Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens

Haoyang Cai[1,2]
Email: haoyang.cai@gmail.com

Nitin Kumar[1,2]
Email: nitbio@gmail.com

Homayoun C. Bagheri[3]
Email: bagheri@ieu.uzh.ch

Christian von Mering[1,2]
Email: mering@imls.uzh.ch

Mark D. Robinson[1,2*]
Email: mark.robinson@imls.uzh.ch

Michael Baudis[1,2*]
Email: mbaudis@imls.uzh.ch

[1] Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

[2] Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

[3] Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

[*] Corresponding authors

Abstract

Background

Chromothripsis is a recently discovered phenomenon of genomic rearrangement, possibly arising during a single genome-shattering event. This could provide an alternative paradigm in cancer development, replacing the gradual accumulation of genomic changes with a “one-off” catastrophic event. However, the term has been used with varying operational definitions, with the minimal consensus being a large number of locally clustered copy number aberrations. The mechanisms underlying these chromothripsis-like patterns (CTLP) and their specific impact on tumorigenesis are still poorly understood.

Results

Here, we identified CTLP in 918 cancer samples, from a dataset of more than 22,000 oncogenomic arrays covering 132 cancer types. Fragmentation hotspots were found to be located on chromosome 8, 11, 12 and 17. Among the various cancer types, soft-tissue tumors exhibited particularly high CTLP frequencies. Genomic context analysis revealed that CTLP rearrangements frequently occurred in genomes that additionally harbored multiple copy number aberrations (CNAs). An investigation into the affected chromosomal regions showed a large proportion of arm-level pulverization and telomere related events, which would be compatible to a number of underlying mechanisms. We also report evidence that these genomic events may be correlated with patient age, stage and survival rate.

Conclusions

Through a large-scale analysis of oncogenomic array data sets, this study characterized features associated with genomic aberrations patterns, compatible to the spectrum of “chromothripsis”-definitions as previously used. While quantifying clustered genomic copy number aberrations in cancer samples, our data indicates an underlying biological heterogeneity behind these chromothripsis-like patterns, beyond a well defined “chromthripsis” phenomenon.

Keywords

Chromothripsis, Human cancer, Array comparative genomic hybridization, SNP array

Background

One consistent hallmark of human cancer genomes are somatically acquired genomic rearrangements, which may result in complex patterns of regional copy number changes [1,2]. These alterations have the potential to interrupt or activate multiple genes, and consequently have been implicated in cancer development [3]. Analysis of genomic rearrangements is essential for understanding the biological mechanisms of oncogenesis and to determine rational points of pharmacological interference [4,5]. Some large-scale efforts have been undertaken to correlate genomic rearrangements to genome architecture as well as to the progression dynamics of cancer genomes [6,7]. At the moment, the stepwise development of cancer with the gradual accumulation of multiple genetic alterations is the most widely accepted model [8].

Recently, using state-of-the-art genome analysis techniques, a phenomenon termed “chromothripsis” was characterized in cancer genomes, defined by the occurrence of tens to hundreds of clustered genomic rearrangements, having arisen in a single catastrophic event [9]. In this model, contiguous chromosomal regions are fragmented into many pieces, via presently unknown mechanisms. These segments are then randomly fused together by the cell’s DNA repair machinery. It has been proposed that this “shattering” and aberrant repair of a multitude of DNA fragments could provide an alternative oncogenetic route [9], in contrast to the step-by-step paradigm of cancer development [8–10]. The initial study reported 24 chromothripsis cases, with some evidence of a high prevalence in bone tumors [9].

Besides human cancers, recent studies have also reported chromothripsis events in germline and non-human genomes [11–13]. However, due to the overall low incidence of this phenomenon, most studies were limited to relatively small numbers of observed events. For example, in a study screening 746 multiple myelomas by SNP arrays, only 10

cases with chromothripsis-like genome patterns were detected [14]. Larger sample numbers are required to gain further insights into features and mechanisms of these events in different cancers.

In contrast to a strict definition of chromothripsis events relying on sequencing based detection of specific genomic rearrangements [15], other studies [7,14,16] have described chromothripsis events based on genomic array analysis without support from whole genome sequencing data. Overall, the minimal consensus of array based studies is the detection of a large number of locally clustered CNA events. In table 1, we provide an overview of studies which so far have reported instances of “chromothripsis” in human cancers [7,9,11,13-14, 16–35].

Here, we present a statistical model for the discovery of clustered genomic aberration patterns, similar to those previously labeled as “chromothripsis” events, from genomic array data sets. For the scope of this article, we introduce the term “chromothripsis-like patterns” (CTLP) when discussing those events.

Applying our methodology to 22,347 genomic arrays from 402 GEO (Gene Expression Omnibus) derived experimental series [36], we were able to detect 918 chromothripsis-like cases, and to determine the frequency and genomic distribution of CTLP events in this dataset. Our collection of oncogenomic array data represents 132 cancer types as defined using the ICD-O 3 (International Classification of Diseases for Oncology) coding scheme, enabling us to estimate the incidence of CTLP in diverse tumor types. Among the CTLP cases, varying distributions of fragmented chromosomal regions as well as an abundance of large non-CTLP copy number aberrations (CNA) regions were found, and the genomic context of chromothripsis-like events was investigated. Finally, we evaluated clinical associations of CTLP cottoning samples, based on the clinical information at hand. Overall, this study characterized heterogeneous features of chromothripsis-like events through a large-scale analysis of oncogenomic

array data sets and provides a better understanding of clustered genomic copy number patterns in cancer development.

Results

Detection of chromothripsis-like patterns from oncogenomic arrays

We collected 402 GEO series, encompassing 22,347 high quality genomic arrays of human cancer samples. A procedure was employed to detect CTLP from these arrays (Figure 1A). The annotated information of the arrays, including normalized probe intensity, segmentation data and quality evaluation, was obtained from our arrayMap database [37] (see Methods for array processing pipeline). After removing technical repeats (e.g. multiple platforms for one sample), a total of 18,394 cases representing 132 cancer types remained. The input data is summarized, at array and case-level, respectively, in Additional file 1: Table S1 and Additional file 2: Table S2. The segmentation data and array profiling can be accessed and visualized through the arrayMap website (http://www.arraymap.org).

According to previous studies, segmental copy number status changes and significant breakpoint clustering are two relevant features of chromothripsis [9,23]. For an automatic identification of CTLP, we developed a scan-statistic based algorithm [38]. We employed a maximum likelihood ratio score, which is commonly used to detect clusters of events in time and/or space and to determine their statistical significance [39] (see Methods). For each chromosome, the algorithm uses a series of sliding windows to identify the genomic region with the highest likelihood ratio as the CTLP candidate. In order to test the performance of the algorithm, 23 previously published chromothripsis cases with available raw array data were collected and used as a training set. This data contained 31 chromothriptic and 475 non-chromothriptic chromosomes that acted as

positive and negative controls, respectively (Additional file 3: Table S3). Comparison of copy number status change times and likelihood ratios showed that chromothriptic chromosomes could reliably be distinguished from non-chromothriptic ones (Additional file 1: Figure S1). We generated a receiver operating characteristic (ROC) curve from the training set results, and selected cutoff values based on this curve (copy number status switch times ≥ 20 and log10 of likelihood ratio ≥ 8) (Figure 1B). Furthermore, the sliding window scan statistic accurately identified the genomic regions involved (Additional file 1: Figure S2). Applying this algorithm to the complete input data set, a total of 1,269 chromosomes from 918 cases passed our thresholds and were marked as CTLP events (Additional file 1: Figure S3, Additional file 4: Table S4).

Chromothripsis-like patterns across diverse tumor types

When evaluating the 1,269 CTLP events, we found a pronounced preference for some chromosomes; this preference showed only limited association with chromosome size (Figure 2A). CTLP occurred more frequently in chromosome 17 than in any other chromosome. This observation is in accordance with data reporting an association between chromothripsis and *TP53* mutations in Sonic-Hedgehog medulloblastoma and acute myeloid leukemia [23]. *TP53* is located in the p arm of chromosome 17, and is involved in cell cycle control, genome maintenance and apoptosis [40,41]. Our dataset showed *TP53* losses in 438 out of 918 (~48%) CTLP cases, compared to 3,274 out of 17,476 (~19%) cases in the non-CTLP group ($p < 2.2\times10^{-16}$; two-tailed Fisher's exact test; Additional file 2: Table S2). 45 of the 438 *TP53* deletions were part of a CTLP, confirming *TP53* mutation as a recurring event with possible involvement in CTLP formation. Other chromosomes with relatively high incidences of CTLP included chromosomes 8, 11 and 12.

In our study, genomic projection of regional CTLP frequencies revealed their heterogeneous distribution in different cancer types (Figure 2B). The total length of fragmented genomic regions (CNA level and interspersed normal segments) accounted for 1%-14% of the corresponding genomes (Figure 2C). The large size of our input data set, resulting in high number of CTLP cases, permitted an investigation of the frequency and genomic distribution of these patterns in different cancer types. Our input samples represented 65 “diagnostic groups”, as defined by a combination of ICD-O morphology and topography codes. The majority of samples (18,238) came from 50 diagnostic groups, each represented by more than 25 arrays. We observed in total of 918 CTLP events across all 18,394 cases, representing an overall ~5% prevalence. The 17 diagnostic groups represented by at least 45 cases, and having frequencies higher than 4% (CTLP high) are listed in Table 2 (full list in Additional file 5: Table S5).

The initial study by Stephens *et al.* hypothesized that chromothripsis has a high incidence in bone tumors [9]. Notably, several soft tissue tumor types appeared in our “CTLP high” frequency set (6 out of 17), including the 3 types with the highest scores. Moreover, the high prevalence of CTLP in soft tissue tumors was reflected in the ICD-O specific frequencies (Additional file 6: Table S6). The genesis and/or effect of multiple localized chromosomal breakage-fusion events may be related to specific molecular mechanisms in those tumor types. Notably, gene fusions are well-documented recurring events in sarcomas [42], in contrast to most other solid tumors, and a local clustering of genomic re-arrangements had been previously reported for liposarcomas [43]. So far, more than 40 fusion genes have been recognized in sarcomas and treated as potential diagnostic and prognostic markers [42]. Possibly, the double-strand breaks and random fragment stitching events in chromothripsis-like events promote the generation of oncogenic fusion genes [9]. Further sequencing-based efforts will be needed to identify the true extent of fusion gene generation and to elucidate their functional impact in

chromothripsis-like cases.

Genomic context of chromothripsis-like events

It has been hypothesized that chromothripsis is a one-off cellular crisis generating a malignant clone in a very short time [9,44]. However, in many of the CTLP samples in our study, highly fragmented chromosomal regions were embedded in larger CNA regions showing variations in patterns and overall extent (Figure 3A). To test whether CTLP generating events are associated with overall genomic instability, we examined the extent of all copy number imbalances detected in our dataset. Comparing the 918 CTLP positive arrays with the remainder of 17,476 CTLP negative arrays, we found that CTLP samples tended to have higher proportions of CNA coverage in their genomes ($p < 2.2\times10^{-16}$; Kolmogorov-Smirnov test) (Figure 3B,C). This indicated that chromothripsis-like events frequently co-occur with other types of copy number aberrations. Plausible and non-exclusive explanations could be that CTLP might frequently arise due to previously established errors in the maintenance of genomic stability, or that chromothriptic aberrations involving genomic maintenance genes may predispose to the acquisition of additional CNA. For those frequent cases exhibiting additional non-CTLP CNA events, their possible contribution to oncogenesis has to be considered when modeling the role of chromothripsis-like events in cancer development.

Potential mechanisms for chromosome shattering

While the mechanism(s) responsible for the generation of chromothripsis remain elusive, a number of studies have proposed hypotheses including ionizing radiation [9], DNA replication stress [45], breakage-fusion-bridge cycles [9,23,46], premature chromosome compaction [47], failed apoptosis [48,49] and micronuclei formation [50]. Some of these

proposed mechanisms are associated with features which could be addressed in our study.

In our dataset, although most (76%) CTLP cases presented single chromosome CTLP events, in approximately 24% CTLP affected at least 2 chromosomes (Figure 4A). For certain candidate mechanisms, e.g. micro-nucleus formation due to mitotic delay, this observation would imply more than one event, whereas the observation appears compatible with e.g. an aborted apoptosis process.

For relating to cytogenetic aberration mechanisms, an additional parameter explored by us was the extent of CTLP regions when normalized to their respective chromosomes. Affected regions were classified into the categories “arm-level” (≥ 90% arm length), “chromosome-level” (≥ 80% chromosome length) or “localized” (Figure 4B). Arm-level CTLP events were observed with a relatively high frequency (~19%). In the arm-level patterns, the CTLP rearrangements were concentrated in one chromosome arm, with the other arm of the same chromosome remaining normal or showing isolated CNA. Since arm-level events involve both peri-centromeric and telomeric regions, cytogenetic events involving these chromosomal structures present themselves as possible causative mechanisms.

Notably, one model that closely conforms to this pattern involves breakage-fusion-bridge cycles [9,23,46,47,51–54]. In general, such cycles start with telomere loss and end-to-end chromosome fusions. When the dicentric chromosomes are formed and pulled to opposite poles during anaphase, a double-strand DNA break acts as starting point for the next cycle. Chromosomal rearrangements would gradually accumulate during the additional cycles, and should be concentrated in one chromosome arm, particularly near the affected telomere. In our dataset, up to 44% of all CTLP chromosomes involved telomere regions. We performed simulations to explore whether this telomere enrichment could be explained by chance. In brief, for each sample, we

retained the location of CTLP region in the genome and shuffled the telomere position of each chromosome while keeping the length of each chromosome constant. In contrast to the actual observations, the simulation did not result in telomeric CTLP enrichment (p < 0.0001; 10,000 simulations; see Methods). CTLP generation through breakage-fusion-bridge cycles would be a viable candidate hypothesis compatible both with the statistically significant telomere enrichment and the high proportion of arm-level pulverization. However, for arm-level CTLP events centromere-related instability mechanisms should also be considered for future discussions.

Clinical implications

Based on clinical associations of "chromothripsis" patterns, it has been claimed that these events may correlate with a poor outcome in the context of the respective tumor type [14,25,55]. In our meta-analysis, we explored a general relation of CTLP with clinical parameters, across the wide range of cancer entities reflected in our input data set. Clinical data was collected from GEO and from the publications of the respective series (Additional file 2: Table S2 and Additional file 1: Table S7) and parameters available for at least 1,000 cases were considered. From our dataset, CTLP seemed to occur at a more advanced patient age as compared to non-CTLP samples (Figure 5A) [23]. CTLP mainly occurred at stage II and III (70%), which was significantly different from the stage distribution of total samples (55.2%) (p = 0.0149; Chi-square test) (Figure 5B). No difference of grade distribution was observed in our dataset (p = 0.425; Chi-square test) where CTLP samples showed a predominance for grades 2 and 3, similar to the bulk of all samples (~80%). We also found that CTLP was overrepresented in cell lines compared to primary tumors ($p < 2.2\times10^{-16}$; two-tailed Fisher's exact test).

For a subset of 1,203 patients, we were able to determine basic follow-up parameters

(follow-up time and survival status). For 72 of these individuals, CTLP was detected in their tumor genomes. Notably, patients with CTLP survived a significantly shorter time than those without this phenomenon (p = 0.0039; log-rank test; Figure 5C). Note that this analysis was based on a sample of convenience averaged over cancers, stages and grades. If we break down this dataset by cancer type, the numbers are not large enough to provide statistical confidence (Additional file 1: Figure S4). While the cancer type independent association of CTLP patterns and poor outcome is intriguing, potential clinical effects of chromothripsis-like genome disruption should be evaluated in larger and clinically more homogeneous data sets.

Sensitivity of array platforms for detection of chromothripsis-like patterns

Presumed chromothripsis events have been reported from genomic datasets generated through different array and sequencing based techniques (see table 1). We performed an analysis of the platform distribution of our CTLP samples, to estimate the detection bias among various genomic array platforms. As the resolution of a platform depends both on type and density of the probes on an array, we divided the platforms into 4 groups according to their probe numbers and techniques (BAC/P1, DNA/cDNA, oligonucleotide ≤ 200K and oligonucleotide > 200K). Although CTLP were detected by all types of genomic arrays, a higher fraction of CTLP samples was found using data from high resolution oligonucleotide arrays (Figure 6), possibly due to increased sensitivity related to higher probe density. Indeed, when performing platform simulations, the sensitivity of CTLP detection improved with increasing probe numbers (Additional file 1: Figures S5 and S6; see Methods). According to these simulations, array platforms consisting of more than 250k probes should be preferred when screening for CTLP events. Since our analysis relied on a variety of array platforms, we can assume that the overall prevalence of CTLP in cancer is higher than our reported 5% of samples.

Discussion

The description of the "chromothripsis" phenomenon has initiated a vital discussion about clustered genomic aberration events and their role in cancer development [52,55,56]. While chromothripsis *senso stricto* has been characterized as a type of focally clustered genomic aberrations generated in a one time cellular event and being limited to a defined set of copy number states [15], other operational definitions have been employed based on clustered aberrations [7,16,23,45,55,57]. It seems likely that some of the previous discussions of "chromothripsis" referred to a number of underlying event types, all resulting in localized genome fragmentation and re-assembly events. For instance, DNA double strand break and end-joining-mediated repair may result in a restricted number of copy number levels, whereas aberrant replication based mechanisms will lead to a more diverse set of copy number aberrations [45,55]. Here, we introduce the term "chromothripsis-like patterns" (CTLP) when referring to clustered genomic events, to accommodate both common labelling and presumed biological variability of clustered genomic copy number aberrations.

At this time, due to the lack of sufficiently large number of cancer data sets from whole-genome sequencing analyses, a meta-analysis of "strict" chromothripsis cases is not feasible. We have followed a pragmatic approach to quantify the occurrence of CTLP from genomic array data sets. In our algorithm, we implemented the two most significant features shared by different operational chromothripsis definitions, namely copy number status changes and breakpoints clustering, which can be well measured by array based technologies. Previous studies provided various algorithms to detect "chromothripsis" events [9,15,58]. However, besides its application to an extensive data set, the specific advantage of our method presented here is its ability to detect regions of shattering with

limited influence from the varying sizes of affected chromosomes. Since the step length of our scanning window is 5 Mb, theoretically the detected CTLP regions are within an accuracy of ±5 Mb. Note that the performance of this algorithm may be influenced by poor quality arrays, especially those with highly scattered and unevenly distributed probe signal intensities.

In this study, we identified 918 CTLP-containing genome profiles, based on an analysis of copy number aberration patterns from 22,347 oncogenomic arrays and representing 132 cancer types. Despite the inherent limitations of such a meta-analysis approach, we were able to provide several new insights regarding the distribution of clustered genomic copy number aberrations and to produce a comprehensive estimate of CTLP incidence in a large range of cancer entities.

In our analysis, CTLP exhibited an uneven distribution along tumor genomes, with disease related local enrichment. These “CTLP dense” chromosomal regions may reveal associations between disease related cancer associated genes and molecular mechanisms behind genome shattering events. This potential correlation is exemplified by the prevalence of mutant *TP53* in “chromothriptic” Li-Fraumeni syndrome associated Sonic-Hedgehog medulloblastomas [23]. As the extent of CTLP related deletions of the *TP53* locus indicates, CTLP related gene dosage changes may predispose to double-hit effects on specific tumor suppressors. In contrast, we found regional enrichment for CTLP with pre-dominant copy number gains on chromosomes 8, 11 and 12. In the initial study, chromosome 8 shattering was found in a small cell lung cancer cell line [9]. This event contained the *MYC* oncogene, which had be shown to be amplified in 10-20% of small cell lung cancers [59]. Moreover, strong overexpression of *MYC* involved in a “chromothripsis” region was also detected in a neuroblastoma sample [25]. In a study of colorectal tumors, chromosomes 8 and 11 were involved in concurrent pulverization events with generation of fusion genes, involving e.g. *SAPS3* and *ZFP91* [18]. In a study

on hepatocellular carcinoma, *CCND1* amplification was embedded within a "chromothriptic" event on chromosome 11 [24]. Therefore, the overall uneven distribution of CTLP may point to specific driver mutations that contribute to CTLP generation, and/or to a class of cancer promoting mutations based on regional genome shattering events.

When comparing cancer types, we observed a high CTLP prevalence in a limited set of entities, particularly in among soft tissue tumors. This finding supports and improves upon a previous prediction of particularly high "chromothripsis" rate in bone tumors [9]. Also, the uneven distribution of CTLP is a strong indicator for a disease related selection of specific genomic aberrations, supporting their involvement in the oncogenetic process.

In the initial study, the authors stated that chromothripsis could be a one-off cataclysmic event that generates multiple concurrent mutations and rearrangements [9]. However, the role of chromothripsis in terms of "shortcut" to cancer genome generation is still elusive. We note that additional and complex non-CTLP genome re-arrangements exist in the majority of CTLP samples. The number and uneven distribution of affected chromosomes in CTLP supports the biological heterogeneity of cancer samples with CTLP containing genome profiles. Furthermore, the normalized spatial distribution of shattered chromosomal regions, as well as the observed significant overlap between telomere and pulverized regions is supportive of breakage-fusion-bridge cycles as one of the mechanisms acting in a subset of samples. Further efforts are needed to investigate the temporal order of chromothripsis and non-chromothripsis events in complex samples, and to substantiate the existence of a dichotomy between "one-off" chromothripsis and other classes of localized genome shattering events, all resulting in clustered genomic copy number aberrations.

In our associated clinical data, CTLP were related to more advanced tumor stages and overall worse prognosis when compared to non-CTLP cases. One possible

explanation is that the numerous concurrent genetic alterations induced by genome shattering events disturb a large number of genes and contribute to more aggressive tumor phenotypes. By themselves, these observations do not differentiate whether CTLP arise as a early events promoting aggressive tumor behavior with fast growth rates and reduced response rates to therapeutic interventions; or whether this observation relates to underlying primary mutations predisposing to genomic instability, aggressive clinical behavior and CTLP as a resulting epiphenomenon. Interestingly, the high rate of *TP53* involvement by itself would support both possibilities for this gene, i.e. chromothripsis as result of *TP53* mutation as well as chromothriptic events with *TP53* locus involvement promoting an aggressive clinical behavior.

From Table 1 we may notice that the array based technologies are, in general, less sensitive than whole-genome sequencing data for calling chromothripsis-like events. This is partly due to the very limited ability of most array platforms to detect balanced genomic aberrations, such as inversions and translocation events. In the future, the accumulation of large-scale sequencing data should be able to provide further insights into localise genome shattering events.

Conclusions

CTLP represent a striking feature occurring in a limited set of cancer genomes, and can be detected from array based copy number screening experiments, using biostatistical methods. The observed clustered genomic copy number aberrations may reflect heterogenous biological phenomena beyond a single class of “chromothripsis” events, and probably vary in their specific impact on oncogenesis. Fragmentation hotspots derived from our large-scale data set could promote the detection of markers associated with genome shattering, or may be used for assigning disease related effects to CTLP-

induced genomic events.

Methods

Genome-wide microarrays and data preparation

In this study, we screened 402 GEO series [36], encompassing 22,347 high quality genomic arrays (Additional file 2: Table S2). All selected arrays were human cancer samples hybridized onto genome-wide array platforms. The normalized probe intensities, segmented data and quality information were obtained from the arrayMap database, which is a publicly available reference database for copy number profiling data [37]. In brief, the annotated data was obtained by the following processing pipeline: for Affymetrix arrays, the aroma.affymetrix R package was employed to generate log2 scale probe level data [60]; for non-Affymetrix arrays, available probe intensity files were processed; CBS (Circular Binary Segmentation) algorithm [61] was performed to obtain segmented copy number data. The probe locations were mapped on the human reference genome (UCSC build hg18). In the case of technical repeats (e.g. one sample was hybridized on multiple platforms), only one of the arrays was considered for analysis (preferably with the highest resolution and/or best overall quality). The array profiling can be visualized through the arrayMap website.

Scan-statistic based chromothripsis-like pattern detection algorithm

To detect chromothripsis-like cases, we formulated an algorithm identifying clustering of copy number status changes in the genome. Several parameters were considered to define the alteration of copy number status:

i) The thresholds of log2 ratio for calling genomic gains and losses. These values

were array specific and stored in arrayMap database. For each array, the thresholds were obtained from related publications or empirically assigned based on the log2 ratio distribution.

ii) The intensity distance between adjacent segments. Due to local correlation effects between probes or the existence of background noise, the segmentation profiles occasionally exhibit subtle striation patterns. This pattern is constituted with a large number of small segments, which is unlikely to be a biological phenomenon. To reduce artificial copy number status change, the distance of signal intensity between adjacent segments was used as a threshold, and defined here as the sum of the absolute values to call gains and losses. If the distance of two adjacent segments differed by less than this threshold, the copy number status change was not considered.

iii) Segment size. The resolution of a platform depends on the density of probes on the array. One of the platforms with the highest density in our dataset is Affymetrix SNP6, which contains 1.8 million polymorphic and non-polymorphic markers with the mean inter-marker distance of 1.7 kb. It provides a practical resolution of 10 to 20 kb. Therefore, in this study, segments smaller than 10 kb were removed.

In order to identify clustering of copy number status changes, a scan-statistic likelihood ratio based on the Poisson model was employed [39]. In our implementation, a fixed-size window was moved along the genome and for each window the likelihood ratio was computed from observed and expected copy number status change times. Let G be the genome represented linearly, and W is a window with fixed size. As the window W moves over G, it defines a collection of zones Z, where $Z \subset G$. Let n_W denotes the observed copy number status change times in window W, and n_G the total number of observed status change in G. μ_W is the expected status change times in window W, and is calculated as $W/G \times n$. The likelihood function is expressed as

$$\lambda = \begin{cases} \left(\frac{n_W}{\mu_W}\right)^{n_W} \left(\frac{n_G - n_W}{n_G - \mu_W}\right)^{n_G - n_W} & if \quad \frac{n_W}{\mu_W} > \frac{n_G - n_W}{n_G - \mu_W} \\ 1 & otherwise \end{cases}$$

This function detects the zone that is most likely to be a cluster.

Due to lack of prior knowledge about the size of *W*, we predefined a series of window sizes from 30 Mb to 247 Mb (Additional file 1: Table S8), which were based on chromosome sizes. The scanning process was repeated for the series of window sizes for each sample. When *W* moved over *G*, the step length was set to 5 Mb, and there was no overlap between different chromosomes in window *W*. In this way, for each genome, the collection of *Z* contained 4,414 windows in various sizes. The window that maximized the likelihood ratio defined the most probable CTLP region. Thus it can detect both the location and the size of the cluster. When analyzing the complete input dataset, the window with the highest likelihood ratio was selected as a candidate of chromothripsis for each chromosome of the 22,347 arrays. The R script for detecting CTLP cases can be provided upon request.

Analysis of fragment enrichment in telomere region

Telomere positions were simulated to test the DNA fragment enrichment. For each case, the CTLP region was kept at its location in the genome. Locations of chromosome terminals were randomly selected while the length of each chromosome was kept. A genomic interval of 5 Mb from the chromosome terminal was considered as the telomere region. The simulation was performed 10,000 times.

Simulation of platform resolution

The 15 Affymetrix SNP6 CTLP chromosomes in the training set were used for simulation.

For each genome, a certain number of probes were randomly chosen from the original probe set. These probes generally represented the profile that the same sample was hybridized on a platform with corresponding resolution. Then the CTLP pattern detection algorithm was applied on the simulated arrays, and the number of cases that passed the thresholds were recorded.

Statistical testing

The significance in the number of CTLP cases with *TP53* loss in comparison to those in non-CTLP cases was assessed using two-tailed Fisher's exact test. We performed a Kolmogorov-Smirnov test to compare the distributions of copy number aberration proportions in the genome between CTLP and the other cases. The Chi-square test was used to assess the significance in the distribution of both patient stage and grade in CTLP and the whole input dataset. The associations between the number of cell lines in CTLP and non-CTLP cases were tested by two-tailed Fisher's exact test. The difference in the survival curves between two subgroups was evaluated by the log-rank test.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

HC, MDR and MB conceived and designed the experiments. HC and NK analyzed the data. HC, NK, HCB, CvM, MDR and MB contributed reagents/materials/analysis tools. All authors contributed to draft the manuscript and approved the final manuscript.

Acknowledgements

The authors would like to thank Henrik Bengtsson and Ni Ai for useful discussions.

References

1. Albertson DG, Collins C, McCormick F, Gray JW: Chromosome aberrations in solid tumors. Nat Genet 2003, 34(4):369–76.

2. Baudis M. (2007). Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. BMC Cancer, 7(1), 226.

3. Yates LR, Campbell PJ: Evolution of the cancer genome. Nature Reviews Genetics 2012, 13(11):795–806.

4. Chen JM, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP: Genomic rearrangements in inherited disease and cancer. Seminars in Cancer Biology 2010, 20(4):222–233.

5. Chin L, Gray JW: Translating insights from the cancer genome into clinical practice. Nature 2008, 452(7187):553–563.

6. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Henry KTM, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M: The landscape of somatic copy-number alteration across human cancers. Nature 2010, 463(7283):899–905.

7. Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ: Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer

genomes. Genome Research 2013, 23(2):217-27.

8. Stratton MR, Campbell PJ, Futreal PA: The cancer genome. Nature 2009, 458(7239):719–724.

9. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, Mclaren S, Lin ML, Mcbride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, Futreal PA, Campbell PJ: Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell 2011, 144:27–40.

10. Kitada K, Taima A, Ogasawara K, Metsugi S, Aikawa S: Chromosome-specific segmentation revealed by structural analysis of individually isolated chromosomes. Genes Chromosom. Cancer 2011, 50:217–27.

11. Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR, Mclaughlan CJ, Bawden CS, Reid SJ, Faull RLM, Snell RG, Hall IM, Shen Y, Ohsumi TK, Borowsky ML, Daly MJ, Lee C, Morton CC, Macdonald ME, Gusella JF, Talkowski ME: Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat Genet 2012, 44(4):390–397.

12. Deakin JE, Bender HS, Pearse AM, Rens W, O'brien PCM, Ferguson-Smith MA, Cheng Y, Morris K, Taylor R, Stuart A, Belov K, Amemiya CT, Murchison EP, Papenfuss AT, Graves JAM: Genomic Restructuring in the Tasmanian Devil Facial Tumour: Chromosome Painting and Gene Mapping Provide Clues to Evolution of a Transmissible Tumour. PLoS Genet 2012, 8(2):e1002483.

13. Kloosterman WP, Guryev V, Roosmalen MV, Duran KJ, Bruijn ED, Bakker SCM, Letteboer T, Nesselrooij BV, Hochstenbach R, Poot M, Cuppen E: Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Human Molecular Genetics 2011, 20(10):1916–1924.

14. Magrangeas F, Avet-Loiseau H, Munshi NC, Minvielle S: Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. Blood 2011, 118(3):675–678.

15. Korbel JO, Campbell PJ: Criteria for Inference of Chromothripsis in Cancer Genomes. Cell 2013, 152(6):1226–1236.

16. Northcott PA, Shih DJH, Peacock J, Garzia L, Morrissy AS, Zichner T, Stuetz AM, Korshunov A, Reimand J, Schumacher SE: Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature 2012, 488(7409):49–56.

17. Le LP, Nielsen GP, Rosenberg AE, Thomas D, Batten JM, Deshpande V, Schwab J, Duan Z, Xavier RJ, Hornicek FJ, Iafrate AJ: Recurrent Chromosomal Copy Number Alterations in Sporadic Chordomas. PLoS ONE 2011, 6(5):e18846.

18. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A, Jing R, Parkin M, Pugh T, Verhaak RG, Stransky N, Boutin AT, Barretina J, Solit DB, Vakiani E, Shao W, Mishina Y, Warmuth M, Jimenez J, Chiang DY, Signoretti S, Kaelin WG, Spardy N, Hahn WC, Hoshida Y, Ogino S, Depinho RA, Chin L, Garraway LA, Fuchs CS, Baselga J, Tabernero J, Gabriel S, Lander ES, Getz G, Meyerson M: Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet 2011, 43(10):964–968.

19. Kloosterman WP, Hoogstraat M, Paling O, Tavakoli-Yaraki M, Renkens I, Vermaat JS, van Roosmalen MJ, van Lieshout S, Nijman IJ, Roessingh W, van 't Slot R, van de Belt J, Guryev V, Koudijs M, Voest E, Cuppen E: Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biology 2011, 12(10):R103.

20. Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, Easton J, Chen X, Wang J, Rusch M, Lu C, Chen SC, Wei L, Collins-Underwood JR, Ma J, Roberts KG, Pounds SB, Ulyanov A, Becksfort J, Gupta P, Huether R, Kriwacki RW, Parker M, Mcgoldrick DJ, Zhao D, Alford D, Espy S, Bobba KC, Song G, Pei D, Cheng C, Roberts S, Barbato MI, Campana D, Coustan-Smith E, Shurtleff SA, Raimondi SC, Kleppe M, Cools J, Shimano KA, Hermiston ML, Doulatov S, Eppert K, Laurenti E, Notta F, Dick JE, Basso G, Hunger SP, Loh ML, Devidas M, Wood B, Winter S, Dunsmore KP, Fulton RS, Fulton LL, Hong X, Harris CC, Dooling DJ, Ochoa K, Johnson KJ, Obenauer JC, Evans WE, Pui CH, Naeve CW, Ley TJ, Mardis ER, Wilson RK, Downing JR, Mullighan CG: The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature 2012, 481(7380):157–163.

21. Kitada K, Aida S, Aikawa S: Coamplification of Multiple Regions of Chromosome 2, Including MYCN, in a Single Patchwork Amplicon in Cancer Cell Lines. Cytogenet Genome Res 2012, 136:30–37.

22. Poaty H, Coullin P, Peko JF, Dessen P, Diatta AL, Valent A, Leguern E, Prévot S, Gombé-Mbalawa C, Candelier JJ, Picard JY, Bernheim A: Genome-Wide High-Resolution aCGH Analysis of Gestational Choriocarcinomas. PLoS ONE 2012, 7:e29426.

23. Rausch T, Jones DTW, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, Jäger

N, Remke M, Shih D, Northcott PA, Pfaff E, Tica J, Wang Q, Massimi L, Witt H, Bender S, Pleier S, Cin H, Hawkins C, Beck C, Deimling AV, Hans V, Brors B, Eils R, Scheurlen W, Blake J, Benes V, Kulozik AE, Witt O, Martin D, Zhang C, Porat R, Merino DM, Wasserman J, Jabado N, Fontebasso A, Bullinger L, Rücker FG, Döhner K, Döhner H, Koster J, Molenaar JJ, Versteeg R, Kool M, Tabori U, Malkin D, Korshunov A, Taylor MD, Lichter P, Pfister SM, Korbel JO: Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. Cell 2012, 148(1-2):59–71.

24. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, Carnevali P, Stinson J, Johnson S, Diao J, Yeung S, Jubb A, Ye W, Wu TD, Kapadia SB, Sauvage FJD, Gentleman RC, Stern HM, Seshagiri S, Pant KP, Modrusan Z, Ballinger DG, Zhang Z: The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Research 2012, 22(4):593–601.

25. Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, Valentijn LJ, van der Ploeg I, Hamdi M, van Nes J, Westerman BA, van Arkel J, Ebus ME, Haneveld F, Lakeman A, Schild L, Molenaar P, Stroeken P, van Noesel MM, Øra I, Santo EE, Caron HN, Westerhout EM, Versteeg R: Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature 2012, 483(7391): 589–593.

26. Lapuk AV, Wu C, Wyatt AW, Mcpherson A, Mcconeghy BJ, Brahmbhatt S, Mo F, Zoubeidi A, Anderson S, Bell RH, Haegert A, Shukin R, Wang Y, Fazli L, Hurtado-Coll A, Jones EC, Hach F, Hormozdiari F, Hajirasouliha I, Boutros PC, Bristow RG, Zhao Y, Marra MA, Fanjul A, Maher CA, Chinnaiyan AM, Rubin MA, Beltran H, Sahinalp SC, Gleave ME, Volik SV, Collins CC: From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in

prostate cancer. J. Pathol. 2012, 227(3):286–297.

27. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, Zhang H, Zeid R, Ren X, Cibulskis K, Sivachenko AY, Wagle N, Sucker A, Sougnez C, Onofrio R, Ambrogio L, Auclair D, Fennell T, Carter SL, Drier Y, Stojanov P, Singer MA, Voet D, Jing R, Saksena G, Barretina J, Ramos AH, Pugh TJ, Stransky N, Parkin M, Winckler W, Mahan S, Ardlie K, Baldwin J, Wargo J, Schadendorf D, Meyerson M, Gabriel SB, Golub TR, Wagner SN, Lander ES, Getz G, Chin L, Garraway LA: Melanoma genome sequencing reveals frequent PREX2 mutations. Nature 2012, 485(7399): 502–506.

28. Natrajan R, Mackay A, Lambros MB, Weigelt B, Wilkerson PM, Manie E, Grigoriadis A, A'hern R, Groep PVD, Kozarewa I, Popova T, Mariani O, Turajlic S, Furney SJ, Marais R, Rodruigues DN, Flora AC, Wai P, Pawar V, Mcdade S, Carroll J, Stoppa-Lyonnet D, Green AR, Ellis IO, Swanton C, Diest PV, Delattre O, Lord CJ, Foulkes WD, Vincent-Salomon A, Ashworth A, Stern MH, Reis-Filho JS: A whole-genome massively parallel sequencing analysis of BRCA1 mutant oestrogen receptor-negative and -positive breast cancers. J. Pathol. 2012, 227:29– 41.

29. Nik-Zainal S, Loo PV, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, Mclaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, Mcbride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jönsson G, Mariani O, Boyault S, Miron P, Fatima A, Langerød A, Aparicio SAJR, Tutt A, Sieuwerts AM, Borg Å, Thomas G, Salomon AV, Richardson AL, Børresen-Dale AL, Futreal PA, Stratton MR, Campbell PJ, of the International Cancer Genome Consortium

BCWG: The Life History of 21 Breast Cancers. Cell 2012, 149(5):994–1007.

30. Kloosterman WP, Tavakoli-Yaraki M, Roosmalen MJV, Binsbergen EV, Renkens I, Duran K, Ballarati L, Vergult S, Giardino D, Hansson K, Ruivenkamp CAL, Jager M, Haeringen AV, Ippel EF, Haaf T, Passarge E, Hochstenbach R, Menten B, Larizza L, Guryev V, Poot M, Cuppen E: Constitutional Chromothripsis Rearrangements Involve Clustered Double-Stranded DNA Breaks and Nonhomologous Repair Mechanisms. Cell Reports 2012, 1(6):648–655.

31. Wu C, Wyatt AW, Mcpherson A, Lin D, Mcconeghy BJ, Mo F, Shukin R, Lapuk AV, Jones SJM, Zhao Y, Marra MA, Gleave ME, Volik SV, Wang Y, Sahinalp SC, Collins CC: Poly-gene fusion transcripts and chromothripsis in prostate cancer. Genes Chromosom. Cancer 2012, 51(12):1144–1153.

32. Jones DTW, Jäger N, Kool M, Zichner T, Hutter B, Sultan M, Cho YJ, Pugh TJ, Hovestadt V, Stütz AM, Rausch T, Warnatz HJ, Ryzhova M, Bender S, Sturm D, Pleier S, Cin H, Pfaff E, Sieber L, Wittmann A, Remke M, Witt H, Hutter S, Tzaridis T, Weischenfeldt J, Raeder B, Avci M, Amstislavskiy V, Zapatka M, Weber UD, Wang Q, Lasitschka B, Bartholomae CC, Schmidt M, Kalle CV, Ast V, Lawerenz C, Eils J, Kabbe R, Benes V, Sluis PV, Koster J, Volckmann R, Shih D, Betts MJ, Russell RB, Coco S, Tonini GP, Schüller U, Hans V, Graf N, Kim YJ, Monoranu C, Roggendorf W, Unterberg A, Herold-Mende C, Milde T, Kulozik AE, Deimling AV, Witt O, Maass E, Rössler J, Ebinger M, Schuhmann MU, Frühwald MC, Hasselblatt M, Jabado N, Rutkowski S, Bueren AOV, Williamson D, Clifford SC, Mccabe MG, Collins VP, Wolf S, Wiemann S, Lehrach H, Brors B, Scheurlen W, Felsberg J, Reifenberger G, Northcott PA, Taylor MD, Meyerson M, Pomeroy SL, Yaspo ML, Korbel JO, Korshunov A, Eils R, Pfister SM, Lichter P: Dissecting the genomic complexity underlying medulloblastoma. Nature 2012, 488(7409): 100–105.

33. Stevens-Kroef M, Weghuis DO, Croockewit S, Derksen L, Hooijer J, Elidrissi-Zaynoun N, Siepman A, Simons A, Kessel AGV: High detection rate of clinically relevant genomic abnormalities in plasma cells enriched from patients with multiple myeloma. Genes Chromosom. Cancer 2012, 51(11):997–1006.

34. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, Maher CA, Fulton R, Fulton L, Wallis J, Chen K, Walker J, Mcdonald S, Bose R, Ornitz D, Xiong D, You M, Dooling DJ, Watson M, Mardis ER, Wilson RK: Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. Cell 2012, 150(6):1121–1134.

35. Zehentner BK, Hartmann L, Johnson KR, Stephenson CF, Chapman DB, Baca MED, Wells DA, Loken MR, Tirtorahardjo B, Gunn SR, Lim L: Array-Based Karyotyping in Plasma Cell Neoplasia After Plasma Cell Enrichment Increases Detection of Genomic Aberrations. American Journal of Clinical Pathology 2012, 138(4):579–589.

36. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A: NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Research 2013, 41(D1):D991–D995.

37. Cai H, Kumar N, Baudis M: arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies. PLoS ONE 2012, 7(5):e36944.

38. Naus JI: The distribution of the size of the maximum cluster of points on a line. Journal of the American Statistical Association 1965, 60:532–538.

39. Kulldorff M: A spatial scan statistic. Commun statist 1997, 26(6):1481–1496.

40. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA:

COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research 2011, 39(Database):D945–D950.

41. Vogelstein DLB, Levine AJ: Surfing the p53 network. Nature 2000, 408:307–10.

42. Mitelman F, Johansson B, Mertens F: The impact of translocations and gene fusions on cancer causation. Nature Reviews Cancer 2007, 7(4):233–245.

43. Taylor BS, Barretina J, Maki RG, Antonescu CR, Singer S, Ladanyi M: Advances in sarcoma genomics and new therapeutic targets. Nat Rev Cancer 2011, 11(8): 541-57.

44. Maher CA, Wilson RK: Chromothripsis and Human Disease: Piecing Together the Shattering Process. Cell 2012, 148(1-2):29–32.

45. Liu P, Erez A, Nagamani SCS, Dhar SU, Kołodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, Bacino CA, Campos-Acevedo LD, Delgado MR, Freedenberg D, Garnica A, Grebe TA, Hernández-Almaguer D, Immken L, Lalani SR, Mclean SD, Northrup H, Scaglia F, Strathearn L, Trapane P, Kang SHL, Patel A, Cheung SW, Hastings PJ, Stankiewicz P, Lupski JR, Bi W: Chromosome Catastrophes Involve Replication Mechanisms Generating Complex Genomic Rearrangements. Cell 2011, 146(6):889–903.

46. Sorzano COS, Pascual-Montano A, de Diego AS, Martinez-A C, van Wely KH: Chromothripsis: Breakage-fusion-bridge over and over again. Cell Cycle 2013, 12(13):1–8.

47. Meyerson M, Pellman D: Cancer Genomes Evolve by Pulverizing Single Chromosomes. Cell 2011, 144:9–10.

48. Fullwood MJ, Lee J, Lin L, Li G, Huss M, Ng P, Sung WK, Shenolikar S: Next-Generation Sequencing of Apoptotic DNA Breakpoints Reveals Association with Actively Transcribed Genes and Gene Translocations. PLoS ONE 2011,

6(11):e26054.

49. Tubio XEJ: When catastrophe strikes a cell. Nature 2011, 24:476–477.

50. Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, Pan Y, Nezi L, Protopopov A, Chowdhury D, Pellman D: DNA breaks and chromosome pulverization from errors in mitosis. Nature 2012, 482(7383):53–58.

51. Artandi SE, Depinho RA: Telomeres and telomerase in cancer. Carcinogenesis 2010, 31:9–18.

52. Holland AJ, Cleveland DW: Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. Nat Med 2012, 18(11):1630– 1638.

53. McClintock B: The Production of Homozygous Deficient Tissues with Mutant Characteristics by Means of the Aberrant Mitotic Behavior of Ring-Shaped Chromosomes. Genetics 1938, 23:315–76.

54. McClintock B: The Stability of Broken Ends of Chromosomes in Zea Mays. Genetics 1941, 26:234–82.

55. Forment JV, Kaidi A, Jackson SP: Chromothripsis and cancer: causes and consequences of chromosome shattering. Nature Reviews Cancer 2012, 12(10): 663–670.

56. Jones MJK, Jallepalli PV: Chromothripsis: Chromosomes in Crisis. Developmental Cell 2012, 23(5):908–917.

57. Malhotra A, Lindberg M, Faust GG, Leibowitz ML, Clark RA, Layer RM, Quinlan AR, Hall IM: Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. Genome Research 2013, 23(5):762–776.

58. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, Getz G, Meyerson M, Beroukhim R: Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013, 45(10): 1134-1140.

59. Sher T, Dy GK, Adjei AA: Small Cell Lung Cancer. Mayo Clinic Proceedings 2008, 83(3):355– 367.

60. Bengtsson H, Simpson K, Bullard J, Hansen K: aroma.affymetrix: A genetic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Tech Report #745 Department of Statistics, University of California, Berkeley 2008.

61. Olshen AB, Venkatraman ES, Lucito R, Wigler M: Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004, 5(4): 557–572.

Figures

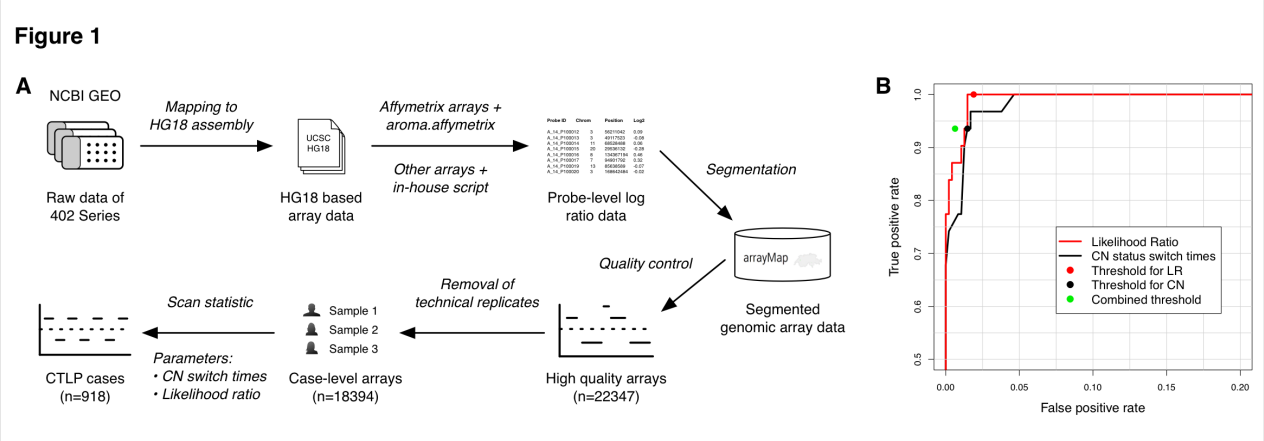

Figure 1 Detection of chromothripsis-like patterns from genomic arrays. (A) Schematic description of the detection procedure. Raw array data of 402 GEO series are first collected and re-analyzed, then annotated and stored in arrayMap database. For high quality arrays, a scan-statistic based algorithm was employed to identify CTLP cases. (B) The ROC curve of the training set and selected thresholds. Two predictors were tested, copy number status change times and the likelihood ratio. Both predictors were integrated into the combined threshold.

Figure 2

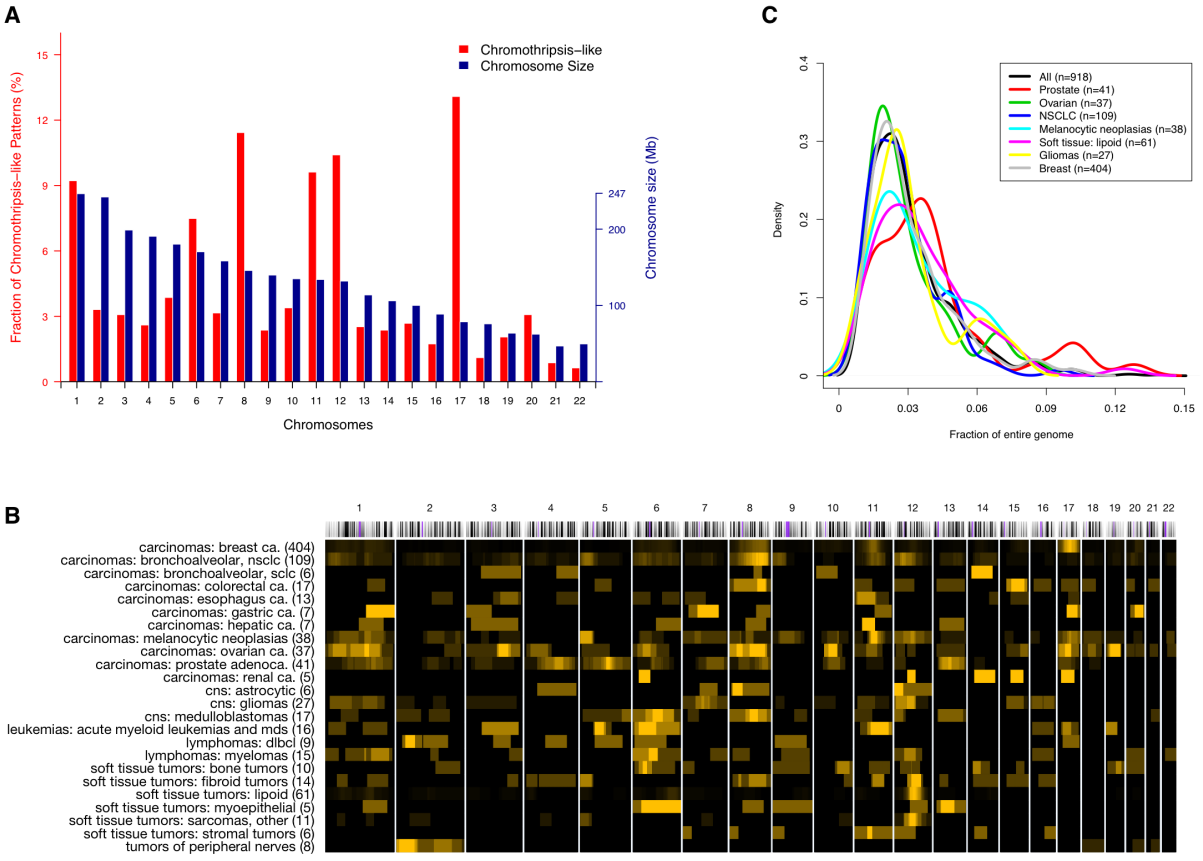

Figure 2 Frequency and CNA coverage length distribution of CTLP regions in the genome. (A) Red and blue bars indicate CTLP frequency in percent of all CTLP and chromosome size in megabases, respectively. (B) Local distribution of CTLP regions among diagnostic groups. Each row represents a cancer type and each column represents a chromosome. We use a black-to-yellow gradient for representing CTLP frequencies ranging from lowest to highest, normalized for each row. The numbers in brackets indicate the number of cases. Groups with at least 5 CTLP cases are shown. (C) Distribution of CTLP events as fractions of the affected genomes, represented as density plots for common cancer types. The fractions for each sample have been calculated as sum of genome bases chromosomes 1-22, divided by the genomic length of CTLP regions as identified through our scanning approach.

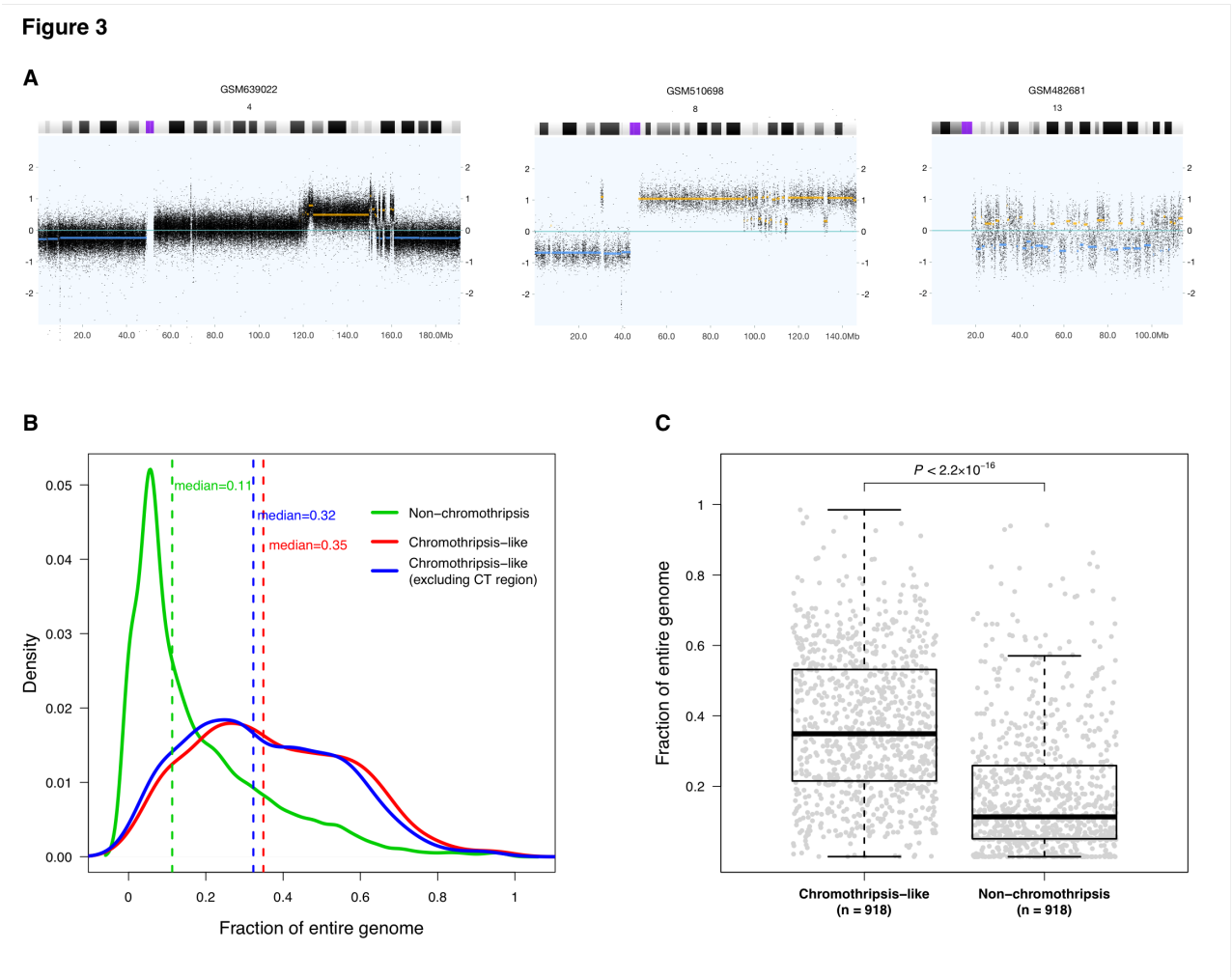

Figure 3 Genomic context of CTLP events. (A) Example copy number profiles of chromosomes with changes suggestive of chromothripsis. In these examples, chromosomal fragmentation events are related to other types of copy number aberrations, and exhibit different combination patterns. The x-axis indicates genomic locations in Mb, and the y-axis is the log2 value of probe signal intensity. Yellow and blue lines represent called genomic gains and losses respectively. (B) Distribution of CNAs as fraction of the genome, compared between CTLP and non-CTLP cases. CT, chromothripsis-like. (C) Distribution of CNA fractions for individual samples. For the non-CTLP group, 918 samples were randomly chosen from the total set of 17,476 cases, to generate an equally sized comparison. The *p*-value, indicating significant difference between the genome fraction distributions of two groups, is based on a Kolmogorov-Smirnov test. The fractions for each sample have been calculated as sum of genome bases chromosomes 1-22 divided by the sum of all CNAs in the sample (with and without CTLP regions).

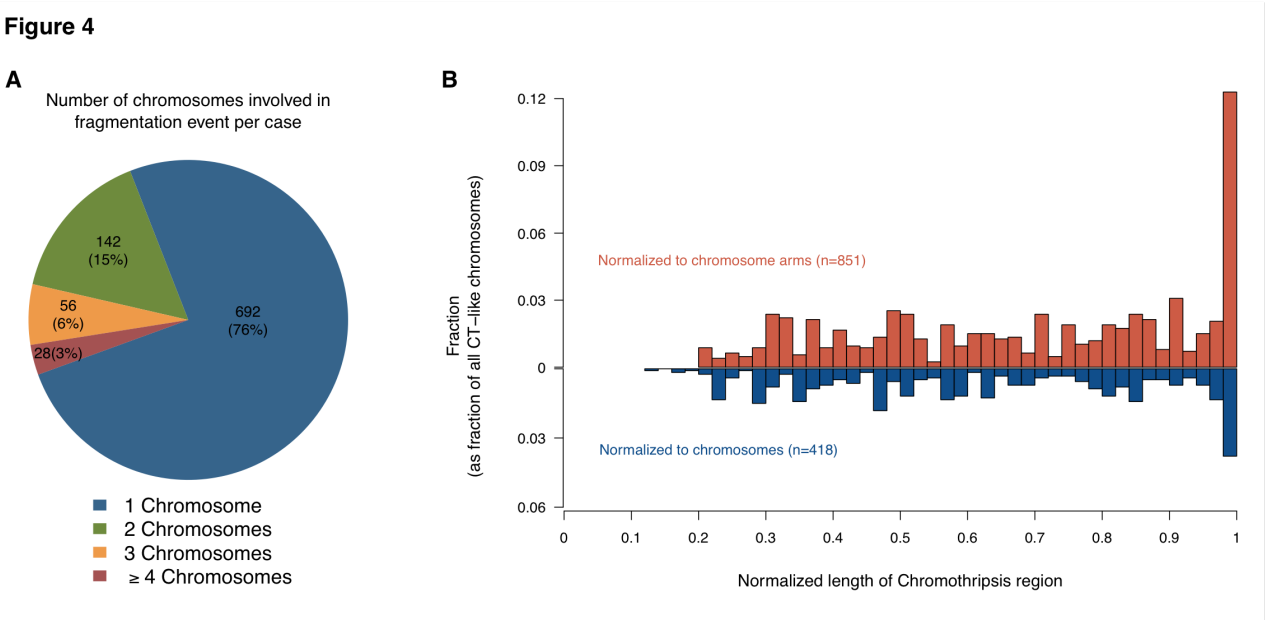

Figure 4 The distribution of CTLP regions in terms of chromosome number and length. (A) The number of chromosomes affected by CTLP per sample. The numbers outside and inside the brackets are number and percentage of CTLP samples respectively. (B) Length distribution of CTLP regions normalized to chromosome or chromosome arm lengths. For each chromosome, regions restricted on a single arm were normalized to arm lengths (red bars), otherwise were normalized to chromosome lengths (blue bars). More than 10% of all CTLP events involve whole chromosome arms.

Figure 5

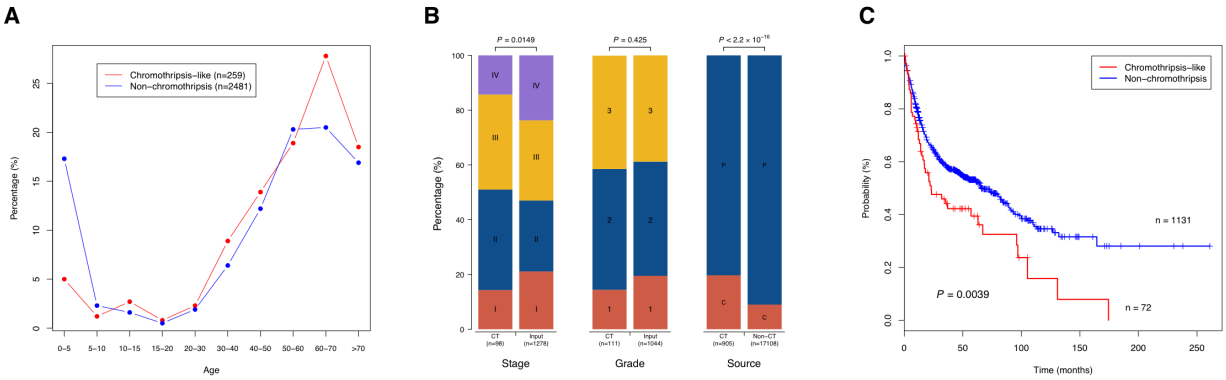

Figure 5 Clinical perspective on CTLP events. (A) Distribution of CTLP percentage versus patient age. (B) Distribution of sample stage, grade and source between CTLP and input dataset or non-CTLP cases. *p*-values are derived from Chi-square test (stage and grade) or two tailed Fisher's exact test (source). P, primary tumor; C, cell line. (C) Kaplan-Meier survival curves for CTLP versus non-CTLP cases. The *p*-value is based on log-rank test.

Figure 6

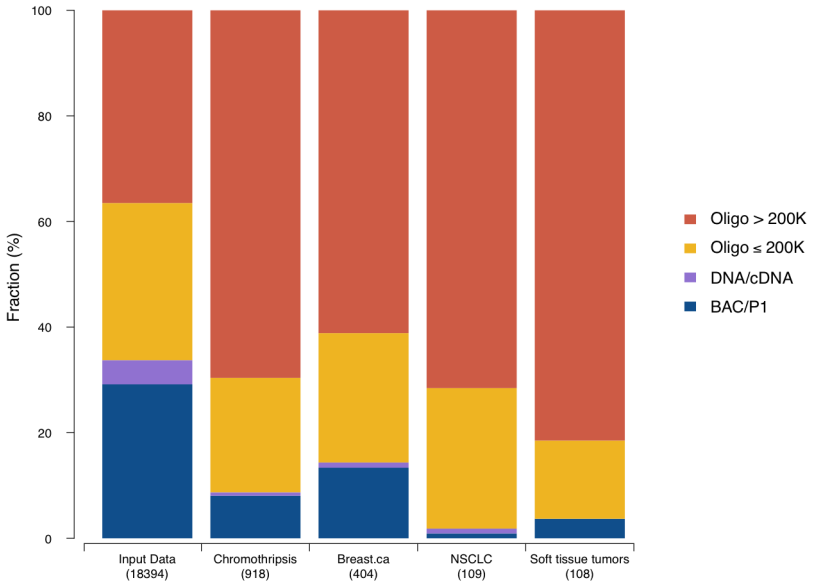

Figure 6 Platform distribution based on different resolutions and technology types. Different analysis groups are shown, including the whole input dataset, inferred CTLP cases and three cancer types. The larger fraction of high-density oligonucleotide arrays in samples with CTLP compared to the overall platform distribution indicates an increased sensitivity of these platforms for CTLP events. Oligo, oligonucleotide; NSCLC, Non-small cell lung cancer.

Table 1 Summary of chromothripsis-like cases identified in previous and current studies

Study[a]	Chromothripsis-like cases[b]	Sample size	Techniques	Cancer/sample types
Stephens et al. [9]	24	776	paired-end sequencing, SNP array	55 cancer types[d]
Kloosterman et al. [13]	1	[c] 3	mate-pair sequencing	germline, congenital defects
Le et al. [17]	1	21	aCGH	chordoma
Magrangeas et al. [14]	10	764	SNP array	multiple myeloma
Bass et al. [18]	3	9	whole-genome sequencing	colorectal adenocarcinoma
Kloosterman et al. [19]	4	4	mate-pair sequencing, SNP array	colorectal cancer
Zhang et al. [20]	3	12	whole-genome sequencing	acute lymphoblastic leukaemia
Kitada et al. [21]	5	150	aCGH	na
Poaty et al. [22]	1	14	aCGH	gestational choriocarcinoma
Rausch et al. [23]	52	605	whole-genome sequencing, SNP array	7 cancer types
Jiang et al. [24]	1	4	paired-end sequencing	hepatocellular carcinoma
Molenaar et al. [25]	16	87	paired-end sequencing, SNP array	neuroblastoma
Chiang et al. [11]	2	52	whole-genome sequencing, aCGH	germline
Lapuk et al. [26]	1	6	whole-genome/ transcriptome sequencing, aCGH	neuroendocrine prostate cancer
Berger et al. [27]	2	25	whole-genome sequencing	melanoma
Natrajan et al. [28]	1	2	whole-genome sequencing	breast cancer
Nik-Zainal et al. [29]	3	21	whole-genome sequencing	breast cancer
Kloosterman et al. [30]	10	10	mate-pair sequencing, SNP array	congenital disease
Wu et al. [31]	3	3	paired-end sequencing, aCGH	prostate cancer
Northcott et al. [16]	na	1087	SNP array, whole-genome sequencing	medulloblastoma
Jones et al. [32]	2	3	whole-genome sequencing	medulloblastoma
Kroef et al. [33]	1	61	SNP array	multiple myeloma
Govindan et al. [34]	1	17	whole-genome sequencing	non-small cell lung cancer
Kim et al. [7]	124	8227	aCGH, SNP array	30 cancer types
Zehentner et al. [35]	1	28	aCGH	plasma cell neoplasia
Current study	918	18394	aCGH, SNP array	132 cancer types[e]

[a] Data up to 21st December, 2012, [b] na, not available, [c] Family trio: father, mother, child, [d] According to site and histology, [e] Classified by ICD-O code

Table 2 Frequency of chromothripsis-like patterns among cancer types

Cancer type	Chromothripsis-like cases				Input cases	Frequency (95% confidence interval)
	Oligo > 200K	Oligo ≤ 200K	BAC or cDNA	Total		
Soft tissue tumors: lipoid	49	12	0	61	114	53.5% (44%-62.8%)
Soft tissue tumors: fibroid tumors	14	0	0	14	59	23.7% (14%-36.9%)
Soft tissue tumors: sarcomas, other	9	0	2	11	48	22.9% (12.5%-37.7%)
Carcinomas: breast ca.	247	99	58	404	3652	11.1% (10.1%-12.1%)
Carcinomas: esophagus ca.	13	0	0	13	135	9.6% (5.4%-16.2%)
Carcinomas: bronchoalveolar, NSCLC	78	29	2	109	1164	9.4% (7.8%-11.2%)
Soft tissue tumors: bone tumors	7	3	0	10	123	8.1% (4.2%-14.8%)
Carcinomas: bronchoalveolar, SCLC	3	3	0	6	90	6.7% (2.7%-14.5%)
Carcinomas: prostate adenoca.	1	40	0	41	653	6.3% (4.6%-8.5%)
CNS: CNS PNET	4	0	0	4	65	6.2% (2%-15.8%)
Carcinomas: melanocytic neoplasias	31	1	6	38	621	6.1% (4.4%-8.4%)
Soft tissue tumors: myoepithelial	3	0	2	5	85	5.9% (2.2%-13.8%)
Carcinomas: ovarian ca.	31	5	1	37	801	4.6% (3.3%-6.4%)
Carcinomas: gastric ca.	1	5	1	7	160	4.4% (1.9%-9.2%)
CNS: gliomas	14	13	0	27	669	4% (2.7%-5.9%)
Soft tissue tumors: stromal tumors	5	1	0	6	151	4% (1.6%-8.8%)
CNS: medulloblastomas	13	4	0	17	430	4% (2.4%-6.4%)

Only cancer types with input cases ≥ 45 and frequency ≥ 4% are shown

Additional Files

Additional file 1: Supplementary figures and tables.

Figure S1: Scatter plot of the training set.

Figure S2: The positive training set and CTLP detection algorithm performances.

Figure S3: Scatter plot of CTLP candidates.

Figure S4: Kaplan-Meier survival curves for CTLP versus non-CTLP cases in specific cancer types.

Figure S5: An example of the platform resolution based simulation using data from an Affymetrix SNP6 array (~1.8 million probes).

Figure S6: CTLP detection sensitivity of simulated platform resolutions.

Table S1: Overview of input dataset.

Table S7: Demographic and clinico-pathologic characteristics of input and CTLP samples.

Table S8: Sizes of sliding windows for the scan-statistic based algorithm.

Additional file 2:

Table S2. Input dataset of high quality arrays with probe level raw data.

Additional file 3:

Table S3. Training set collected from published chromothripsis cases.

Additional file 4:

Table S4. Detected chromothripsis-like cases.

Additional file 5:

Table S5. Frequency of chromothripsis-like pattern among diagnostic groups.

Additional file 6:

Table S6. Frequency of chromothripsis-like pattern among ICD-O codes.