6

Cancer Genomics

This chapter aims to introduce the reader to an important problem in biology and in our own lives: cancer. Massive efforts have been dedicated to providing large-scale cancer-related molecular data to the scientific community. In this chapter, we provide the reader with some of the background material and concepts necessary for starting work in this area. First, we will walk the reader through a brief history of cancer genetics, from its origins, through the molecular biology revolution and, finally, the modern age of genomics. This history will illustrate the path leading to our current understanding of cancer as a molecular disease caused by mutations. We will also explain the most common types of genomic alterations found in cancer. Then, we will go through some recent examples on how topological data analysis techniques have been used in cancer research: identifying molecular markers associated with patients in breast cancer, distinguishing benign moles from melanomas, and studying the response of various cancers to different drugs.

6.1 A Brief History of Cancer

In March 1953, Carl O. Nordling, a Finnish architect with an inclination to statistical problems, published an article in the *British Journal of Cancer* [386]. He noted a common observation that most of us have unfortunately experienced through our relatives and friends: the incidence of cancer increases with age. Apart from some specific pediatric tumors, cancer is uncommon in children and adolescents (with a typical incidence of $\sim$ 18/100,000 per year), and becomes more common in adults (increasing to $\sim$ 500/100,000 per year) [472].

Nordling plotted the logarithm of cancer-associated death rates versus the logarithm of age using data from males from four different countries: the United States, United Kingdom, France and Norway. Interestingly, the mortality data fitted similar

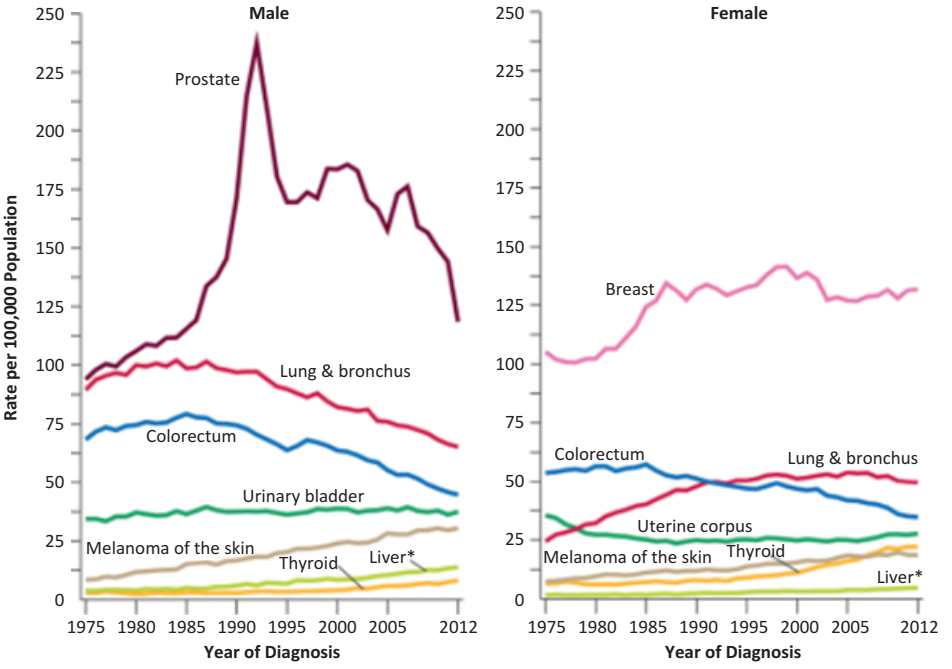

Figure 6.1 Age-adjusted incidence rates of different tumors for males (left) and females (right). Source: Surveillance, Epidemiology, and EndResults (SEER) Program, National Cancer Institute. The large fluctuations in the prostate incidence are due to over-diagnosis associated to changes in diagnosis procedures, including the PSA (Prostate Specific Antigen) test. General trends can be associated with change of habits (such as in lung cancers). The incidence of other tumors, including melanomas, seems to be steadily increasing. Source: [457]. From Rebecca L. Siegel, Kimberly D. Miller, Ahmedin Jemal, *Cancer Statistics, CA: A Cancer Journal for Clinicians*, Volume 66, Issue 1, pp 7-30, Jan 2016 © 2016. Reprinted with Permission from John Wiley and Sons.

straight lines in all four data sets (Figure 6.2), suggesting that the mortality of cancer followed a rule $\sim t^{\alpha}$, where t is the age and $\alpha \sim 6$. Nordling's work was carried out in a time where the main causes of tumors were still unclear, and conflicting evidence suggested both exogenous and endogenous elements contributing to carcinogenesis.

Several lines of work in the beginning of the twentieth century suggested that particular chemical compounds could induce cancers in laboratory animals. Katsusaburō Yamagiwa and Kōichi Ichikawa induced skin carcinomas in rabbits by painting their ears with coal tar [542], showing its carcinogenic properties. In the middle of the twentieth century, chemists were systematically cataloging compounds by their ability to cause tumors in mice. In addition to chemical compounds, it was clear that radiation could also increase the incidence of tumors.

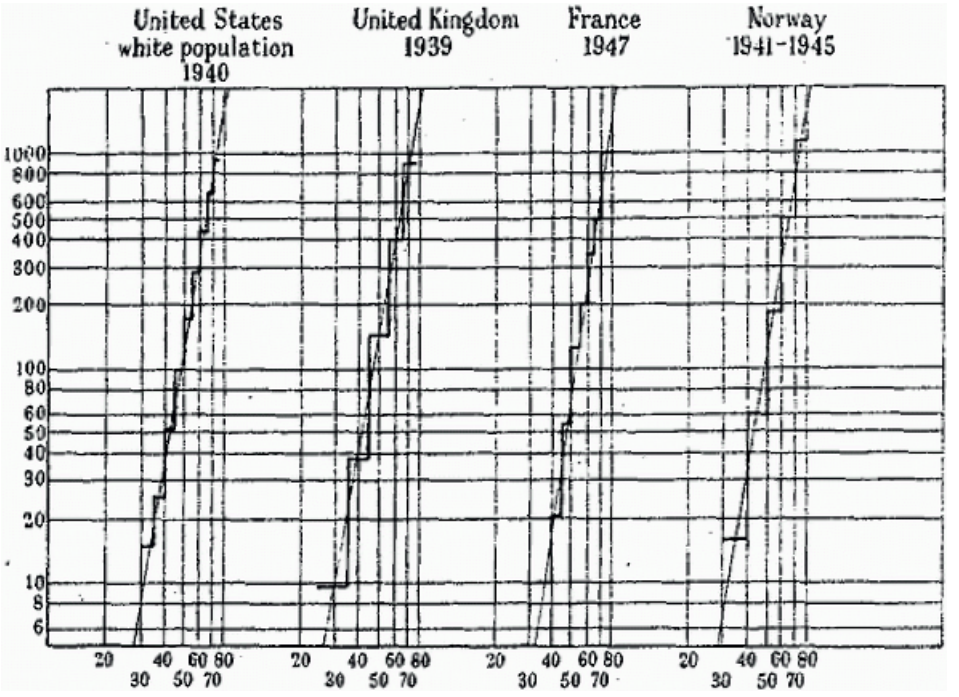

Figure 6.2 Nordling analysis of cancer mortality rates in males from four different countries. Source: [386]. Reprinted by permission from Springer Nature: *Nature, British Journal of Cancer*, A new theory on the cancer-inducing mechanism, C. O. Nordling, 7.1, 1953. © 1953.

The beginning of the atomic era was marked with scientists, health workers, and survivors of nuclear attacks exposed to high doses of radiation, leading to a remarkable increase in the incidence of unusual tumors. How did these diverse classes of agents all lead to cancer?

Theodor Boveri was a German biologist who studied the organization of cellular genomic material into chromosomes. In 1902 he hypothesized that cancers originate from alterations in the genomic material of a single normal cell [68] (see Figure 6.3). In 1927, Hermann J. Muller received a Nobel Prize for showing that X-rays could induce mutations in flies [364] (see Figure 6.4). Before publishing his observations in 1953, Nordling was also aware that "*the original cancerous cell is nothing but an ordinary cell affected by genetic mutation of some kind*." Nordling reasoned that, if a single mutation was sufficient to cause cancer, and there was a constant rate of mutations, then tumor incidence should be independent of age. However, if multiple mutations were needed, the incidence should increase with age. Thus, he proposed a simple model that explained the "universal" $\alpha \sim 6$

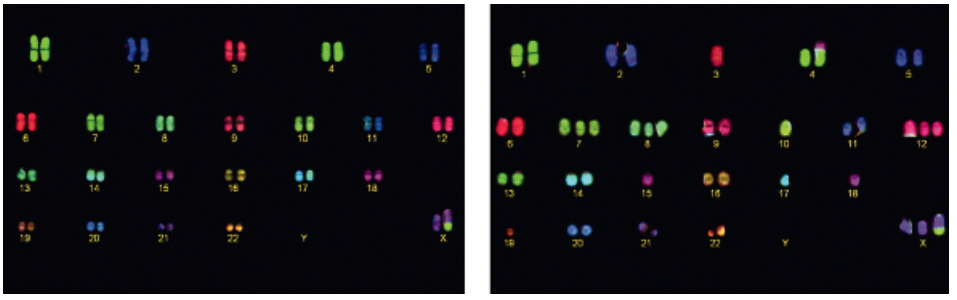

Figure 6.3 Cancer genomes can be unstable. Instead of 23 pairs of similar chromosomes, the chromosomes in cancer cells come in different copy numbers, with amplified and deleted regions. Source: [270]. From Aniek Janssen et al., Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations, Science 333.6051 (2011): 1895-1898. © 2011. Reprinted with permission from AAAS.

Figure 6.4 A few pioneers in the history of cancer genetics. On the left, Theodor Boveri hypothesized that cancer could originate from alterations of the genomic material of normal cell. (The History Collection / Alamy Stock Photo). In the center, Katsusaburo Yamagiwa showed that some chemical compounds could cause tumors in laboratory animals. (Pictorial Press Ltd / Alamy Stock Photo.) On the right, Hermann Muller showed in 1927 that X-rays could cause mutations in flies. (INTERFOTO / Alamy Stock Photo.)

by assuming that tumors were caused by around seven independent mutations. Nordling discussed variations on the universal factor, including hormonal-related tumors in women and childhood tumors. These observations were replicated in greater detail and different mathematical models were proposed in subsequent years [16].

6.2 Cancer in the Era of Molecular Biology

Some of the first breakthroughs in the molecular biology of cancer came through an unexpected route: viruses. As we showed in the previous chapter, viruses are small particles that replicate inside cells. Some viruses replicate and kill the host cell, while others lead to uncontrolled cell proliferation. At the turn of the twentieth century, sporadic observations reported transmissible tumors in animals. It was the work of Peyton Rous in 1909 at Rockefeller University that opened the field of tumor virology. Studying a sarcoma in a hen of "light color and pure blood," Rous showed that it was able to transmit the tumor to other chickens [438]. In 1911, Rous filtered the tumor cells and showed that inoculation of the cell-free filtrate in other chickens was sufficient to recreate the tumor [437]. The Rous sarcoma virus (RSV) was the first definite oncovirus (tumor causing virus) identified, playing a crucial role in the successive developments in the molecular understanding of cancer. Peyton Rous was awarded the Nobel Prize in 1966 for his work, more than 50 years after the original discovery.

This is where the modern history of molecular biology and cancer research starts. The work on RSV was revisited at the end of the 1950s by Renato Dulbecco, Harry Rubin, and Howard Temin. They observed that normal chicken cells could be "transformed" into tumor cells in a Petri dish. The transformed cells resembled tumor cells but could be studied in a simpler environment than a living animal. Although the genomic material of RSV is RNA, Temin showed that the viral genetic material persisted in the transformed cells in the form of DNA [493]. Temin proposed that after RSV infected a cell, it generates DNA that is able to replicate as the cell's normal DNA. The enzyme responsible for the RNA to DNA conversion (reverse transcriptase) was simultaneously identified by Temin and David Baltimore [29, 492] in 1970. Viruses like RSV that carry reverse transcriptase are called *retroviruses*. Dulbecco, Baltimore and Temin shared the Nobel Prize in 1975. (See Figure 6.5.)

The possibility of studying tumors in vitro and the development of new technologies led to a young generation of molecular biologists that expanded on the relation between oncoviruses and cancer. RSV is a small virus that contains only four genes, one gene more than other related, non-transforming retroviruses. This led to the speculation that the extra gene encoded a protein src (for sarcoma), that was somehow responsible for the transformation. The puzzle was finally solved in 1976 by Michael Bishop and Harold Varmus, when they found that homologous copies of the transforming gene were present in different bird genomes [484]. (See Figure 6.6.) The discovery that the genes that were responsible for transforming normal cells were present in untransformed cells was surprising. However, it was not a new idea. Several researchers, including Robert Huebner and George Todaro, had previously suggested that cancer could be the result of the activation of silent

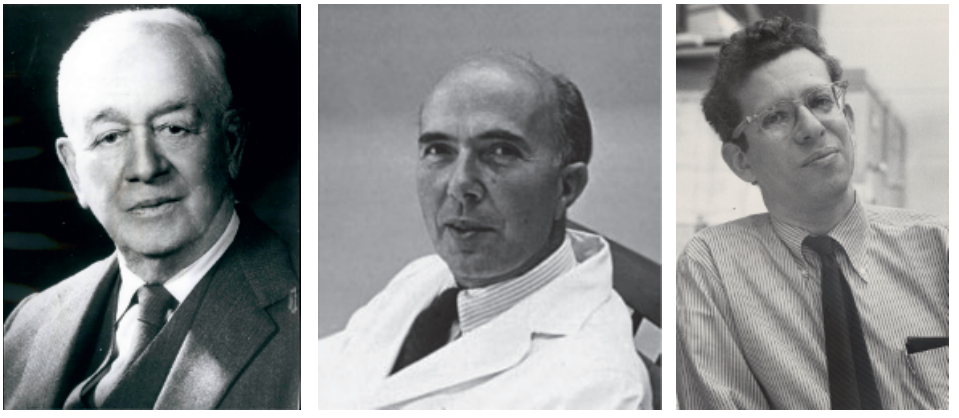

Figure 6.5 The beginning of the understanding of the molecular biology of cancer through oncoviruses. From left to right: Peyton Rous (Keystone Pictures USA / Alamy Stock Photo), Renato Dulbecco (Science History Images / Alamy Stock Photo), and Howard Temin. (Courtesy of the University of Wisconsin-Madison Archives (ID 16555))

genes present in normal cells. These activated genes, or oncogenes, were derived from our own normal genes. The original gene, called the proto-oncogene, is activated through diverse genetic and exogenous mechanisms leading to malignant transformation. The work of Bishop and Varmus established that a normal cellular gene, in this case *c-src* (c- for cell), was the ancestor of the transformed gene found in the RSV oncovirus, *v-src* (v- for virus). This work demonstrated that cancers could be due to alterations in our own genes, and that some viruses, like RSV, could hijack these genes leading to transformation. This discovery was awarded a Nobel Prize in 1989. This was a fascinating discovery that led to the quest for oncogenes in our own cells.

The findings of Bishop and Varmus led to speculation that perhaps other mechanisms could activate oncogenes as well. In 1982 three independent investigators, Robert Weinberg, Michael Wigler and Mariano Barbacid, cloned the first oncogene, *RAS* [203, 398, 444, 456].

A related line of research was carried out on a different type of virus, SV40 (Simian virus 40). SV40 is a double-stranded DNA virus (unlike RSV). The genome of this virus is small, containing only five different genes. Two of these genes, called the T-antigens, are expressed early after infection. SV40 was first identified in primary cells isolated from kidneys of monkeys that were used for the production of the Salk poliovirus vaccine. [1] Although monkey cells die when infected with the virus, occasionally cells from other species (mice and hamsters)

[1] The story of the discovery of this virus is fascinating but outside the scope of this book. We recommend reading chapter 5 of the book of A. Levine on viruses [328].

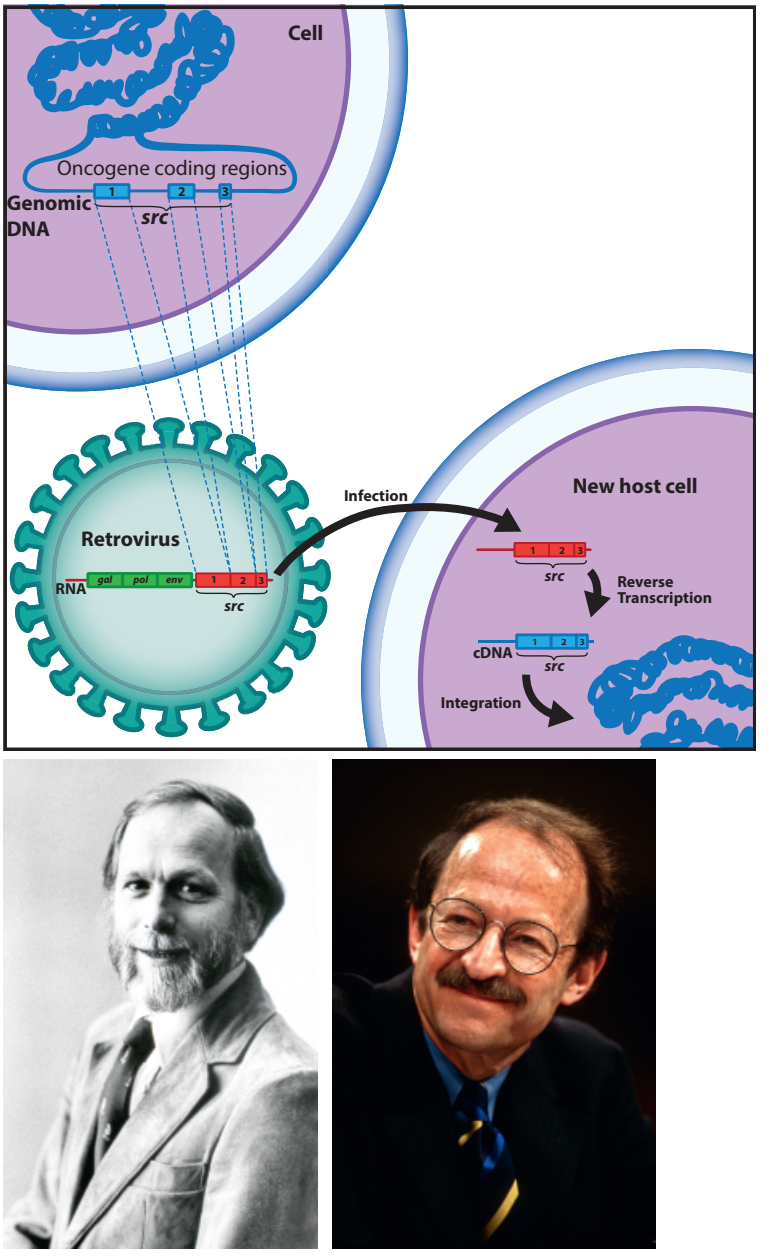

Figure 6.6 Viruses can cause cancer. Retroviruses can take genes from host cells and integrate them into other cells. Some of these host genes can lead to the transformation of infected cells. For the discovery of the cellular origin of retroviral oncogenes, Michael Bishop (left, Science History Images / Alamy Stock Photo) and Harold Varmus (right, Richard Ellis / Alamy Stock Photo) were awarded the Nobel Prize in 1989. Source: Nobel Prize webpage.

are transformed. In the rare case where SV40 genomic material integrates into the host cell genome, it transcribes the T-antigens, generating a cancerous cell. In 1979, Arnold Levine, Lionel Crawford, David P. Lane, and Lloyd Old identified a human 53 kilodalton protein (named p53) bound to the large T protein [139, 313, 331]. In the late 1980s, it became clear that p53 inhibited cell proliferation. It was found that p53 was inactivated in a large fraction of human cancers (near 50%), and that mice with inactivated copies of the *p53* gene grew a variety of tumors. These proteins that become inactivated in tumors are called tumor supressors. These results suggested that tumors can be associated to oncogenes, like *RAS*, becoming activated in tumors, and to tumor suppressor genes, like *p53*, becoming inactivated.

6.3 The Standard Model of Tumor Evolution

We now briefly summarize the standard model of tumor evolution as put forth by Peter Nowell in 1976 [388]. We have many cells in our body, around 4×10^{13}. As time passes and cells divide, mutations accumulate randomly. If any of these cells acquires the right combination of mutations, it could lead to uncontrolled cell proliferation. In this model, all cancer cells are derived from a single cell, the original clone (see Figure 6.7). Chemical carcinogens or radiation could accelerate the processes of tumorigenesis and progression, leading to earlier onset. Some of these mutations could lead to activation of oncogenes or inactivation of tumor suppressors. These mutations confer selective advantage, and subsequent clones replace ancestral ones, accelerating the progression.

Mutations contribute to cancer formation and progression by activating and inactivating key fundamental cellular processes [227]. These altered processes include indefinite replicative potential (cells can replicate without apparently diminishing their replication potential), evasion of programmed cell death (*apoptosis*), stimulation of self-sufficient growth signals (cells can replicate without receiving external inputs), inactivation of antigrowth signals, abnormal metabolism, activation of signals for blood vessel formation (*angiogenesis*), invasion, evasion of immune response, and spread of tumor cells to other organs (*metastasis*).

Mutations accumulated in our bodies throughout our lives are termed *somatic*. These are not all the mutations that could contribute to cancer. Mutations that are inherited (*germline mutations*) could increase the risk of developing certain cancers. For instance, Frederick Li and Joseph Fraumeni characterized families with mutations in *p53*, whose members developed multiple tumors at an early age. Other familial diseases have also helped in the identification of different genes implicated in cancer.

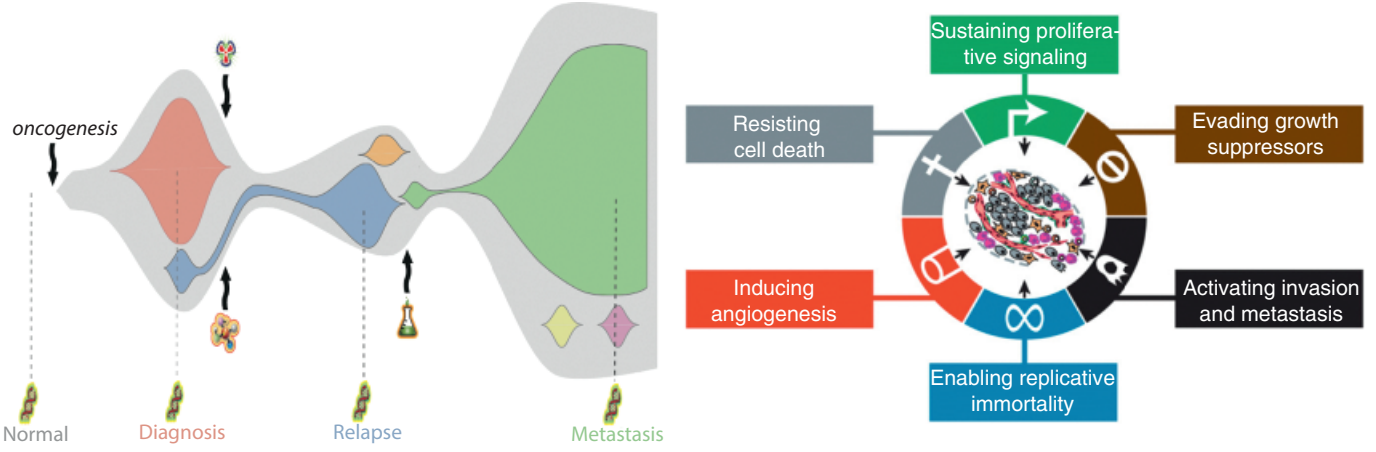

Figure 6.7 The origins and evolution of cancers. Left: Cancer understood as a clonal process. A single clone starts acquiring mutations that lead to uncontrolled growth. The cartoon represents subsequent nested clonal expansions that characterize different phases of tumor evolution (time runs from left to right). We have more than 10 trillion cells in our bodies that accumulate mutations over time. These mutations lead to clonal expansions. Tumors are diagnosed when the number of cancerous cells is sufficient to be detected with current technologies, or when associated clinical symptoms become apparent. Tumors are treated, but other clones could emerge, leading to resistance to therapy and/or metastasis. Source: [546]. Reprinted by permission from Springer Nature: Springer: Zairis S., Khiabanian H., Blumberg A. J., Rabadán R. (2014) Moduli Spaces of Phylogenetic Trees Describing Tumor Evolutionary Patterns. In: Ślęzak D., Tan A. H., Peters J. F., Schwabe L. (eds) *Brain Informatics and Health*. BIH 2014. Lecture Notes in Computer Science, vol 8609. Springer, Cham. Right: Some of these mutations activate and deactivate important key steps necessary for oncogenesis and tumor progression. The figure on the right represents the summary by Hanahan and Weinberg of the main mechanisms that are altered in tumor proliferation. Source: [227]. From Douglas Hanahan and Robert A. Weinberg, Hallmarks of Cancer: The Next Generation, *Cell*, Volume 144, Issue 5, pp 646-674. Reprinted with permission from Elsevier.

Although primitive, the model of clonal evolution through somatic mutations is a first approximation to how tumors evolve. However, it is far from comprehensive, as many other factors are known to play a crucial role, including surrounding cells (microenvironment), epigenetic mechanisms, and immune response.

6.4 Cancer in the Era of Genomic Data

The process of identifying oncogenes and tumor suppressors, so laborious and painstaking in the 1980s, has been changed dramatically by the development of high-throughput sequencing techniques in the first decade of the twenty-first century. Genomic material from samples obtained from tumors could be sequenced and compared to the normal tissue. Studies of cohorts of patients with different tumors started to show the distribution of somatic alterations associated with particular tumors. The process of finding somatic alterations in cancer starts with the collection of tumor samples and matched normal tissue from patients. Figure 6.8 shows a standard procedure for how these mutations are read. The DNA from each sample is extracted, fragmented, and enriched for certain genomic regions of interest. For instance, one could be interested in sequencing some genomic region with an oncogene or a tumor suppressor. Some popular procedures involve the selection of all genomic regions that contain coding genes, also known as the exome. This is called whole-exome sequencing or WES. The capture of the genomic regions

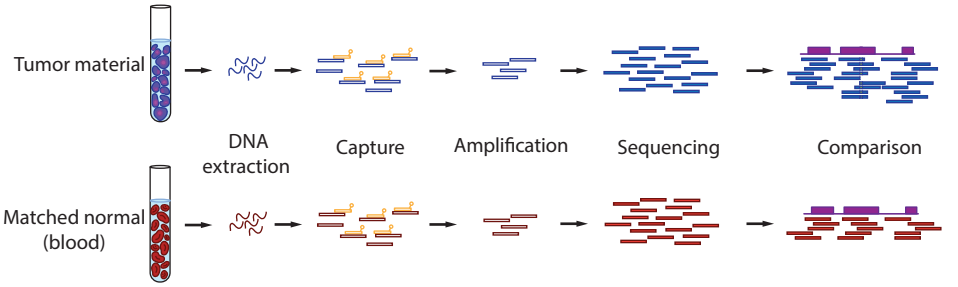

Figure 6.8 High-throughput sequencing allows us to read the somatic alterations in tumor cells. Samples are obtained from tumor and matched normal tissue from the same individual. The DNA is extracted and fragmented. In the case of whole-exome sequencing, different regions are captured using complementary probes. The captured material is then amplified and sequenced. The final results are files containing reads (small sentences of nucleotides) that can be aligned to a reference genome and annotated for diverse variants. The results from the tumor and normal tissue are compared for the identification of diverse alterations present in the tumor tissue but not in the matched control.

of interest is achieved through various techniques usually relying on complementary oligonucleotides. Other protocols aim for whole-genome sequencing (WGS) without enrichment for particular regions. Finally the captured DNA is sequenced, generating a large collection of sequences consisting of the four characters (A, C, G, and T), each roughly 100 nucleotides in length. These sequences are referred to as reads.

The next procedure is the alignment of these reads to a reference (haploid) genome. Sequence alignment is the procedure of finding the best match between two different genomic sequences. Similarity between the two sequences should account for the possibility of one sequence not matching some of the nucleotides in another sequence. In our particular case of interest, reads from the tumor and matched normal DNA are aligned into the human genome. The comparison between the tumor and normal alignments allows the identification of diverse genetic alterations (see Figure 6.9).

For instance, the reads from the tumor sample could report a particular base change that are not present in the matched normal tissue. This is an example of a *somatic point mutation*. The difference could be a small number of bases in the tumor that are not present in the normal (small insertion) or present in the normal but not in the tumor (small deletion). Sometimes, large portions of the genome (from thousands of bases to whole chromosomes) are lost or amplified in the tumor (Figure 6.9). These events can be identified by either an unusually low or high number of reads in the tumor sample, respectively. Other types of alterations such as translocations involve the generation of new genes by the joining of two distant genomic regions. There could also be genomic material that maps into viruses or bacteria that could be present in the tumor sample.

In the sections that follow we will briefly walk the reader through some of the most common somatic alterations in cancer.

6.4.1 Point Mutations

Somatic point mutations are on average the most common somatic alterations across many different tumor types, although there is a large range of variation in the absolute number (see Figure 6.10). Pediatric tumors have in general a lower number of mutations than adult tumors, in concordance with the observations by Nordling in the 1950s [386]. Some tumors like melanoma or lung tumors present large numbers of somatic point mutations associated with the exposure to different carcinogenic environmental factors, like UV radiation or tobacco smoke, respectively. There is a fraction of tumors with mutations inactivating the DNA damage repair pathways that accumulate large numbers of mutations, sometimes referred to as hypermutant tumors.

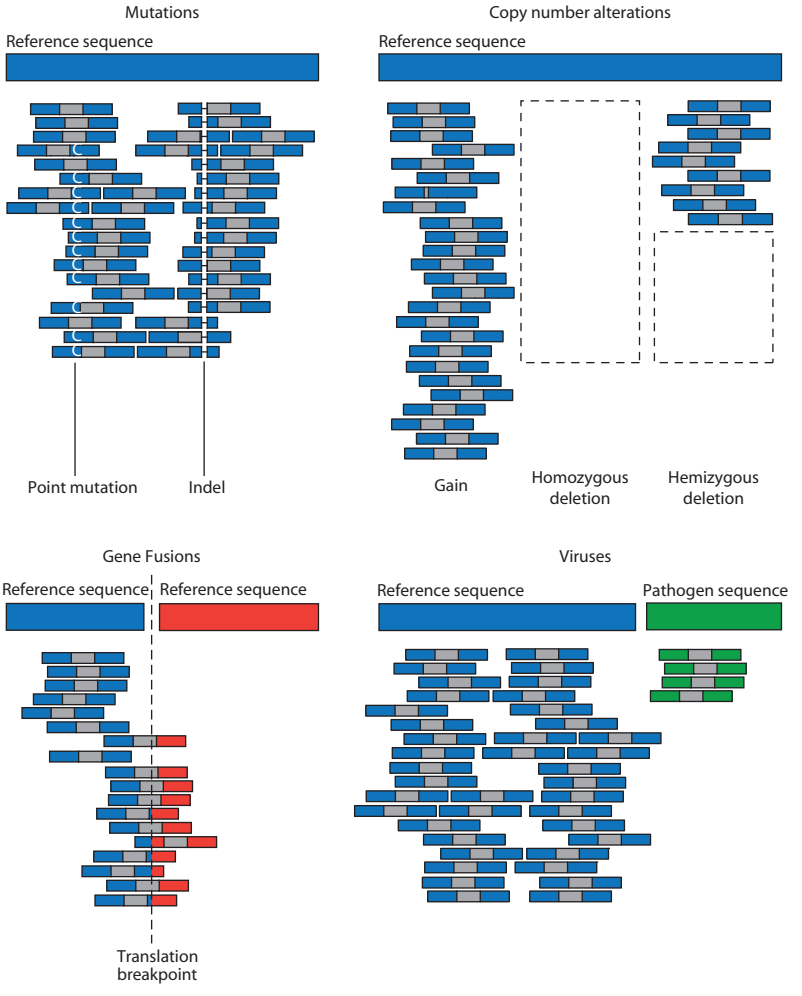

Figure 6.9 Diverse alterations that can be read using genomic technologies. Point mutations are identified when reads aligned to a particular genomic location report a different base than that in the reference genome and the matched normal. Indels, or small insertions and deletions, are reported when the best alignment of a set of reads mapped to a genomic locus skips or inserts a few bases that are not present in the reference and matched controls. Copy number variations can be identified by two complementary methods: statistically significant difference in number of reads in the tumor versus normal, and loss of heterozygosity. A gain (a few extra copies) or amplification (more than 10 extra copies) can be inferred when there are more reads mapped to a particular genomic locus in tumor than normal. Heterozygous (1 copy) or homozygous (both copies) losses occur when there are fewer reads. Loss of heterozygosity (LOH) is the loss of some variants that are heterozygous (50% one allele and 50% the other) in the normal but change the allele frequency in the tumor. Translocations can be identified through pairs of reads mapping to distant locations in reference locations. Finally, there can be reads that do not map to the human genome. These can be the result of the presence of other organisms in the tumor.

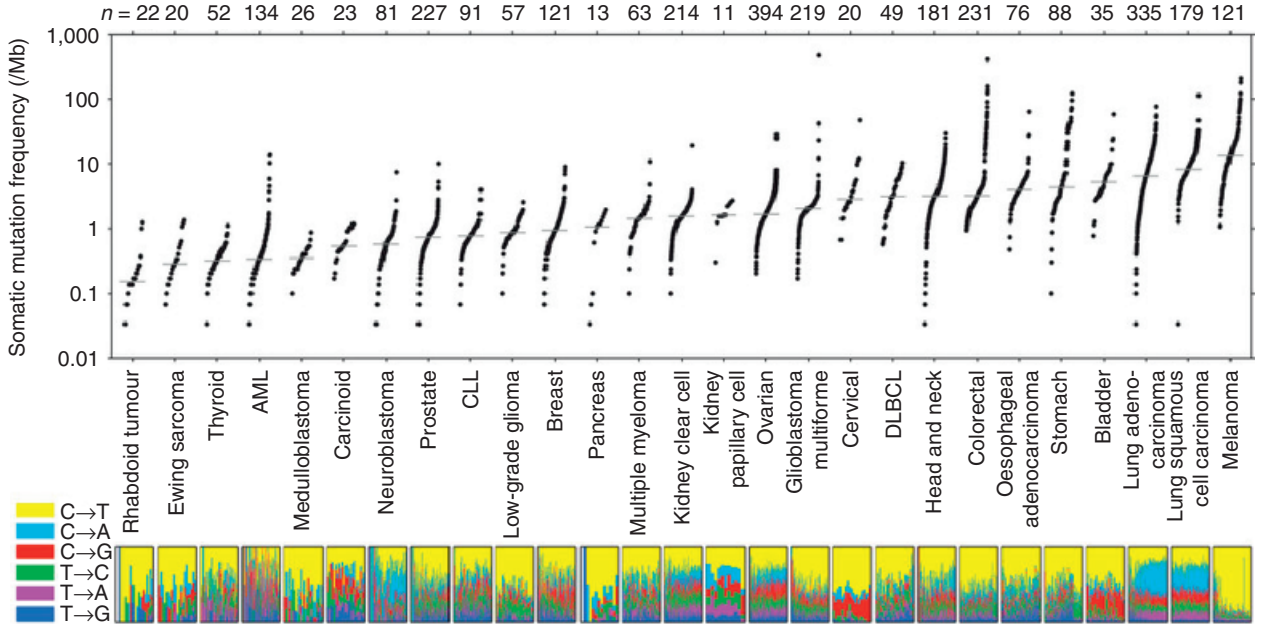

Figure 6.10 This figure (from [316]) represents the rate of somatic mutations per million genome bases across many different tumors. The number of point mutations differs very dramatically across different tumor types. In general, pediatric tumors such as rhabdoid tumors show very few mutations, while some tumors associated to exposure to different carcinogenic environmental factors (such as melanoma and UV radiation) present large numbers of somatic point mutations. Source: [316]. Reprinted by permission from Springer Nature: *Nature*, Mutational heterogeneity in cancer and the search for new cancer-associated genes, Lawrence, Michael S., et al., 499.7457 (2013): 214-218. © 2013.

Case	Genomic Position	Ref/Var	gene	AA	Tumor Frequency	Depth	tf_lower	tf_upper	Normal Frequency	Depth
TCGA-14-3477	chr17:7518257-7518257	G/A	TP53	P250L	98	241	92	99	0	271
TCGA-06-0747	chr7:55221823-55221823	C/T	EGFR	A289A,A289	97	3987	95	97	1	123
TCGA-14-1453	chrX:76760970-76760970	C/T	ATRX	R1803H,R17	96	250	89	99	0	133
TCGA-06-2563	chr7:55189204-55189204	C/T	EGFR	R252C,R252	95	2784	93	96	1	163
TCGA-06-2565	chr7:55178574-55178574	G/A	EGFR	R108K,R108	94	4239	93	95	0	142
TCGA-06-0129	chr17:7518263-7518263	C/T	TP53	R248Q	94	84	78	99	0	84
TCGA-06-6389	chr17:7518915-7518915	T/C	TP53	Y220C	93	30	54	100	0	52
TCGA-12-0618	chr17:7577568-7577568	C/T	TP53	C238Y,C238	90	103	74	99	0	81
TCGA-02-2483	chr17:7517845-7517845	C/T	TP53	R273H	90	59	64	99	0	60
TCGA-14-3477	chr4:54831633-54831633	G/A	PDGFRA	D400N	88	1813	85	91	0	160
TCGA-14-0865	chr4:54834499-54834499	A/G	PDGFRA	N468S	88	541	81	93	0	77
TCGA-06-2570	chr17:7519249-7519249	G/C	TP53	Q136E	88	75	67	99	0	70
TCGA-14-3477	chr4:54831657-54831657	G/C	PDGFRA	E408Q	87	1180	82	90	0	110
TCGA-06-0237	chr7:55210075-55210075	T/G	EGFR	L62R,L62R,L6	85	1282	80	89	0	200
TCGA-06-0221	chrX:76944376-76944376	G/A	ATRX	Q177*,Q139	85	88	63	96	0	119
TCGA-06-0221	chr17:7578512-7578513	TC/-	TP53	TK(140-139)	83	68	55	98	0	78
TCGA-06-1804	chr10:89643813-89643813	G/A	PTEN	G44D	83	16	28	99	0	118
TCGA-14-1456	chrX:76758774-76758774	C/G	ATRX	Q1843H,Q18	76	41	30	96	0	77
TCGA-06-0125	chr7:55200537-55200537	G/T	EGFR	G598V,G598	73	4429	71	76	0	148
TCGA-26-5132	chr13:47851731-47851731	C/T	RB1	R445*	73	44	37	93	0	67
TCGA-02-2483	chrX:76735029-76735029	C/-	ATRX	W2001-,W19	71	42	35	93	0	36
TCGA-12-0670	chr7:55189237-55189237	A/C	EGFR	T263P,T263P	70	3733	67	73	0	589
TCGA-26-5133	chr17:26691654-26691654	G/-	NF1	S2309-,S228	70	82	43	89	0	156
TCGA-12-0619	chr17:7578440-7578440	T/C	TP53	K164E,K164E	70	64	41	89	0	120
TCGA-12-0656	chr1:115058052-115058052	T/A	NRAS	Q61L	69	1151	63	74	0	1361
TCGA-19-2619	chr1:155100785-155100785	G/A	NTRK1	Q46Q,Q76Q,	68	84	42	86	0	66
TCGA-06-2558	chr17:7517822-7517822	C/G	TP53	D281H	65	55	35	87	0	68
TCGA-06-2559	chr10:89682973-89682973	G/T	PTEN	R159S	64	103	41	81	0	137
TCGA-06-2558	chr10:89682884-89682884	C/T	PTEN	R130*	63	129	42	79	0	269
TCGA-06-2561	chr12:25289551-25289551	C/T	KRAS	G12D,G12D	60	105	38	77	0	50
TCGA-12-0707	chr17:26700287-26700287	A/G	NF1	I2405V,I2384	57	21	15	89	0	119
TCGA-12-0656	chr10:89710832-89710832	C/T	PTEN	R335*	56	715	48	64	0	912
TCGA-14-1455	chr4:54825917-54825917	T/C	PDGFRA	C235R	54	3684	50	57	0	385
TCGA-06-5417	chr17:7518988-7518988	G/A	TP53	R196*	54	87	32	72	1	98
TCGA-06-5417	chrX:76665474-76665474	T/C	ATRX	H2254R,H22	51	185	36	65	0	230
TCGA-19-5950	chr11:107677665-107677665	A/G	ATM	Y1753C,Y40S	51	123	33	67	0	134
TCGA-06-5417	chr3:180418785-180418785	G/A	PIK3CA	E545K	48	99	29	67	0	111
TCGA-12-0692	chr13:28597589-28597589	T/A	FLT3	K772N	48	108	29	66	0	39
TCGA-26-1442	chr2:208821357-208821357	C/T	IDH1	R132H	47	55	23	71	0	88
TCGA-06-2570	chr2:208821357-208821357	C/T	IDH1	R132H	46	155	30	60	0	132
TCGA-19-5954	chr3:180390317-180390317	C/T	PIK3CA	R4*	46	50	21	70	0	63
TCGA-06-0128	chr2:208821357-208821357	C/T	IDH1	R132H	42	145	26	58	0	146
TCGA-06-0140	chr13:47849145-47849145	C/T	RB1	Q436*	42	140	26	59	0	141
TCGA-12-0656	chr17:26583330-26583330	T/G	NF1	L1104R,L110	41	292	29	53	0	270
TCGA-76-4925	chr1:155116543-155116543	G/A	NTRK1	K689K,K719	41	247	28	54	0	216
TCGA-12-0616	chr4:55592080-55592080	G/A	KIT	P468P,P468	40	131	24	58	0	106

Figure 6.11 An example of somatic point mutations in glioblastomas from The Cancer Genome Atlas. This is the most common brain tumor type in adults. Each row reports a particular somatic mutation. Different columns annotate the mutations: the patient identifier, the genomic position, the type of mutation, gene, amino acid change, frequency of the mutation in the sample, how many reads were mapped to the location, lower posterior estimates of frequency in tumor, upper posterior estimates of frequency in tumor, frequency of the alteration in the matched normal sample, and number of reads in the normal mapped to the position.

The typical genomic information associated with point mutations can be displayed in a table similar to Figure 6.11. Each row represents a single point mutation. The first column contains the (de-identified) code of a patient. The second column captures the chromosomal location where the somatic mutation is found.

For instance, the first row tells us there is a mutation in position 7,518,257 of chromosome 17. This mutation changes the amino acid 250 of the *TP53* gene from a proline (P) to a leucine (L). We all have two copies of chromosome 17 in our cells; however this particular mutation, despite not occurring in normal cells (0% frequency), was at a frequency of 98% in tumor cells. The columns referred to as depth represent the number of sequence reads that cover that genomic position, a proxy of the relative amount of DNA in a particular genomic locus. A mutation

that occurs in one of the copies of *TP53* and was found in all tumor cells should be present in 50% of reads as there are two copies in each cell. The fact that this number is close to 100% indicates that the non-mutated allele is probably lost (see the next paragraph). This is one of the most common mechanisms of inactivating tumor supressors where both alleles are inactivated through point mutations or deletions.

The second row is also illustrative. It reports a mutation in position 55,221,823 of chromosome 7. This mutation does not change the amino acid where it is located (alanine) but despite being absent in the normal tissue is reported in 97% of the sequence reads covering this particular region. Closer inspection of the depth shows an extraordinary increase from 123 reads to 3987 in tumor, suggesting a 60 fold amplification of the mutated allele. It is known that this genomic region contains an oncogene (*EGFR*) that is frequently amplified in glioblastomas. The synonymous mutation (a mutation that did not change the amino acid) was carried over in the amplification process.

A close inspection of this list reveals several important features. There are several genes recurrently mutated in glioblastomas (*TP53*, *ATRX*, *EGFR*, *PTEN*, *PDGFRA*, among others). Some of these mutations appear amplified (*EGFR* or *PDGFRA*) while others are accompanied with loss of the non-mutated allele (*TP53*). Other genes present inactivating mutations, suggesting that they could function as tumor suppressors (*ATRX*, *RB1*, *PTEN*).

6.4.2 Copy Number Alterations

A second type of somatic alteration, common across many different tumors, is *copy number alterations*. As cells replicate, genomic regions can be lost (deletion) in one allele (hemizygous) or both alleles (homozygous); on the other hand, extra copies of a genomic region can be incorporated (amplification). These alterations can range from a few bases to the whole chromosome. Chromosomal regions that are amplified or deleted that contain oncogenes or tumor suppressors could be selected in the process of tumor development.

Two types of information are usually considered here for the characterization of copy number alterations. The first is the loss or increase of genomic material in tumor sample versus the normal counterpart, indicating a deletion or amplification, respectively. The second is the changes of allele frequencies of heterozygous positions. Let us explain what we mean. Across the genome of our normal cells there are positions that differ between the two chromosomes. For instance, it could be that in one position in our chromosome we observe a C while in the same position in the other chromosome we observe a T. These positions are called heterozygous. If one of the two regions in the chromosomes is lost or amplified, that will generate an imbalance between the two alleles (C or T), one becoming more frequent than

the other. This phenomenon is called loss of heterozygosity (LOH). Copy number losses or amplifications can be identified by changes in the amount of genomic material covering a particular genomic region and a corresponding LOH.

Figure 6.12 shows two examples of these phenomena in chronic lymphocytic leukemia patients. On the left, there is a loss of one of the regions of chromosome 17 (in the p arm) reflected by loss of tumor material in that region and LOH in polymorphic position in this area. On the right, there is an example of a different phenomenon where no significant difference of genomic material is present in chromosome 20, but a significant LOH is observed across a whole chromosomal arm. This is an example of a phenomenon called copy neutral loss of heterozygosity, where one of the genomic regions is lost and the corresponding region of the other chromosome is duplicated.

6.4.3 Gene Fusions and Translocations

Gene fusions are the result of translocations, where two distant genomic regions, within or between chromosomes, are joined together. Translocations can take two different genes and generate a new one containing features from both parental genes. The most famous example of gene translocation is the Philadelphia chromosome, the result of a reciprocal translocation between chromosomes 9 (q34) and 22 (q11); see Figure 6.13. First described in 1960 by Peter Nowell [387], these two chromosomal regions contain two genes, the Abelson murine leukemia viral oncogene homolog 1 (or *ABL1*) (in chromosome 9) and the breakpoint cluster region gene (or *BCR*). *ABL1* is a potent tyrosine kinase whose activity is highly regulated. When fused to the new gene, *BCR-ABL* becomes a potent oncogene with uncontrolled activity. This is a very common translocation in chronic myelogenous leukemia. *BCR-ABL* fusions are the target of imatinib, one of the major successes of targeted cancer therapy [152].

A second type of translocation involves the juxtaposition of a very activating promoter, a region in the genome that controls the transcription of a given gene, next to a potent oncogene without affecting its coding domains. One example of this phenomenon is the *MYC* oncogene, a potent oncogenic transcription factor suspected to be active in more than 50% of all tumors. This activation occurs in a variety of ways, including translocations activating the expression of *MYC* in different B-cell lymphomas, leading to dysregulated expression of a normal protein [131].

Genomic approaches such as the ones described above are mapping the landscape of translocations across many different tumor types. When reads or pairs of reads are partially aligned to two different genomic regions, one can infer that these two regions have been joined in the tumor. Sequencing the RNA instead of the DNA allows one to read which novel expressed transcripts have appeared in

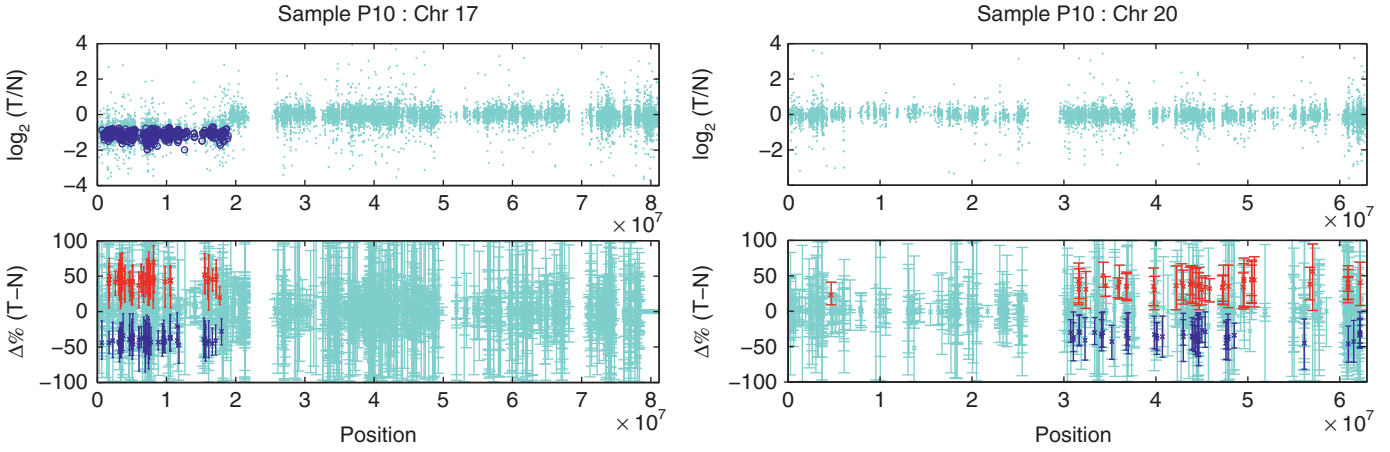

Figure 6.12 Amplifications and losses of chromosomal regions can be read as gains or losses of genomic material across the region and loss of heterozygosity. Top figures represent the logarithm in base 2 of the ratio between the amount of genomic material covering different genomic sections in the tumor versus the normal control. The lower part is an estimate of the allele frequency of the heterozygous position in the tumor samples. Red and blue bars represent posterior estimates of frequencies that differ from the expected 50%. On the left, we have an example of a loss of the p arm of chromosome 17 in a chronic lymphocytic leukemia patient. On the right, we have an example of a copy neutral loss of heterozygosity, where no significant loss or gain of material can be found but there is a systematic LOH in a whole chromosomal arm.

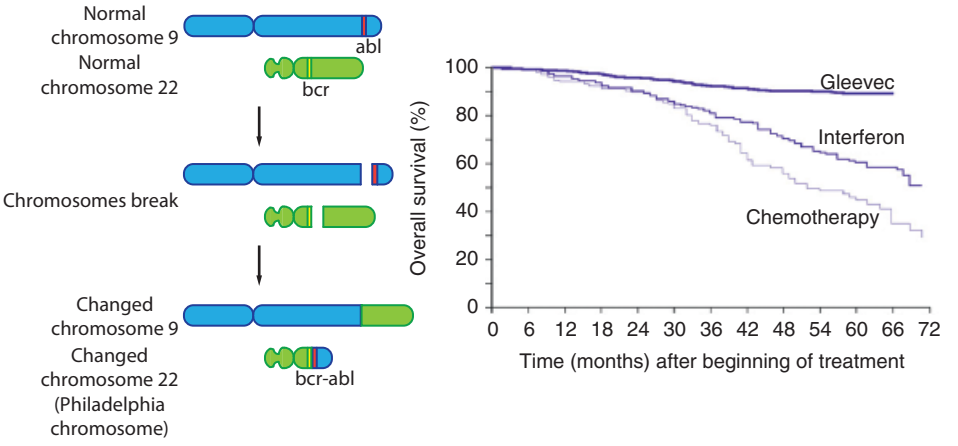

Figure 6.13 Left: The Philadelphia chromosome is the result of a reciprocal translocation between chromosomes 9 and 22 bringing together two genes *BCL* and *ABL1*. The new fusion protein is a potent oncogene and a common alteration in B-cell acute lymphoblastic leukemias, and virtually all chronic myelogenous leukemia. Right: *BCR-ABL* fusions are the target of imatinib (Gleevec or Glivec), one of the most effective targeted therapies in the recent history of cancer. Source: Reprinted by permission from Springer Nature: Springer Nature, *Nature Medicine*, Perspectives on the development of imatinib and the future of cancer research, Brian J. Druker, 15(10): 1149-52 © 2009.

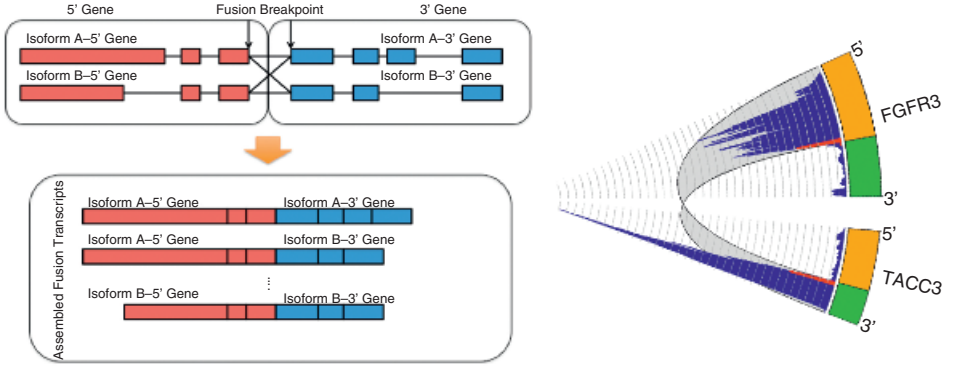

Figure 6.14 Left: Identifying gene fusions by reconstructing novel transcripts. Right: Unbiased characterization of gene fusion events are allowing us to map and catalog many potent oncogenic fusions in different cancers. For instance, *FGFR-TACC* fusions appear in 3% to 5% of glioblastomas.

the tumor (left panel of Figure 6.14). Using this information, one can characterize the most common fusion events across many different tumors. For instance, *FGFR-TACC* fusions [461] are found in nearly 5% of all glioblastomas (right panel of Figure 6.14), but also in many bladder and lung tumors.

6.4.4 Viruses

Somatic mutations accumulate during the entire lifespan of each individual. However, they are not the only important contributors to cancers. The relationship between viruses and cancer has been the subject of a long and tortuous investigation. As we explained in the introduction to this chapter, some of the major discoveries in the understanding of the molecular mechanisms of cancers came through the study of transforming viruses. However, it was not until the 1960s that the relation between some viruses and specific human cancers became evident. The World Health Organization recently estimated that nearly 20% of human cancers are caused by or associated with infections [399]. At present seven viruses have been shown to be strongly associated to human cancers

1. **Epstein-Barr virus** (Human Herpes Virus 4, HHV-4 or EBV). In 1958, Denis P. Burkitt, a surgeon working in Uganda, described a fast growing tumor type affecting the jaws of children with a median age of five years [84]. In 1963 a specimen of this tumor (Burkitt's lymphoma) was sent to London where Michael A. Epstein and Yvonne Barr identified a new virus, now known as the Epstein-Barr virus [165]. EBV is a common virus that for some reason is associated with Burkitt's lymphomas in children from Equatorial Africa but rarely in the rest of the world. Figure 6.15 represents the somatic mutations and virus

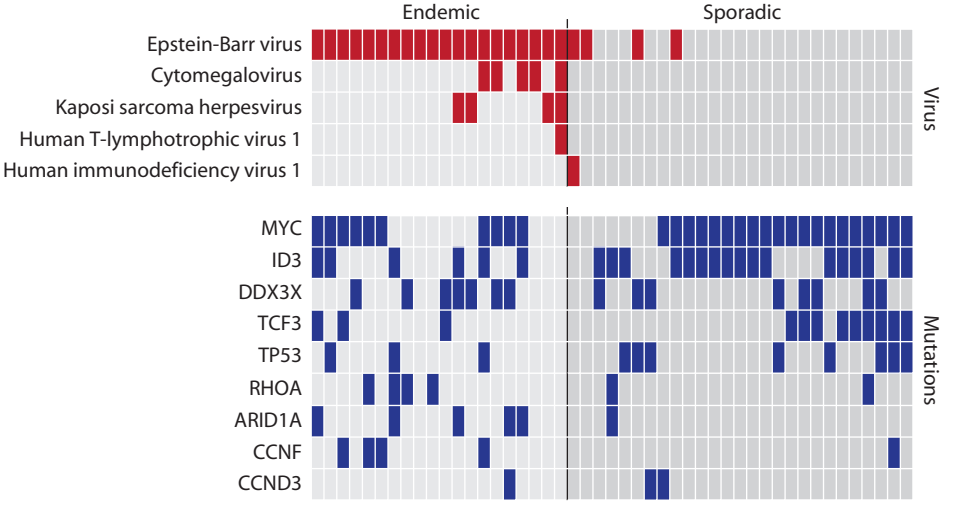

Figure 6.15 Burkitt's lymphomas occur in children in Equatorial Africa and rarely in other parts of the world. Each column is a separate patient; the colored rectangles indicate the presence of particular viruses or mutation. The endemic tumors are always associated with the EBV virus while the association is rare in sporadic cases. In addition to viruses, somatic point mutations occur in key oncogenes and tumor suppressors. Source: [2].

associated with endemic Burkitt's lymphomas from Uganda (left) in comparison to sporadic cases from the United States (right) [2]. Endemic Burkitt's lymphomas are associated with somatic mutations in distinct genes and the EBV virus. Since then, EBV has been found in many other tumors, including nasopharyngeal carcinomas, gastric cancers and specific types of peripheral T-cell lymphomas.

2. **Human T-cell lymphotropic virus type 1** (HTLV1) is a retrovirus that has been associated with a rare type of T-cell lymphoma (adult T-cell lymphoma or ATL), discovered in Japan at the end of the 1970s [243]. The virus is rare in most populations in the world, with higher prevalence in Japan, the Caribbean, and some populations in South America. Transmission is through contaminated blood products or direct mother to child transmission.
3. **Human papillomavirus** (HPV) is a type of non-enveloped double-stranded DNA virus (a Group I virus in Baltimore's classification) with a genome size of around 8000 nucleotides. HPV has been associated to a significant fraction of oral, cervical, vaginal, vulvar, penile and anal cancers. Harald zur Hausen received the Nobel Prize in 2008 for the discovery that human papilloma viruses cause cervical cancer. It has been estimated that more than half a million people get HPV-related cancers every year. Fortunately, the development of the HPV vaccine could prevent the development of these cancers.
4. **Hepatitis B virus** (HBV) is a member of the Hepadnaviridae family of viruses. HBV is one of the smallest enveloped viruses that infect mammals (viral particle of diameter of 40 nm). The genome, of only 3200 nucleotides, is made of partly double-stranded and partly single-stranded DNA. Chronic hepatitis, caused by HCV or HBV infection, can lead to hepatocellular carcinomas. Fortunately, a HBV vaccination has been available since 1981.
5. **Hepatitis C virus** (HCV) is a positive-sense single-stranded RNA virus that we briefly encountered in the previous chapter. Like HBV, chronic infection is associated to high risk of developing liver cancer.
6. **Kaposi's sarcoma-associated herpesvirus** (HHV-8) is a double-stranded DNA virus belonging to the herpeviridae family, with a large genome of 170,000 bases. It is associated to rare lymphoproliferative disorders, such as primary effusion lymphoma, multicentric Castleman's disease, and Kaposi's sarcomas, a rare tumor in immunosuppressed patients. Kaposi's sarcomas were first reported in 1872 by a Hungarian physician working in Vienna, Moritz Kaposi, as a rare condition in some Mediterranean populations. During the AIDS epidemics in the 1980s, near 50% of AIDS patients reported Kaposi's sarcomas. It was identified in 1994 by Yuan Chang and Patrick S. Moore [101].
7. **Merkel cell carcinoma virus** (MCCV) is a small (5400 bases), non-enveloped, double-stranded DNA virus. It was identified in 2008 by the same team that

identified HHV-8 [175]. The integration of the genomic sequence of this virus has been found in 80% of a rare, highly aggressive type of skin cancer, Merkel cell carcinomas.

6.5 Differential Gene Expression Analysis in Cancer

Each cell generates a different number of RNA copies for each gene, and many of these RNAs (the coding RNAs) will be translated into proteins. The RNA expression of a cell, the number of RNA copies of each gene, is correlated then with the amounts of proteins that are being produced. This relation is not linear, as different RNAs can be translated at different rates and proteins can have very different lifetimes. In addition, many RNAs are not translated into proteins; these are the so-called non-coding RNAs. Non-coding RNAs play many functions in the cell, such as regulation of other transcripts or as scaffolds for proteins, but many of these functions remain uncharacterized.

The activation of different pathways in tumors is reflected in changes of expression of different genes. For instance, there are important oncogenes that are transcription factors, such as *MYC*, which activates transcription of a large number of genes involved in the cell cycle. Looking at the RNA expression of transcriptional targets of *MYC* informs us about the activity of this protein. The expression of all the genes in a cell provides extremely useful information about transcriptional programs that are active in the cell.

Gene expression in cancer has been used for many purposes, for instance, to examine which transcriptional programs are activated (like cell proliferation) or inactivated (like apoptosis) in tumors. From the transcriptionally active genes one can infer the activity of different transcription factors. Another standard technique is to study the mechanisms of action of a particular drug or the effect of a mutation by comparing the expression profiles of cells before and after treatment. A different use that we will return to is the classification of patients based on the transcriptional programs of the tumor cells. These studies proceed by collecting RNA from large collections of tumors and classifying patients according to their transcriptional profile. These clusters then can be associated to other phenotypes, such as drug response or survival.

Most of the RNA expression data in tumors considers a population of cells (bulk), including tumor cells and surrounding cells (microenvironment) that are associated with the tumor. The measured total transcription abundance is then the sum of the transcription abundance of the cells of the sample, which may be a complex mixture. Many tumors present cells that are different at the genomic,

transcriptomic or cellular level; this is referred to as tumor heterogeneity. In addition, stromal cells, the cells surrounding the tumor, may be quite diverse. Cells also vary their transcriptional state over time, as they differentiate, replicate, or engage with the environment. In sum, bulk transcriptional data reflects a mixture of cells and transcriptional programs. Single cell sequencing techniques are now used to disentangle these effects. In the next chapter, we will introduce the reader to several of these techniques and some examples of their use in the context of cancer.

6.6 The Space of Glioblastomas

As tumors evolve, their genomes accumulate mutations. Sequencing tumors provides a way of understanding the molecular mechanisms driving this process, as well as the mechanisms of resistance to therapies and potential therapeutic alternatives tailored to the genetic background of specific tumors. Glioblastoma (GBM) is one of the most common and most aggressive types of brain tumors. Median survival after initial diagnosis is little bit more than a year. The standard of care consists of surgery followed by radiotherapy and an alkylating agent, temozolomide (TMZ). Tumors invariably recur, leading to a fatal outcome. How these tumors evolve, the effect of the therapies, and the mechanism of relapse in these tumors is unclear.

To study how GBM evolves, Wang et al. sequenced longitudinal tumor samples in 114 GBM patients, both at relapse and at diagnosis [520]. They also sequenced tumor-matched normal samples. Comparison of the mutational profile in the three samples provides mutations that are in common (founder mutations), those specific to diagnosis and those specific to recurrence. Mutations that are specific to diagnosis could be associated to sensitivity to therapy, and mutations associated to recurrence could inform us about the mechanisms of resistance. From each of the 114 samples, we have three numbers corresponding to the three different types of mutations. From each triplet one can draw a simple phylogenetic tree of three branches: one representing the mutations in common, another branch representing the mutations specific to diagnosis, and the final branch representing the mutations acquired in the recurrent tumor. The mutational story of each patient is then represented by a tree, and the genomic information of the 114 patients is a forest (Figure 6.16). Now the question of how tumors in different patients evolve can be understood in terms of the metric spaces of phylogenetic trees described in Section 4.7.

We can now show the forest representing our data project as points in $\mathbb{P}\Sigma_3$ (Figure 6.16), as described in Section 4.7.2. Here, the upper corner represents the

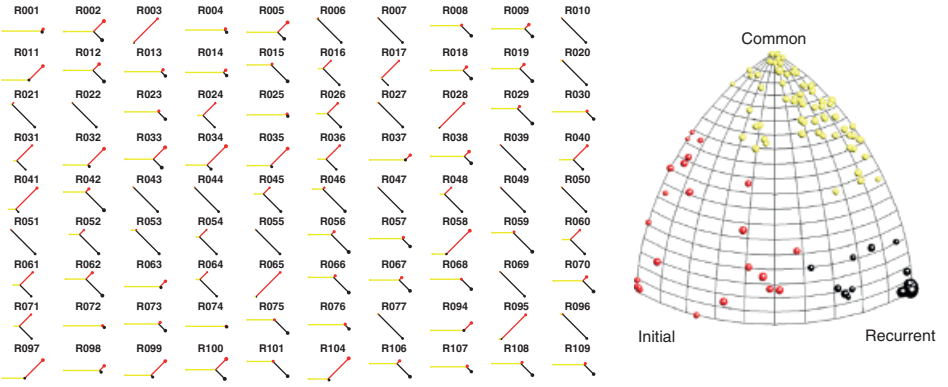

Figure 6.16 Genomic information of tumors from a patient generates a tree representing the different phases of tumor evolution. Cohorts of patients can be represented by a forest (left). In yellow are mutations that are in common in diagnosis and relapse, in red the ones that are specific to diagnosis and in black those that are specific to relapse. The forest can be mapped to points in evolutionary moduli spaces (right). Machine learning and statistical techniques can be applied to classify patient histories and to associate different mutational profiles or clinical outcomes. Source: [520]. Reprinted with permission from Springer Nature: Wang, Jiguang, et al. "Clonal evolution of glioblastoma under therapy." Nature Genetics 48.7 (2016): 768-776.

fraction of mutations that are common to both samples, the left corner represents the fraction exclusive to the untreated sample, and the right corner represents the fraction exclusive to recurrence.

If we want to see if there are different patterns of how tumors evolve in different patients, we can perform clustering. We describe the application of three clustering algorithms to this metric space: k-means clustering, spectral clustering, and density-based spatial clustering (DBSCAN). In order to ensure stability of the results, they were cross-validated using Monte Carlo simulations. Unsupervised clustering of the different phylogenies identifies three clusters. The yellow group represents the limiting case where few mutations are lost from diagnosis. This is similar to the classical model of linear tumor evolution, where mutations accumulate in clones that drive recurrence. The abundance of points far from the right edge of the diagram suggests that in most patients, the dominant clones prior to treatment appear to be replaced by new clones that do not share many of the same mutations. If many mutations in the initial sample are lost at recurrence, this suggests that the clone dominant at recurrence originated (i.e., diverged from the clone dominant at diagnosis) a relatively long time before the initial sample was taken. This is an interesting finding as it suggests that a different clone to the one that caused the initial tumor is responsible for the recurrent tumor.

Interestingly, patient histories identified in the black cluster correspond to particular trees with very long branches associated to relapse tumors. These long branches, associated with a phenomenon called hypermutation, were present only in patients treated with TMZ, and patients in this cluster are associated with longer survival (more than two years). All these patients harbor mutations in the mismatch repair pathway, mostly inactivating mutations in *MSH6*. The MSH6 protein plays an essential role in repairing damaged DNA, by fixing potential mistakes in the replication of DNA. These tumors cannot effectively repair the damage caused by the therapy (TMZ), accumulating many more mutations in branches of the tree associated to the relapse. These mutations are also different from the other mutations. Hypermutated recurrent tumors are highly enriched with C to T (and G to A) transitions, occur in a CC/GG motif, and are associated with the expression of the hypermutated genes.

6.7 Cross-Sectional Data in Cancer and Patient Stratification Using Expression Data

One important question both from the basic biology and clinical points of view is how molecular data, mutations, fusions, and expression can classify patients into different subtypes. From the basic biology point of view, this is important because common molecular features could tell us about the molecular patterns that are associated to particular patients. These molecular patterns can in turn reveal the specific pathways that are activated or deactivated in tumors, and the specific alterations that are related to these pathways. From the clinical point of view, the problem of patient stratification, or classifying patients into different sets, is an extremely relevant one, as molecular data can tell us whether patients could be sensitive or resistant to a particular therapy, and what molecular features are associated to progression or metastasis. Clinical questions can be translated into a problem of understanding the shape or structure of molecular data associated to a large cohort of patients. Information on many patients is usually referred to as cross-sectional data. The "dual or transpose problem," looking at different genes in the space of patients, is probably the more interesting biologically, as genes differentially regulated in a group of patients reveal common deregulated pathways (Figure 6.17).

All of the molecular data discussed in the previous sections could be used to classify patients or genes. The data can be discrete and sparse, such as somatic point mutations, with a binary value that indicates if the mutations are present or not. Expression data, in contrast, is a very rich source of information as it associates each transcript with a real value that corresponds to the copies of mRNA present

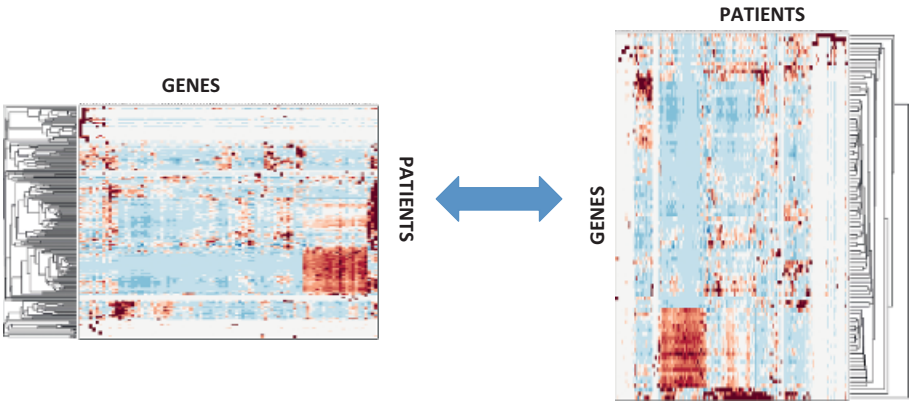

Figure 6.17 Hierarchical clustering of expression point cloud data corresponding to peripheral T-cell lymphoma patients. The data can be seen as two "dual" point clouds. In the space of genes each point is a patient and clustering of points corresponds to clustering of patients. On the dual space, the space of patients, each point is a gene and clustering of points corresponds to clustering of genes.

in the tumor. Most of the data corresponds to the ensemble of cells present in the sample and as such represents a complex mixture of expression levels from different tumor cells, and even different types of non-tumor cells that are also present in the sample. The typical structure of expression data is a point in a very high-dimensional real vector space, as there are typically on the order of 22,000 potential transcripts. Each patient represents a point in this space, and the cross-sectional data corresponds to a point cloud. The question of stratification is usually posed as a clustering problem: how many groups of patients are there presenting similar expression profiles?

Expression-based classification of tumors has been a dominant theme for research since the first microarray experiments and there is an extensive literature on the topic [236, 512] that we do not have the space to discuss. All these earlier approaches are in some way or another based on the idea of clustering patients and genes (see Figure 6.18). It could be, however, that the point cloud data does not have a nice cluster structure. Indeed, that is generally the case due to many biological and technical factors. Not every tumor activates or suppresses different pathways with the same strength, resulting in a more continuous structure from suppression to activation. There is also a common phenomenon of non-tumor cells infiltrating the tumor sample. These and other factors contribute to generating large continuous structures that sometimes are not correctly represented by clusters.

In [383], the authors studied the point cloud data associated with 295 breast cancers, given microarray gene expression data and normal breast tissue. Expression data was normalized and represented by Mapper. As we saw in Section 2.8, Mapper represents the cluster structure of the inverse images of a function on the data. In

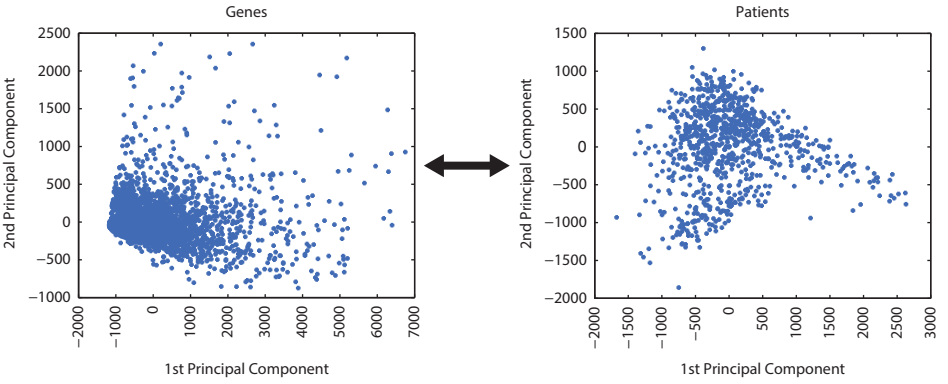

Figure 6.18 Expression point cloud data corresponding to 881 breast cancer patients from The Cancer Genome Atlas Consortium.

this study the function used was provided by a measure of the deviation of the expression data of the tumor samples compared to the expression in normal controls. Clusters of points of overlapping intervals in the image of this function were represented as nodes, and edges corresponded to shared points between different clusters (see top left, Figure 6.19). Blue colors correspond to samples with close similarity to normal tissues (left part of the figure). On the right hand side the samples diverge into two branches. The lower branch, named in the study as c-MYB+ tumors, constitutes 7.5% of the cohort (22 tumors). These tumors are most distinct from the normal tissues and are characterized by the high expression of particular genes, including *c-MYB*, *ER*, *DNALI1* and *C9ORF116*. Hierarchical clustering fails to identify this particular subset of tumors (see bottom left Figure 6.19), and segregates these tumors into separate clusters with low confidence. Interestingly, these tumors do not correspond to a previously reported breast cancer expression subtype. This new class of tumors, c-MYB+ tumors, is characterized by very good survival and no metastasis.

To validate these observations in an independent cohort, we looked at samples with high expression of *DNALI1* and *C9ORF116* (more than a 2-fold overexpression) in 960 breast invasive carcinomas from The Cancer Genome Atlas. Of these 960 tumors, 32 have expression in these genes and show excellent survival (right Figure 6.19), confirming the observation of [383]. The tumors do not contain *TP53* mutations and deletions, and are associated with *GATA3* mutations, suggesting a distinct mutational and expression subtype.

A different approach was taken by L. Seemann, J. Schulman and G. Gunaratne [450]. They looked at the expression data of 202 glioblastoma patients from The Cancer Genome Atlas. Expression-based clustering has identified four groups of glioblastoma expression profiles: classical, mesenchymal, proneural and neural [71, 515]. This partition into four groups has been questioned on several grounds.

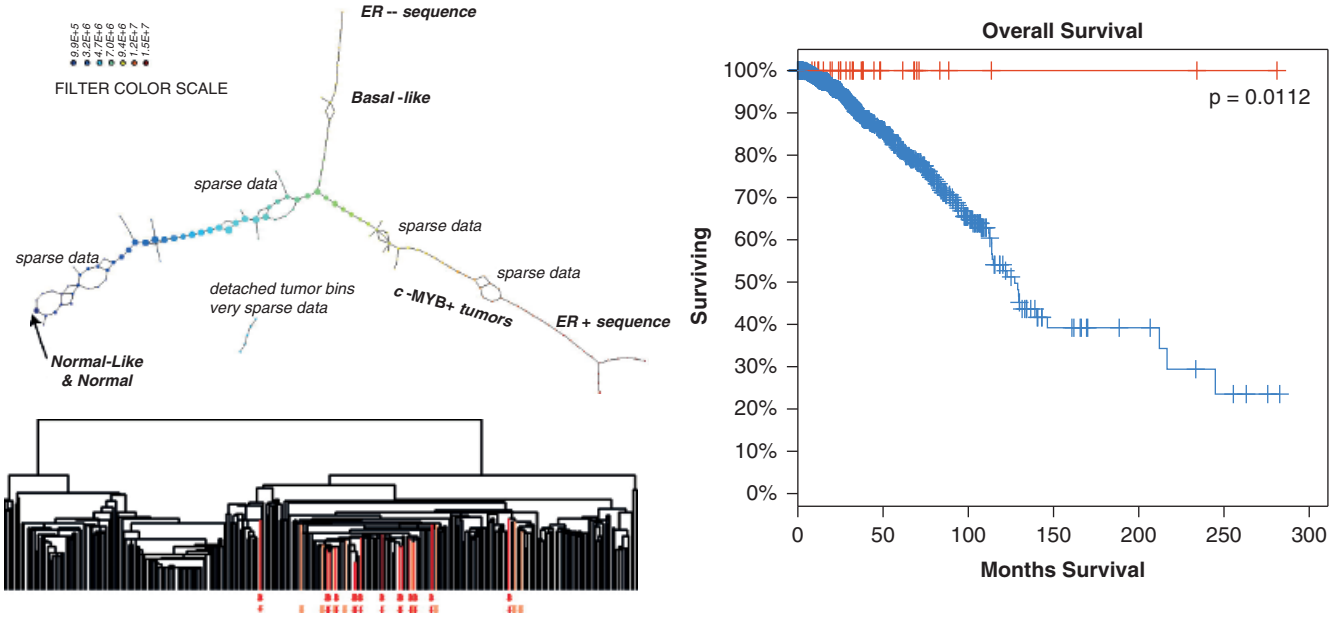

Figure 6.19 The structure of the space of expression of breast tumors. Top left: Mapper representation of the gene expression data from 295 breast tumors. The filter function was a measure of deviation from expression in normal breast tissue controls. Blue colors correspond to samples with close similarity to normal tissues. Source: [383] From Monica Nicolau, Arnold J. Levine, Gunnar Carlsson, *Proceedings of the National Academy of Sciences* Apr 2011, 108 (17) 7265-7270. Reprinted with Permission from Proceedings of the National Academy of Sciences. Tumors with expression profiles significantly different from normal tissue are represented by the two arms on the right hand side. The upper arm is characterized by low expression of the estrogen receptor (ER−). The lower branch contains samples with high expression of c-MYB+. These c-MYB+ tumors cannot be identified using standard clustering (in lower left figure hierarchical clustering split c-MYB+ tumors, represented in red). Independent validation using 960 breast invasive carcinomas from The Cancer Genome Atlas of two of the highest expressed genes in c-MYB+ tumors, DNALI1 and C9ORF116, shows very good prognosis for these tumors.

First, several groups have shown that sampling from different genomic regions does generate different profiles [197, 478]. This heterogeneity has also been seen using single cell expression data [401]. The expression profile also changes over time, before and after therapy [520]. Thus, there are significant concerns about the reliability of classifications of tumors based on expression profiles; such classification might not even make sense in highly heterogeneous tumors. This motivates the search for different approaches that can recover more complex structures.

An approach using persistent homology was explored in [450] to stratify patients is based on a hierarchical partition of samples. A first step was dimensionality reduction. A common problem in all expression-based studies is that the number of genes whose expression is considered is usually much larger than the number of samples, or patients in this case. However, many genes are not expressed, and many others have a similar pattern of variation. This suggests that the dimensionality of the gene space is effectively much lower. Seemann and colleagues propose to reduce first the number of genes. In particular, zeroth dimensional persistent homology was used to cluster genes with similar expression profile. Similarity between different patient profiles was computed using Pearson correlation, or more specifically, $d_{ij} = 1 - \text{corr}(e_i, e_j)$, where e_i is the expression profile of patient i. Using the persistent homology of the associated Vietoris-Rips filtration, 30 genes were identified as characterizing the two long-lived clusters. Patients were then represented by projection onto these genes. The resulting clustering analysis produced novel predictions for genes implicated in GBM. This work represents just the first step towards the use of topological methods for more nuanced classification of tumors.

6.8 Cross-Sectional Data in Cancer and Identifying Driver Genes in Cancer

We have described how different genomic alterations could contribute to cancer formation and progression. We have also shown how we can identify these alterations using genomic technologies. However, not every alteration in a tumor plays a role in its clonal history. For instance, many tumors contain tens of thousands of point mutations, but only a handful of those have a role in oncogenesis and progression (termed the *driver alterations*). How can we identify a few driver alterations within the large background of other irrelevant alterations (the so-called passengers)? Most of the ideas for identifying the most relevant players are based on the concept that, if a gene is found mutated more than would be expected from random variation in a sufficiently large cohort of patients, these alterations have probably been selected along the history of the tumor. Figure 6.20 shows a common representation, known as circos plot, of somatic alterations from 150 glioblastoma exomes. The external annotation refers to different chromosomes. In the interior,

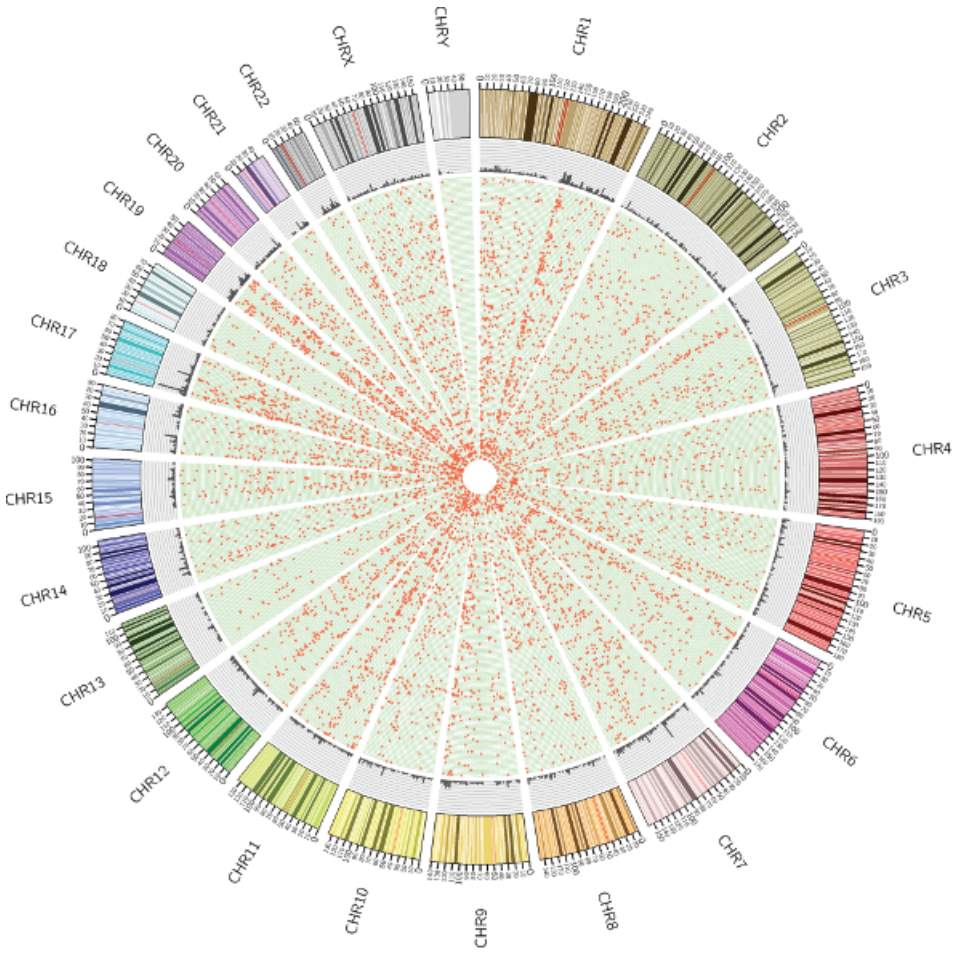

Figure 6.20 A representation of the mutations that occur in 150 glioblastoma patients. Each patient is represented by a concentric circle and the angle represents the chromosomal position where the mutation occurs. Genes frequently mutated across many patients are captured by the histogram in the external part of the representation.

in light green there is information on the 150 tumors represented on concentric circles [183]. Protein changing mutations are represented as red dots. Finally, between the external depiction of chromosomes and the mutations there is a histogram, representing the number of times that a particular gene has been mutated in the cohort. Recurrent mutations occur in chromosome 7, containing an oncogene (*EGFR*), and in chromosome 17, containing a tumor suppressor (*TP53*).

Recurrence-based methods are the standard approach for identifying driver genes in cancer. However, we know that there are alterations that are not very

frequent but could have a strong impact on the tumor development. An alternative approach for identifying genes with a clear phenotypic effect is to look for alterations that share a common phenotypic profile, such as expression. Obviously not every driver gene alteration should have a strong impact in expression, and not every alteration with strong impact in expression is a driver alteration. But the association of the genetic and phenotypic effects is strong indication that the alteration has an effect. Using the expression profiles of many patients and the correlation as a similarity measure, one can apply Mapper, using the distances to the first two nearest neighbors as a filter function. Now, one can see if patients with a particular alteration present a similar expression profile. Figure 6.21 represents expression data from 512 low grade glioma samples.

The Mapper representation finds three distinct groups with strong statistical association with mutated genes.

- A group enriched in *CIC*, *NOTCH1* and *IDH1* mutations.
- A group enriched in *IDH1* and *TP53* mutations.
- A group enriched in *EGFR* and *PTEN* mutations.

These subtypes capture the recent glioma classification [381] into *IDH* wild-type cases, *IDH* mutant with a co-deletion of chromosomal arms 1p and 19q, called *IDH* mutant-codel, and finally patients with mutations in *IDH1* without the co-deletion 1p/19q (non-codel). Codel tumors have a good prognosis and are associated with mutations in *IDH1*, *CIC*, *FUBP1*, and *NOTCH1*. The non-codel tumors harbor mutations in *TP53*. The lower grade gliomas without an *IDH1* mutation clinically resemble high grade gliomas (glioblastomas), and are associated to *EGFR* and *PTEN* deletions.

In summary, by studying the localization of different mutations in the expression space, one is able to identify genes that are associated with specific expression profiles. These expression profiles are associated with different tumor subtypes that have well-defined clinical characteristics.

6.9 The Tissue of Origin of Melanomas

Melanoma is the most aggressive form of skin cancer, with a five-year survival rate of 98% for early-stage tumors compared to 63% and 16% for regional lymph node and distant metastasis, respectively [472]. A strong risk factor for melanoma is UV exposure, which typically causes an abundance of somatic mutations on the order of 10 mutations per megabase [12]. This mutational burden far exceeds that of many other solid tumors and complicates the process of separating the passenger from the driver mutations. Interestingly, it has recently been shown

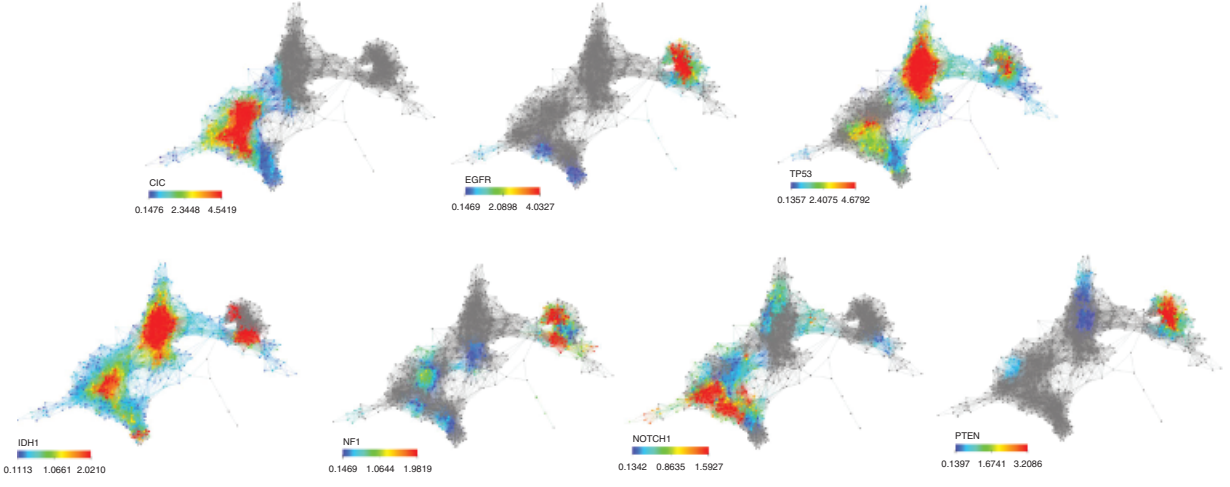

Figure 6.21 Identifying cancer driver genes using Mapper. If certain genes induce a common expression profile across patients, this should be reflected as similarity in the Mapper representation.

that higher mutational loads in melanoma actually lead to greater immunogenicity, better response to immune checkpoint inhibitors like CTLA-4 antibodies, and improved survival [471].

A hallmark feature of melanoma tumorigenesis is the phenomenon of oncogene-induced senescence, in which a typically strong oncogenic mutation – usually *BRAF* or *NRAS* in melanoma – triggers a "fail-safe mechanism" through the p19-p53 or p16-pRb pathways leading to senescence (stop proliferation) rather than cell proliferation. The result is a benign growth called a nevus, known in lay terms as a mole. The most prevalent types of acquired nevi are the common acquired nevus and the dysplastic nevus, both of which have an increased risk of progression to cancer, generally through the loss of tumor suppressor genes like *PTEN*.

A number of genomic studies have explored the mutational and regulatory landscape of metastatic melanoma [245, 514], and The Cancer Genome Atlas (TCGA) has made publicly available an abundance of genomic, transcriptomic, epigenetic, and clinical data for primary and metastatic melanoma. Despite this progress, the key transcriptional events that govern the development from normal skin to nevus to primary and finally metastatic melanoma remain uncharacterized.

The continuous nature of the transformation and the fact that tumor samples are "contaminated" with normal cells of different kinds leads to continuous point cloud data with structures that are difficult to discern with standard clustering techniques. We now describe the results of applying Mapper to transcriptional data from patient-derived samples from melanomas to identify substructures that were predictive of a number of features, including survival, tumor attributes, and gene modules. Comparison of these results to other standard methods of unsupervised exploratory data analysis, including hierarchical clustering and principal component analysis (PCA), is illuminating [334].

The full spectrum of transcriptional changes throughout the progression of melanoma has yet to be fully elucidated. Expression data was collected from 122 punch biopsies of four different subtypes of tissue: 51 primary melanoma (PM), 27 common acquired nevus (CAN), 15 dysplastic nevus (DN), and 29 normal skin (NS). 13 DN were also matched to 13 NS. Initial processing of data included mapping and aligning, calculating counts per transcript using subread, and normalizing using trimmed mean of M-values [434]. An initial approach was to identify differentially expressed genes (DEG) distinguishing PM, CAN, DN, and NS using a standard linear model [469]. This analysis identified the molecular signature of 4862 genes. We then calculated Z-scores derived from the normalized log counts per million for the set of DEG. We next calculated the pairwise distance matrix samples using a distance associated to the correlation ($d = 2(1 - r)$, where r is the Pearson correlation). We then provided this pairwise distance matrix as input for unsupervised analysis with

(1) principal component analysis (PCA),
(2) hierarchical clustering, and
(3) Mapper, using the Euclidean distance as the metric and distances to the first two nereast neighbors as the filter function.

PCA of these genes confirmed general separation of these tissue types (Figure 6.22).

Hierarchical clustering was also performed with the Euclidean distance metric and average linkage (Figure 6.22). This method also demonstrated overall effective clustering between all subtypes except for DN and NS. While 13 DN and 13 NS derived from the same patient, this fact alone did not explain their clustering. A separate analysis showed that DN clustered with NS likely due to a similarly low melanocytic content in both tissue types. Interestingly, two subclusters of PM separated far from the rest of the PM cohort, one low-thickness subcluster that clustered with DN and NS and one high-thickness subcluster.

The Mapper representation provided a richer analysis of the data that shared some features of the PCA and hierarchical clustering stories, but also contrasted in other unexpected ways. We were able to capture a rich topological network that not only demonstrated the separation of tumor subtypes, but also suggested CAN as a further outgroup from the rest of the other subtypes (Figure 6.23).

Interestingly, there were two general subclusters of DN and NS, one to the right and to the bottom of the PM subnetwork. On closer inspection of the members comprising these two subclusters, the right subcluster contained 7 DN and 4 NS, while the bottom subcluster contained 3 DN and 24 NS, suggesting that Mapper was better able to distinguish between DN and NS. Of the 13 matched DN, 4 in the right and 3 in the bottom subcluster were found next to their matched NS. However, 6 of the matched DN were placed in the right subcluster and were separated from their matched NS in the bottom subcluster. These findings suggest that Mapper, to some extent, was able to distinguish the tumor subtype.

Although the Mapper representation was built based on differential expression of subtypes, the resulting structure reflects biological and phenotypic attributes beyond just subtype. Coloring by tumor thickness of PM, we can see that there is a coherent progression of tumor thickness away from the bottom and right DN and NS subclusters where the outer flares approach a higher tumor stage (Figure 6.23).

Recall that hierarchical clustering identified a subcluster of PM that was closer to the DN and NS clade but apart from the rest of the PM. Mapper resolved this inconsistency by readily demonstrating that these lower thickness PM were closer to the rest of the PM subnetwork but were close to the DN and NS subclusters, as well.

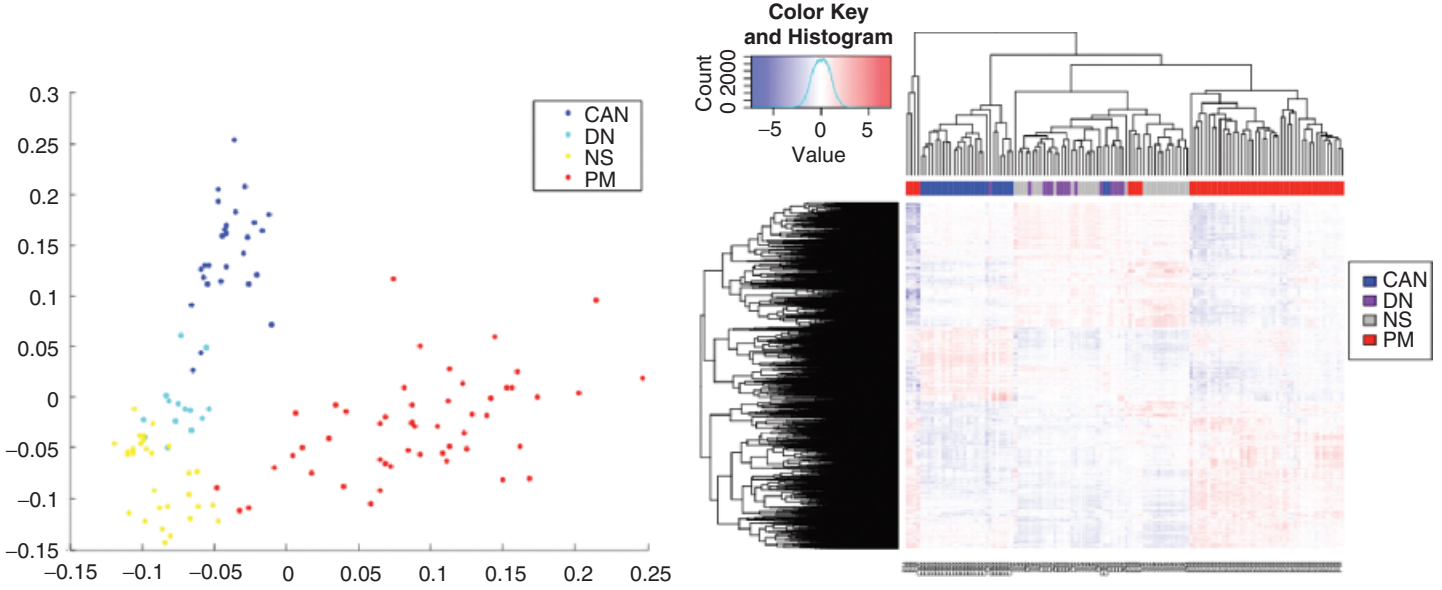

Figure 6.22 Expression analysis of skin, nevi, and melanomas. Left: Principal component analysis of expression of 51 primary melanoma (PM), 27 common acquired nevus (CAN), 15 dysplastic nevus (DN), and 29 normal skin (NS). The analysis shows two flares corresponding to primary melanomas and nevi emanating from normal type core. Right: Hierarchical clustering (Euclidean distance metric with average linkage) artificially separates some primary melanomas and fails to separate normal tissue from dysplastic nevi.

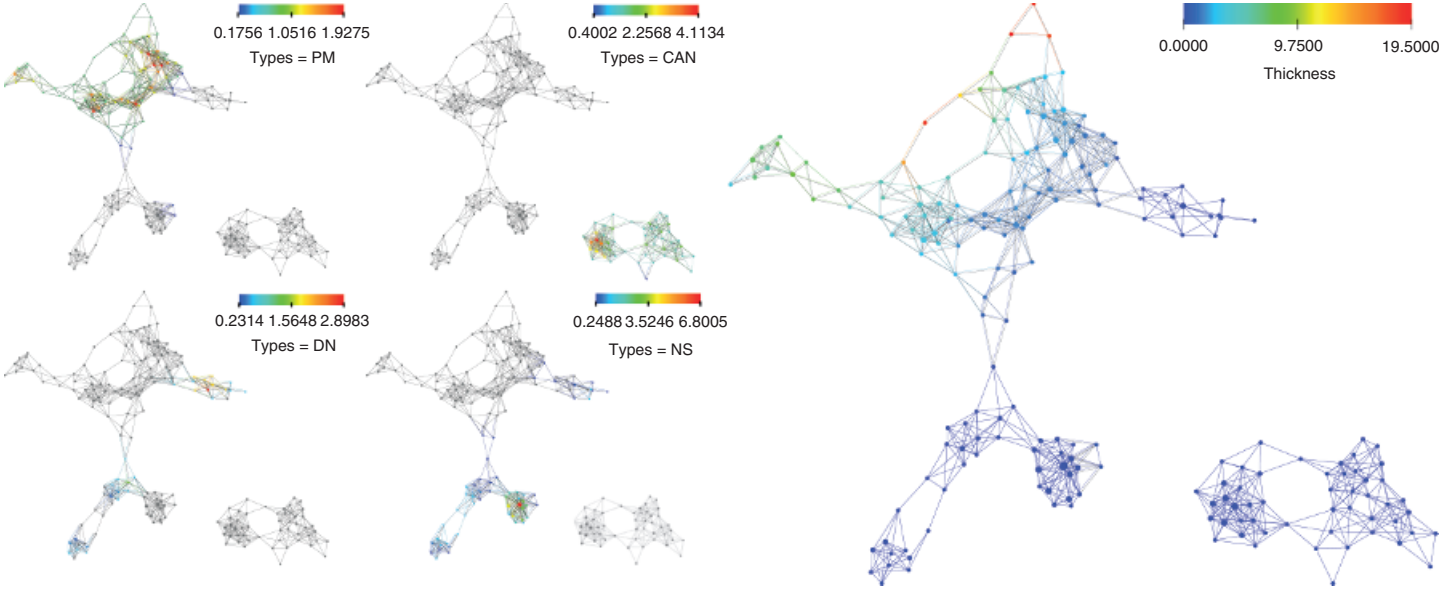

Figure 6.23 Mapper representation of skin, nevi, and melanomas. Distance metric was Euclidean and the filter function was the distances to the first two nearest neighbors. Left: Mapper graphs of tumor subtypes primary melanoma (PM, upper left), common acquired nevus (CAN, upper right), normal skin (NS, lower left), and dysplastic nevus (DN, lower right). Right: Mapper graphs of subtypes colored by tumor thickness of primary melanomas.

The Cancer Genome Atlas (TCGA) also provided a ready source of public RNA sequencing data of both 93 primary (PM) and 352 metastatic tumors (MM). The metastatic tumors included 72 regional cutaneous/in-transit/satellite metastasis (RCM), 215 regional lymph node metastasis (RLNM), and 65 distant metastasis (DM). Similar to the previous analysis, we first applied a standard linear model [469] to identify 695 genes that were differentially expressed between primary and metastatic melanoma. PCA of the resulting DEG data showed some separation of PM away from all subtypes of MM along both the first and second components. Unsurprisingly, subtypes of MM are not clustered separately as DEG analysis was performed with the label MM versus PM as the primary covariate (Figure 6.24).

Hierarchical clustering provided a similar picture as PCA. Three major clades were identified, including a predominantly PM cluster and a predominantly RLNM cluster. The third cluster was heterogeneous (Figure 6.24). Again, Mapper was able to provide a portrait of the shape of the data in richer detail. Not only was Mapper able to separate PM from MM, but it was also able to identify distinct clusters of RCM and RLNM. Interestingly, DM and PM occupied distinct domains of the same cluster (Figure 6.25).

Beyond subtype differentiation, the inherent structure of the topological network reflected underlying biological structure as well. Coloring by time to death after diagnosis as well as by living status at end of study, we identified two subclusters with better survival, one among the PM cohort with greater number of survivors at the end of study and a subgroup within the RLNM cohort with longer time to death among individuals that died at end of study (figure 6.25).

6.10 Association between Drug Sensitivity and Genomic Alterations

In the previous chapters we have described different kinds of somatic alterations and how these alterations stratify patients and define prognostic markers. We have seen that cancers in different patients are the result of different evolutionary histories and environmental exposures. The phenotypic evolution of tumors, their growth, how they metastasize, and how they respond to therapies will depend on these factors. The overall goal of precision cancer approaches is to find ways of linking the genomic and environmental data of a tumor to specific therapies.

We would like to end this chapter with a description of some applications of topological data analysis to understanding how genetic information could be used in connection to drug sensitivity. Methods that model and predict therapeutic sensitivity of cancer can be extremely useful in the development of more effective treatments. Somatic genetic alterations in cancer have been linked with the aberrant

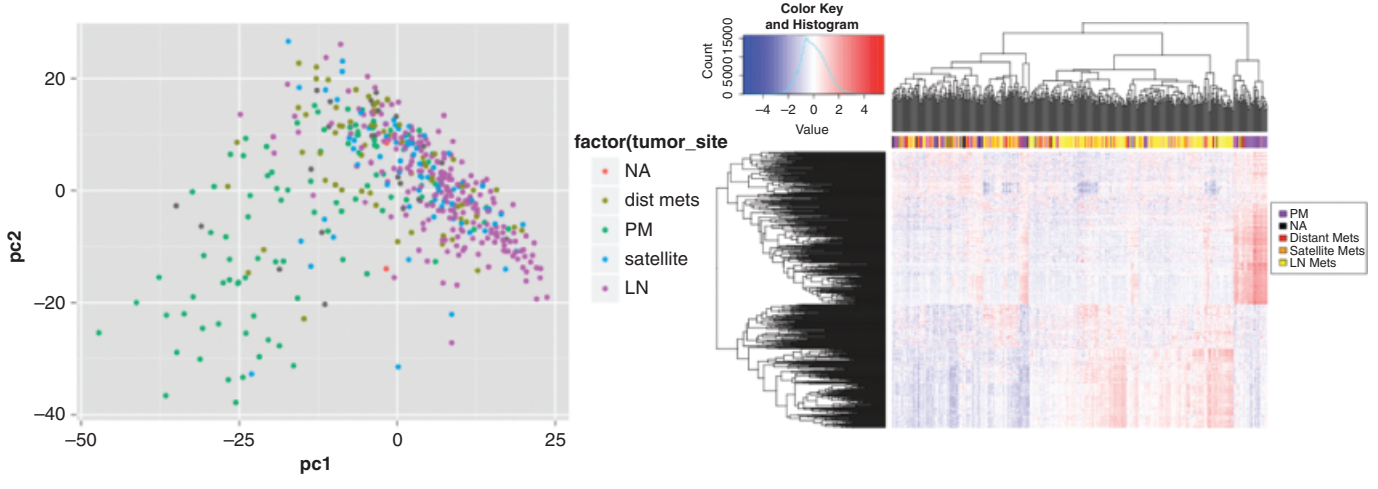

Figure 6.24 Expression analysis of primary and metastatic melanomas. Left: PCA of primary and metastatic melanomas. Right: Heatmap of differentially expressed genes for primary and metastatic melanomas. Hierarchical clustering of samples along the horizontal axis and hierarchical clustering of genes along the vertical axis. Euclidean distance metric with average linkage was used.

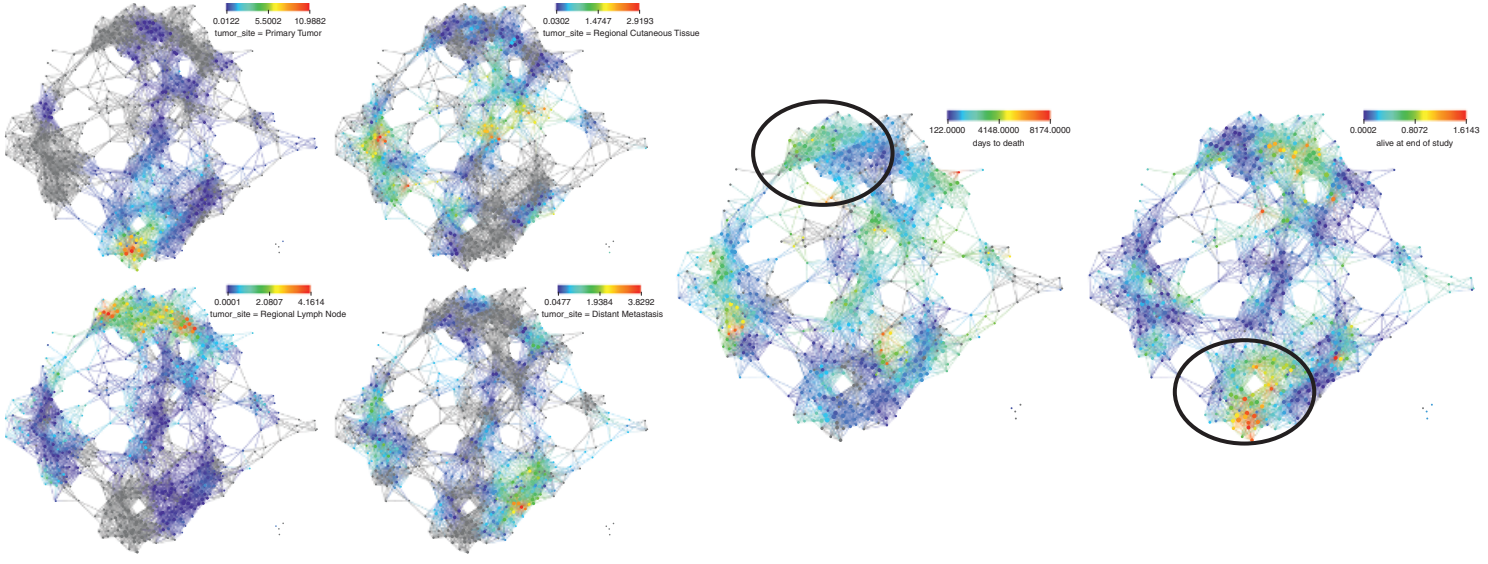

Figure 6.25 Mapper characterization of melanoma expression data from The Cancer Genome Atlas. The distance to the first two nearest neighbors was used as a filter function. Left: The four figures that the physical site location is related to expression in the Mapper graph. Upper left: primary melanoma (PM). Lower left: regional lymph node metastases (RLNM). Upper right: regional cutaneous tissue metastases (RCM). Lower right: distant metastases (DM). Right: Mapper graph of TCGA showing time until death after diagnosis (left) and whether the patient was alive at the end of the study (right). Circled are two subclusters of longer survival.

behavior of signaling pathways, which has led to the development of therapies targeted at these pathways.

While recent advances in sequencing technology make it possible to obtain a wealth of data on the genomic and transcriptomic profiles of specific tumors, our ability to translate this information into improvements in clinical outcomes is limited by our lack of understanding in two areas. First, we do not understand the function of most mutations. Second, we do not know which drugs are best suited for targeting the pathways those mutations affect or participate in. We use Mapper in a computational approach for genome-based drug sensitivity prediction to determine the therapeutic impact of recurrent gene alterations and the role of tumor heterogeneity in drug resistance across a range of cancers.

We have performed initial analysis of the Cancer Cell Line Encyclopedia (CCLE), which contains genomic characterizations of a large panel of cancer cell lines. In the left part of Figure 6.26, we have analyzed expression across cell lines using Mapper. Each node is a set of cancer cell lines that share similar expression profiles, and gene expression is used to construct a similarity metric. The filter function was the map to $\mathbb{R}^2$ specified by the distances to the first two nearest neighbors. Using this approach we first identify the overall network structure of cancer cell lines. Then, we identify specific genes that are characteristic of certain cell lines. For example, we identify that *PDGFRA* is expressed with high specificity in glioblastoma and neuroblastoma.

In the right part of Figure 6.26, we perform an analysis on the same data set, transposed. Here, each node is a gene, and the expression across different cell lines is used to construct a distance metric for the graph. We color genes to show the average expression across central nervous system cell lines. We observe a large cohort of genes centrally expressed across most cell lines. Flares emanating from the main set of genes correspond to genes with unique patterns of expression in particular cell lines. For example, central nervous system specific genes are localized in the flare on the right side of the network. Furthermore, we can localize specific genes within the network to identify neighbors with correlated expression, as we do with *EGFR* in the figure inset.

These networks can be used in a useful manner not only for representing these data sets, but also for predicting drug sensitivity based on genomic alterations. For example, we can perform clustering and feature selection based on these representations of the sample space. Using drug sensitivity information across each cell line from CCLE, specific neighborhoods of sensitivity can be identified. Then, these neighborhoods can be modeled for enrichment of specific alterations. In this way, we incorporate genetic background into the prediction of drug sensitivity. In Figure 6.27, we show a representation in the cell line space, highlighting a common mutation across many tumors in *BRAF*. We showed in the previous section that the

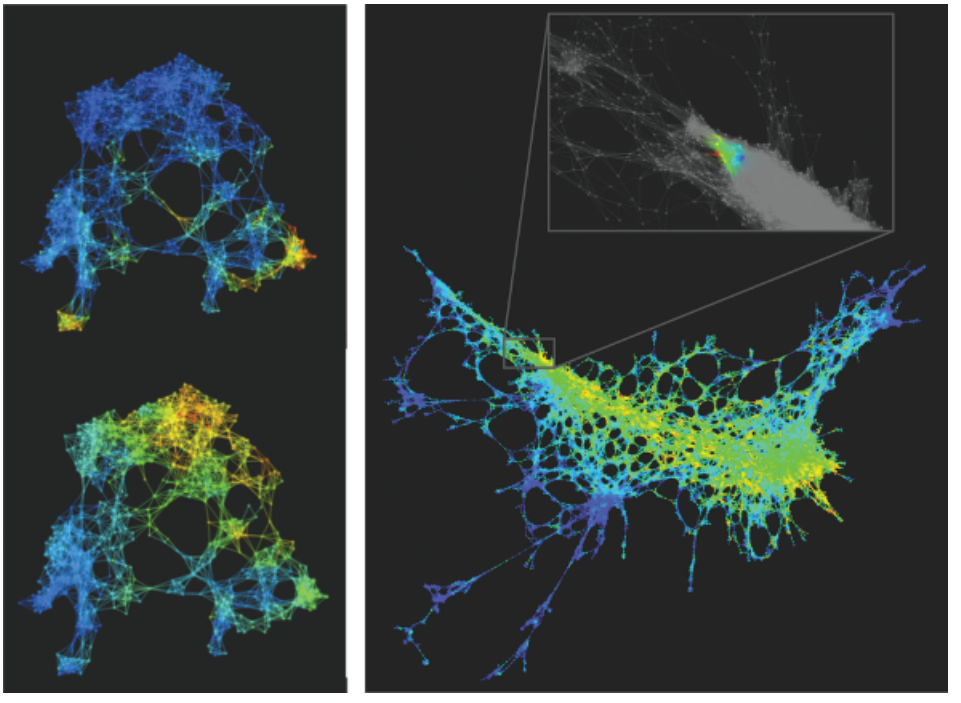

Figure 6.26 Mapper graphs of expression data from the Cancer Cell Line Encyclopedia (CCLE). Left: These images show two distinct patterns of gene expression of targets across a large variety of human cancers. Each node is a set of cell lines sharing a similar expression profile. On the top left, *PDGFRA* is expressed with high specificity in glioblastoma and neuroblastoma (warm colors in the uppermost and lower right nodes). The bottom left image shows the expression of *EGFR* across several tumors (warm colors in nodes in the left portion of the image). Right: Dual representation, where nodes are composed of genes that share expression across CCLE cell lines. Normalized correlation was used to generate the metric, and the filter functions are the first two principal components. Coloring shows average expression in central nervous system cell lines. Inset localizes *EGFR* within the larger network.

BRAF mutation V600E is frequently found in a large fraction of melanomas, but specific point mutations in *BRAF* are also found across many other tumors including 100% of hairy cell leukemias [498], 57% of Langerhans cell histiocytosis [26], and 36% of thyroid papillary cancers [297]. In the representation, we see that *BRAF* mutant cell lines co-localize in a specific region of the Mapper representation. Recently, different drugs have been developed to target specific alterations in *BRAF*. We represent on the right panel of Figure 6.27 the cell lines that are sensitive to PLX4720, a specific V600 mutant *BRAF* inhibitor. PLX4720 is the precursor of PLX4032 (Vemurafenib), a specific *BRAF* inhibitor approved by FDA for treatment of late-stage melanoma.

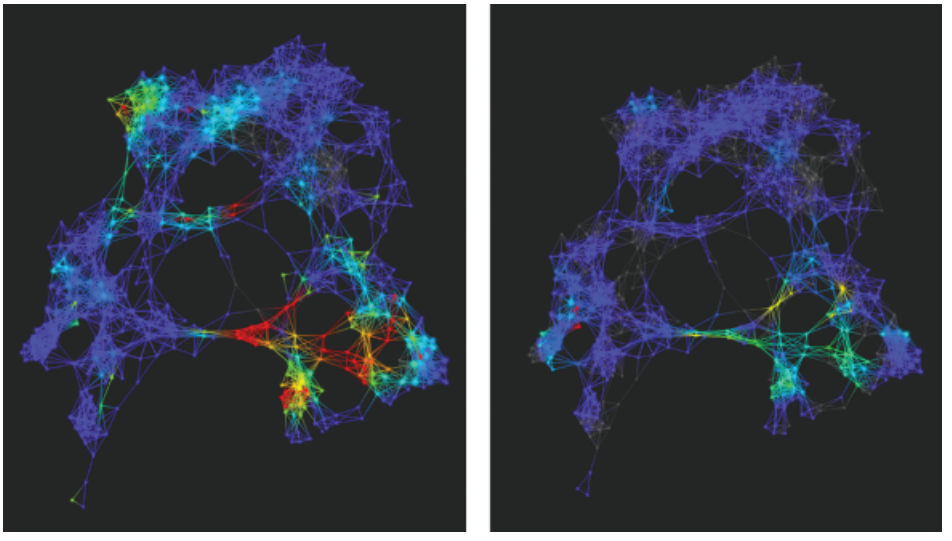

Figure 6.27 Left: Mapper graphs of expression data from the Cancer Cell Line Encyclopedia (CCLE). Nodes are composed of genes, features are the expression across CCLE cell lines. Norm correlation metric with principal component lenses is used. Each node is a set of cell lines with similar expression profile. For representation purposes, the network is then colored based on the expression of specific genes of interest. In this example, coloring shows cell lines with *BRAF* mutations. *BRAF* is a gene mutated in different tumors, including melanomas. Right: Coloring shows cell lines sensitive to PLX4720, a compound with specific action on mutant *BRAF*.

Cellular heterogeneity reflects both clonal heterogeneity and genetic instability; thus, it can be impacted by anticancer therapy on several levels. First, new selective pressures are expected to favor relatively treatment-resistant clonal subpopulations over sensitive ones, therefore limiting clonal diversity. Second, genotoxic treatments may elevate genomic instability, thereby potentially increasing cellular genetic diversity. Despite its clinical importance, the potential impact of cancer therapy on cellular genetic heterogeneity is largely unknown. Topological data analysis (TDA) methods to model phenotypic and genetic determinants of drug resistance *in silico* can generate testable hypotheses for addressing drug resistance. Very recently, the role of clonal heterogeneity in tumors and the impact on therapy has been studied using topological data analysis in [321, 322].

6.11 Summary

- Cancer is the result of mutations in our cells. Nowell first formalized the notion of cancer as an evolutionary process in which genetic instability creates variation that is sieved by natural selection [388]. This instability leads to a menagerie

of somatic mutations, including substitutions, indels, copy number variants, methylation aberrations, translocations, and gene fusions.

- Cancer evolution is largely a clonal process, where an initial tumor cell population proliferates, often through abnormal mitosis, and different cell lineages are created by sequential mutations that confer greater or lesser fitness.
- Fitness is determined by the "hallmarks of cancer": the tumor cell's abilities of evasion of apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals, sustained angiogenesis, limitless replicative potential, and tissue invasion and metastasis [227].
- The mutations at each step of a clonal expansion endow the cancer cell with variable fitness. Indeed, these mutations can be divided into driver lesions that further tumor progression and passenger lesions that are byproducts of the mutagenic environment of the cancer [222]. Driver genes provide potential candidates for oncogene addiction, in which the survival of a tumor cell becomes increasingly dependent on the continuing function of a lesion [536].
- Recent technological developments are generating large scale molecular data from many different tumor types in many patients. This data includes systematic characterization of mutations, expression and methylation profiles, and many other kinds of information. Problems in studying the molecular mechanisms of cancer initiation and progression, prognosis, or sensitivity to therapies can be translated into understanding the structure of these data sets.
- Mapper has been used to study molecular data in cross-sectional studies to understand the transcriptional similarity between the tumors in different patients. These studies classify patients based on transcriptional similarity and identify subsets of patients with differential survival.
- Molecular data can be collected along the progression of a tumor or at different locations in a metastatic process. Mapper has been used to understand the relation between molecular data and physical or temporal information, and progression of the disease.
- Molecular characteristics and responses to drugs vary enormously across patients (patient heterogeneity) and also within a tumor of a single patient (tumor heterogeneity). Integrating molecular and drug response data is one of the main challenges in developing precision therapeutics for cancer patients. Topological data analysis can uncover the structure of these data sets, linking molecular features to drug response characteristics.
- As we will see in the next chapter, single cell technologies are generating molecular data across many cells in different biological systems, including tumors. How these cells are different and how these differences relate to tumor evolution and drug response constitute a fascinating problem that is now beginning

to be explored. The structure of these single cell tumor data sets remains poorly characterized; unsupervised techniques, including topological data analysis, can potentially identify fundamental molecular mechanisms driving tumor initiation progression. As these technologies improve and larger single cell cancer data sets become available, we expect a greater need for these methods.

6.12 Suggestions for Further Reading and Databases

There are a few books that we recommend to the neophyte.

- *The Emperor of all Maladies* [363] is a Pulitzer prize winning book that narrates the history of our understanding and treatment of cancer.
- Robert Weinberg's *The Biology of Cancer* [531], provides a comprehensive overview of cancer research, explaining the main molecular mechanisms that have been identified.
- Arnold Levine's book on viruses [328], is a nice introduction to viruses, including historical accounts on the discovery and mechanisms of oncoviruses.

There are extensive databases that provide a large variety of different data sets associated with multiple cancers.

- The Cancer Genome Atlas constitutes a large US-based effort between the National Cancer Institute and National Human Genome Research Institute to characterize genomic/transcriptomic/epigenetic changes together with clinical annotation in 33 types of cancer. (http://cancergenome.nih.gov)
- The International Cancer Genome Consortium constitutes a worldwide effort to generate a comprehensive genomic, transcriptomic and epigenomic description in 50 different major tumor types. https://dcc.icgc.org.
- The Cancer Cell Line Encyclopedia (CCLE) project provides a detailed genetic characterization of a large panel of human cancer cell lines together with responses to different drugs. www.broadinstitute.org/software/cprg/?q=node/11
- The Genomics of Drug Sensitivity in Cancer (GDSC) is a resource for therapeutic and genomic characterizations of a large panel of cell lines. www.cancerRxgene.org