase region on chromosome 5 (Figure 13.1).

The pileup plots show that stacking regions are not biased to specific regions in the human genome – the stacking regions do not overlap between different single cells. When stacking occurs, the overrepresented regions are usually contained to regions of less than 1000bp. This size correlates to the size range of PCR amplification, and thus we assume that each stack represents a

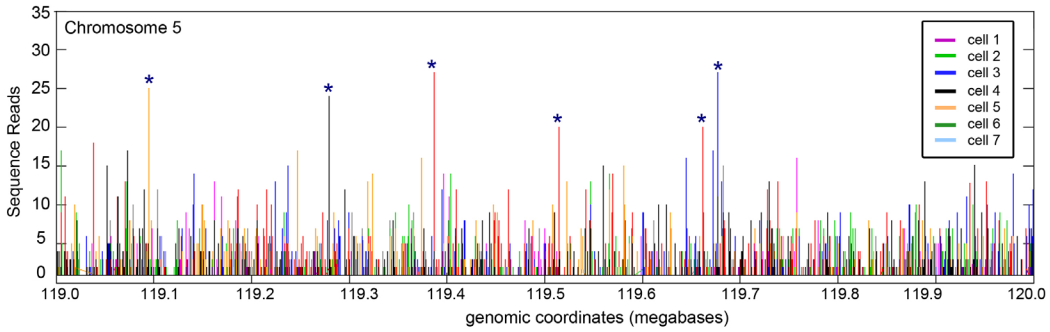


Figure 13.1 – WGA Stacking
Pileup plots showing the distribution and stacking of sequence reads in seven single fibroblast cells. Thie region shown is a one megabase region on chromosome five. Asterisks denote regions of stacking, in which more than 20 reads cover a single WGA fragment.

single WGA fragment. Moreover, the frequency of each stack is generally less than 1 stack per megabase, and thus will not greatly affect estimates of copy number estimation at the resolution we have choosen (50kb). Additionally, these plots show that most sequence reads are randomly distributed across the genome with frequencies of no more than one.

13.2 Sequence Read Counting in Variable Bins

To determine copy number from sequence data other studies have calculated read density in intervals with fixed length (Alkan et al., 2009; Chiang et al., 2009; Yoon et al., 2009). However, we use an alternative method, using variable intervals, counting only reads with unique mapping to the genome. Since the density of unique mapping sites is not uniform in the genome, we use genomic bins of variable length but with uniform expected read density. To create these bins, we randomly sampled 200 million sequences *in silico* of length 48bp from the UCSC reference genome, introduced single nucleotide errors with the frequency encountered during Illumina sequencing, and mapped the reads back to the reference genome using Bowtie (Langmead et al., 2009). We established boundaries for 50,009 genomic bins such that the expected number of mapped reads in each bin was equal.

Variable bins have the advantage of avoiding repetitive elements in the human genome, including LINES, SINE, LTRs, microsatellites and simple repeat, as well as centromeric and telomeric regions. In repetitive regions, the

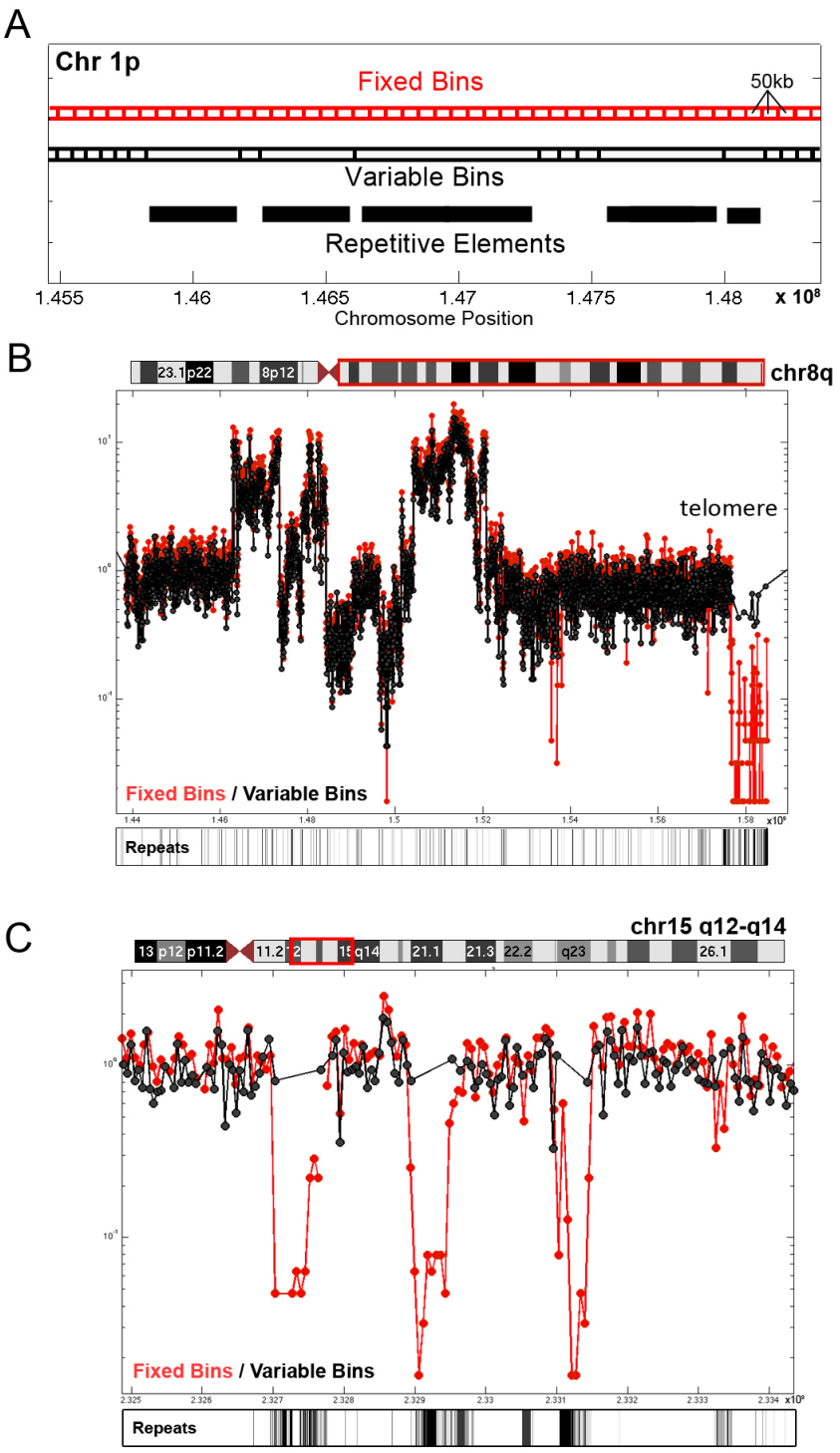


Figure 13.2 – Variable Binning
Fixed and variable bin counts were calculated from a single SK-BR-3 cell (A) Fixed bin intervals are compared to variable bins across a region with repetitive elements on chromosome 1p (B) Fixed and variable bins counts are shown on chromosome 8q with repetitive elements annotations below. (C) Fixed and variable bin counts are shown on chromosome 15q12-q14 containing three highly repetitive regions shown below.

size of each variable bin is adjusted to compensate for regions with low or high mappable read counts, thereby maintaining a consistent mean value for each bin (Figure 13.2A). To illustrate this difference, we calculated both fixed and variable bins from the sequence read counts of a single SK-BR-3 cell, and show a closer view of two regions on chr8q and chr15q12-q14. On the telomeric region of chr8q there is a higher density of repetitive elements, which appear as a large chromosomal deletion in the fixed bin profile (Figure 13.2B). In contrast, the variable bin profile shows the expected ground state copy number. On chr15q12-q14 we also show an intrachromosomal region with three highly repetitive areas. In the fixed interval profiles they appear as homozygous deletions, whereas the variable bins show the expected ground state copy number (Figure 13.2C). In summary, variable bins have a great advantage of not reporting erroneous chromosome deletions that are commonly calculated by fixed interval algorithms.

13.3 Absolute Copy Number Quantification

Genetic theory predicts that single cells will have integer values for chromosome states, suggesting that absolute copy number can be measured in single cells. To do this, we first eliminated reads with identical start coordinates to avoid PCR duplicates. We then counted sequence reads using variable bins resulting in a linear array of bin counts. The bin counts were segmented, each segment having a distribution of bin counts significantly different from its adjacent segments, as judged by a Kolmogorov-Smirnov statistic (Grubor et al., 2009). The result for a single nucleus from SK-BR-3 cells is shown in Figure 13.3A. To obtain a better sense of the detail in the data, we show a region of chromosome 8q near the *MYC* locus in Figure 13.3B. with the segmentation indicated by the red lines, and the variable bin counts in blue dots.

Many bin count distributions are recurrent in non-adjacent segments, and the median counts of non-adjacent segments are separated by steps. These steps are roughly uniform on a linear scale and likely correspond to integer differences in copy number. To present the evidence for integer differences more clearly, we display a Gaussian kernel smoothed density of the absolute values of the difference between median bin counts for all pair-wise combinations of bins from different segments (Figure 13.3C-E). The uniform steps between groups are very

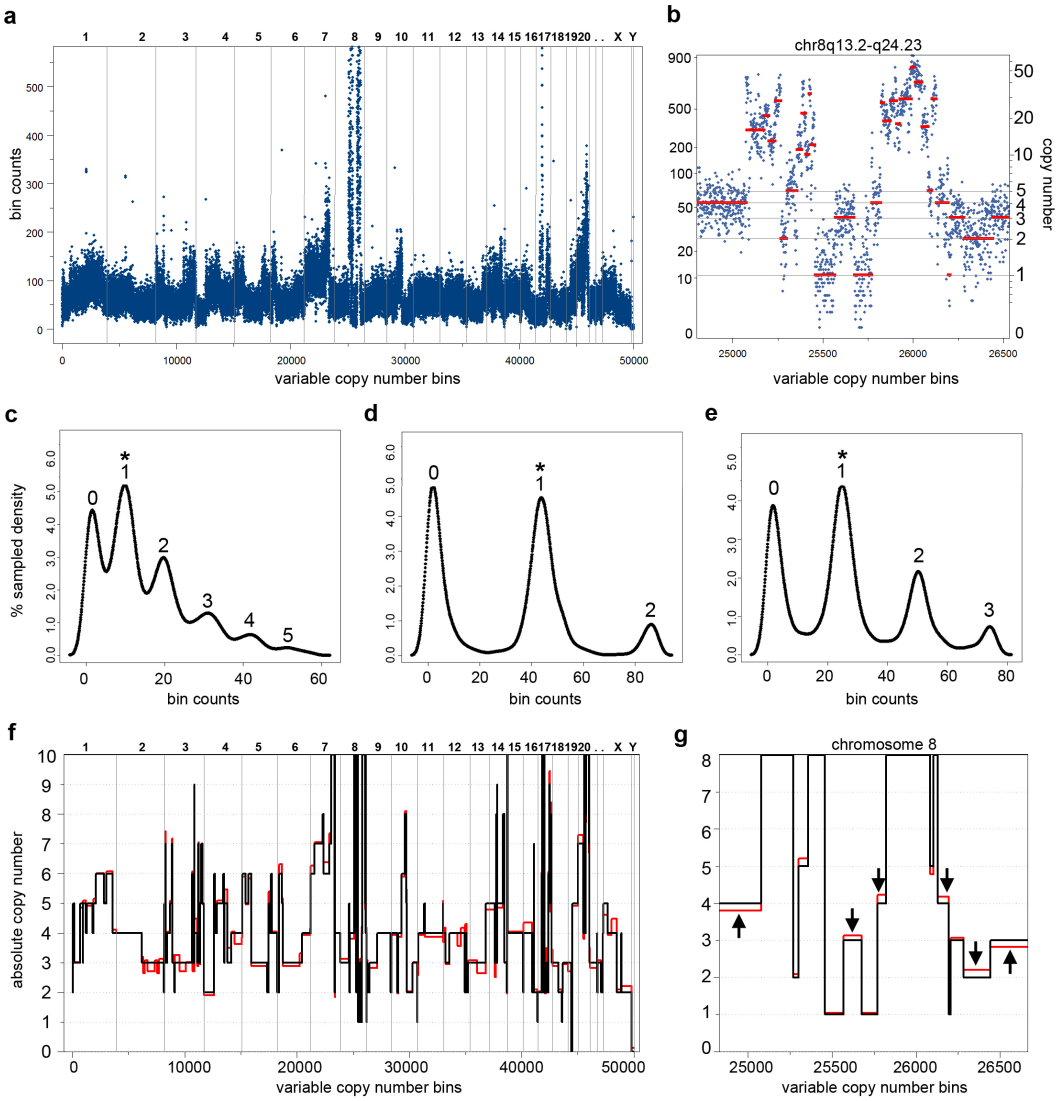


Figure 13.3 - Absolute Copy Number Quantification from Read Density
(A-C, F-G) Absolute copy number calculations are shown for a single SK-BR-3 cell. (A) Mapped sequence reads are counted in variable bins of uniform expected read density and plotted in genomic order (B) Variable bin counts in blue are plotted on a log scale for an amplified region of chromosome 8 and KS-segments are plotted in red. Thin horizontal lines indicate integer copy number estimates. (C-E) Gaussian kernel smoothed density plots with stars denoting the first increment peak for (C) SK-BR-3, (D) a hypodiploid tumor cell, and (E) an AA tumor cell. (F) A KS-segmented profile in black is compared to the absolute copy number profile in red, and (G) this region is shown for chromosome 8 with absolute copy number on the ordinate.

apparent. This is a general property of the data from single nuclei from cultures or from tumors, provided enough steps are present, and shows that the data is highly quantile.

We convert our KS-segmented data into profiles of absolute copy number as follows. We take the differential bin count of the second peak, denoted by an asterisk in Figure 13.3C-E, to represent a copy number "increment" of 1. We then divide every bin count in the profile by the increment and round to infer the absolute copy number. By plotting the original KS segmented profile against the transformed absolute copy number profile, we see that they are in close agreement, but differ in that decimal values have been converted to integers (Figure 13.3F-G). However, for diploid or near diploid cells there are generally few steps from which to observe the increment, and we use a different method, taking the increment as the median bin count on the autosomes divided by two.

CHAPTER 14

Genomic Variation in Cell Culture and Validation of SNS

We validated the SNS method by comparing the absolute copy number profiles from a single cell to a million. We selected a breast cancer cell line, SK-BR-3, which contains a complex aneuploid profile with many genomic amplifications and deletions of cancer genes, which we expected to detected in both profiles. We also selected a normal fibroblast cell line, SKN1, which contains no chromosome aberrations outside of normal copy number variants (CNVs). The diploid genome of the fibroblasts would show us if any random or biased amplifications of the genome were introduced by the WGA method. We also analyzed and compared seven single cells from both cultures to test the hypothesis that cell cultures are genetically clonal.

14.1 Single versus Million Cell Profiles

We applied SNS to a single SK-BR-3 cell (Figure 14.1A) and compared it to an absolute copy number profile measured from a million cells (Figure 14.1B). Overall, the profile of a single cell closely resembles the profile from a million cells (Figure 14.1C-D). The SK-BR-3 genome contains many major amplifications of oncogenes (*RD2, TPD52, NBS1, EXT1, HAS2, MYC, ERBB2, BCAS1*) and a homozygous deletion of a tumor suppressor (*DCC*), all of which could be detected with similar breakpoints in the profile of a single SK-BR-3 cell. To obtain a better sense of the data, we show the copy number profiles and actual bin counts from a single nucleus (Figure 14.1E) and from $>10^6$ SK-BR-3 cells (Figure 14.1F) for a complex region on chromosome 8q13.2-q24.23 containing several oncogenes including *MYC*. The fine-scale amplification pattern in the two samples is highly similar, showing that WGA does not introduce a consistent bias when using bins on the scale we have chosen. The main difference lies in the bin count data of single cell, showing a higher standard deviation, but this does not greatly affect the absolute copy number profile since the data is segmented.

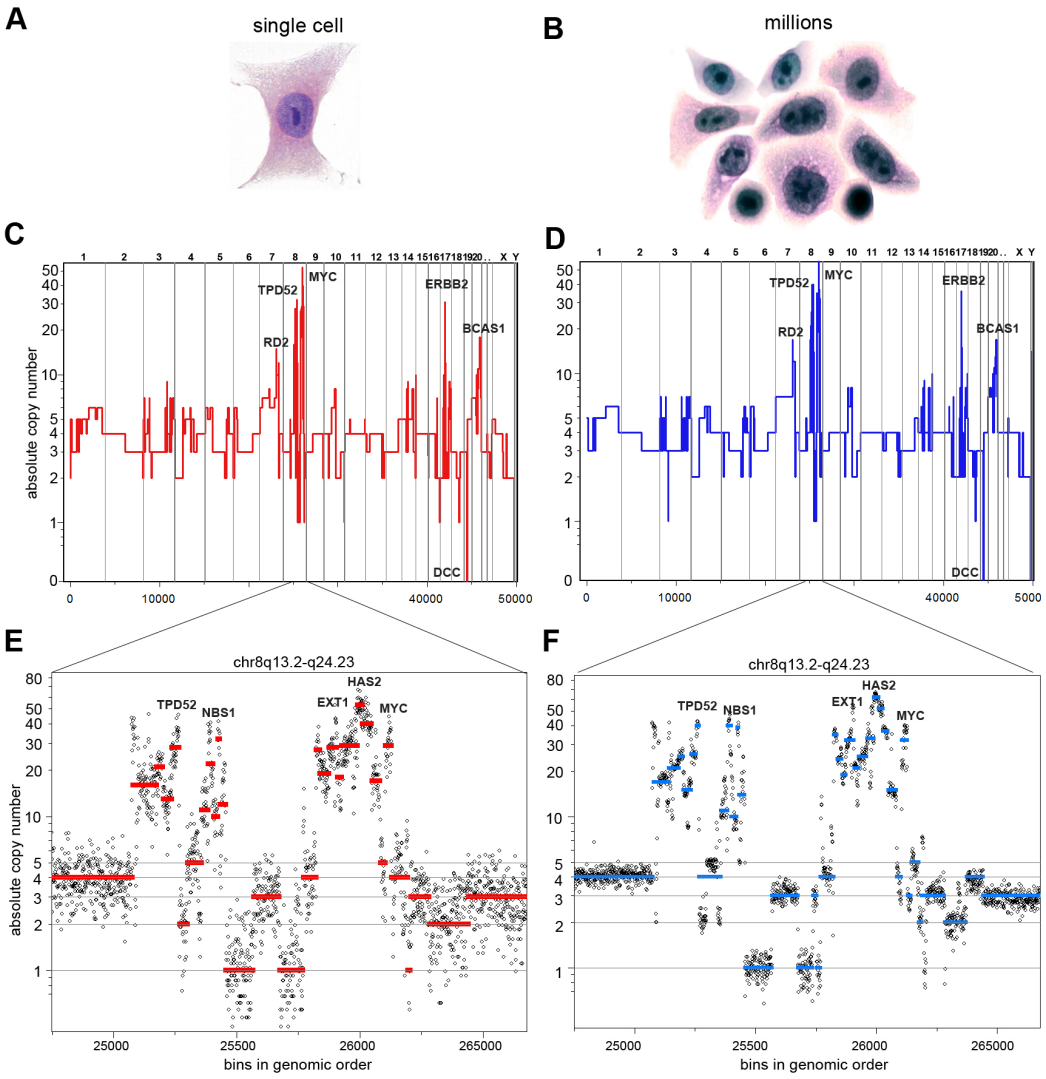


Figure 14.1 – Absolute Copy Number Profile of a Single Cell Compared to Millions of Cells
(A) Single SK-BR-3 cell stained with H&E (B) Millions of SK-BR-3 cells are stained with H&E (C) The absolute copy number profile for a single SK-BR-3 cell is shown compared to (B) millions of cells (E-F) A region on chromosome 8q13.2-q24.23 is plotted showing the absolute copy number profile in red and a ratio of raw bin counts in black for (E) a single cell , and (F) a sample of a million cells

14.2 Genetic Variation in Cell Cultures

After validating the SNS method, we decided to investigate the relative stability of genomes in cell culture. We analyzed and compared seven single SK-BR-3 breast cancer cells and seven single human fibroblast cells by SNS. To compare the genomes, we calculated absolute copy number profiles and used one-dimensional hierarchical clustering with a Euclidean distance metric to group the profiles. We display the profiles for the fibroblast and SK-BR-3 culture using a heatmap with genomic order on the y-axis, showing amplifications in green, deletions in red and ground state copy number in black (Figure 14.2).

Only minor genomic variation is seen among the individual cells in these cultures. In the fibroblast culture, the seven individual cells (F1-F7) are very similar to each other and to the million cell sample (FM), showing mainly diploid copy number (Figure 14.2A). Two major CNVs appear as amplifications in all single cells and in the million cell sample. There are also a few regions that are consistently deleted near telomeres and centromeres, which are likely to be artifacts from the inability of the alignment software to map sequence reads in these regions. Similarly, the SK-BR-3 culture shows only minor genetic variation between the seven individual cells (S1-S7) and in comparison to the million cell sample (SM). All of the major amplifications and deletions in this aneuploid genome are detected in all of the seven single cells (S1-S7).

We conclude from the fibroblast culture that little if any random events or biases are introduced during WGA. From the SK-BR-3 culture we see that major amplifications and deletions are detected in every single cell, validating our method. Moreover, these experiments answer an interesting question regarding genomic stability in cell culture, namely that individual cells have very clonal genomes. However, our measurements are based on copy number variation, so we cannot exclude the possibility that single cells in culture show significant variation in point mutations or epigenetic patterns. Future experiments will also need to determine if genomic stability is a common property of all cell cultures, particularly in cancer cell lines that have multiple ploidy distributions (ALAB, BT-483, BT-549, UACC-893), which we know from our tumor studies to be indicative of genetically divergent subpopulations.

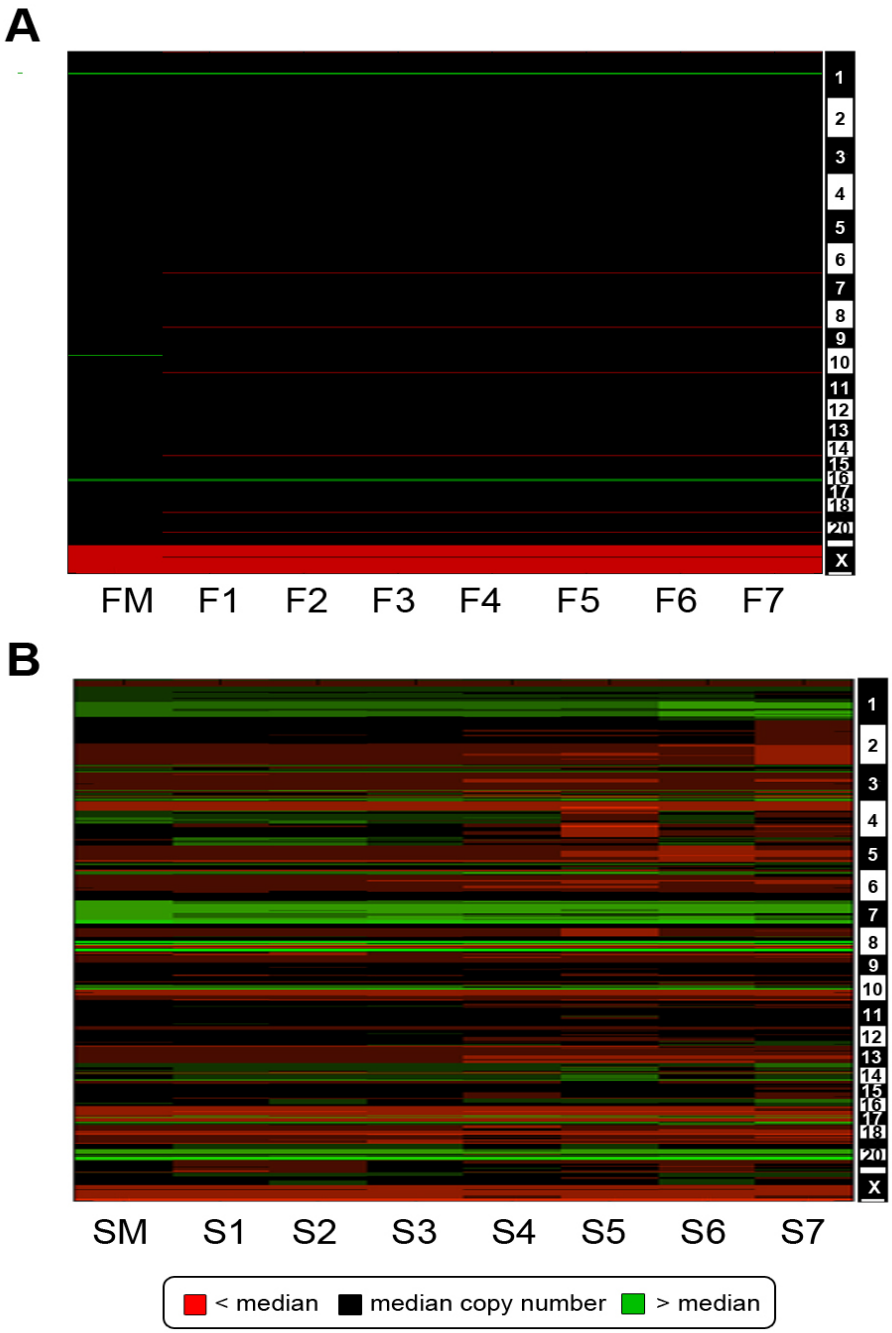


Figure 14.2 – Heatmaps of Single Cells in Cultures
(A) Heatmap showing the absolute copy number profiles of seven single SK-BR-3 cells (S1-S7) compared to a million cell profile (SM) (B) A heatmap showing the absolute copy number profiles of seven single fibroblast cells (F1-F7) compared to a million cell sample (FM). Profiles are plotted in genomic order on the y-axis with amplifications shown in green, deletions in red and ground state copy number in black.

CHAPTER 15

Analysis of 100 Single Cells from a Heterogeneous Breast Tumor

We sought to study the genomes of 100 single cells from a heterogeneous breast tumor to understand its genetic substructure and infer tumor progression. To do this, we selected a highly aneuploid breast tumor (T10), which was previously shown by SPP to be genetically heterogeneous (Navin et al., 2010). T10 is a basal-like ductal carcinoma, a particularly aggressive subtype of breast cancer that is associated with poor survival. By histopathology, T10 was shown to be poorly differentiated, high grade (III) and have triple negative receptor status (ER-, PR- and Her2-). Our theory is that much can be learned by studying numerous cells from a single polygenomic tumor, instead of a conducting a longitudinal analysis of many tumors. This is would increase our chances of detecting rare or intermediate cells that may play an important role in tumor progression.

15.1 Isolation of 100 Single Cells by FACS

In order to preserve anatomical information on cell location, we macro-dissected T10 into twelve sectors, and isolated 100 nuclei from six sectors (S1-S6) as shown in Figure 15.1. Four major subpopulations were resolved as peaks by FACS: a hypodiploid fraction (F1, 1.7N), a diploid or Ψdiploid fraction (F2, 2N), and two sub-tetraploid fractions (F3, 3.1N and F4, 3.3N). In the upper three sectors only diploid and hypodiploid fractions were observed, while the lower three sectors contained the two subtetraploid fractions (F3 and F4) in addition to the diploid cells. We deposited 100 single cells from various fractions and sectors (Figure 15.2, lower panel) into individual wells on a 96-well plate and used SNS to quantify absolute copy number. As quality control, we analyzed only nuclei that had greater than one million sequence reads. To insure that our sequence data derived from single cells, and not multiple nuclei that were incorrectly sorted, we kept statistics on total number of reads (depth) and the proportion of the genome covered (breadth). We discarded seven outliers with sequence profiles and read statistics that clearly indicated mixtures due to higher than expected genome coverage, leaving us with 93 cells for analysis.

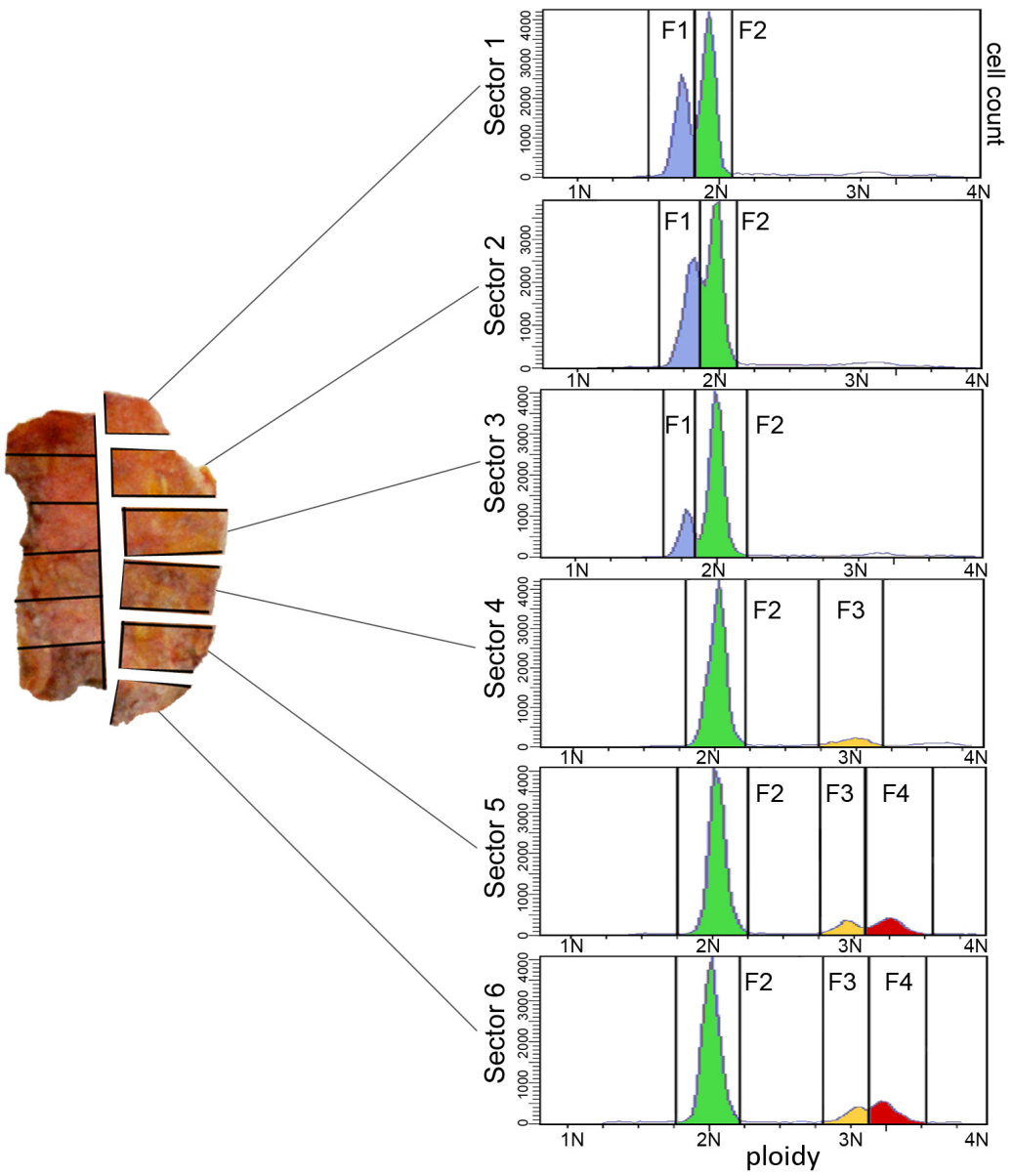


Figure 15.1 – Isolation of 100 Single Tumor Cells by FACS
A ductal carcinoma was macro-dissected into 12 sectors, and nuclei were isolated from six sectors. The nuclei were stained with DAPI and flow-sorted by FACS to generate histograms of ploidy. The FACS profiles from the six sectors shows four distributions of ploidy (F1-F4), which were gated to isolate 93 single cells from different distributions and sectors.

15.2 Cluster Analysis of 93 Single Cell Profiles

To understand the genetic relationship between the single cells, we used hierarchical clustering with a Euclidean metric applied to the absolute copy number profiles. The clustering results are displayed as a heatmap with genomic position preserved on the Y-axis (Figure 15.2) with amplifications colored in green and deletions in red relative to the ground state copy number in black. Below, we display a region matrix that shows the anatomic sector from which each cell was originally collected. By this analysis, the remaining 93 profiled cells from the tumor, regardless of the sector-of-origin, could be clustered into five subpopulations we call D, P, H, AA and AB. Three of the major tumor subpopulations (H, AA and AB) are highly clonal and comprise slightly less than half the cells of the tumor. These cells were isolated from the hypodiploid (F1) and two sub-tetraploid (F3 and F4) FACS fractions, respectively.

In our previous study by SPP, we identified some of these subpopulations (D, H, AA and AB) by profiling millions of cells by array CGH (Navin et al., 2010), but could not determine if they were composite mixtures of different tumor clones. Here, we clearly show that each subpopulation is very clonal - composed of cells that share highly similar copy number profiles. Each subpopulation (H, AA and AB) is related to the others by many shared genomic alterations but have also diverged and show distinct attributes. The AB cells, for example, all have 50-fold amplification of the *KRAS* oncogene, while the H cells display the characteristic 'sawtooth' pattern comprising broad chromosomal deletions (Hicks et al., 2006). In this tumor, the H clones are anatomically segregated in the sectors of the tumor (S1-S3), while the AA and AB clones are intermixed and occupy the other sectors (S4-S6). These results agree with our previous cytological studies showing anatomic segregation of the *KRAS* tumor clones in T10.

15.3 Ψ Pseudodiploid Subpopulation

The cells of the diploid gate comprise slightly greater than half the cells of the tumor and are found in all sectors. Hierarchical clustering divides them into two groups. The majority (34/42) have a normal profile and we call them D for diploids. Unexpectadly, the remainder of the cells isolated from the 2N FACS gate (8/42) contained broad chromosomal deletions and amplifications. We

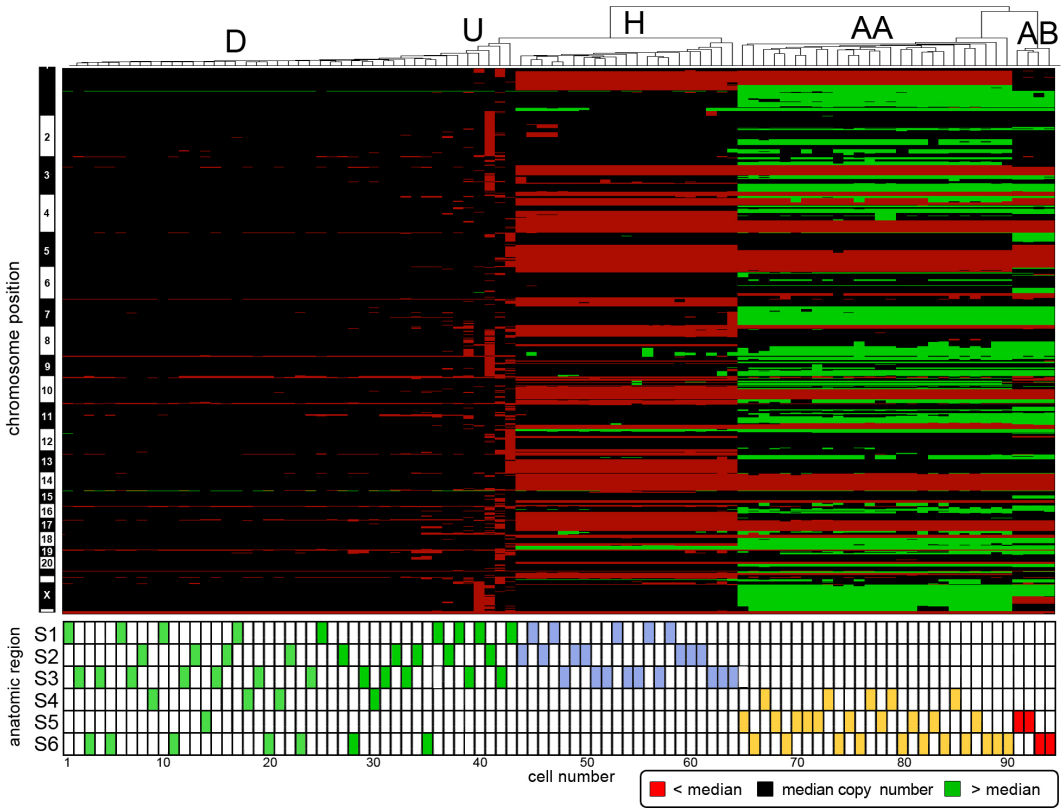


Figure 15.2 – Heatmap of 93 Single Cells and Position Matrix
Upper Panel; the absolute copy number profiles from 93 single cells were hierarchically clustered and displayed as a heatmap with genomic position preserved on the x-axis. Clustering shows the major subpopulations which are labeled as D, U, H, AA and AB. Lower Panel the anatomic location of each single cell in the six sectors (S1-S6) is displayed in a position matrix that corresponds to the heatmap above. The colors represent the ploidy peaks from which they were isolated (F1 in blue, F2 in green, F3 in orange and F4 in red).

call this subpopulation 'pseudodiploid' (Ψdiploid). Unlike the other clonal subpopulations, which contain highly similar genome profiles, individual Ψdiploid cells did not share the majority of chromosome aberrations, nor did not share any chromosome breakpoints with the major tumor subpopulations (Figure 15.3A). Thus, they are likely to represent an unstable population of precursor cells, one of which (we did not detect) may have further evolved into the major subpopulations.

The majority of chromosome aberrations were not shared between individual Ψdiploid cells, however we did identify one exception: a common region that was deleted in five out of eight Ψdiploid cells. Interestingly, the specific breakpoints surrounding the deletion varied from cell to cell, suggesting convergent evolution (Figure 15.3B). This hemizygous deletion eliminated one copy of the *RASSF1* tumor suppressor in addition to several normal genes.

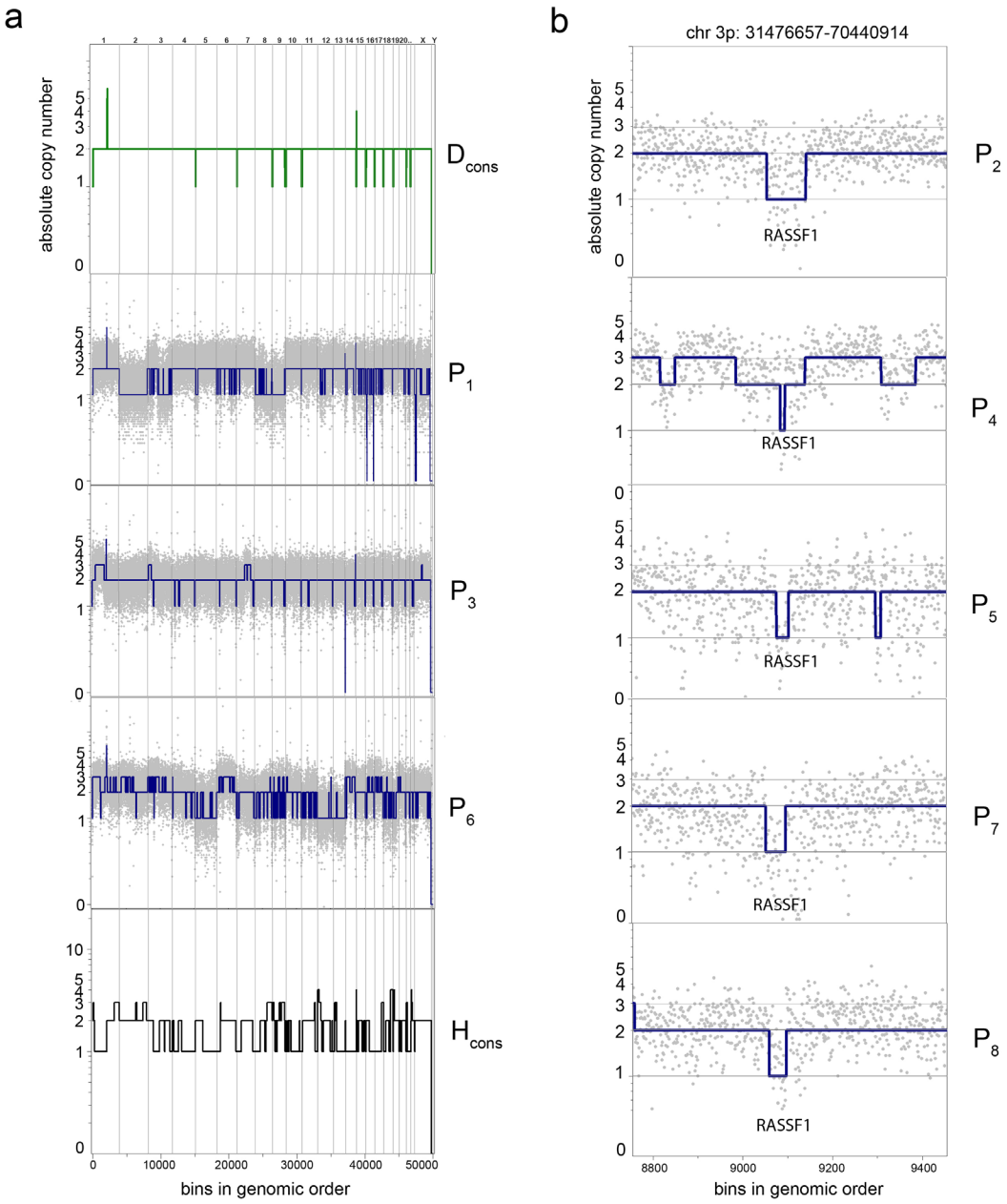


Figure 15.3 – Genomic Profiles of Pseudodiploid Cells
(A) The consensus copy number profiles of the diploid (Dcons) and hypodiploid (Hcons) subpopulations are compared to the absolute copy number profiles of three Ψdiploid cells (P1, P3, P6) and their bin count ratios showing that they are very divergent profiles, sharing almost no chromosome aberrations. (B) However, a single hemizygous deletion of the RASSF1 tumor suppressor is shared between 5/8 Ψdiploids cells on chromosome 3p21.31.

RASSF1 is a particularly potent tumor suppressor, functioning in both DNA damage repair and cell cycle arrest (Hamilton et al., 2009) and may contribute to the stochastic phenotype of these cells. We also have preliminary data from a second primary ductal carcinoma, showing a significant number of Ψdiploid cells, suggesting that this subpopulation may play a broad role in tumor progression.

CHAPTER 16

Phylogenetic Analysis of Single Tumor Cells

16.1 Absolute Copy Number Tree of 100 Single Cells

To understand the evolutionary history of the T10 breast tumor we constructed a neighbor-joining tree using the single cell profiles. Our approach involves calculating the Euclidean distance between 100 absolute copy number profiles and applying the neighbor-joining algorithm (Saitou and Nei, 1987). Euclidean distance is a density metric when applied to copy number profiles and is justified biologically, since large chromosome aberrations carry more weight than focal events. This algorithm reflects the biological consequences of losing or gaining broad chromosomal regions, by affecting gene dosage of more genes than focal events. The single cell copy number tree is shown in Figure 16.1, with the subpopulations color coded, showing five major branches of evolution that correspond to FACS gates. The overall grouping of individual cells is similar to the hierarchical heatmap shown in the previous chapter, however the genetic distance between groups is now evident. In this tree we can see very little genetic variation within each major subpopulation (D, ΨD, H, AA, AB), suggesting that they are very clonal. The tree also shows a close genetic distance between the Ψdiploids and diploids, and it is clear that the Ψdiploid are a genetically diverse group, each cell having diverged by a different distance from the diploid cells. The next closest group to the diploids is the hypodiploid cells, while the AA and AB tumor subpopulations have evolved by a considerable distance. Most important, it is clear that all cells, outside of the Ψdiploids, share a common genetic lineage and are likely to have evolved from a single progenitor cell. In other words, we did not identify any single cells that appeared as an out-group with a distinct genetic history.

16.2 Chromosome Breakpoint Tree of 93 Single Cells

An alterative approach to inferring the evolutionary history of a tumor by copy number is to construct a phylogenetic tree based on chromosome breakpoints.

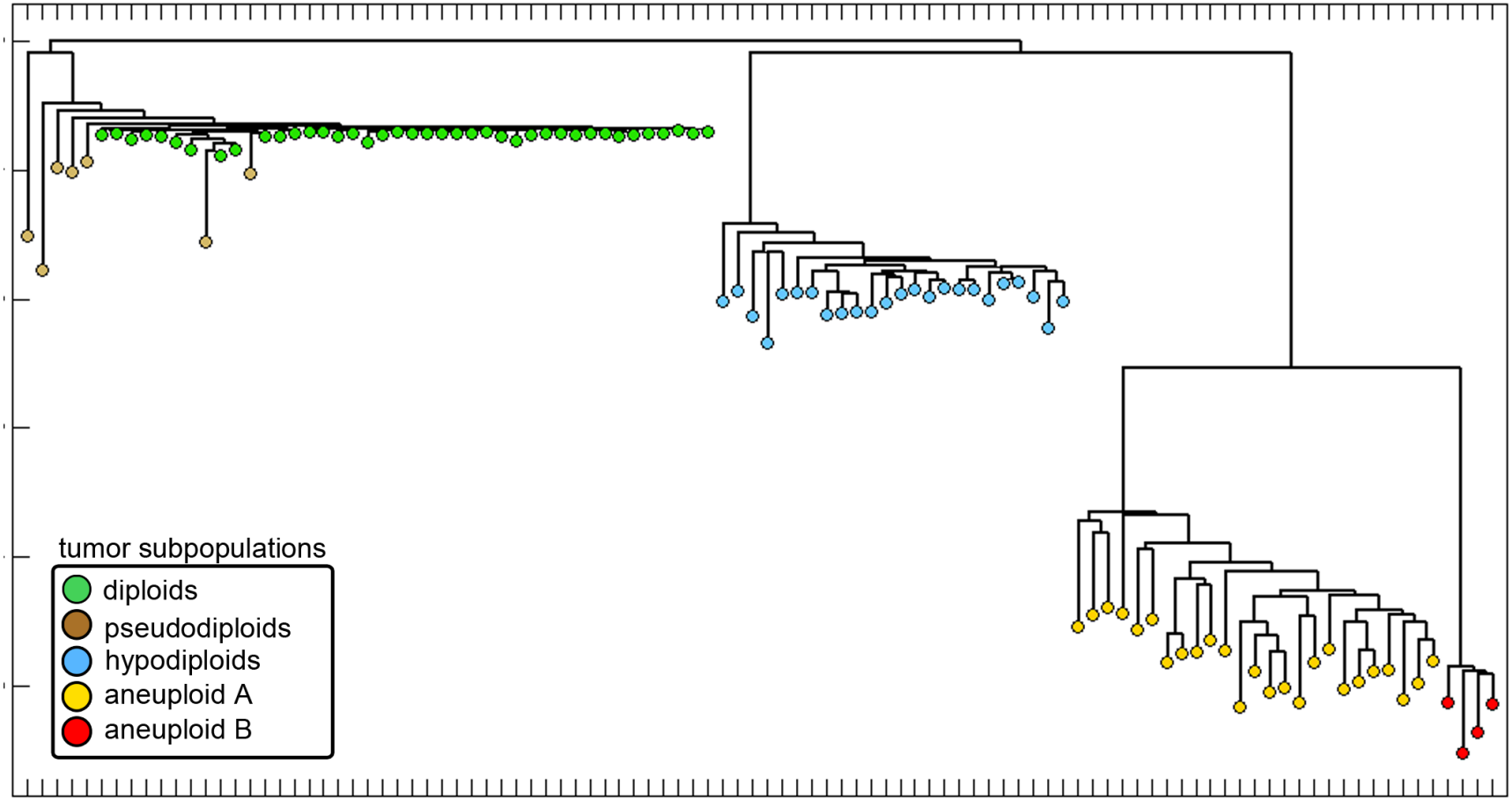


Figure 16.1 – Absolute Copy Number Tree of 100 Single Cells
A neighbor-joining tree was calculated from the absolute copy number profiles of 93 single cells using a Euclidean distance. The tree shows five major branches of evolution representing clonal subpopulations: diploid (green), Ψdiploid (brown), hypodiploid (blue), aneuploid A (orange) and aneuploid B (red).

Chromosome breakpoint markers have the advantage of not weighing large chromosome aberrations more than focal events, and thus provide an orthogonal approach to understanding the evolutionary divergence between single cells. In theory, a chromosome breakpoint tree will exaggerate the differences within clonal subpopulation, whereas a copy number tree will show a greater distance between subpopulations. To calculate a common set of chromosome breakpoints we eliminated breakpoints events with a high standard deviation and limited our analysis to breakpoint regions covering no more than seven adjacent bins. This resulted in 657 chromosome breakpoints observed in at least two cells and reduced each copy number profile to a binary string: 1 if it displayed the breakpoint, 0 otherwise. We used a neighbor-joining algorithm and Hamming distance to build a phylogenetic tree and rooted it by the parental diploid node (Figure 16.2A). The overall structure of the tree is highly similar showing the four major branches of evolution. However, the Ψdiploid cells are now intermixed with the diploid cells and span a larger genetic distance from this group. The major subpopulations (D, H, AA and AB) form highly similar groups, however the genetic variation within each subpopulation is more evident using chromosome breakpoint patterns.

16.3 Inheritance of Chromosome Breakpoints Between Subpopulations

Chromosome breakpoint markers enable us track breakpoints they are inherited between subpopulations, or diverge to form new subpopulations. To analyze these events, we used biclustering (two-dimensional clustering) to group the 657 chromosome breakpoints and single cells. The results are plotted in a heatmap with the columns ordered according to the order in the breakpoint tree (Figure 16.2B), allowing us to visually identify breakpoints that correspond to subpopulations. Each of the three major tumor subpopulations, H, AA, and AB, clearly contains shared breakpoints that distinguish them individually. There are also ample numbers of breakpoints that all three populations share but that are not abundant in the D + P subpopulations, evidence of their descent from a common ancestor (n_1 in Figure 16.2A). Less abundant but also evident are breakpoints shared by AA and AB but not by H, indicating their descent from an ancestor (n_2) after H diverged from the common path. However, the distance between the inferred common ancestor n_1 and the common ancestor n_2 is very small, so we can infer that the three subpopulations emerged when the tumor was much smaller.

By contrast, the divergence of the subpopulations after n_1 and n_2 is very large, with AB showing the greatest phylogenetic distance from the diploids. These results allow us to order the single cell profiles and infer the evolutionary pathway of this breast tumor, the focus of the next chapter.

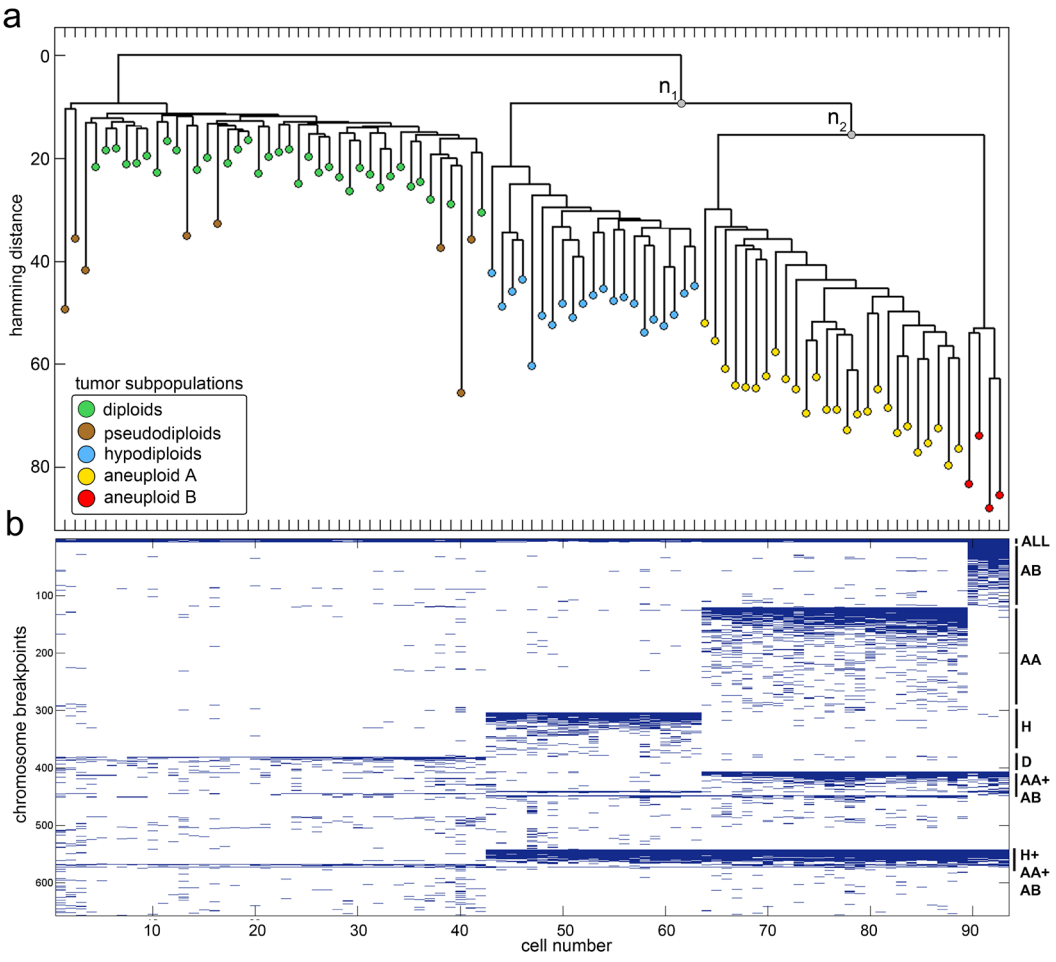


Figure 16.2 – Chromosome Breakpoints Tree and Heatmap from 93 Single Cells
(A) A neighbor-joining tree was constructed by calculating hamming distances from 657 chromosome breakpoint patterns from 93 single cells. This tree shows four major subpopulations: hypodiploids (blue), aneuploid A (orange), aneuploid B (red) and an intermixed group of diploids and Ψdiploids. This tree was rooted by the parent dipoid node. (B) The 657 common chromosome breakpoints were biclustered and ordered to correspond to the neighbor-joining tree, showing which of the breakpoints are shared or divergent in the subpopulations.

CHAPTER 17

Evolution by Clonal Expansions in T10

Our single cell analysis of T10 shows that it evolved by a series of clonal expansions that share a common genetic history. To estimate the evolutionary distance between common ancestors and better understand the evolution of the major subpopulations in T10, we calculated and compared consensus profiles. For all of the single cells within clonal subpopulations, we calculated the most frequent absolute copy number value (majority rules) to generate consensus profiles that represent each subpopulation. Assuming that mutational complexity increases with time, we ordered the consensus profiles. We present this data in a summary figure showing both the consensus copy number profiles and the corresponding FACS ploidy histograms (Figure 17.1).

T10 is composed mainly of diploid cells (34 cells) (Figure 17.1A), some unknown number of which underwent copy number changes resulting in a significant population of Ψdiploid cells (8 cells). The Ψdiploid cells did not achieve prominence and thus represent terminal nodes in the evolutionary lineage (Figure 17.1B). However, a single precursor cell (gray), which we did not detect, is likely to have lost many broad chromosomal regions and progressed into the hypodiploid subpopulation (H). This subpopulation was very successful and was the first to undergo a large clonal expansion to form a significant mass of the tumor (21 cells). The H subpopulation correlates with a large downward shift in ploidy to 1.7N (Figure 17.1C). A common ancestor of this subpopulation eventually evolved into the aneuploid A subpopulation, acquiring a number of focal amplifications and deletions, which correlated with a large upward shift in ploidy to 3.1N. This upward shift in ploidy may be the result of either cell fusion with a neighbor or endoreduplication of the genome, events which cannot be distinguish by analyzing tumor progression in single cells *a posteriori.* The AA subpopulation was highly successful and underwent the second and largest clonal expansion (26 cells). Finally, a common ancestor of AA evolved into the highly malignant AB subpopulation, by acquiring a massive amplification (over 50 fold) of the *KRAS* oncogene and homozygous deletions of the *EFNA5* and *COL4A5* tumor suppressors. This was a relatively small clonal expansion (4

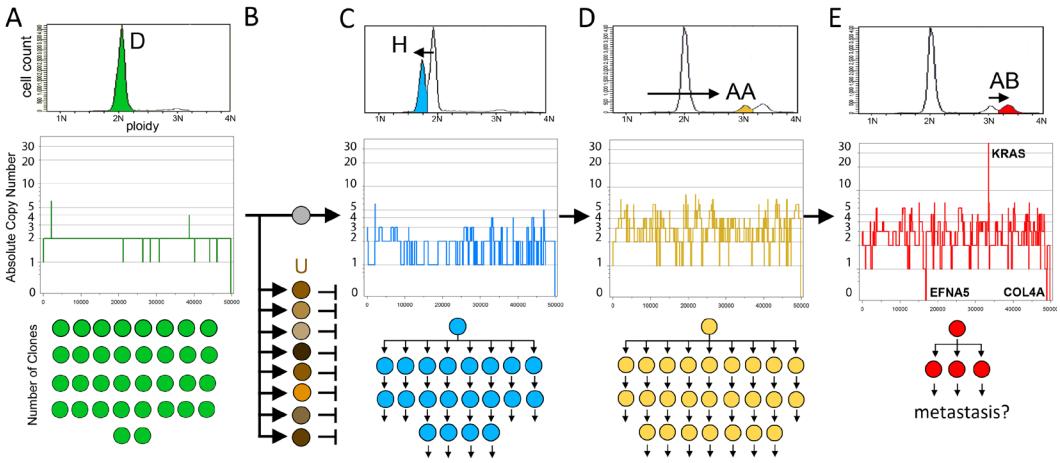


Figure 17.1 – Evolution by Clonal Expansions in T10
Consensus profiles were ordered according to their phylogenetic distance in the chromosome breakpoint tree. (A) Most normal cells from the 2N gate (34 cells) show diploid copy number profiles (B) Some of these cells progressed to form an unstable Ψdiploid subpopulation (8 cell), which did not undergo further evolution. However, one of these cells was a precursor (grey), which evolved into the hypodiploid subpopulation (C) The hypodiploid genome is characterized by a large downward shift in ploidy (1.7N) and broad chromosome deletions. This subpopulation underwent the first major clonal expansion (21 cells) (D) Eventually, a common ancestor evolved into the aneuploid A subpopulation, by aquiring additional amplifications and deletions, resulting in a large upward shift in ploidy (3.1N) and a second major clonal expansion (26 cells) (E) A common ancestor evolved into the AB subpopulation by acquiring an amplification of the *KRAS* oncogene and homozygous deletions of the *EFNA5* and *COL4A* tumor suppressors. This subpopulation underwent a small clonal expansion (4 cells) and may have migrated away from the primary site.

cells), suggesting that these cells may have migrated away from the primary site to metastasize, since the overexpression of *KRAS* has been shown to lead to cell migration by *in vitro* overexpression experiments (Fotiadou et al., 2007).

To more rigorously estimate the evolutionary divergence of the subpopulations from their common ancestors (n_1 and n_2) we applied phylogenetic inference. We calculated common chromosome breakpoint patterns from the consensus copy number profiles and applied neighbor-joining to construct a phylogenetic tree (Figure 17.2). The consensus breakpoint tree clearly shows a non-linear progression from D to H to AA to AB through a series of common ancestors (n_1 and n_2). We excluded the Ψdiploid profiles from this analysis, since consensus profiles could not be calculated from such highly divergent copy number profiles. Pie charts were also calculated to show the percentage of cells that were sampled from each subpopulation relative to the total number (93 cells).

From this analysis it is clear that the hypodiploid cells were the first subpopulation to evolve from the diploid cells, however by the time we have measured their genomes, they diverged a significant distance from their common ancestor (n_1). Importantly, this tree shows that there is a very short phylogenetic distance between the divergence of the n_1 and n_2 common ancestors. This suggests that the AA and AB subpopulations have evolved independently for a long time relative to the total evolutionary time of the tumor. Moreover, this tree shows that the AB subpopulation has diverged the longest evolutionary distance from the diploid cell, which is consistent with our finding that the AB absolute copy number profiles contain the largest number of chromosome aberrations.

Our data show that the tumor mass evolved by sequential clonal expansions (SCE) through common ancestors. In our previous analysis by SPP it was unclear if the subpopulations were single clones or composite

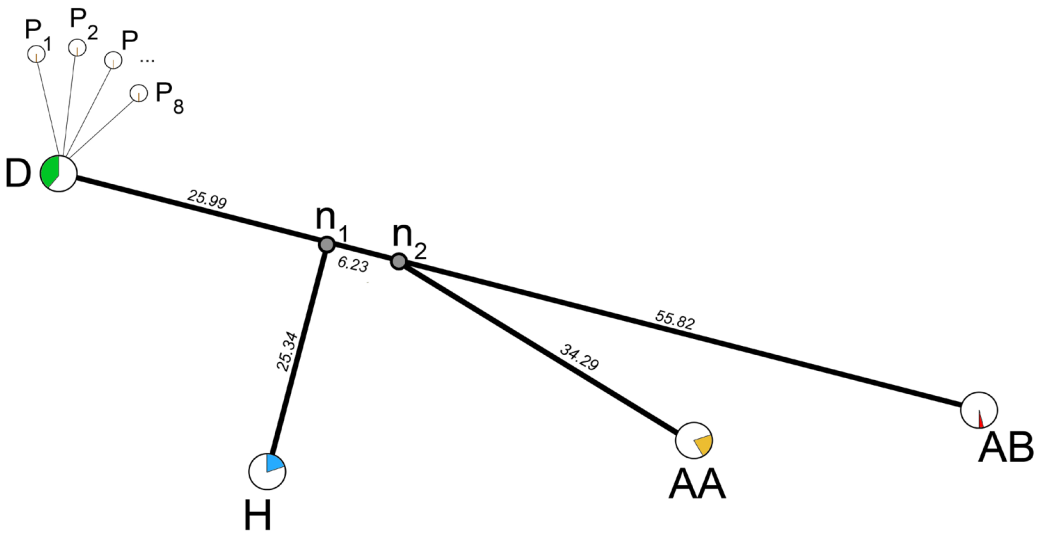


Figure 17.2 – Phylogenetic Inference of Common Ancestors in T10
The common ancestors (n1 and n2) in T10 were inferred from a neighbor-joining tree of consensus profiles. Evolutionary distance is shown between each node, and the pie charts show the relative proportion of cells that constituted the subpopulation. In this tree, the H subpopulations was the first to evolve from D by the common ancestor n1. After only a short evolutionary distance, the n2 common ancestor emerged and the AA and AB subpopulations diverged. The AB subpopulation shows the longest evolutionary distance from diploid.

mixtures, however at single cell resolution we show strong evidence for clonal subpopulations. Moreover, we show that the majority of breakpoints are inherited and persistent through successive subpopulations, suggesting that they share a common genetic lineage. The exception is the Ψdiploid subpopulation, which are likely to be a population of unstable precursor cells, the majority of which do not undergo further evolution. In the next chapter, we present SCE as a general model for tumor progression and discuss the biological and clinical implications.

CHAPTER 18

Sequential Clonal Expansion Model

18.1 SCE Model

Our results from the SPP and SNS study suggest a general model for tumor progression by Sequential Clonal Expansions (SCE). Both the progression of monogenomic and polygenomic tumors can be explained by one or more rapid bursts of genomic instability followed by the stable expansion of clonal subpopulations (Figure 18.1). In monogenomic tumors our model assumes a brief period of genomic instability (shaded in grey) resulting in the generation of a dominant aneuploid subpopulation that undergoes a stable expansion to form the tumor mass (Figure 18.1A). We show the A1 line with a positive slope, rising from 100 to 110 genetic events, to indicate that tumor cells are not perfect clones, but that minor genomic variation exists within the subpopulation. In polygenomic tumors the generation of new clonal subpopulations occur through successive rounds of genomic instability followed by the stable expansion of clones. We assume that the generation of intermediates occurs within a short evolutionary time, since we did not observe these cells in our studies (Figure 18.1B). In the polygenomic tumors we also represent each clonal subpopulation (A1, A2, A3) by lines with positive slopes to indicate minor genomic variation within each subpopulation. Our model suggests that periods of genomic instability are relatively short compared to the total growth of the tumor. The vast majority of the tumor's growth involves expanding highly stable genomes through numerous mitoses, analogous to the expansion of adaptive immune cells in response to an infection.

18.2 Intermediates are Rare

Our model assumes that major rearrangements in the tumor genome occur within one or more short burst of evolutionary time. This assumption is based on the lack of gradual intermediates that we observe in the progression of tumor subpopulations. In our initial studies by SPP we thought that intermediates were likely to be present in small numbers, and thus masked by the overwhelming signal from major subpopulations, since we used samples consisting of millions of

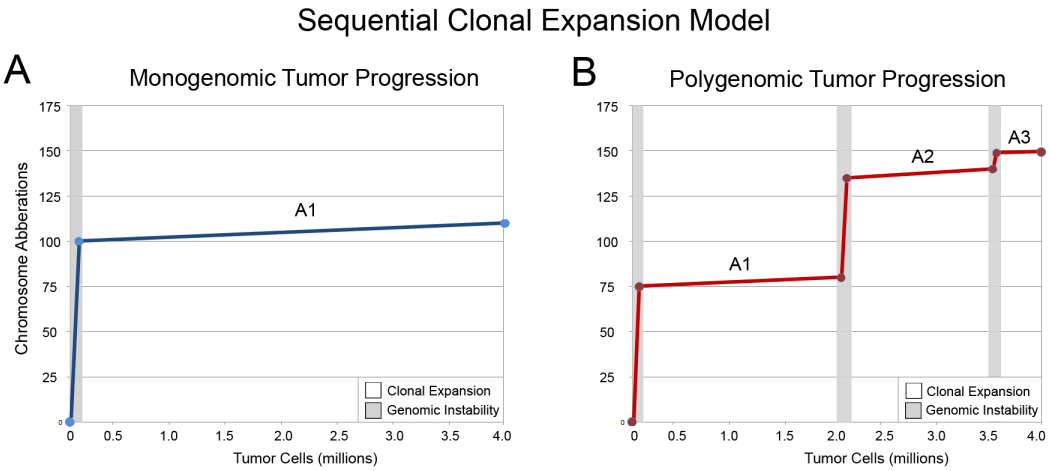


Figure 18.1 –Sequential Clonal Expansion (SCE) Model
Chromosome aberrations are plotted against the total number of tumor cells, showing the mutation rate as the tumor grows. White areas correspond to clonal expansions, while grey shading depicts periods of genomic instability. (A) In monogenomic tumors an early period of genomic instability (grey) results in the generation of a stable aneuploid tumor cell. This cell undergoes a stable expansion to form the major clonal subpopulation (A1) which dominates the tumor mass. Minor genomic variation is seen as A1 undergoes a stable expansion to form the tumor mass (B) In polygenomic tumors several burst of genomic instability are followed by stable expansions, forming the major subpopulations (A1, A2, A3). Within each subpopulation, only minor genomic variation is seen during the clonal expansions.

cells. We expected, therefore, to find rare intermediate genomes in our single cell analysis of a polygenomic tumor, but this was not the case. Despite sampling a hundred single cells, we did not detect any intermediate genome profiles between the major subpopulations in T10. This observation can be explained in two ways: (1) intermediate cells are present, but we did not gate these subpopulations by FACS, or (2) intermediate cells are extremely rare and could not be detected in our sampling of 100 cells. We believe in the latter explanation, since we gated distributions broadly by FACS and placed gates from neighboring distributions directly adjacent to one another. Thus, we think that thousands of single cells may need to be sampled in order to observe intermediate genomes.

18.3 Biological Explanations for Rare Intermediates

The scarcity of intermediate tumor genomes may be explained by catastrophic biological events such as cell fusion. In normal cells, cell fusion is believed to be a rare and tightly regulated process that takes place in the fertilization of egg and

sperm, formation of placenta, fusion of myoblasts to form muscle cells and in the formation of megakaryocytes in bone tissues (Lu and Kang, 2009). However, *in vivo* animal models of cancer have shown that the frequency of cell fusion can rise up to 1% in the tumors (Duelli and Lazebnik, 2003; Rachkovsky et al., 1998). The consequence of cell fusion in programmed normal cells often results in cells with multiple nuclei, however in non-programmed accidental fusions (cancers) the two nuclei may fuse to form a tetraploid hybrid. The hybrid tetraploid cells may then undergo multipolar divisions, leading to chromosome missegregation and following cytokenisis, the generation of cells with supernumerary chromosomes. Several DNA repair and cell cycle pathways have been implicated in regulating genomic instability, through tumor suppressors genes such *TP53, BRCA1, BRCA2, p16 ink4a/ARF,* and *ATM* (Negrini et al., 2010). If these genes (or combinations thereof) are lost in missegregated genomes, then they may lead to genomic instability, and possibly drive tumor progression.

Our data shows possible evidence for cell fusion in the basal-like tumors (T10 and T12). In these tumors we observe sectors with hypodiploid tumor cells that have lost many broad chromosome regions, and correlate with a downward shift in ploidy to 1.7N in their FACS histograms. In the same tumors, we also see sectors showing a large upward shift in ploidy to subtetraploid (3.1N and 3.3N), an almost perfect duplication of DNA content. The copy number profiles from the subtetraploid fractions clearly evolved from the hypodiploid cells as evident in the numerous chromosome breakpoints they have inherited. In T10, the duplication of the genome is particularly evident at single cell resolution, when the copy number profile of a hypodiploid cell to an aneuploid cell. However, we cannot exclude that possibility of endoreduplication through a mitotic defect in cytokenisis, resulting in an internal duplication of total DNA content in a single cell. In tumor samples, where the genome has already progressed, it is very difficult to distinguish between cell fusion and endoreduplication. Both mechanisms are reasonable theoretical explanation for the lack of intermediate genomes that we observe between the clonal tumor subpopulations.

A more radical, biological explanation is that the tumor genome evolves gradually off-site at a distant metastasis, acquiring a dramatically altered profile and then returns to the primary tumor to greatly expand its mass. Such evidence would support the Self-Seeding hypothesis, which assumes that metastatic tumor cells also enhance the growth of the primary tumor by reseeding (Norton

and Massague, 2006). Recently, an animal model for reseeding was generated, providing an ideal system to study tumor cell migration, particularly when combined with single genome methods such as SNS (Kim et al., 2009). In humans, the offsite development of tumor cells would be supported if more intermediates were found in distant metastases, than in the primary tumor.

Another plausible biological explanation for the lack of intermediates is telomere attrition. There is overwhelming evidence that telomerase is frequently inactivated in tumors leading to telomere shortening and aneuploidy, in a process that has been referred to as 'episodic telomere crisis' (DePinho and Polyak, 2004). In this model, the uncapping of telomeres leads to breakage-fusion-bridge (BFB) cycles, generating double stranded breaks and highly aneuploid tumor genomes. Recent evidence has suggested that telomere-based BFB may occur in short bursts, driving benign cancers to malignancy (Chin et al., 2004). In polygenomic tumors we may find several 'episodes' of telomere crisis occurring, followed by the subsequent restabilization and expansion of the tumor genome. This hypothesis could be tested if intermediate cells could be isolated, by correlating telomerase expression with periods of genomic stability and instability.

18.4 Punctuated Equilibrium vs. Gradualism

A similar model to SCE has been proposed in the field of evolutionary biology to explain the apparent gaps in the fossil records. In 1972 Gould and Eldredge proposed the theory of 'Punctuated Equilibrium' (Gould and Eldredge, 1972), challenging the established model of Phyletic Gradualism (Avise, 1977; Sheldon, 1987) to explain the lack of intermediate species. In this model, species experience very little evolutionary change for the majority of their geological history and remain in a state of stasis. Then, in sudden evolutionary bursts, cladogenesis occurs in which species split into two reproductively isolated groups (Figure 18.2B) (Gould and Eldredge, 1993). This model contrasts with Phyletic Gradualism, in which speciation occurs steadily over a long evolutionary time, eventually transforming species into reproductively isolated clades (Figure 18.2A). The lack of intermediate genomes that we observe in polygenomic tumor subpopulations parallels Punctuated Equilibria, by assuming that species remain phenotypically static for long periods of evolutionary time. We often see large numbers of tumors cells with highly similar genomes, representing long periods of evolutionary time of carcinogenesis. Thus, when an advantageous genotype is

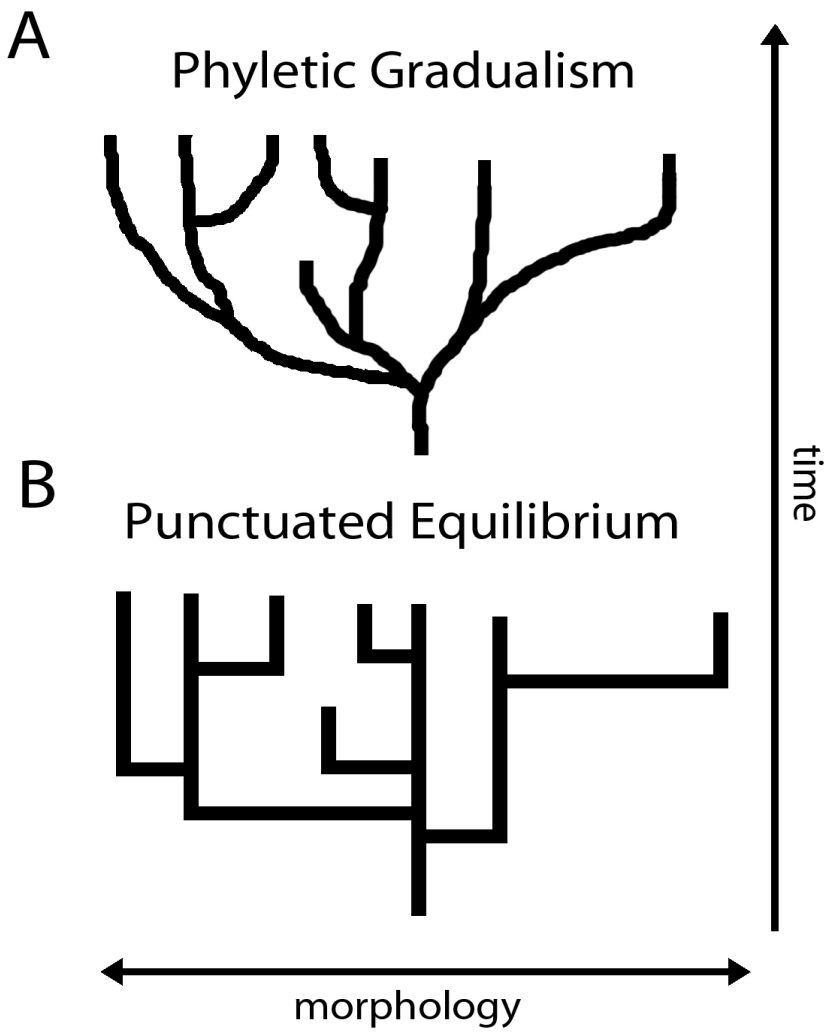


Figure 18.2 – Punctuated Equilibria
(A) Phyletic Gradualism assumes a long gradual tree in which speciation occurs over long period of evolutionary time, slowly forming new clades of reproductively isolated species. (B) Punctuated Equilibrium implies a tree in which species are in stasis for many generations, and then in short evolutionary bursts (represented by rectangular lines) form new clades of species.

achieved through mutation in an environment, strong selection causes the genome to remain stable for extended period of time, while near-perfect copies are made. Such an evolutionary perspective would suggest that sudden changes in the genomes of tumor cells might reflect dramatic environmental changes in selection factors, such as hypoxia, necrosis, angiogenesis or chemotherapy in the tumor microenvironment.

18.5 Evidence for Clonal Evolution

In principle, SCE shares assumptions with the clonal evolution model, originally proposed by Peter Nowell in 1976 (Nowell, 1976). Specifically, the monogenomic tumors that we studied are consistent with 'monoclonal evolution', while polygenomic tumors are consistent with 'polyclonal evolution' models for tumor progression. In the 1980's there was much debate on which of these models applied to various cancer types. In our studies of ductal carcinomas, we see evidence for both models, with an approximately equal frequency. The underlying assumption of these models is that the majority of tumor cells can continue to proliferate to form the tumor mass, rather than undergoing a continuous regeneration from rare precursor cell, such as a cancer stem cell (Clarke et al., 2006). In polygenomic tumors we see evidence that the majority of tumor cells continue to proliferate, expanding the tumor mass through sequential clonal subpopulations. However, a major difference between clonal evolution and SCE relates to the intermediate cells. SCE implies a scarcity of gradual intermediates, while clonal evolution assumes that many intermediate cells are generated while the tumor evolves. Specifically, polyclonal evolution assumed that a long series of gradual intermediates would drive tumor growth. SCE may also be consistent with the Self-seeding hypothesis, however more work will be needed to assess whether tumor clones show a peripheral organization at single cell resolution.

18.6 Evidence Against Stochastic Models for Tumor Progression

Several models for tumor progression, including the mutator phenotype, predict that the genome is highly unstable, containing a large diversity of non-expanded mutations in heterogeneous tumors. These models assume that the random accumulation of non-expanded mutations drive tumor progression. In no case did we observe evidence for such a model in our analysis of genomic copy number variation in breast tumors. In every tumor analyzed, we found that copy number profiles share a common genetic lineage, falling into one or more homogeneous groups. Even at single cell resolution we found that the majority of tumor cells (85/93) share a genetic lineage falling into three major subpopulations, with the only exception being a small subpopulations of Ψdiploids cells (8/93). In addition to our studies, many experiments have shown that when multiple samples are taken from single heterogeneous tumors, and compared, they share the majority of genetic mutations (Aubele et al., 1999; Shipitsin et al., 2007; Teixeira

et al., 1996; Teixeira et al., 1995; Torres et al., 2007). Thus we conclude that a stochastic model is unlikely to explain tumor progression in ductal carcinomas.

18.7 Clinical Implications of SCE

Biological models are by definition built upon incomplete information. At best, these explicit models for tumor progression provide guideposts for further exploration and strategies for clinical therapy. The original model for Clonal Evolution, predicts that a myriad of gradual intermediate tumor cells with different genomes need to be targeted by therapy to eliminate the tumor. This would be a formidable task, as the numerous intermediates would require individual targeting. SCE, on the other hand, predicts that targeting a few tumor genotypes will eradicate the majority of tumor cells and cure the disease. These models commonly assume that the majority of tumor cells have the capacity for continued proliferation, and thus imply that all of the tumor cells must be eliminated to cure the disease. Thus both models warn that missing even a single tumor cell with therapy, could lead to relapse and regrowth of the entire tumor mass.

In stark contrast, the cancer stem cell model predicts that targeting a small subpopulation of cells (CD44+/CD24- in breast tumors) would effectively treat the disease, irrespective of the rest of the tumor cells. This prediction has led to an intense study of eradicating cancer stem cells with chemical inhibitors – so far with limited success. Thus, these general models imply drastically different approaches to targeting tumor cells with therapies. Resolving which models apply to which cancer types will undoubtedly lead to better patient treatments.

The single cell methods that we have developed provide an exceptionally useful tool to study the relationship between therapy and the regression or expansion of tumor subpopulations. The SCE model may help understand how tumor subpopulations respond to therapy. We may find that monogenomic tumors respond better to therapy, since targeting a single tumor genome should eradicate the majority of tumor cells, and reduce tumor mass. In contrast, therapies in polygenomic tumors may only eradication single subpopulations, allowing the other subpopulations to expand in its place. Alternatively, if tumor subpopulations cooperate, then eliminating a single subpopulation could also lead to the demise of other subpopulations.

CHAPTER 19

Future Directions

Overview

As with any good study, we have generated more questions than have been answered. Some of the obvious questions can be answered by simple applications of technology, while other questions will require designing intricate experiments. The obvious questions include: Do other cancer types show clonal subpopulations? Do monogenomic and polygenomic tumors correlate with survival or clinical parameters? In this study we developed a single cell method and applied it to a polygenomic tumor, however we have not yet examined any monogenomic tumors. We may find that the monogenomic tumor class does not exist at single cell resolution, but that they are instead composite mixtures of tumor cells. Some of the more compelling biological questions in the primary tumor involve understanding the role of Ψdiploids in tumor progression and determining if cooperation between tumor subpopulations can accelerate tumor growth. Moreover, the development of a robust single cell method will enable us to track tumor cell migration from the primary tumor to the metastasis, and perhaps allow us to answer the fundamental question: Is metastasis a unidirectional process?

19.1 Genomic Variation in Monogenomic Tumors

Are monogenomic tumors really ‘monogenomic’? With the development of SNS we deliberately sought to analyze 100 single cells from a polygenomic tumor so that we could infer progression. In future experiments we would like to apply the same approach to one or more monogenomic tumors. Our working hypothesis is that monogenomic tumors contain highly similar genome profiles in the majority of tumor cells. However, at single cell resolution, we may find that there is no monogenomic tumor class, but rather that they all consist of composite mixtures of divergent genomes. Alternatively, we may find that monogenomic tumors do contain highly stable aneuploid genomes throughout the entire tumor mass, which would confirm our hypothesis. In these experiments we may also discover an abundance of Ψdiploid cells, showing that they are also common in monogenomic

tumors. Such studies will shed light into the relative stability of aneuploid genomes in tumors.

19.2 Clinical Correlations with Genomic Heterogeneity

Our samples consisted of twenty ductal carcinomas with limited clinical information beyond grade and receptor status. These tumors were mainly high grade (III) and consisted of various combinations of estrogen, progesterone and Her2 receptor types. Initially we wanted to test the hypothesis that polygenomic tumors would show a change in grade between tumor sectors. To do this we stained tissue sections from each sector with hematoxylin and eosin, and graded the sections with help from a pathologist (Dr. Anders Zetterberg). Using the Fisher's Exact test, we found no significant correlations. We did, however, find a clinical correlation between triple-negative receptor status in the basal-like T10 and T12 tumors, and polygenomic progression by transitioning from hypodiploid to aneuploid genome patterns. A larger study of basal-like, triple negative breast tumors is need to determine if this pattern of progression is a common phenomenon in this subtype.

We also hypothesize that monogenomic tumors correlate well with patient survival, but lack such clinical data on our collection of tumors. We would like to conduct another study on a larger group of patients with detailed clinical information about survival. This will allow us to construct Kaplan-Meier curves to compare survival in monogenomic and polygenomic tumors. We anticipate that polygenomic tumors will show poor survival, since, in theory, targeted therapies may not eliminate all of the tumor subpopulations, allowing other subpopulations to expand in their place.

Single cell analysis may also allow us to investigate genomic heterogeneity in the early stages of cancers. For example in early stage breast cancer, such as DCIS, tissue is often limited to less than 100 cells. Using Laser Capture Microdissection, we can isolate single cells from breast ducts to investigate whether significant genomic heterogeneity exists in these early cancers, or if a single dominant clones has already begun to expand.

19.3 Clinical Applications of Single Nucleus Sequencing

A major advantage of single cell analysis, is that minute tissue samples can be extracted from a tumor, allowing less invasive procedures to be performed on patients. For example, a fine-needle could slowly be dragged through the tumor mass to aspirate cells from multiple regions. The aspirate could then be analyzed to quantify genomic copy number in hundreds of tumor cells to estimate genomic heterogeneity. Fine needle aspirates have the advantage of being far less invasive to the patient than surgical biopsies, which generally extract large portions of the tumor and surrounding tissues.

Single cell analysis may also have clinical applications in detecting circulating tumor cells. During angiogenesis tumors often construct 'leaky' vasculature, which results in the shedding of many cells into the circulatory system. These circulating tumor cells often occur in frequencies of less than 1 in a million in the blood. Using epithelial surface markers such as cytokeratins, it may be possible to isolate a few tumor cells for single cell analysis. If aneuploid genomes can be detected in the blood, then this appearance may serve as an early warning sign of cancer, a procedure that is even less invasive procedure than fine needle aspiration. One day it may even be possible for primary care physicians to take blood samples during routine patient checkups for detection of early signs of circulating tumor cells.

19.4 Elucidating the Role of ΨDiploid Cells in Tumor Progression

In our analysis of 100 single cells in a polygenomic breast tumor, we identified an unexpected subpopulation of Ψdiploid cells with random chromosome aberrations. These cells were flow-sorted along with normal diploid cells from a 2N gated distribution. The stochastic amplifications and deletions that were detected in their genomes were not shared with the major aneuploid tumor subpopulations, nor were they shared between other Ψdiploid cells. This subpopulation constituted a significant proportion of the tumor mass 8/93 cells and may represent an unstable precursor subpopulation, from which one cell may eventually evolve into the major tumor subpopulations.

An important question to address is whether Ψdiploid cells are unique to T10, or alternatively, are commonly found in all breast tumors. Studies addressing this questions will require the analysis of large numbers of cells in other tumors, since they occur at a very low frequency in T10 (8/93). We have

begun to address this question in a second breast tumor, AST5, and have found that they are also present (3/14 cells), suggesting that they may play a broad role in breast tumor progression. More research will be needed to see if they are common to other cancer types.

A possibility may be that the genomic events that we observe in Ψdiploid cells are an artifact of the SNS method, and therefore do not represent a biological precursor. In theory random chromosome deletions could be explained by nuclei that were shaved during the mincing of tissues, or lost during the transfer of nuclei. We do not think this to be the case, since we also observe amplifications (albeit less commonly) in these cells. Nevertheless, we cannot exclude the possibility that Ψdiploid cells are artifacts of the methodology. To distinguish between these possibilities we can design experiments using normal breast tissues that can be obtained from reduction mammoplasties. If no Ψdiploid cells can be detected in these normal tissues, then we would conclude that they are not artifacts of the SNS method.

19.5 Investigating Cooperation Between Tumor Subpopulations

In monogenomic tumors, single clone expand and dominate the tumor mass, presumably having outcompeted other tumor cells. Polygenomic tumors, however, maintain multiple clonal subpopulations that coexist within the tumor. In an environment with limited resources, natural selection would predict the fittest population to have outcompeted the others, particularly when intermixing occurs in the same tissues. This selection would not occur if the clones are anatomically segregated, because they occupy different 'environments' and thus may not compete for the same resources. In theory clones may occupy different territories because they are better outfitted to deal with selection pressures such as hypoxia, necrosis, angiogenesis or chemotherapy that exist in their respective microenvironments. However, when genetically distinct tumor clones are intermixed within tissues, it implies a cooperative effect. In our detailed FISH analysis of clone organization in T10 we observed that A1 and A2 were stochastically intermixed. This raises the question: does the A1 subpopulation support the growth of A2? In theory the nature of their interactions could be commensal, mutualistic or even parasitic. It is difficult to test such hypothesis through experimentation in human tissue samples.

Our FACS analysis showed that T10 was not the only tumor to contain multiple aneuploid subpopulations within single tumor sections. Many polygenomic tumors showed this pattern, and it would be interesting to conduct additional detailed FISH experiments using subpopulation-specific markers to see if tumor clones are commonly intermixing in tissues, or, alternatively, cluster into discrete domains. The latter would suggest that they do not cooperate directly, but we cannot exclude the possibility of indirect paracrine signaling across long distances in the tumor mass.

We have identified a number of breast cancer cell lines (ALAB, BT-483, BT-549, UACC-893) that show multiple aneuploid peaks in their FACS histograms . Such 'polygenomic' cell cultures may serve as a good model for studying human tumor cell cooperation. Using these cultures we could design experiments to flow-sorted or subclone subpopulations into separate cultures. We could then assay the growth rates of the individual 'monogenomic' cultures and compare them to the original 'polygenomic' culture, to see if their coexistance potentiates their growth. Studying the interactions of tumor clones has clinical significance, since targeting a subpopulation with therapy could lead to the rise or the demise of neighboring subpopulations.

19.6 Analyzing DNA Sequence Mutations in Single Cells

SNS can also be used to measure DNA sequence mutations in single cells when sufficient read density is achieved to call heterozygous or homozygous events. However, our current methods impose a major limitation, because less than 10% of the genome is randomly amplified from a single cell. This makes the interrogating of specific loci or cancer genes very difficult, since the probability of having sufficient overlapping reads between single cells will be very low. The coverage limitation is likely to be imposed by the initial amplification of the genome by the Φ29 polymerase, and thus increasing the efficiency of this reaction is imperative to single cell sequence analyses. By optimizing sequence coverage in single cells, we may also be able to use targeted approaches such as microarray capture (Hodges et al., 2007) or solution capture (Gnirke et al., 2009) methods to investigate the inheritance of point mutations in specific cancer genes. These mutations will enable the reconstruction of single cell lineages based purely on sequence, and allow us compare them to the lineages we have constructed from copy number data. With single cell sequence data, we can also investigate if

monogenomic and polygenomic tumors classifications are supported by DNA sequence analysis, an orthogonal approach.

19.7 Investigating Metastasis With Single Genome Analysis

In theory, tumor lineages can be traced all the way to the final step of progression: metastasis. While seemingly an obvious extension of the studies on primary tumors, metastatic studies are few because the material is rare. Metastases are seldom excised or biopsied in late stage patients unless part of a dedicated study, and recurrence – sometimes years after the surgery – is often treated by different physicians at different institutions. Therefore matching the correct primary and metastatic tumor samples from the same patient is often formidable. Distant metastasis, however, is nearly always the direct cause of patient mortality, and understanding its relationship to the primary tumor is of paramount importance.

A major question revolves around determining which cells are capable of initiating metastasis and how they can be identified. Also, which subpopulations in polygenomic tumors have the ability to metastasize? Do tumor clones from the primary and metastatic tumors share the majority of chromosome aberrations or do they acquire new mutations that confer metastatic potential? Is metastasis a unidirectional process, or do tumor cells return to the primary site to reseed the primary tumor? To address these questions we can apply SNS analysis to multiple tissues samples that have been collected from a single patient. In breast cancer, we would ideally collect primary tumors, lymph nodes, circulating tumor cells and multiple distant metastases to track the migration of tumor cells with genomic markers from single cells. A good source for such samples would be human cadavers, since pathologists generally remove the tumor tissues at the same time - directly after death. Thus the tumor samples would have a high probability of belonging to the same patient.

Several studies have measured copy number aberrations with microarrays to detect changes occurring between primary and metastatic tumors. In studies of various cancers types, they report that metastatic profiles are highly similar to primaries and diverge by few, if any, genetic events. (Bockmuhl et al., 2004; Hovey et al., 1998; Israeli et al., 2004; Jiang et al., 2005; Liu et al., 2009b). These studies may have analyzed monogenomic tumors, in which the dominant clones have metastasized. This suggests a remarkable stability of the tumor genome as

it migrated from the primary tumor, to the lymph nodes, the blood and finally the metastases. An alternative explanation is that the profiles represent mixtures of composite clones. If we find at single cell resolution that primary and metastatic tumors indeed contain composite mixtures of the same clones, then we can assume that tumor clones are trafficking in equilibrium between both anatomical tumor sites.

Single cell metastatic studies may also teach us more about the cooperation of clones in the tumor microenvironment. If we find that the same two clonal subpopulations are present in both the primary and metastatic tumors, it would imply a co-dependence for survival or growth. We may also find that only a single subpopulation has metastasized from a polygenomic primary tumor, suggesting that it had metastatic potential, and was not dependant on the other subpopulations. Moreover, we can compare inert subpopulations (that have not metastasize) to subpopulations that are present in both the primary tumor and the metastasis. Such direct comparisons may reveal specific mutations that are associated with conferring metastatic potential to cells. Another interesting question revolves around the presence of Ψdiploids in the metastasis. If we find by SNS analysis that Ψdiploid cells are exclusive to the primary tumor tissues, then we can assume this location to be the site of origin. In summary, tracing genomic markers in single cells has the potential to illuminate the metastatic progression of tumor cells.

CHAPTER 20

Final Remarks

In our study of 20 breast tumors using SPP, we identified two classes of genomic structural variation: monogenomic and polygenomic tumors, that occurred with an approximately equal frequency. Monogenomic tumors consisted of an apparently single dominant subpopulation of tumor cells with a highly stable genome structure. We could not infer progression from these tumors, because genetic time points were not present. These monogenomic tumors showed remarkable stability in their genome structure. We cannot exclude the possibility that such a class does not exist, and that all monogenomic tumors are composite mixtures of cells, however, this is unlikely, since our single cell analysis of a polygenomic tumor (T10) revealed that the major subpopulations were genetically very homogeneous.

In the polygenomic tumors we found that heterogeneity could be ascribed to relatively few (2-3) homogeneous subpopulations. These subpopulations were either intermixed or anatomically segregated in the tumor, the latter of which has important clinical implications for diagnostic sampling. Assuming that mutational complexity increases with time, we used these subpopulations as genetic time points to reconstruct the evolutionary history of each tumor. The clearest examples were observed in the basal-like breast tumors (T10 and T12), which progressed from diploid, to hypodiploid to aneuploid subpopulations, acquiring genetic events as they evolved. By analyzing these and other polygenomic tumors it became apparent that evolution occurred by the clonally expansion of highly similar genomes. However, despite our efforts to isolate subpopulations by region and ploidy, these analyses were based on samples from millions of cells. Thus, we could not exclude the possibility that each subpopulation was actually a composite mixture of different clones.

To more clearly understand the clonal composition and patterns of progression, we developed a high-resolution method to quantify copy number in single cells called SNS. To validate our method we analyzed single cells from cultures (SKN1 and SK-BR-3) that were presumably, very clonal. Our results not only validated our method, but also showed that there is only minor genomic copy number variation in these cell cultures - a finding that will bring great

comfort to scientists working with these cell culture. We also show that absolute copy number profiles can be quantified in single cells, which provides a major advantage over estimating copy number ratios by array CGH experiments. To our knowledge, SNS is the first method to quantify genome-wide copy number in single cells with a high resolution (50kb).

We applied SNS to 100 single cells from a polygenomic tumor (T10) to investigate the clonal composition and infer a detailed lineage of progression. Our analysis clearly shows that the tumor was composed of three clonal tumor subpopulations, rather than composite mixtures. At single cell resolution, we did observed minor genomic variation within each subpopulation, however, these differences were relatively small compared to the phylogenetic distance between subpopulations. By reconstructing the evolutionary history of this tumor using single cells, we show strong evidence that tumor growth occurred by sequential clonal expansions of highly stable aneuploid genomes.

We also identified an unexpected minor subpopulation of Ψdiploid cells in the diploid fraction that contained random chromosome aberrations. Unlike the major subpopulations, these cells were genetically diverse - sharing almost no genetic events - and may represent a stochastic precursor subpopulation. The role and prominence of these cells in tumor progression remains to be investigated.

Our data suggests a general model for tumor progression by sequential clonal expansions (SCE), which relates to the clonal evolution model. Both models assume that the majority of tumor cells have the capacity for continued proliferation, however differ in the number of gradual intermediates that they assume. Our model challenges models of continuous regeneration from a precursor cell, or expansion by the accumulation of non-expanded random mutations. Future experiments will determine if our model is supported by orthogonal data, such as epigenetic or DNA sequence analysis.

Our investigation of tumor heterogeneity and progression has led to the development of two new approaches, SPP and SNS, to measure genomic copy number aberrations within tumors. The latter method has opened up new avenues for studying single cells in cancers and will be useful for investigating other human diseases. Currently, estimating copy number in single cells is expensive ($1000.00 per genome), however we expect this cost to drop by the end of the

year to about the cost of a microarray ($100.00), as technological developments continue to drive down the expenses of massively parallel sequencing. Moreover, the development of multiplexing methods such as 'barcoding' of libraries from single cells will allow multiple single cells to be sequenced together on a single flowcell lane, further driving down the overall costs, and perhaps permitting the analysis of hundreds of single tumor cells in a single sequencing run.

As technology continues to evolve, more analyses of complex mixtures will give way to methods aimed at the individual cell. Single genome methods will give us a clearer picture of how cells develop in rapidly evolving populations, such as tumors. As we bring the magnifying glass closer, we will learn more about the rare subpopulations that play a role in tumor progression, giving us a clearer picture of genomic instability in cancer. While our studies have focused on the development of primary tumors, our methods will further enable us to track single cancer cells as they migrate into the circulation and seed metastatic tumors. We may find that metastasis occurs very early in tumor progression and that it is not a unidirectional process, as current dogma assumes. Understanding these models and the clonal composition of tumors will undoubtedly lead to improvements in clinical diagnosis and patient treatment, and perhaps even cures.

CHAPTER 21

Detailed Methods

21..1 Patient Samples

Twenty frozen primary ductal carcinomas were obtained from the Cooperative Human Tissue Network (T1–T6,T9-T11), Peggy Kemeny at North Shore University Hospital (T7–T8), Asterand Corporation (T16–T17), Larry Norton at Memorial Sloan-Kettering Cancer Center (T12–T14), and from Hanina Hibshoosh at Columbia University (T19–T20). The frozen ductal carcinoma T10 (CHTN0173) was obtained from the Cooperative Human Tissue Network. Pathology shows that this tumor was poorly differentiated and high grade (III) as determined by the Bloom-Richardson score, and triple-negative (ER-, PR- and Her2/Neu-) as determined by immunohistochemistry. Cell lines used in this study include a normal male immortalized skin fibroblast (SKN1) and a breast cancer cell line (SK-BR-3).

21.2 Sector-Ploidy-Profiling (SPP)

Macro-dissection of Tumor Sectors

The 1–2-cm^2 frozen tumors were macro-dissected into eight to 16 sectors of equal size using surgical scalpels. Half of the sectors from each tumor were used to prepare tissue sections at 6 μm in size using a cryomicrotome. The other half of the adjacent tumor sectors were used to isolate nuclei for SPP.

Isolation of Subpopulations by FACS

Nuclei were isolated from tumor samples by finely mincing a tumor sector in a Petri dish in 1.0–2.0 mL of NST-DAPI buffer (800 mL of NST [146 mM NaCl, 10 mM Tris base at pH 7.8, 1 mM CaCl2, 21 mM MgCl2, 0.05% BSA, 0.2% Nonidet P-40]), 200 mL of 106 mM MgCl2, 10 mg of DAPI, and 0.1% DNase-free RNase A using two no. 11 scalpels in a cross-hatching motion. Minced tissue was stored on wet ice for 15 min. Before flow cytometric analysis, samples were filtered through 37-μm plastic mesh. In all LSRII and FACS Vantage analysis, a small amount of prepared nuclei from each tumor sample was mixed with a

diploid control sample (derived from a lymphoblastoid cell line of an apparently normal person) to accurately determine the diploid peak position within the tumor DNA content distribution and establish FACS collection gates. Nuclei were sorted with a Becton Dickinson FACS Vantage DiVa Flow Cytometer and Cell Sorter by gating cellular distributions with differences in their total genomic DNA content according to DAPI intensity. Additionally, a small sample of cells ($n < 5000$) from the adjacent sectors (that were used for histology) had nuclei isolated and stained with DAPI for analysis by a Becton Dickinson LSRII flow cytometer to generate a histogram of the DNA distributions in order to determine if they were consistent with the flow-sorted tumor sectors.

Representational Oligonucleotide Microarray Analysis (ROMA)
DNA was isolated from the flow-sorted nuclei using the QIAGEN Genomic DNA Isolation Kit. A total of 200ng of DNA was used to make complexity-reducing representations of genomic DNA for whole-genome copy number analysis by ROMA as described by Grubor et al. (2009). ROMA greatly increases signal-to-noise ratios and diminishes the amount of sample required for analysis; therefore, no additional whole-genome amplification step was required from the tumor sectors. Samples were hybridized on two array platforms: 85K arrays based on BglII representations (samples T1–T14), and 390K arrays based on DpnII representations, depleted of DpnII fragments containing AluI sites (T15–T20). The microarrays were custom designed with probes complementary to the complexity-reducing representations and manufactured by NimbleGen. Hybridizations of the 85K experiments were performed in color reversal to prevent color bias and ensure data quality, while 390K experiments were performed without a dye swap. All tumor samples were cohybridized with a reference genome from fibroblast DNA.

21.3 Processing of ROMA Microarray Experiments

The ROMA experiments were scanned, gridded, and normalized with a Lowess curve-fitting algorithm followed by a local normalization as described by Hicks et al. (2006). The data were imported and analyzed using Splus (Insightful) and Matlab (Mathworks), and the geometric mean ratio was computed from each color channel. In color-reversal experiments, the geometric mean of two log

ratios was calculated. The data were then segmented to define nonoverlapping genomic regions that vary in copy number across the human genome using both the Kolmogorov-Smirnov algorithm (Grubor et al. 2009) and the circular binary segmenter (Venkatraman and Olshen 2007). The segmented genomic copy number profiles from each sector were then used for the statistical analysis.

21.4 Fluorescence in situ hybridization

FISH probes were constructed by one of two methods. The *KRAS* and *ETNK* probes were designed using the PROBER algorithm and pooled from PCR products 500–1400 bp in length (Navin et al. 2006). The LCTR and RCTR probes were designed using bacterial artificial chromosomes from the UCSC Genome Browser. FISH analysis was conducted on interphase cells in 10-μm frozen tissue sections. These probes were hybridized to frozen tissue sections that were fixed in methanol overnight and moved to 70% ethanol. The FISH experiments were performed as reported by Hicks et al. (2006) with DAPI staining to visualize the nucleus. Selected cells were photographed in a Zeiss Axioplan 2 microscope equipped with an Axio Cam MRM CCD camera and Axio Vision software.

In order to mitigate the analysis of shaved nuclei, we employed three precautionary steps. First, we cut relatively large (7 μm) tissue sections using a cryomicrotome in order to encompass whole nuclei. Second, we captured Z-planes that contained 40–50 images from each 63× objective microscope using a mechanical stage. Using Axiovision Software, we generated Z-plane images of the DAPI-stained nuclei, which we used to exclude any partially shaved nuclei in the quantification of FISH probe signals. Third, we hybridized two diploid control probes to all nuclei (RCON and LCON) that surround the *KRAS* amplification on chromosome 12p12.1 and a MYC control probe on chromosome 8. These control probes served as indicators that the nucleus was not shaved on chromosome 12p12.1. When we did not observe two copies of each control probe in the nucleus, it was not scored for copy number. Using these three criteria, we observed that the majority of cells that we scored (89.69%) showed copy number signals consistent with one of three subpopulations: D, A1, or A2. However, some nuclei (10.31%) did report patterns of copy number that were inconsistent with the predicted subpopulations. We cannot distinguish if these nuclei represented a minor subpopulation or if they were shaved nuclei. Finally, in order to avoid

probe artifacts, we did not score any nuclei where the probes did not overlap the DAPI channel.

21.5 Statistical Analysis of ROMA Profiles

In order to identify highly similar copy number profiles in single tumors for profile coalescing, we calculated a matrix of Pearson correlations between profiles and used a neighbor-joining algorithm (Saitou and Nei 1987). The neighbor-joining algorithm was used in place of an ultrametric method because we did not assume an equal distance from each copy number profile to the root node. In our calculations of correlation matrices, we used segmented data from the autosomes in order to exclude extraneous correlations from the sex chromosomes, and since our reference sample was male. The correlation matrix was converted to a distance matrix using (1-correlation). Clusters of highly similar copy number profiles were then "coalesced" into mean segmented profiles to represent each subpopulation in a single tumor. The pairwise difference between coalesced profiles was then calculated to identify subpopulation-specific amplifications and deletions. Each genomic lesion was annotated to identify UCSC genes (Hsu et al. 2006) and cancer genes. Cancer genes were identified using a compiled database from the cancer gene consensus (Futreal et al. 2004) and the NCI cancer gene index (Sophic Systems Alliance Inc., Biomax Informatics A.G). Distance trees were calculated using the same methods for coalescing profiles (1-Pearson correlations and neighbor-joining). A single distance tree was calculated for each tumor. Additionally, the minimum correlation between all tumor profiles is reported as the clonal correlation (cc), a measure of intratumor heterogeneity. In a separate analysis, we used the same methods to construct a distance tree using all tumor copy number profiles. In this analysis, we clustered the 85K (T4–T14) and 390K (T15–T20) tumor profiles separately and did not use any diploid profiles as a root node.

21.6 Single Nucleus Sequencing (SNS)

Macro-dissection of Tumor Sectors

Nuclei were isolated from cell lines and from the frozen tumor using an NST-DAPI buffer (800 mL of NST [146 mM NaCl, 10 mM Tris base at pH 7.8, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 0.05% BSA, 0.2% Nonidet P-40]), 200 mL of 106 mM

$MgCl_2$, 10 mg of DAPI, and 0.1% DNase-free RNase A. The frozen tumor was first macro-dissected into 12 sectors of equal size using surgical scalpels and nuclei were isolated from six sectors for FACS by finely mincing a tumor sector in a Petri dish in 1.0–2.0 mL of NST-DAPI buffer using two no. 11 scalpels in a cross-hatching motion. The cell lines were lysed directly in a culture plate using the NST-DAPI buffer, after first removing the cell culture media. All nuclei suspensions were filtered through 37-µm plastic mesh prior to flow-sorting.

Isolation of Single Nuclei by FACS

Single Nuclei were sorted by FACS using the BD Biosystems Aria II flow cytometer by gating cellular distributions with differences in their total genomic DNA content (or, ploidy) according to DAPI intensity. First a small amount of prepared nuclei from each tumor sample was mixed with a diploid control sample (derived from a lymphoblastoid cell line of a normal person) to accurately determine the diploid peak position within the tumor and establish FACS collection gates. Before sorting single nuclei, a few thousand cells were sorted to determine the DNA content distributions for gating. A 96-well plate was prepared with 10ul of lysis solution in each well from the Sigma-Aldrich GenomePlex© WGA4 kit. Single nuclei were deposited into individual wells in the 96-well plate along with several negative controls in which no nuclei were deposited.

Whole Genome Amplification

Whole genome amplification was performed on single flow-sorted nuclei as described in the Sigma-Aldrich GenomePlex WGA4 kit kit (cat # WGA4-50RXN) protocol. WGA fragments from the frozen breast tumor and SK-BR-3 single cells were used directly for Single-read library construction using the Illumina Genomic DNA Sample Prep Kit (cat # FC-102-1001) and following standard protocol with a gel purification size range of 300-250bp. WGA fragments from the fibroblast cell line were first sonicated using the Diagenode Bioruptor© using the following program: 2 times, 7 minutes with 30 seconds high on/off mode in ice cold water. Sonication removes a specific 28bp adapter sequence that is added on during WGA, and improves the total number of sequencing reads per lane.

Construction of Sequencing Libraries

Single-read libraries from single nuclei were sequenced on individual flow-cell

lanes using the Illumina GA2 analyzer for 76 cycles. Data was processed using the Illumina GAPipeline-1.3.2 to 1.6.0 Sequence reads were aligned to the human genome (HG18/NCBI36) using the Bowtie alignment software with the following parameters: 'bowtie –S –t –m 1 –best –strata –p16' to report only top scoring unique mappings for each sequence read. To eliminate PCR duplicates, we removed sequences with identical start coordinates.

21.7 Read Depth Counting in Variable Bins

Since the human genome contains many repetitive elements, we measured copy number in genomic bins of uniform expected read density, instead of fixed intervals. Specifically, we simulated sequence reads by sampling 200 million sequences of length 48 from the human reference genome (HG18/NCBI36) and introduced single nucleotide errors with a frequency encountered during Illumina sequencing. These sequences were mapped back to the human reference genome using Bowtie[16] with parameters as described above. We assigned a number of bins to each chromosome based on the proportion of simulated reads mapped. We then divided each chromosome into bins with an equal number of simulated reads. This resulted in 50009 genomic bins with no bins crossing chromosome boundaries. The median genomic length spanned by each bin is 54kb. For each cell the number of reads mapped to each variable length bin was counted.

21.8 Absolute Copy Number Quantification

Vectors of read counts in the variable bins were segmented into intervals, each segment having a distribution of bin counts significantly different from adjacent segments as determined by the Kolmogorov-Smirnov statistic[17]. For the aneuploid cells a Gaussian kernel smoothed density of the lowest 95% of the absolute values of the segmented bincount differences of all pairs of segments was computed using the density function in S-PLUS (S-PLUS 2000, MathSoft, Inc.). It is assumed that the first peak represents the mode of the distribution of segmented bincounts in identical copy number states. The second peak represents the mode of the distribution of segmented bincounts in regions differing by a copy number of one, which we refer to as the copy number one increment. For the diploid cells the median segmented bincount represents a copy number increment of two. For each cell the segmented bincount is divided by the copy number one increment and rounded to the nearest integer to give the copy number estimate.

Consensus profiles were calculated from the absolute copy number from all profiles within a subpopulation, by taking the majority value but rounded down when equal.

21.9 Gene Annotations

Amplifications and deletions identified in the single cell copy number profiles were annotated to identify UCSC genes and cancer genes. Cancer genes were identified using a compiled database from the cancer gene consensus and the NCI cancer gene index (Sophic Systems Alliance Inc., Biomax Informatics A.G).

21.10 Heatmap Clustering

Vectors of absolute copy number profiles were hierarchically clustered using average linkage and a Euclidean distance metric. The heatmap was configured to show values below the copy number median, or deletions) as red, median copy number as black, or values above the copy number median (or amplifications) as green.

21.11 Common Breakpoint Detection

Breakpoints are defined as bins with a copy number different than the previous bin in genome order. A transition from a lower copy number to a higher copy number (in genome order) is considered to be a different event than the opposite transition. To find breakpoint regions we count each breakpoint in each cell and the immediately neighboring bins. A contiguous set of bins with counts greater than 1 is designated a breakpoint region. This results in a set of 786 breakpoint regions. Each cell is then scored for the occurrence of each of these 786 events, a one meaning the cell has a copy number transition of that type (low to high or high to low) in that genomic region and a zero meaning no copy number transition of that type in that region.

21.12 Neighbor-Joining Tree of Chromosome Breakpoints

We used chromosome breakpoints patterns to build a neighbor-joining tree. To eliminate breakpoints events with a high standard deviation, we limited our analysis to breakpoint regions covering no more than seven adjacent bins (N = 657). Using the city-block (i.e., manhattan or hamming) metric, we calculated a distance matrix from the binary chromosome breakpoint patterns identified in

the 100 single cells using Matlab (Mathworks). From this distance matrix we constructed a tree using the neighbor-joining algorithm. The use of neighbor-joining is justified, in that it is an ultrametic method and does not assume an equal distance from each single cell to the root node, nor a fixed rate of mutation during tumorigenesis.

21.13 Heatmap of Chromosome Breakpoints

The heatmap is based on the same set of breakpoints (N = 657) used to build the neighbor-joining tree. Blue indicates the presence of an event, white means no event. The columns are ordered as in the tree. The rows are ordered to show clearly which of the subsets of the four main groups in the tree share which events. The groups are ordered D, H, AA, AB. A four dimensional binary vector represents each of the 16 possible subsets of these groups (subset vector). Each breakpoint is represented by a four dimensional vector of the percent of cells in each group having an event at that breakpoint (the "breakpoint vector"). The angle from each breakpoint vector to each subset vector is computed as well as the length of each projection vector. If the length of the projection vector is less than 0.05 the breakpoint vector is assigned to the empty (0,0,0,0) subset, otherwise it is assigned to the subset vector with the smallest angle to the breakpoint vector. The rows are ordered by subset vector in the following order: (1,1,1,1), (0,0,0,1), (0,0,1,0), (0,1,0,0), (1,0,0,0), (0,0,1,1), (0,1,0,1), (1,0,0,1), (0,1,1,0), (1,0,1,0), (1,1,0,0), (0,1,1,1), (1,0,1,1), (1,1,0,1), (1,1,1,0), (0,0,0,0). Within each subset the rows are in descending order by the number of cells in that subset having an event and then in ascending order by the number of cells not in that subset having an event.

References

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., and Clarke, M.F. (2003). Prospective identification of tumorigenic breast cancer cells. Proc Natl Acad Sci USA *100*, 3983-3988.

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O.*, et al.* (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet *41*, 1061-1067.

Allred, D.C., Wu, Y., Mao, S., Nagtegaal, I.D., Lee, S., Perou, C.M., Mohsin, S.K., O'Connell, P., Tsimelzon, A., and Medina, D. (2008). Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. Clin Cancer Res *14*, 370-378.

Aubele, M., Mattis, A., Zitzelsberger, H., Walch, A., Kremer, M., Hutzler, P., Höfler, H., and Werner, M. (1999). Intratumoral heterogeneity in breast carcinoma revealed by laser-microdissection and comparative genomic hybridization. Cancer Genet Cytogenet *110*, 94-102.

Aubele, M., and Werner, M. (1999). Heterogeneity in breast cancer and the problem of relevance of findings. Anal Cell Pathol *19*, 53-58.

Avise, J.C. (1977). Is evolution gradual or rectangular? Evidence from living fishes. Proc Natl Acad Sci U S A *74*, 5083-5087.

Bachtiary, B., Boutros, P.C., Pintilie, M., Shi, W., Bastianutto, C., Li, J.H., Schwock, J., Zhang, W., Penn, L.Z., Jurisica, I.*, et al.* (2006). Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. Clin Cancer Res *12*, 5632-5640.

Benetkiewicz, M., Piotrowski, A., Diaz De Stahl, T., Jankowski, M., Bala, D., Hoffman, J., Srutek, E., Laskowski, R., Zegarski, W., and Dumanski, J.P. (2006). 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.

Bergamaschi, A., Kim, Y.H., Wang, P., Sorlie, T., Hernandez-Boussard, T., Lonning, P.E., Tibshirani, R., Borresen-Dale, A.L., and Pollack, J.R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer *45*, 1033-1040.

Bielas, J.H., Loeb, K.R., Rubin, B.P., True, L.D., and Loeb, L.A. (2006). Human cancers express a mutator phenotype. Proc Natl Acad Sci U S A *103*, 18238-18242.

Bielas, J.H., and Loeb, L.A. (2005). Mutator phenotype in cancer: timing and perspectives. Environ Mol Mutagen *45*, 206-213.

Bilke, S., Chen, Q.R., Westerman, F., Schwab, M., Catchpoole, D., and Khan, J. (2005). Inferring a tumor progression model for neuroblastoma from genomic data. J Clin Oncol *23*, 7322-7331.

Bockmuhl, U., You, X., Pacyna-Gengelbach, M., Arps, H., Draf, W., and Petersen, I. (2004). CGH pattern of esthesioneuroblastoma and their metastases. Brain Pathol *14*, 158-163.

Bonnet, D., and Dick, J.E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat Med *3*, 730-737.

Brown, T.M., and Fee, E. (2006). Rudolf Carl Virchow: medical scientist, social reformer, role model. Am J Public Health *96*, 2104-2105.

Calza, S., Hall, P., Auer, G., Bjohle, J., Klaar, S., Kronenwett, U., Liu, E.T., Miller, L., Ploner, A., Smeds, J.*, et al.* (2006). Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. Breast Cancer Res *8*, R34.

Campbell, L.L., and Polyak, K. (2007). Breast tumor heterogeneity: cancer stem cells or clonal evolution? Cell Cycle *6*, 2332-2338.

Campbell, P.J., Pleasance, E.D., Stephens, P.J., Dicks, E., Rance, R., Goodhead, I., Follows, G.A., Green, A.R., Futreal, P.A., and Stratton, M.R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci USA *105*, 13081-13086.

Carey, L.A., Perou, C.M., Livasy, C.A., Dressler, L.G., Cowan, D., Conway, K., Karaca, G., Troester, M.A., Tse, C.K., Edmiston, S.*, et al.* (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA *295*, 2492-2502.

Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M.J., Zhao, X., Carter, S.L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E.S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods *6*, 99-103.

Chin, K., de Solorzano, C.O., Knowles, D., Jones, A., Chou, W., Rodriguez, E.G., Kuo, W.L., Ljung, B.M., Chew, K., Myambo, K.*, et al.* (2004). In situ analyses of genome instability in breast cancer. Nat Genet *36*, 984-988.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R., Kuo, W.L., Lapuk, A., Neve, R.M., Qian, Z., Ryder, T.*, et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell *10*, 529-541.

Chin, S.F., Wang, Y., Thorne, N.P., Teschendorff, A.E., Pinder, S.E., Vias, M., Naderi, A., Roberts, I., Barbosa-Morais, N.L., Garcia, M.J.*, et al.* (2007). Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. Oncogene *26*, 1959-1970.

Clarke, M.F., Dick, J.E., Dirks, P.B., Eaves, C.J., Jamieson, C.H., Jones, D.L., Visvader, J., Weissman, I.L., and Wahl, G.M. (2006). Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells. Cancer Res *66*, 9339-9344.

Cole, K.A., Krizman, D.B., and Emmert-Buck, M.R. (1999). The genetics of cancer--a 3D model. Nat Genet *21*, 38-41.

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

Corver, W.E., Middeldorp, A., ter Haar, N.T., Jordanova, E.S., van Puijenbroek, M., van Eijk, R., Cornelisse, C.J., Fleuren, G.J., Morreau, H., Oosting, J.*, et al.* (2008). Genome-wide allelic state analysis on flow-sorted tumor fractions provides an accurate measure of chromosomal aberrations. Cancer Res *68*, 10333-10340.

DePinho, R.A., and Polyak, K. (2004). Cancer chromosomes in crisis. Nat Genet *36*, 932-934.

Duelli, D., and Lazebnik, Y. (2003). Cell fusion: a hidden enemy? Cancer Cell *3*, 445-448.

Eldredge N, Gould J. 1972. Punctuated equilibria: An alternative to phyletic gradualism. In Models in paleobiology (ed. TJM Schopf), pp. 82–115. Freeman, San Francisco.

Endoh, Y., Tamura, G., Kato, N., and Motoyama, T. (2001). Apocrine adenosis of the breast: clonal evidence of neoplasia. Histopathology *38*, 221-224.

Farabegoli, F., Santini, D., Ceccarelli, C., Taffurelli, M., Marrano, D., and Baldini, N. (2001). Clone heterogeneity in diploid and aneuploid breast carcinomas as detected by FISH. Cytometry *46*, 50-56.

Fialkow, P.J. (1974). The origin and development of human tumors studied with cell markers. N Engl J Med *291*, 26-35.

Fitzgerald, P.J. (1986). Homogeneity and heterogeneity in pancreas cancer: presence of predominant and minor morphological types and implications. Int J Pancreatol *1*, 91-94.

Fotiadou, P.P., Takahashi, C., Rajabi, H.N., and Ewen, M.E. (2007). Wild-Type NRas and KRas Perform Distinct Functions during Transformation. Molecular and Cellular Biology *27*, 6742-6755.

Fuhrmann, C., Schmidt-Kittler, O., Stoecklein, N.H., Petat-Dutter, K., Vay, C., Bockler, K., Reinhardt, R., Ragg, T., and Klein, C.A. (2008). High-resolution array comparative genomic hybridization of single micrometastatic tumor cells. Nucleic Acids Res *36*, e39.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Geigl, J.B., Obenauf, A.C., Waldispuehl-Geigl, J., Hoffmann, E.M., Auer, M., Hormann, M., Fischer, M., Trajanoski, Z., Schenk, M.A., Baumbusch, L.O*., et al.* (2009). Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays. Nucleic Acids Res *37*, e105.

Giaretti, W., Monaco, R., Pujic, N., Rapallo, A., Nigro, S., and Geido, E. (1996). Intratumor heterogeneity of K-ras2 mutations in colorectal adenocarcinomas: association with degree of DNA aneuploidy. Am J Pathol *149*, 237-245.

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

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C*., et al.* (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol *27*, 182-189.

Gould, S.J., and Eldredge, N. (1993). Punctuated equilibrium comes of age. Nature *366*, 223-227.

Grubor, V., Krasnitz, A., Troge, J.E., Meth, J.L., Lakshmi, B., Kendall, J.T., Yamrom, B., Alex, G., Pai, D., Navin, N.*, et al.* (2009). Novel genomic alterations and clonal evolution in chronic lymphocytic leukemia revealed by representational oligonucleotide microarray analysis (ROMA). Blood *113*, 1294-1303.

Hamilton, G., Yee, K.S., Scrace, S., and O'Neill, E. (2009). ATM regulates a RASSF1A-dependent DNA damage response. Curr Biol *19*, 2020-2025.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.

Healy, J., Thomas, E.E., Schwartz, J.T., and Wigler, M. (2003). Annotating large genomes with exact word matches. Genome Research *13*, 2306-2315.

Heim, S., Mittleman, F. (Eds.), 2009. Cancer Cytogenetics, third ed.John Wiley and Sons ISBN 10:0-470-18179-6.

Heng, H.H., Bremer, S.W., Stevens, J., Ye, K.J., Miller, F., Liu, G., and Ye, C.J. (2006a). Cancer progression by non-clonal chromosome aberrations. J Cell Biochem *98*, 1424-1435.

Heng, H.H., Stevens, J.B., Liu, G., Bremer, S.W., Ye, K.J., Reddy, P.V., Wu, G.S., Wang, Y.A., Tainsky, M.A., and Ye, C.J. (2006b). Stochastic cancer progression driven by non-clonal chromosome aberrations. J Cell Physiol *208*, 461-472.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N.E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V.*, et al.* (2006a). Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome Research *16*, 1465-1479.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N.E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V.*, et al.* (2006b). Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome Res *16*, 1465-1479.

Hicks, J., Muthuswamy, L., Krasnitz, A., Navin, N., Riggs, M., Grubor, V., Esposito, D., Alexander, J., Troge, J., Wigler, M.*, et al.* (2005). High-resolution ROMA CGH and FISH analysis of aneuploid and diploid breast tumors. Cold Spring Harb Symp Quant Biol *70*, 51-63.

Hirsch, F.R., Ottesen, G., Podenphant, J., and Olsen, J. (1983). Tumor heterogeneity in lung cancer based on light microscopic features. A retrospective study of a consecutive series of 200 patients, treated surgically. Virchows Arch A Pathol Anat Histopathol *402*, 147-153.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J.*, et al.* (2007). Genome-wide in situ exon capture for selective resequencing. Nat Genet *39*, 1522-1527.

Hoglund, M., Frigyesi, A., Sall, T., Gisselsson, D., and Mitelman, F. (2005). Statistical behavior of complex cancer karyotypes. Genes Chromosomes Cancer *42*, 327-341.

Hoglund, M., Gisselsson, D., Hansen, G.B., Sall, T., and Mitelman, F. (2002). Multivariate analysis of chromosomal imbalances in breast cancer delineates cytogenetic pathways and reveals complex relationships among imbalances. Cancer Res *62*, 2675-2680.

Hovey, R.M., Chu, L., Balazs, M., DeVries, S., Moore, D., Sauter, G., Carroll, P.R., and Waldman, F.M. (1998). Genetic alterations in primary bladder cancers and their metastases. Cancer Res *58*, 3555-3560.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. Bioinformatics *22*, 1036-1046.

Hu, Z., Fan, C., Oh, D.S., Marron, J.S., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L.*, et al.* (2006). The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics *7*, 96.

Ignatiadis, M., and Sotiriou, C. (2008). Understanding the molecular basis of histologic grade. Pathobiology *75*, 104-111.

Imle, A., Polzer, B., Alexander, S., Klein, C.A., and Friedl, P. (2009). Genomic instability of micronucleated cells revealed by single-cell comparative genomic hybridization. Cytometry A *75*, 562-568.

Israeli, O., Gotlieb, W.H., Friedman, E., Korach, J., Goldman, B., Zeltser, A., Ben-Baruch, G., Rienstein, S., and Aviram-Goldring, A. (2004). Genomic analyses of primary and metastatic serous epithelial ovarian cancer. Cancer Genet Cytogenet *154*, 16-21.

Jiang, J.K., Chen, Y.J., Lin, C.H., Yu, I.T., and Lin, J.K. (2005). Genetic changes and clonality relationship between primary colorectal cancers and their pulmonary metastases--an analysis by comparative genomic hybridization. Genes Chromosomes Cancer *43*, 25-36.

Johann, D.J., Rodriguez-Canales, J., Mukherjee, S., Prieto, D.A., Hanson, J.C., Emmert-Buck, M., and Blonder, J. (2009). Approaching solid tumor heterogeneity on a cellular basis by tissue proteomics using laser capture microdissection and biological mass spectrometry. J Proteome Res *8*, 2310-2318.

Kallioniemi, O.P. (1988). Comparison of fresh and paraffin-embedded tissue as starting material for DNA flow cytometry and evaluation of intratumor heterogeneity. Cytometry *9*, 164-169.

Khalique, L., Ayhan, A., Weale, M.E., Jacobs, I.J., Ramus, S.J., and Gayther, S.A. (2007). Genetic intra-tumour heterogeneity in epithelial ovarian cancer and its implications for molecular diagnosis of tumours. J Pathol *211*, 286-295.

Kim, M.Y., Oskarsson, T., Acharyya, S., Nguyen, D.X., Zhang, X.H., Norton, L., and Massague, J. (2009). Tumor self-seeding by circulating cancer cells. Cell *139*, 1315-1326.

Klein, C.A., Schmidt-Kittler, O., Schardt, J.A., Pantel, K., Speicher, M.R., and Riethmüller, G. (1999). Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. Proc Natl Acad Sci USA *96*, 4494-4499.

Komaki, K., Sano, N., and Tangoku, A. (2006). Problems in histological grading of malignancy and its clinical significance in patients with operable breast cancer. Breast Cancer *13*, 249-253.

Kruger, S., Thorns, C., Bohle, A., and Feller, A.C. (2003). Prognostic significance of a grading system considering tumor heterogeneity in muscle-invasive urothelial carcinoma of the urinary bladder. Int Urol Nephrol *35*, 169-173.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lapidot, T., Sirard, C., Vormoor, J., Murdoch, B., Hoang, T., Caceres-Cortes, J., Minden, M., Paterson, B., Caligiuri, M.A., and Dick, J.E. (1994). A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. Nature *367*, 645-648.

Le Caignec, C., Spits, C., Sermon, K., De Rycke, M., Thienpont, B., Debrock, S., Staessen, C., Moreau, Y., Fryns, J.P., Van Steirteghem, A.*, et al.* (2006). Single-cell chromosomal imbalances detection by array CGH. Nucleic Acids Res *34*, e68.

Li, C., Heidt, D.G., Dalerba, P., Burant, C.F., Zhang, L., Adsay, V., Wicha, M., Clarke, M.F., and Simeone, D.M. (2007). Identification of pancreatic cancer stem cells. Cancer Res *67*, 1030-1037.

Linder, D., and Gartler, S.M. (1965). Glucose-6-phosphate dehydrogenase mosaicism: utilization as a cell marker in the study of leiomyomas. Science *150*, 67-69.

Lips, E.H., van Eijk, R., de Graaf, E.J., Doornebosch, P.G., de Miranda, N.F., Oosting, J., Karsten, T., Eilers, P.H., Tollenaar, R.A., van Wezel, T.*, et al.* (2008). Progression and tumor heterogeneity analysis in early rectal cancer. Clin Cancer Res *14*, 772-781.

Liu, J., Bandyopadhyay, N., Ranka, S., Baudis, M., and Kahveci, T. (2009a). Inferring progression models for CGH data. Bioinformatics *25*, 2208-2215.

Liu, W., Laitinen, S., Khan, S., Vihinen, M., Kowalski, J., Yu, G., Chen, L., Ewing, C.M., Eisenberger, M.A., Carducci, M.A.*, et al.* (2009b). Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nat Med *15*, 559-565.

Loeb, L.A., Springgate, C.F., and Battula, N. (1974). Errors in DNA replication as a basis of malignant changes. Cancer Res *34*, 2311-2321.

Lu, X., and Kang, Y. (2009). Cell fusion as a hidden force in tumor progression. Cancer Res *69*, 8536-8539.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J.*, et al.* (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Research *13*, 2291-2305.

Maley, C.C., Galipeau, P.C., Finley, J.C., Wongsurawat, V.J., Li, X., Sanchez, C.A., Paulson, T.G., Blount, P.L., Risques, R.A., Rabinovitch, P.S.*, et al.* (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet *38*, 468-473.

Marusyk, A., and Polyak, K. (2009). Tumor heterogeneity: Causes and consequences. Biochim Biophys Acta.

Matsumoto, T., Fujii, H., Arakawa, A., Yamasaki, S., Sonoue, H., Hattori, K., Kajiyama, Y., Hirose, S., and Tsurumaru, M. (2004). Loss of heterozygosity analysis shows monoclonal evolution with frequent genetic progression and divergence in esophageal carcinosarcoma. Hum Pathol *35*, 322-327.

Mitelman, F., Johansson, B., Mandahl, N., and Mertens, F. (1997). Clinical significance of cytogenetic findings in solid tumors. Cancer Genet Cytogenet *95*, 1-8.

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

Nafe, R., Glienke, W., Burgemeister, R., Gangnus, R., Haar, B., Pries, A., and Schlote, W. (2004). Regional heterogeneity of EGFR gene amplification and nuclear morphology in glioblastomas. An investigation using laser microdissection and pressure catapulting. Anal Quant Cytol Histol *26*, 65-76.

Navin, N., Grubor, V., Hicks, J., Leibu, E., Thomas, E., Troge, J., Riggs, M., Lundin, P., Månér, S., Sebat, J.*, et al.* (2006). PROBER: oligonucleotide FISH probe design software. Bioinformatics *22*, 2437-2438.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V.*, et al.* (2010). Inferring tumor progression from genomic heterogeneity. Genome Res *20*, 68-80.

Negrini S, Gorgoulis VG, Halazonetis TD (2010) Genomic instability--an evolving hallmark of cancer. Nat Rev Mol Cell Biol 11:220-8.

Noguchi, S., Motomura, K., Inaji, H., Imaoka, S., and Koyama, H. (1992). Clonal analysis of human breast cancer by means of the polymerase chain reaction. Cancer Res *52*, 6594-6597.

Noguchi, S., Motomura, K., Inaji, H., Imaoka, S., and Koyama, H. (1994). Clonal analysis of predominantly intraductal carcinoma and precancerous lesions of the breast by means of polymerase chain reaction. Cancer Res *54*, 1849-1853.

Norton, L. (2008). Cancer stem cells, self-seeding, and decremented exponential growth: theoretical and clinical implications. Breast Dis *29*, 27-36.

Norton, L., and Massague, J. (2006). Is cancer a disease of self-seeding? Nat Med *12*, 875-878.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23-28.

O'Brien, C.A., Pollett, A., Gallinger, S., and Dick, J.E. (2007). A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. Nature *445*, 106-110.

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

Paget, S. (1889). The Distribution of secondary growths in cancer of the breast. Lancet *133*, 571-573

Paik, S., Kim, C., and Wolmark, N. (2008). HER2 status and benefit from adjuvant trastuzumab in breast cancer. N Engl J Med *358*, 1409-1411.

Pandis, N., Jin, Y., Gorunova, L., Petersson, C., Bardi, G., Idvall, I., Johansson, B., Ingvar, C., Mandahl, N., Mitelman, F.*, et al.* (1995). Chromosome analysis of 97 primary breast carcinomas: identification of eight karyotypic subgroups. Genes Chromosomes Cancer *12*, 173-185.

Pantou, D., Rizou, H., Tsarouha, H., Pouli, A., Papanastasiou, K., Stamatellou, M., Trangas, T., Pandis, N., and Bardi, G. (2005). Cytogenetic manifestations of multiple myeloma heterogeneity. Genes Chromosomes Cancer *42*, 44-57.

Pathare, S., Schaffer, A.A., Beerenwinkel, N., and Mahimkar, M. (2009). Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression. Int J Cancer *124*, 2864-2871.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A.*, et al.* (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752.

Pugh, T.J., Delaney, A.D., Farnoud, N., Flibotte, S., Griffith, M., Li, H.I., Qian, H., Farinha, P., Gascoyne, R.D., and Marra, M.A. (2008). Impact of whole genome amplification on analysis of copy number variants. Nucleic Acids Res *36*, e80.

Rachkovsky, M., Sodi, S., Chakraborty, A., Avissar, Y., Bolognia, J., McNiff, J.M., Platt, J., Bermudes, D., and Pawelek, J. (1998). Melanoma x macrophage hybrids with enhanced metastatic potential. Clin Exp Metastasis *16*, 299-312.

Roka, S., Fiegl, M., Zojer, N., Filipits, M., Schuster, R., Steiner, B., Jakesz, R., Huber, H., and Drach, J. (1998). Aneuploidy of chromosome 8 as detected by interphase fluorescence in situ hybridization is a recurrent finding in primary and metastatic breast cancer. Breast Cancer Res Treat *48*, 125-133.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

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

Sawada, M., Azuma, C., Hashimoto, K., Noguchi, S., Ozaki, M., Saji, F., and Tanizawa, O. (1994). Clonal analysis of human gynecologic cancers by means of the polymerase chain reaction. Int J Cancer *58*, 492-496.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M.*, et al.* (2004). Large-scale copy number polymorphism in the human genome. Science *305*, 525-528.

Selvarajah, S., Yoshimoto, M., Ludkovski, O., Park, P.C., Bayani, J., Thorner, P., Maire, G., Squire, J.A., and Zielenska, M. (2008). Genomic signatures of chromosomal instability and osteosarcoma progression detected by high resolution array CGH and interphase FISH. Cytogenet Genome Res *122*, 5-15.

Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J.*, et al.* (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809-813.

Sheldon, P.R. (1987). Parallel gradualistic evolution of Ordovician trilobites. Nature *330*, 561-563.

Shen, Q., and Singh, P. (2004). Identification of a novel SP3 binding site in the promoter of human IGFBP4 gene: role of SP3 and AP-1 in regulating promoter activity in CaCo2 cells. Oncogene *23*, 2454-2464.

Shipitsin, M., Campbell, L.L., Argani, P., Weremowicz, S., Bloushtain-Qimron, N., Yao, J., Nikolskaya, T., Serebryiskaya, T., Beroukhim, R., Hu, M.*, et al.* (2007). Molecular definition of breast tumor heterogeneity. Cancer Cell *11*, 259-273.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.D., van de Rijn, M., Jeffrey, S.S.*, et al.* (2001). Gene expression patterns of carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A *98*, 10869-10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S.*, et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A *100*, 8418-8423.

Talseth-Palmer, B.A., Bowden, N.A., Hill, A., Meldrum, C., and Scott, R.J. (2008). Whole genome amplification and its impact on CGH array profiles. BMC Res Notes *1*, 56.

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

Teixeira, M.R., Pandis, N., Bardi, G., Andersen, J.A., Mandahl, N., Mitelman, F., and Heim, S. (1994). Cytogenetic analysis of multifocal breast carcinomas: detection of karyotypically unrelated clones as well as clonal similarities between tumour foci. Br J Cancer *70*, 922-927.

Teixeira, M.R., Pandis, N., Bardi, G., Andersen, J.A., Mitelman, F., and Heim, S. (1995). Clonal heterogeneity in breast cancer: karyotypic comparisons of multiple intra- and extra-tumorous samples from 3 patients. Int J Cancer *63*, 63-68.

Torres, L., Ribeiro, F.R., Pandis, N., Andersen, J.A., Heim, S., and Teixeira, M.R. (2007). Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. Breast Cancer Res Treat *102*, 143-155.

Trent, J.M. (1985). Cytogenetic and molecular biologic alterations in human breast cancer: a review. Breast Cancer Res Treat *5*, 221-229.

Tsarouha, H., Pandis, N., Bardi, G., Teixeira, M.R., Andersen, J.A., and Heim, S. (1999). Karyotypic evolution in breast carcinomas with i(1)(q10) and der(1;16)(q10;p10) as the primary chromosome abnormality. Cancer Genet Cytogenet *113*, 156-161.

van der Poel, H.G., Oosterhof, G.O., Schaafsma, H.E., Debruyne, F.M., and Schalken, J.A. (1997). Intratumoral nuclear morphologic heterogeneity in prostate cancer. Urology *49*, 652-657.

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

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

Zojer, N., Fiegl, M., Mullauer, L., Chott, A., Roka, S., Ackermann, J., Raderer, M., Kaufmann, H., Reiner, A., Huber, H.*, et al.* (1998). Chromosomal imbalances in primary and metastatic pancreatic carcinoma as detected by interphase cytogenetics: basic findings and clinical aspects. Br J Cancer *77*, 1337-1342.