On the frequency of genome rearrangement events in cancer karyotypes

Michal Ozery-Flato and Ron Shamir

School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel
{ozery,rshamir}@post.tau.ac.il

Abstract. Chromosomal instability is a hallmark of cancer. The results of this instability can be observed in the karyotypes of many cancerous genomes, which often contain a variety of aberrations. In this study we introduce a new approach for analyzing rearrangement events in carcinogenesis. This approach builds on a new effective heuristic for computing a short sequence of rearrangement events that may have led to a given karyotype. We applied this heuristic on over 40,000 karyotypes reported in the scientific literature. Our analysis implies that these karyotypes have evolved predominantly via four principal event types: chromosomes gains and losses, reciprocal translocations, and terminal deletions. We used the frequencies of the reconstructed rearrangement events to measure similarity between karyotypes. Using clustering techniques, we demonstrate that in many cases, rearrangement event frequencies are a meaningful criterion for distinguishing between karyotypes of distinct tumor classes. Further investigations of this kind can provide insight on the scenarios by which particular cancer types have evolved.

1 Introduction

It is well known that many cancerous genomes exhibit abnormal karyotypes. The abnormalities found in these karyotypes include *numerical aberrations*, i.e. changes in chromosome copy number, and *structural aberrations*, i.e. rearrangements within the genome (see Fig. 1). Some of the malignancies, mostly hematological ones, are associated with specific patterns of aberrations. A classical example of such association is between the "Philadelphia chromosome" abberation (a specific translocation between chromosomes 22 and 9) and chronic myelogenous leukemia [17, 19]. This translocation leads to the formation of the oncogene BCR-ABL [5].

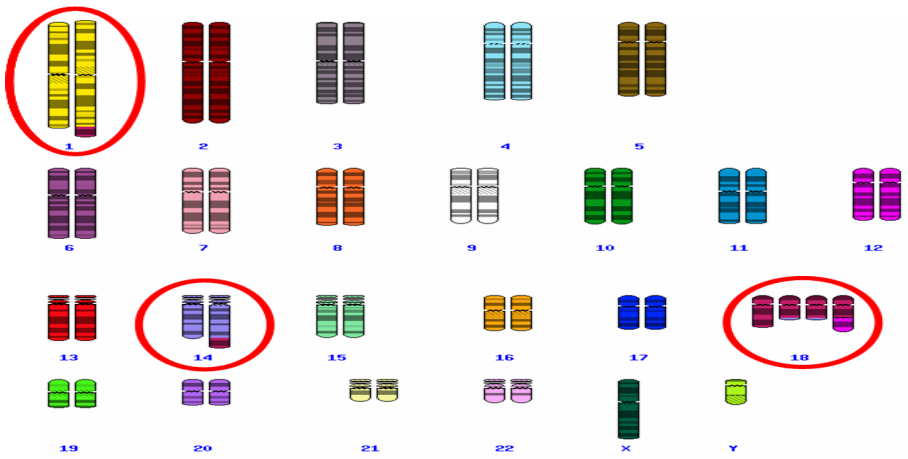

Fig. 1. A schematic view of an aberrant karyotype (produced by the SKYGRAM converter tool [1]). Chromosomes 1,14, and 18 show structural aberrations, and chromosome 18 shows a numerical aberration. (An ISCN description of this karyotype is 47,XY,der(1)t(1,18)(p36;q21),t(14,18)(q32;q21),+der(18)t(12;18)(p11;q21),+der(18)t(14;18).)

Over the last few decades, intensive research on chromosomal abberations in cancer has led to the accumulation of large amount of data on cancerous karyotypes. The largest available public depository of

such data is the Mitelman database [15], which contains over 50,000 karyotypes collected from over 8,000 publications. In this study we analyze this database. Our goal is to understand the main abberation types and their frequency in different cancers. Our hope is that such studies will provide insights and better understanding of the evolution of karyotypes in specific cancer types.

Traditionally, karyotypes have been constructed using chromosome staining methods, mostly G-banding. SKY [22] and M-FISH [25] are relatively new molecular cytogenetic techniques that permit the simultaneous visualization of all the chromosomes in different colors, considerably improving the detection of material exchange between chromosomes. The Mitelman database contains primarily karyotypes based on G-banding. The resolution and the detectable level of details in such karyotypes is lower than what can be observed with SKY and M-FISH or with novel high throughput methods (e.g. array-based CGH [24] and ESP [26]). Nevertheless, we chose to focus on the Mitelman database since it is the largest collection of cancerous karyotypes.

Karyotypes are usually described using the ISCN nomenclature [14]. In this system, every aberrant chromosome is described using specific rearrangement and numerical events, e.g., translocations, inversions, deletions, and duplications. Although ISCN attempts to describe the correct set of events leading to the observed karyotypes, it has almost no ability to do so when there are overlapping rearrangements, e.g. a chromosome involved in two translocations, each at a different position. Moreover, while the inference of the events is an easy task for many modestly rearranged karyotypes of hematological disorders, it can be a computationally hard task when the karyotypes are complex, as often happens in solid tumors.

There are many computational studies analyzing large data sets of cancerous genomes. Most of these analyses consider a cancerous genome as a collection of chromosomal abberations easily computed from the data. For example, in a series of studies, reviewed in [12], Högland et al. analyzed cytogenetic data from individual tumor types, by inspecting various parameters, including the number of gains or losses of genomic fragments, the number of aberrations, and the frequency at which bands are involved in breaks. In another study [21], Sankoff et al. compared the distributions of cancer-related breakpoints, derived from the Mitelman database, and evolutionary breakpoints, derived from a human-mouse comparative map. Another important branch of computational studies searches for statistical dependencies between chromosomal aberrations, usually in the form of tree or directed acyclic graph, such as [6, 7, 12, 11].

Chromosomal aberrations observed in cancer are by and large somatic and thus non-inheritable. When a rearrangement occurs in a genome of a germ-line cell, it can be inherited by offsprings. Indeed, the comparison of genomes of related species reveals that genome rearrangements play a significant role during the evolution of species. In a pioneering paper [20], Sankoff raised the problem of computing a shortest sequence of rearrangement operations between two given genomes, when genomes are represented by linear orders of oriented genes. Over the last fifteen years, this problem was intensively studied for many types of rearrangement events and their combinations, including inversions, translocations, block exchanges, deletions and insertions (see [4] for a review). All these studies ignored the *ploidy* in the genomes, i.e., the number of copies of each chromosome. Since numerical aberrations are prevalent in cancer, every model of cancer rearrangements must contain both numerical and structural events. This makes the reconstruction task more complicated and prevents direct use of results from the rich algorithmic literature on germ-line rearrangements.

The main purpose of this study was to estimate the prevalence of specific types of genome rearrangement events in cancer karyotypes. For this purpose we developed a new efficient heuristic for reconstructing a sequence of events that best explain the transformation from the normal karyotype into a given cancer karyotype. We applied this algorithm to over 40,000 karyotypes published in scientific literature, and collected statistics on event frequency across cancer types. The algorithm is deliberately simplistic, mimicking the process of detecting obvious events and "undoing" them, going back from the given karyotype towards the normal. As such, it does not guarantee finding the shortest solution or finding any solution. However, we reasoned that most reported karyotypes are of limited complexity and thus may be amenable to such approach. Reassuringly, over 98% of the karyotypes were solved by this method. Our study provides for the

first time a broad picture of event frequency in hematological and solid cancers. Our analysis shows that chromosome gains and losses, reciprocal translocations, and terminal deletions, dominate the evolution of cancer karyotypes. By using the event frequencies in each karyotype as its profile, we show that many different cancer types have clearly distinguishable profiles, which can be meaningful for further understanding of the cancers.

This paper is organized as follows. In Section 2 we provide a short background on chromosome aberrations in cancer. In Section 3 we present some basic statistics regarding the complexity of cancer karyotypes. In Section 4 we describe our heuristic for reconstructing genome rearrangement events for a given karyotype. The analysis of the reconstructed events is reported in Section 5. For lack of space, some details are deferred to an appendix.

2 Background

2.1 Mechanisms for chromosomal aberrations

Many molecular mechanisms are involved in the formation of chromosomal aberrations. The following mechanisms are reviewed in [2, 9, 16, 18].

A *double strand break* (DSB) is one of the frequent lesions in DNA. The repair of DSBs in eukaryotic cells is carried out by two main pathways: *non-homologous end joining* (NHEJ) and *homologous recombination* (HR). NHEJ repairs DSBs by directly re-ligating DNA ends, which may create a deletion if sequences surrounding the lesion were lost. Another potential risk of NHEJ is the ligation of two non-matching broken ends, leading to genome rearrangements. HR repairs breaks through interaction of a free DNA end with an intact homologous sequence, which is used as a template to copy missing information prior to religation. Because of the ability to fill in gaps by copying information from a sister chromatid or homologous chromosome, HR runs the risk of generating rearrangements through interaction of similar sequences on non-homologous chromosomes or regions. In particular, HR may extend to the end of a chromosome, resulting in a duplication of the whole "tail" of that chromosome.

Another possible lesion to the DNA is the loss of a telomere. The telomeres protect the ends of chromosomes from fusion with other ends. Thus a chromosome end that lacks a functioning telomere tends to be adhesive and may initialize a *breakage-fusion-bridge* process [13]. Stabilization of the genome occurs only through the net gain of a telomere, either through duplications of protected chromosome ends, or by direct telomere addition. Indeed, telomerase activity has been detected in the majority of malignant epithelial tumors [8].

A direct cleavage through a centromere generates two *telocentric* (i.e. single-arm) chromosomes, each containing a portion of the kinetochore (the functional component of an active centromere). Non-disjunction of sister chromatids of a telocentric chromosome results in the formation of an *isochromosome* or *isoderivative*, i.e. a chromosome with two identical, mirror-image arms.

As elaborated above, DSBs, telomeres dysfunction and centric fissions may lead to structural aberrations. Numerical aberrations may occur when genes involved in chromosome segregation or cytokinesis are deregulated. In particular, failure in cytokinesis (e.g. endomitosis) and multipolar mitoses may alter the ploidy of the genome.

2.2 The Mitelman database

The "Mitelman database of chromosome aberrations in cancer" [15] (henceforth abbreviated MD) contains the description of cancer karyotypes manually culled from the literature over the last twenty years. For our analysis we used the version of March 27, 2007, which contained 53,573 cancerous karyotypes, collected from 8748 published studies. The karyotypes in the database are represented in the ISCN format and can be automatically parsed and analyzed by the software package CyDAS [10]. We shall use here a simplified

version of ISCN for representing karyotypes (see Appendix A). We refer to a karyotype as *valid* if it can be parsed by CyDAS without any errors. According to our processing, 47,045 (87.8%) of the records were valid karyotypes.

2.3 Complex karyotypes

When the cytogeneticist analyzes a sample, several cells are checked. Each abberation described in a cancerous karyotype must be present in at least two cells in the described sample. In some cases the cell population may be non-homogeneous, and contain cells with several distinct karyotypes, resulting from evolution of the cell population during the development of the cancer. A homogeneous cell sample is described by a *simple karyotype*, and a non-homogeneous one has a *complex karyotype*, which consists of several karyotype species. In this study we derive simple karyotypes from complex karyotypes and analyze each of them independently.

About 17% of all valid karyotypes in MD are complex. The total number of simple (valid) karyotypes that we deduced from MD is 57941 (33% of which originate from complex karyotypes). For the rest of this paper we assume that every analyzed karyotype is simple.

3 Basic statistics on karyotype complexity

In this section we present some simple statistics based on the MD regarding the complexities of cancerous karyotypes. Human malignancies can be divided into two main categories: hematological disorders and solid tumors. Our first step was to distinguish between hematological malignancies and solid tumors. The type of neoplasia can be identified by its *morphology*, i.e. the cancer classification based on neoplasm histology, and its *topography*, i.e. the tumor site (applicable only for solid tumors). Based on the morphology and topography descriptors of each karyotype, we partitioned the karyotypes in the database into three categories:

- HEMA: hematological neoplasms, e.g.: leukemia, myeloma, lymphoma.
- BENIGN: solid benign tumors, e.g.: meningioma, leiomyoma, lipoma.
- SOLID: solid malignant tumors, e.g.:adenocarcinoma, Wilms tumor, malignant melanoma.

The HEMA category covers 71.2% of the valid simple karyotypes derived from the MD, while SOLID and BENIGN cover only 22.9% and 5.9% respectively. In the following, we compare the distributions of simple variables defined on karyotypes between these categories. We define a chromosome as *abnormal* if it does not match any chromosome in the standard normal karyotype. As expected, the distribution of the number of abnormal chromosomes per karyotype had the longest tail for solid tumors, while benign and hematological karyotypes seldom have more than five abnormal chromosomes (Fig. 5-a). The number of fragments (maximal contiguous interval in the normal) per an abnormal chromosome (Fig. 5-b) had a similar distribution across categories, with less than 1% of the abnormal chromosomes having four or more fragments. We defined *karyotype ploidy level* as $\lfloor \frac{n+11}{23} \rfloor$, where n is the total number of chromosomes. As expected, solid tumors tended to have higher ploidy, reflecting their higher complexity (Fig. 5-c). Multicentric chromosomes (i.e. chromosomes with more than one centromere) are considered non-stable, as each of the centromeres in these chromosomes may be passed to opposite poles in the mitotic anaphase. Interestingly, all three categories had some 2-4% of karyotypes with multicentric chromosomes (Fig. 5-d). Overall, the difference between the categories are quite subtle. Karyotypes of solid tumors, in particular malignant solid tumors, tend to have more complex abnormal chromosomes and ploidy changes, in comparison to hematological malignancies.

Do the statistics above - as well as those we shall report later - reflect the distributions of properties in cancer karyotypes "in the real world"? The answer is probably no. For example, although up to 80% of all human malignancies are solid, most of the karyotypes in MD belong to hematological malignancies.

One major reason for this bias is the difficulty in cytogenetically analyzing solid tumors. Solid tumor genomes often demonstrate poor visual quality during metaphase. Moreover, the karyotypes of solid tumors are often much more complex and thus more difficult to interpret. In addition, the database contains *reported* karyotypes from the literature, and there is a bias in this reporting. For example, the hematological karyotypes in MD are probably of higher complexity than those simple cases seen regularly in the clinic, which are not deemed publish-worthy as they are too simple or fully understood. While this means that the statistics we are collecting should be interpreted with caution, we believe they can still be useful in understanding how to model cancer evolution on the karyotype level and how different classes and subclasses differ.

4 A sorting algorithm

In this section we describe an algorithm, which we call SKS (Simple Karyotype Sorter), for reconstructing the sequence of rearrangement events (structural and numerical) that have led from the normal karyotype to a given cancer karyotype. We call this process *sorting* the karyotype. The SKS algorithm aims to mimic the intuitive way a cytogeneticist would perform this task, i.e., starting with the cancer karyotype and going backwards towards the normal karyotype one event at a time, taking the simplest and most evident step whenever possible. The SKS algorithm is a heuristic and does not guarantee finding an optimal or even finding any solution sequence when one exists. In Section 5 we shall report on the performance of this heuristic on the MD karyotypes.

4.1 An abstract data structure of a karyotype

A chromosome is *indefinite* if its description includes unknown items. For example, ?→? and 1pter→1p? are indefinite chromosomes. Note that a definite chromosome may contain uncertain items, e.g. 1pter→1p?12. Similarly, a karyotype is *definite* if it contains only definite chromosomes. In what follows we analyze only definite karyotypes, and ignore any uncertainties, e.g. 1p?12 will be considered as 1p12. As can be expected, the percentage of indefinite karyotypes in malignant solid tumors (39.6%) is higher than in hematological neoplasms (28%), and is the lowest for benign tumors (24.2%). Hence, the overall number of karyotypes we analyze here is 40,298.

We represent a karyotype K by the following abstract data structure:

- *Abnormal_Chrs*(K): A set of distinct, orientation-less, abnormal chromosomes. For each abnormal chromosome in *Abnormal_Chrs*(K) we maintain its multiplicity and list of fragments.
- *multiplicity*: a mapping assigning to each normal chromosome id (i.e. 1, ... ,22, X, Y) its multiplicity in K.

4.2 Orphan fragments

Denote by *Frags*(K) the multiset of fragments found in *Abnormal_Chrs*(K). A fragment in *Frags*(K) is *orphan* if there is no other fragment in *Frags*(K) from the same normal chromosome. For example, suppose *Abnormal_Chrs*(K) = {9pter → 9q32::1p36 → 1pter, 14qter → 14p21::9q32 → 9qter, 14p21 → 14qter} then *Frags*(K) = { 9pter → 9q32, 9q32 → 9qter, 14qter → 14p21 × 2, 1p36 → 1pter} and K contains exactly one orphan fragment: 1p36→1pter.

The easiest way to explain an occurrence of an orphan fragment is by a translocation event followed by a loss of one of the two resulting abnormal chromosomes. For an acentric orphan fragment there is an alternative, less conservative explanation: The orphan fragment resulted from a duplication during a process of HR DSB-repair (recall Section 2.1). In Section 5.2 we describe some statistics regarding acentric orphan fragments that suggest the latter explanation is more likely for many cases.

4.3 Algorithm SKS

The SKS algorithm computes a sequence of events $S = \rho_1, \ldots, \rho_t$ that transforms a normal karyotype into a given (cancerous) karyotype K. Starting from K and applying the corresponding inverse operations $S^{-1} = \rho_t^{-1}, \ldots, \rho_1^{-1}$ generates a normal karyotype. The SKS algorithm works in two phases. First, all the abnormal chromosomes are sorted. Then, simple numerical operations "correct" the multiplicities of the normal chromosomes.

We need a few definitions first. A fragment is *centric* if it contains a centromere, and *acentric* otherwise. Let f and g be two fragments from the same normal chromosome. The concatenation $f::g$ is an *adjacency* if f and g have exactly one shared band - which is their fused ends. For example, 1pter→1p11::1p11→1q22 is an adjacency. In this case, f and g are said to be *complementing*. Fragments $f, g \in \mathit{Frags}(K)$ are *uniquely* complementing if no other fragment $h \in \mathit{Frags}(K)$ is complementing to f or g. The types of rearrangement events that we consider will be introduced in the description of algorithm.

Initialization. We first detect simple changes in the karyotype ploidy as follows. Let μ and g be the the median and greatest common divisor of all distinct chromosome multiplicities (both normal and abnormal) respectively. Clearly, $\mu \geq g$. Suppose $g > 1$. In this case we divide all chromosome multiplicities by $d = g$. A single exception is when $\mu = g$ and g is even - in this case we divide by $d = g/2$ (instead of by g). If the chromosome multiplicities were changed (i.e. $d > 1$) - we set $S = \{\rho\}$, where ρ is a corresponding PLOIDY CHANGE event.

Phase I: Sorting the abnormal chromosomes. The abnormal chromosomes are sorted by repeatedly detecting and undoing one of the following events. The phase ends successfully if there are no more abnormal chromosomes, and ends with failure if there are still abnormal chromosomes but no additional event is detected.

- **CHR GAIN**: A *chromosome gain* is a duplication of a complete chromosome. To detect such event, seek an abnormal chromosome, *chr*, whose multiplicity, m, is greater than 1. Perform the inverse operation, i.e., the removal of one copy of *chr*, decreasing its multiplicity to $m - 1$.
- **ISOCHROMOSOME CREATION**: Detect any iso-chromosome or iso-derivative (see Sec. 2). Perform the inverse operation, by removing one of the identical arms.
- **TRANSLOCATION** and **FISSION**: A *translocation* is the exchange of tails between two chromosomes; a *fission* is the split of one chromosome into two contiguous segments. Let f and g be two uniquely complementing fragments found on different chromosomes. Then there are two possible cases. In the first case, the complementing ends of both f and g correspond to chromosome ends. In this case, a FISSION event is detected and the inverse operation is a simple fusion of f and g in their complementing ends (i.e. chromosome fusion). The latter case is when at least one of the complementing ends of f and g is fused to another fragment. In this case, a TRANSLOCATION event is detected and the inverse translocation that fuses the complementing ends of f and g is applied to K.
- **INVERSION**: An *inversion* is the reversal of a DNA segment within a chromosome. This event is detected for a pair of uniquely complementing fragments, f and g, on the same chromosome, that have different orientation. The inverse operation is an inversion that fuses the complementing ends of f and g. For example, suppose the chromosome containing f and g is of the form $f::h_1::-g::h_2$, where $-g$ is the inverse of g and $f :: g$ is an adjacency. In this case, the detected INVERSION event inverts the segment $h_1::-g$.
- **TANDEM DUP**: A *tandem duplication* creates two identical consecutive fragments on the same chromosome creating $h \equiv f_1 :: f_2 :: f_2 :: f_3$. For example, 1pter→1q44::1q31→1qter is a tandem duplication since 1pter→1q44 ≡ 1pter→1q31::1q31→1q44 and 1q31→1qter ≡ 1q31→1q44::1q44→1qter. When identifying such a repetition, simply remove it, forming $h \equiv f_1 :: f_2 :: f_3$.

- **INTERNAL DELETION**: An internal deletion of a fragment within a chromosome is discovered as follows. Detect a non-adjacency pair of concatenated fragments, f::g, for which there exists a fragment h such that *(i)* f::h and h::g are adjacencies, and *(ii)* h does not contain in its span any fragment in *Frags*(K). Replace f::g by fragment $f' \equiv f$::h::g.
- **TAIL DELETION**: A deletion of a chromosome tail (acentric end fragment) is detected by identifying an abnormal chromosome end lacking a pter or a qter, and whose complementing fragment, f, is *(i)* acentric and *(ii)* does not contain in its span any fragment in *Frags*(K). To undo the operation, concatenate f to the chromosome's end such that a new adjacency is formed.
- **ACENTRIC ORPHAN TAIL**: Detect an acentric orphan fragment f that is found on one end of an abnormal chromosome. Eliminate this aberration by a removal of f.
- **CENTRIC ORPHAN FUSION**: Detect a multicentric chromosome *chr* containing a centric orphan f. To undo the operation, perform a fission of *chr* near f such that each of the resulting two chromosomes contains a centromere.

Phase II: Gain/loss events and ploidy changes. If this phase is reached the current karyotype K satisfies *Abnormal_Chrs*(K) $= \emptyset$. Define $\mu(K)$ as the median multiplicity of all chromosomes in K (for gain/loss computations we consider the sex chromosomes as homologs). For any chromosome *chr* whose multiplicity differs from $\mu(K)$, adjust its ploidy to $\mu(K)$ by CHR LOSS or CHR GAIN events. Then, when the ploidy of all chromosomes is $\mu(K)$, adjust the ploidy globally to 2 by prepending a corresponding PLOIDY CHANGE event to S.

5 Experimental results

We ran algorithm SKS on each of the 40,298 definite simple karyotypes derived from MD. We say that a karyotype is *sortable* if SKS transforms it successfully to the normal karyotype. Table 1 shows that the vast majority (>98%) of the karyotypes are sortable. Hence, our rather naive heuristic, which makes only straightforward moves, performs very well on the MD karyotypes.

Table 1. Sortability of MD karyotypes. Numbers are percent out of the karyotypes in each category.

	HEMA	**BENIGN**	**SOLID**	**ALL**
Sortable - numerical aberration only	21.8%	41.1%	43.8%	27.4%
Sortable - with structural aberrations	76.7%	56.7%	54.3%	71.0%
Not sortable	1.5%	2.2%	1.9%	1.7%

5.1 Event rates

Figure 2-a presents the average number of each type of event per karyotype in our reconstruction. The most prevalent reconstructed events in all categories are chromosome gains and losses, tail deletions and translocations. In contrast, most other events are relatively rare, occurring in a tenth of the karyotypes or even less. For example, the translocation rate is 0.54 per karyotype, while inversion rate is only 0.06[1]. Note that while the events of chromosome gain and loss and tail deletion are dominant in the arrangement of malignant solid tumor karyotypes, translocations are relatively more frequent in hematological karyotypes.

Translocations are called *reciprocal* of both of the exchanged fragments are non-empty. Our analysis shows that most (>96%) reconstructed translocations are reciprocal (Fig. 2-b). Additional support to this observation is obtained by analyzing the breakpoint graphs of karyotypes (Appendix B). Interestingly, non-reciprocal translocations are more than twice as common in solid tumors than in hematological karyotypes.

[1] The surprisingly low inversion rate should be taken with caution: clearly, only relatively long inversions covering several bands are detectable in G-banded karyotypes in MD.

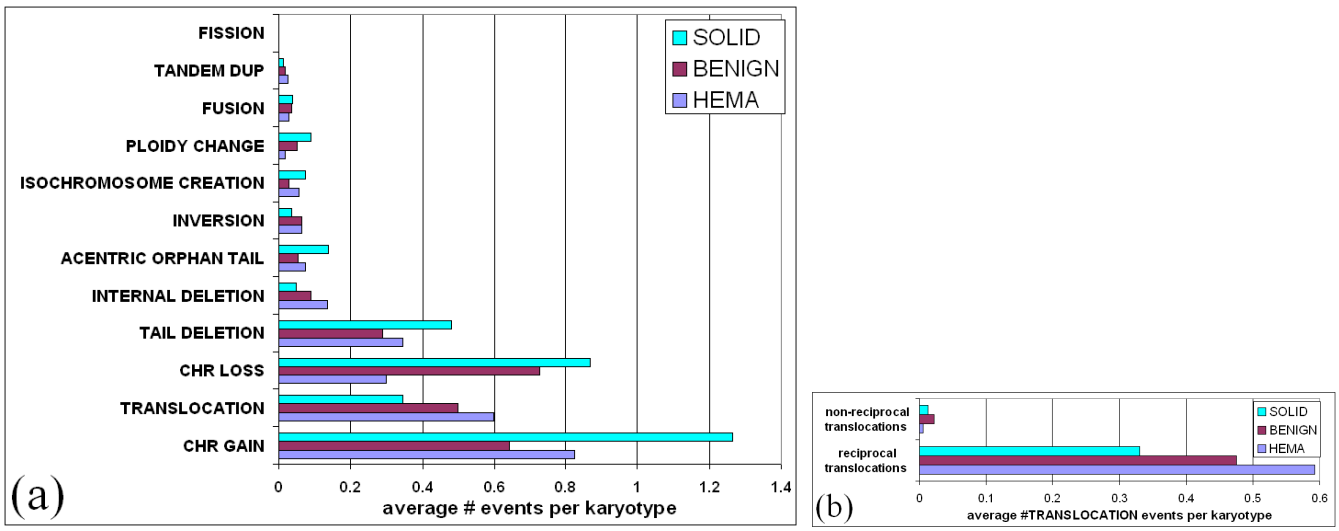

Fig. 2. Frequencies of each rearrangement event. Numbers are based on applying the sorting algorithm to all valid simple karyotypes in the database. (a) The average number of events per karyotype. (b) Average number of reciprocal and non-reciprocal translocations.

5.2 The origin of ACENTRIC ORPHAN TAILs

For a fragment $f \in Frags(K)$, let $chr(f)$ be the normal chromosome of f. Figure 3 presents the distributions $multiplicity(chr(f))$, for centric orphan fragments and for acentric orphan tail fragments. For comparison, we include the distribution of $chr(i)$, $i \in \{1, \ldots, 22\}$, after all abnormal chromosomes have been sorted (i.e. at the completion of Phase I of SKS algorithm). As can be expected, the ploidy of normal autosomal chromosomes is mostly 2. The ploidy of the normal chromosome of centric orphan fragments is usually 1. Thus the most reasonable explanation is that centric fragments evolved from normal chromosomes by translocations or tail deletions. Surprisingly, the ploidy of the normal chromosomes of acentric tail orphans is mostly 2. Since most (98%) of these acentric orphan fragments have one complete end (i.e. pter or qter), this suggests that many of these acentric orphan fragments are the result of a tail duplication event, caused by the HR DSB repair mechanism (see Section 2.1). The alternative scenario is a translocation event, and an additional event of chromosome gain. The latter explanation is more complex and hence less likely.

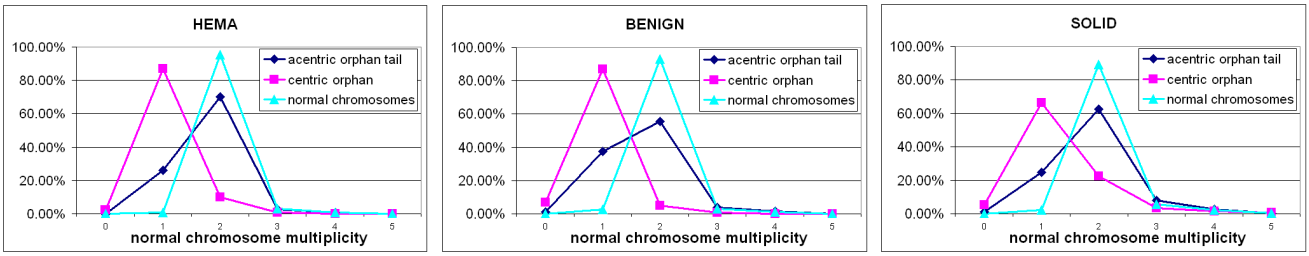

Fig. 3. Orphans and their parent chromosomes. The plots show the distributions of the multiplicity of normal chromosomes corresponding to acentric orphan tail fragments, and to centric orphan fragments. For comparison, each plot also includes the multiplicity of normal (autosomal) chromosomes, after all abnormal chromosomes have been sorted. The distributions are computed separately for categories HEMA, BENIGN and SOLID.

5.3 Rearrangement events as characteristics of cancer classes

Are the events that constitute the history of karyotypes, as reconstructed by the SKS algorithm, meaningful to understanding and distinguishing the different cancer types? To answer this question, we defined several similarity measures between distinct karyotypes, using the event rates reconstructed by the algorithm, and used them to compare cancer classes. Our analysis focused on karyotypes from 14 cancer classes, containing 60–885 karyotypes each (See Tables 2 and 3 for the class descriptions and detailed results). In our tests below we called a test *significant* if it attained p-value $< .0001$, after Bonferroni correction for multiple testing.

Clustering cancer classes by their event profiles. For a karyotype K we define its *event profile*, $\bar{v}(K)$, as a vector whose entries are the frequencies of each event in K (event order is as in Fig. 2a, bottom to top). For example, $\bar{v}(K) = (2, 0, 2, 1, 0, 1, 0, 0, 0, 0, 0, 0)$ for the karyotype K in Fig. 6. Given a set of karyotypes we define the *average event profile* as the coordinate-wise average of the event profiles of the karyotypes. Using Pearson correlation as a similarity measure, we applied an average linkage hierarchical clustering algorithm [23] on the average profiles of the 14 classes. As can be seen in Fig. 7, related cancers tend to cluster close to each other, implying they have similar average event profiles.

Partitioning karyotypes by event profiles. Let C_1 and C_2 be two distinct cancer classes, and let $\Omega = C_1 \cup C_2$. Can the karyotypes in Ω be distinguished, as to which belongs to C_1 and which belongs to C_2, by their event profiles? We partitioned Ω into two clusters, D_1 and D_2 ($\Omega = D_1 \cup D_2$), by applying k-means clustering [23], with $k = 2$, on the event profiles in Ω, and using Pearson correlation as the similarity measure. We measured the p-value of the correspondence between the new partition, $\{D_1, D_2\}$, and the original one, $\{C_1, C_2\}$, using the hypergeometric distribution (see Appendix C for details). We performed this test for all $\binom{14}{2} = 91$ pairs of classes. 26 (28.6%) of the tested pairs were significant.

Partitioning karyotypes by total event frequency. We define *NEvents* as the total number of reconstructed events for the karyotype (i.e., the sum of the entries in $\bar{v}(K)$). Given $\Omega = C_1 \cup C_2$ as before and an integer t, let $D_1^{(t)} = \{K \in \Omega : \text{NEvents}(K) \leq t\}$ and $D_2^{(t)} = \{K : \text{NEvents}(K) > t\}$. We computed the p-value of the match between $\{D_1^{(t)}, D_2^{(t)}\}$ and the original partition, for $t = 0, \ldots, 9$. 45 of the 91 pairs (49.5%) had a significant NEvents-based partition. We repeated the same test with the *NAPT* score [12], which is the number of aberrations in the karyotype's ISCN description[2]. NEvents and NAPT are different indicators of a karyotype's complexity. Interestingly, although NAPT is much less exact than NEvents, 53.8% of the tested pairs had a significant NAPT-based partition. A possible explanation is that the relatively large differences between the classes are captured better by a cruder measure. On the other hand, there is meaningful additional information in individual events. For example, 76.9% of the significant partitions based on event profiles had p-values lower than the corresponding partitions based on NEvents and NAPT, and 6 (14.3%) of the non-significant NAPT-based partitions had corresponding significant partitions based on event profiles.

Partitioning karyotypes using a single type of event. For each type of event, e, let *SEvent(e)* be the number of reconstructed events from type e. For example, SEvent(CHR GAIN) is the number of CHR GAINs (i.e. the first entry in the event profile). Our last test was to partition Ω using SEvent(e), for each type of event e, in the same fashion as above. Due to the relatively low values, we checked only five thresholds ($t = 0 \ldots 4$) for each type of event. Surprisingly, 81.3% of the tested pairs had a significant SEvent-based partition. The lowest p-values were achieved for partitions based on TRANSLOCATIONs (35.6%), CHR LOSSes (27.4%), and CHR GAINs (16.9%).

[2] The NAPT score is calculated by simply counting the number of comma-separated tokens in the ISCN description, disregarding the first two tokens that correspond to the total number of chromosomes and the sex chromosomes description. For example, the NAPT score for the karyotype in Fig. 1 is 5.

6 Conclusion

In this paper we presented novel methods for analyzing and comparing aberrant karyotypes observed in hematological malignancies and in solid tumors cells. We presented a simple yet effective heuristic (the SKS algorithm) for sorting aberrant karyotypes. On over 40,000 karyotypes of the Mitleman database, the algorithm attained a very high success rate (98%) in sorting the karyotypes. We believe that this shows that on such karyotypes of moderate complexity, the set of rearrangement events reconstructed by our algorithm (though not necessarily their order) is a close approximation of the actual gross chromosomal rearrangements that occurred in their evolution. Our analysis implies that the evolution of aberrant karyotypes in somatic cells is dominated by four events: chromosome gains and losses, reciprocal translocations and terminal deletions. The prevalence of chromosome gains and losses is expected, since these events are more easily detected than other more local events, e.g. inversions. Nevertheless, these results emphasize that duplication and deletion events must play a key role in any computational modeling of genome rearrangements in cancer.

By using clustering techniques, we demonstrated that karyotypes belonging to the same cancer class have characteristic event rates, since they often have more similar event frequencies than karyotypes belonging to different classes. Moreover, this suggests that carcinogenesis involves different pathways of gaining chromosomal aberrations for different cancer classes, and further analysis may shed light on the events characterizing different pathways.

One of the goals of this study was to lay the factual foundations for proposing a mathematical model of somatic genome rearrangements that will allow an accurate, non-heuristic systematic analysis of aberrant karyotypes. The simplest model that can generate the spectrum of the aberrations observed in cancerous karyotypes includes four types of events: chromosome gain and loss, breakage, and fusion. For example, a reciprocal translocation can be mimicked by two breaks followed by two fusions. While this simplistic model favors non-reciprocal translocations over reciprocal ones, our study observed the opposite preference in the MD karyotypes. Thus, a more realistic model should consider reciprocal translocations as atomic operations, to reflect the increased probability of their occurrence. Another operation that is worth considering is the duplication of a segment in an existing chromosome (see Section 5.2). Our hope is that a computational investigation of many reconstructed rearrangement sequences will help in pointing out the dominant scenarios through which chromosomal aberrations evolve in specific types of cancer.

Acknowledgments

We are grateful to Igor Ulitsky for his tremendous help in analyzing the event rate profiles, and to Gideon Rechavi, Luba Trakhtenbrot, and Chaim Linhart for helpful discussions and insightful comments. We thank Felix Mitelman and John Wiley & Sons, Inc. for granting us permission to analyze the data in the Mitelman database of chromosome aberrations in cancer.

References

1. NCI and NCBI's SKY/M-FISH and CGH Database, 2001. `http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi`.
2. D.G. Albertson, C. Collins, F. McCormick, and J. W. Gray. Chromosome aberrations in solid tumors. *Nature Genetics*, 34:369–376, 2003.
3. V. Bafna and P. A. Pevzner. Genome rearragements and sorting by reversals. *SIAM Journal on Computing*, 25(2):272–289, 1996.
4. G. Bourque and L.Zhang. Models and methods in comparative genomics. *Advances in Computers*, 68:60–105, 2006.
5. A. de Klein et al. A cellular oncogene is translocated to the philadelphia chromosome in chronic myelocytic leukaemia. *Nature*, 300:765–767, 1982.
6. R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitrou, and A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, 6:37–51, 1999.

7. R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitrou, and A. Schäffer. Distance-based reconstruction of tree models for oncogenesis. *Journal of Computational Biology*, 7:789–803, 2000.
8. G. Krupp et al. Telomerase, immortality and cancer. *Biotechnology Annual Review*, 6:103–140, 2000.
9. D.O. Ferguson and W.A. Frederick. DNA double strand break repair and chromosomal translocation: Lessons from animal models. *Oncogene*, 20(40):5572–5579, 2001.
10. B. Hiller, J. Bradtke, H. Balz, and H. Rieder. CyDAS: a cytogenetic data analysis system. *BioInformatics*, 21(7):1282–1283, 2005. `http://www.cydas.org`.
11. M. Hjelm, M. Höglund, and J. Lagergren. New probabilistic network models and algorithms for oncogenesis. *Journal of Computational Biology*, 13(4):853 –865, 2006.
12. M. Höglund, A. Frigyesi, T. Säll, D. Gisselsson, and F. Mitelman. Statistical behavior of complex cancer karyotypes. *Genes, Chromosomes and Cancer*, 42(4):327–341, 2005.
13. B. McClintock. The stability of broken ends of chromosomes in zea mays. *Genetics*, 26(2):234–282, 1941.
14. F. Mitelman, editor. *ISCN (1995): An International System for Human Cytogenetic Nomenclature*. S. Karger, Basel, 1995.
15. F. Mitelman, B. Johansson, and F. Mertens (Eds.). Mitelman Database of Chromosome Aberrations in Cancer, 2007. `http://cgap.nci.nih.gov/Chromosomes/Mitelman`.
16. J.P. Murnane and Laure Sabatier. Chromosome rearrangements resulting from telomere dysfunction and their role in cancer. *BioEssays*, 26:1164–1174, 2004.
17. P.C. Nowell and D.A. Hungerford. A minute chromosome in human chronic granulocytic leukemia. *Science*, 132:1497, 1960.
18. J. Perry, H.R. Slater, and K.H.A Choo. Centric fission simple and complex mechanisms. *Chromosome Research*, 12(6):627–640, 2004.
19. J.D. Rowley. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243:290–293, 1973.
20. D. Sankoff. Edit distance for genome comparison based on non-local operations. *Lecture Notes in Computer Science*, 644:121–135, 1992.
21. D. Sankoff, M. Deneault, P. Turbis, and C. Allen. Chromosomal distributions of breakpoints in cancer, infertility, and evolution. *Theoretical Population Biology*, 61(4):497–501, 2002.
22. E. Schröck, S. du Manoir, T. Veldman, B. Schoell B, J. Wienberg, M.A. Ferguson-Smith, Y. Ning Y, D.H. Ledbetter, I. Bar-Am, D. Soenksen D, Y. Garini, and T. Ried. Multicolor spectral karyotyping of human chromosomes. *Science*, (5274):494–497, 1996.
23. R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon. Expander: an integrative suite for microarray data analysis. *BMC Bioinformatics*, 6(232), 2005.
24. A. M. Snijders and N. Nowak et al. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29:263–264, 2001.
25. M.R. Speicher, S.G. Ballard, and D.C. Ward. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, 12(4):368–375, 1996.
26. S. Volik and S. Zhao et al. End-sequence profiling: Sequence-based analysis of aberrant genomes. *Proceedings of the National Academy of Science USA*, 100:7696–7701, 2003.

Appendices

A Formal representation of karyotypes

A chromosome is divided by its centromere into two *arms*: a short arm, denoted p, and a long arm, denoted q. Every chromosome arm is partitioned into bands. The bands in each arm are numbered, starting from the centromere, whose assigned to the number 10. The symbol *ter* indicates the (normal) end of a chromosome arm. A position in the chromosome is identified by three fields: *(i)* chromosome, *(ii)* arm, and *(iii)* band designation (either a number or ter). For example, 1p11 corresponds to band 11 in the long arm of chromosome 1; 2p10 and 2q10 both refer to the centromere of chromosome 2; 3pter is the (normal) end of the short arm of chromosome 3.

We refer to a chromosome as *abnormal* if its structure is abnormal. Abnormal chromosomes are defined by their band composition. In the following, we describe abnormal chromosomes in a similar (but not identical) manner to the *detailed system* of ISCN [14]. The term *fragment* refers to a continuous interval of a normal chromosome, identified by the positions of its two ends. When a fragment appears in a chromosome

it has an orientation, denoted by an arrow symbol → between its two ends. For example, 2p12→2qter is a fragment of chromosome 2 that starts in band 2p12 and ends in band 2qter. Two fragments are *identical* if the corresponding chromosome intervals are identical (disregarding orientation). A double colon (::) indicates a concatenation of two fragments. For example, a concatenation of 1p36→1pter to the end of 9pter→9q32 is denoted as 9pter→9q32::1p36→1pter. An abnormal chromosome is presented as a list a concatenation of fragments[3].

The description of a karyotype may contain question marks (?) to indicate uncertainties or unknown items. A question mark may be placed either before an uncertain item, or it may replace an unknown chromosome, arm, or band designation. For example, 1p?12 indicates a questionable identification of band number; 5p? represents an unknown band designation.

B Using cycles and paths for analyzing translocation types

For a cancerous karyotype K we define its *breakpoint graph*, $G(K)$, similarly to [3], as follows. The vertices of $G(K)$ are the ends of the fragments in *Frags*(K). The edges in $G(K)$ are colored either black or gray. Black edges correspond to fused ends in K. Grey edges correspond to complementing ends. For an example, see Fig. 6-c-1.

Let S be a sequence of events reconstructed for K by SKS. Each of the inverse operations for INVERSION, TRANSLOCATION, and FISSION events, forms one or two new adjacencies by fusing complementing ends. Let $G(K, S)$ be the subgraph of $G(K)$ induced by *(i)* the set of black edges, and *(ii)* the grey edges corresponding to pairs of fused complementing ends during the reconstruction of INVERSION, TRANSLOCATION, and FISSION events in S. See Fig. 6-c-2 for an example. It follows that $G(K, S)$ is composed of simple cycles and paths. The *length* of a cycle or path in $G(K, S)$ is the number of grey edges in it. Note that while a path of size l corresponds to l reconstructed events, a cycle of the same length corresponds only to $l - 1$ events. We define the *caliber* of a path or cycle to be the number of corresponding events. A path or a cycle with caliber greater than 1 imply a *breakpoint reuse*, i.e. a break of a formerly created fusion. Figure 4 depicts the average numbers of cycles and paths in a karyotype, for each caliber. It is quite clear that cycles are much more prevalent than paths, even in solid tumors, which indicates that reciprocal translocations are indeed more favored than non-reciprocal ones. Moreover, both structures, cycles and paths, usually have a small caliber.

C Measuring the significance of a partition

In this section we describe the standard hypergeometric score that was used for evaluating the match of two partitions. Let $\{C_1, C_2\}$ and $\{D_1, D_2\}$ be two partitions of Ω. Let $n = |\Omega|$, $n_1 = |C_1|$, $m = |D_1|$, and $k = |C_1 \cap D_1|$. Hence $k \leq \min\{n_1, m\}$. The significance of the correspondence between $\{D_1, D_2\}$ and $\{C_1, C_2\}$ can be evaluated by the probability of having $|C' \cap D_1| \geq k$ where $C' \subset \Omega$ is randomly chosen and $|C'| = n_1$. This probability is given by:

$$p(n, m, n_1, k) = \sum_{i=k}^{\min\{n_1, m\}} \frac{\binom{m}{i}\binom{n-m}{n_1-i}}{\binom{n}{n_1}}$$

The smaller $p(n, m, n_1, k)$, the more significant the correspondence between D_1 and C_1. To compare D_1 with C_2, we compute $p(n, m, n - n_1, m - k)$. The final p-value for the partition $\{D_1, D_2\}$ is thus

$$p\text{-value}(\{D_1, D_2\}, \{C_1, C_2\}) = 2\min\{p(n, m, n_1, k), p(n, m, n - n_1, m - k)\}.$$

(The multiplier 2 is due to Bonferroni correction for multiple testing.)

[3] The exception for this are homogenously staining regions (HSRs), which are regions that contain multiple copies of small DNA fragments. Thus a stained HSR is uniform in appearance (no bands) and its content cannot be identified by cytogenetic methods.

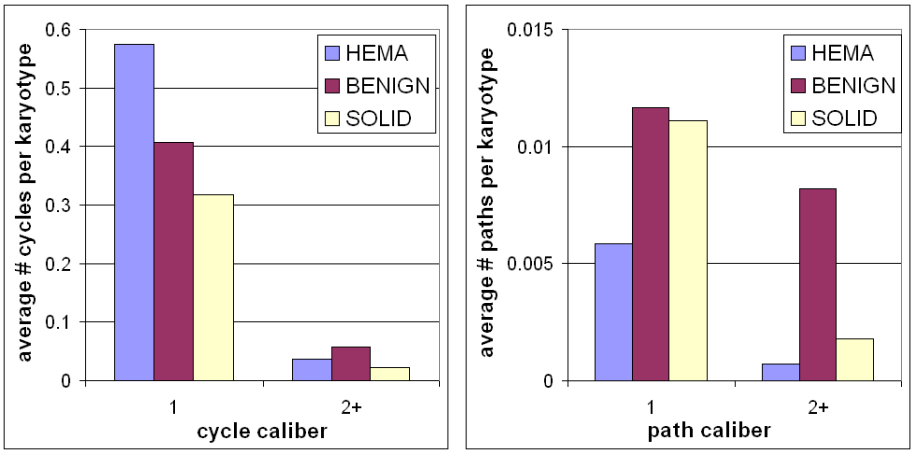

Fig. 4. The distributions of the average numbers of cycles and paths in a karyotype.

Table 2. Cancer classes.

class ID	class name	#karyotypes
27	HEMA-Acute monoblastic leukemia without differentiation (FAB type M5a)	332
28	HEMA-Refractory anemia with excess of blasts	885
31	HEMA-Refractory anemia	875
34	HEMA-Refractory anemia with ringed sideroblasts	230
36	HEMA-Acute myeloblastic leukemia with minimal differentiation (FAB type M0)	286
43	SOLID-Adenocarcinoma-Breast	590
52	HEMA-Acute monoblastic leukemia with differentiation (FAB type M5b)	196
58	HEMA-Refractory anemia with excess of blasts in transformation	424
70	SOLID-Adenocarcinoma-Kidney	859
111	BENIGN-Benign epithelial tumor special type-Breast	97
112	SOLID-Adenocarcinoma-Large intestine	208
118	BENIGN-Adenoma-Large intestine	149
143	SOLID-Adenocarcinoma-Ovary	119
577	BENIGN-Benign epithelial tumor NOS-Breast	60

Table 3. Partition p-values for pairs of cancer classes in Table 2. The p-values presented are after the Bonferroni correction for multiple testing.

Class 1	Class 2	event profile	NEvents	NAPT	SEvent
27	28	1.00E+00	3.11E-03	7.11E-03	3.32E-69
27	34	1.13E-04	7.99E-03	5.15E-03	1.85E-46
27	43	4.50E-13	4.52E-03	2.05E-05	8.04E-37
27	58	1.18E-06	1.00E+00	1.00E+00	3.49E-30
27	111	2.82E-01	3.38E-01	5.89E-02	1.01E-10
27	118	1.48E-31	5.12E-05	8.54E-05	4.73E-43
27	577	8.92E-23	1.02E-14	5.15E-18	1.02E-24
28	34	1.00E+00	1.00E+00	1.00E+00	1.00E+00
28	43	1.00E+00	3.33E-06	1.07E-02	3.41E-16
28	58	1.00E+00	4.17E-01	7.66E-01	1.97E-03
28	111	1.00E+00	1.00E+00	3.35E-02	6.97E-03
28	118	1.36E-01	3.33E-07	9.73E-08	1.58E-23
28	577	1.00E+00	4.47E-18	8.04E-25	5.11E-19
31	36	2.57E-02	1.49E-04	2.84E-06	8.75E-21
31	52	1.00E+00	2.48E-01	1.00E+00	1.72E-50
31	70	1.56E-15	7.16E-74	1.05E-92	1.96E-92
31	112	1.06E-08	2.49E-22	5.80E-22	8.68E-22
31	143	1.00E-13	6.67E-22	7.74E-27	1.92E-20
34	36	2.59E-01	6.59E-01	1.00E+00	7.52E-08
34	52	1.00E+00	3.78E-01	1.00E+00	2.48E-26
34	70	1.90E-01	1.21E-25	2.68E-25	1.76E-29
34	112	8.69E-04	8.19E-07	8.94E-07	1.31E-08
34	143	1.93E-03	1.13E-09	2.63E-10	6.71E-09
36	43	1.00E+00	3.29E-02	5.67E-01	5.60E-01
36	58	1.00E+00	1.00E+00	1.00E+00	2.54E-02
36	111	3.60E-01	1.00E+00	3.15E-01	1.00E+00
36	118	6.02E-09	1.24E-04	1.69E-04	2.70E-10
36	577	1.00E+00	4.91E-14	6.17E-17	4.32E-17
43	58	1.17E-01	7.66E-01	1.26E-01	3.42E-10
43	111	1.00E+00	1.00E+00	1.00E+00	1.00E+00
43	118	1.00E+00	2.17E-02	1.32E-04	4.09E-14
43	577	3.10E-10	2.27E-10	5.85E-16	1.36E-15
52	70	2.39E-63	9.88E-31	1.96E-31	1.46E-53
52	112	7.15E-40	8.01E-10	1.61E-08	1.10E-30
52	143	1.01E-19	1.02E-13	7.03E-13	7.22E-15
58	70	1.00E+00	1.82E-28	2.61E-30	1.15E-31
58	112	1.00E+00	6.64E-05	1.45E-04	6.24E-08
58	143	1.00E+00	3.72E-07	2.73E-08	5.12E-09
70	111	1.00E+00	9.84E-10	1.01E-10	5.51E-14
70	118	6.20E-11	4.41E-09	2.25E-05	2.97E-21
70	577	2.46E-02	5.29E-02	3.37E-06	2.78E-08
111	118	1.00E-06	2.51E-01	2.38E-02	1.09E-12
111	577	1.00E+00	7.50E-08	1.02E-10	9.46E-09
112	143	1.00E+00	1.59E-01	8.58E-01	1.00E+00
118	143	3.01E-01	2.80E-01	1.00E+00	8.93E-02
143	577	1.85E-04	1.79E-02	5.83E-04	7.26E-04

Class 1	Class 2	event profile	NEvents	NAPT	SEvent
27	31	1.00E+00	2.63E-09	1.78E-10	1.45E-98
27	36	2.69E-09	1.00E+00	1.00E+00	4.50E-17
27	52	1.00E+00	7.68E-02	1.10E-02	1.00E+00
27	70	1.00E+00	1.18E-25	2.95E-27	3.21E-83
27	112	3.42E-17	1.12E-07	4.47E-07	3.93E-51
27	143	6.47E-31	6.05E-10	1.42E-09	7.13E-26
28	31	1.00E+00	1.47E-02	1.78E-06	1.60E-10
28	36	1.00E+00	1.00E+00	1.00E+00	2.68E-07
28	52	1.00E+00	4.78E-02	5.89E-02	6.32E-32
28	70	1.00E+00	1.36E-55	1.33E-55	2.83E-45
28	112	1.00E+00	2.06E-12	2.36E-11	8.86E-11
28	143	9.20E-01	1.97E-13	3.61E-13	4.90E-11
31	34	1.00E+00	7.67E-01	1.67E-01	1.90E-01
31	43	9.10E-03	1.55E-14	1.17E-12	2.06E-38
31	58	4.16E-01	4.92E-07	5.52E-08	6.22E-15
31	111	1.00E+00	1.23E-03	6.85E-07	2.26E-09
31	118	1.88E-22	5.32E-14	3.94E-17	1.31E-33
31	577	1.00E+00	1.52E-26	7.82E-34	6.39E-30
34	43	1.00E+00	9.69E-03	3.03E-01	6.82E-09
34	58	1.00E+00	1.00E+00	1.00E+00	1.17E-04
34	111	1.00E+00	1.00E+00	9.73E-02	2.77E-03
34	118	2.80E-15	3.86E-05	5.25E-05	1.23E-18
34	577	1.00E+00	2.19E-14	3.24E-17	8.09E-13
36	52	2.22E-06	9.93E-02	6.79E-03	6.37E-06
36	70	1.00E+00	8.34E-23	3.47E-24	3.47E-39
36	112	4.51E-01	8.74E-07	2.73E-06	2.37E-11
36	143	3.13E-05	2.72E-09	5.01E-09	1.26E-08
43	52	1.61E-03	5.65E-06	6.87E-03	6.59E-15
43	70	1.00E+00	1.13E-33	1.36E-39	4.04E-66
43	112	1.41E-03	1.05E-02	3.59E-04	5.87E-12
43	143	1.00E+00	1.08E-04	1.69E-07	7.25E-06
52	58	1.30E-13	9.73E-04	3.28E-02	2.68E-12
52	111	1.00E+00	3.56E-02	1.08E-02	1.40E-04
52	118	4.17E-33	4.07E-08	2.86E-07	5.38E-27
52	577	5.81E-20	2.43E-17	3.39E-18	2.11E-23
58	111	1.00E+00	1.00E+00	6.47E-01	2.08E-02
58	118	2.52E-03	9.20E-03	4.91E-03	3.66E-15
58	577	1.00E+00	6.84E-12	2.43E-15	4.37E-15
70	112	5.15E-01	5.92E-12	3.21E-07	3.82E-16
70	143	2.90E-01	3.03E-01	2.42E-01	1.62E-13
111	112	5.61E-01	4.46E-01	6.85E-01	8.46E-05
111	143	1.00E+00	4.56E-03	2.72E-02	1.33E-04
112	118	2.77E-03	1.00E+00	1.00E+00	6.93E-05
112	577	5.07E-05	4.23E-07	2.92E-07	4.15E-05
118	577	4.25E-12	8.22E-04	7.80E-05	1.43E-10

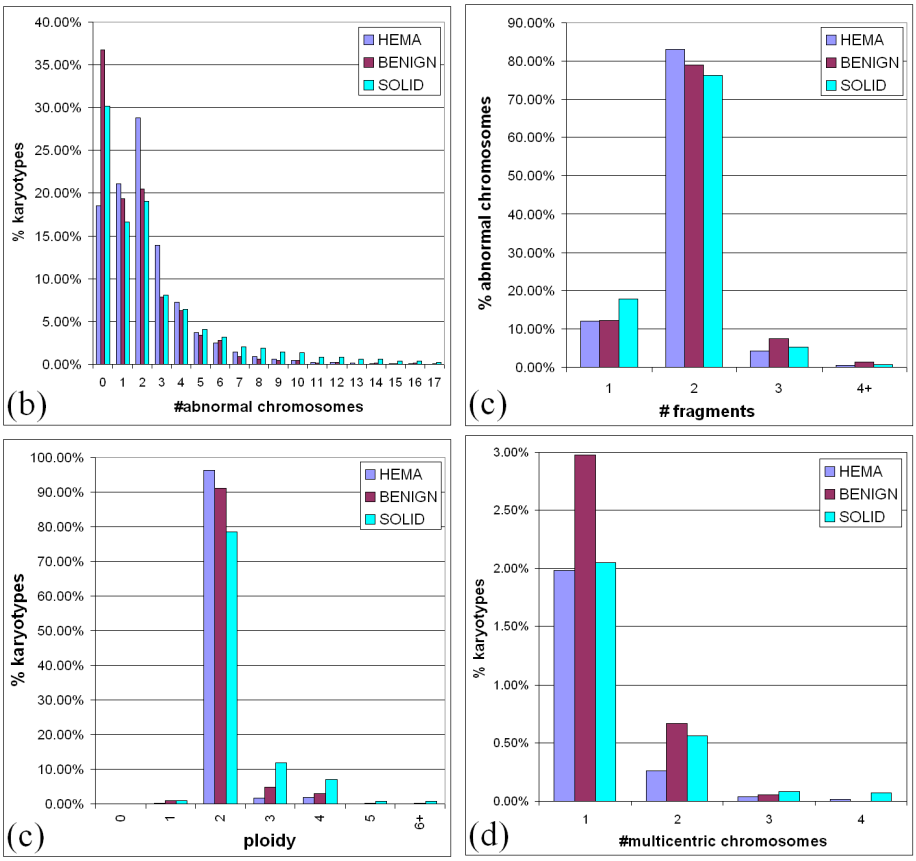

Fig. 5. Basic statistics on karyotype complexity in the Mitelman database. (a) The distribution of the number of abnormal chromosomes per karyotype. (b) The number of fragments per abnormal chromosome. (c) The distribution of karyotype ploidy. (d) The distribution of number of multicentric chromosomes per karyotype. More than 97% of all the karyotypes have no multicentric chromosomes.

(a) An abstract data structure for a karyotype K:

$$Abnormal\_Chrs = \left\{ \begin{array}{c} 18\text{pter} \rightarrow 18\text{q21}::12\text{p11} \rightarrow 12\text{pter}, \\ 1\text{qter} \rightarrow 1\text{p36}::18\text{q21} \rightarrow 18\text{qter}, \\ 14\text{pter} \rightarrow 14\text{q32}::18\text{q21} \rightarrow 18\text{qter}, \\ 18\text{pter} \rightarrow 18\text{q21}::14\text{q32} \rightarrow 14\text{qter} \times 2 \end{array} \right\}$$

$multiplicity[1] = multiplicity[14] = multiplicity[18] = 1$, $multiplicity[i] = 2$ for $i \notin \{1, 14, 18\}$

(b) A sequence of reconstructed events S:

1. ACENTRIC ORPHAN TAIL:	12p11→12pter,
2. CHR GAIN:	18pter→18q21::14q32→14qter,
3. TRANSLOCATION(reciprocal):	14pter→14q32, 14q32→14qter,
4. TRANSLOCATION(non-reciprocal):	18pter→18q21, 18q21→18qter,
5. TAIL DELETION:	1p36→1pter
6. CHR GAIN:	18

(c) The breakpoint graph $G(K)$ (1) and its induced subgraph $G(K, S)$

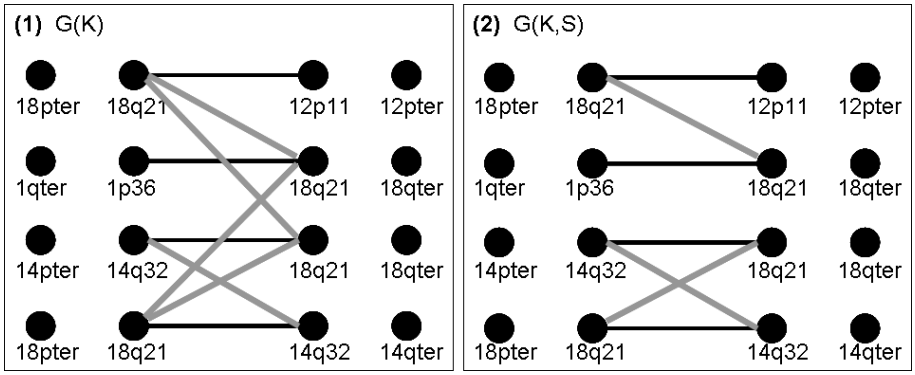

Fig. 6. An analysis of the karyotype in Fig. 1.

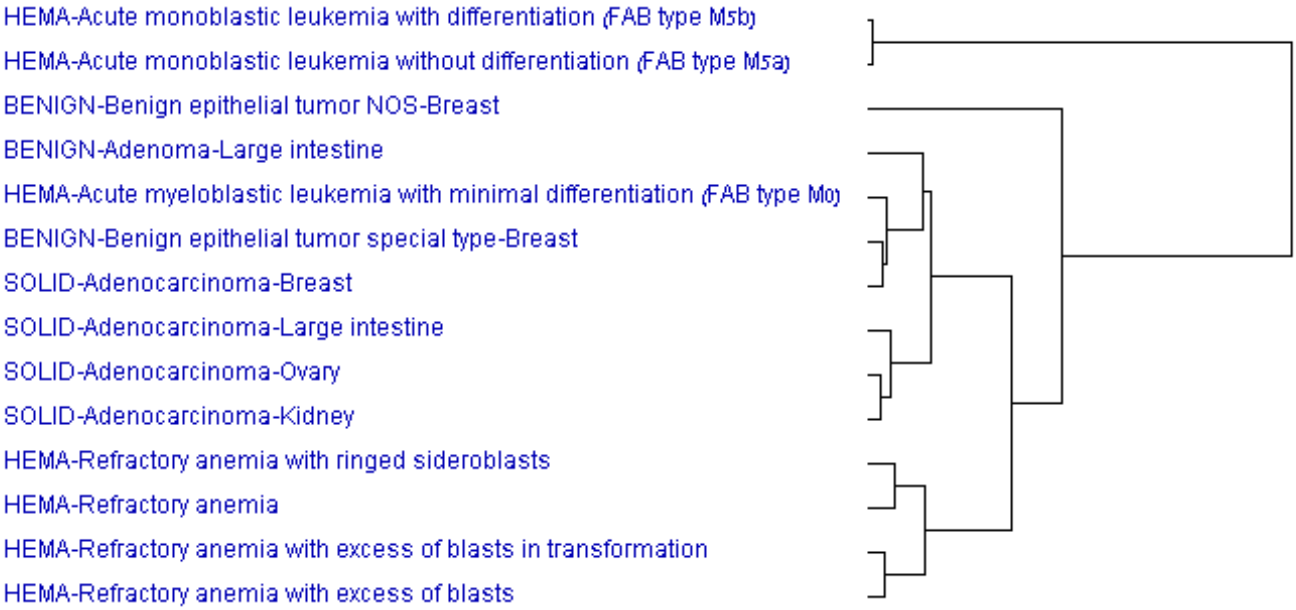

Fig. 7. An hierarchical clustering of different cancer classes based on their average event profiles, using Pearson correlation as similarity function. Each cancer is identified by its category, morphology, and topography (if it is a solid tumor).