Received: 5 November 2018 | Accepted: 23 November 2018

DOI: 10.1002/gcc.22713

WILEY

RESEARCH ARTICLE

Cancer chromosome breakpoints cluster in gene-rich genomic regions

Felix Mitelman[1] | Bertil Johansson[1,2] | Fredrik Mertens[1,2] | Tommy Schyman[3] | Nils Mandahl[1]

[1]Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, Lund, Sweden

[2]Division of Laboratory Medicine, Department of Clinical Genetics and Pathology, Lund, Sweden

[3]Clinical Studies Sweden – Forum South, Skåne University Hospital, Lund, Sweden

Correspondence
Felix Mitelman, Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, SE 221 24 Lund, Sweden.
Email: felix.mitelman@med.lu.se
and
Nils Mandahl, Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, SE 221 24 Lund, Sweden.
Email: nils.mandahl@med.lu.se

Funding information
The Swedish Childhood Cancer Foundation, Grant/Award Number: PR2017-0031; The Swedish Cancer Society, Grant/Award Number: CAN 2015/248

Abstract

Cancer cells are characterized by chromosome abnormalities, of which some, in particular balanced rearrangements, are associated with distinct tumor entities and/or with specific gene rearrangements that represent important steps in the carcinogenic process. However, the vast majority of cytogenetically detectable structural aberrations in cancer cells have not been characterized at the nucleotide level; hence, their importance and functional consequences are unknown. By ascertaining the chromosomal breakpoints in 22 344 different clonal structural chromosome abnormalities identified in the karyotypes of 49 626 cases of neoplastic disorders we here show that the distribution of breakpoints is strongly associated ($P < 0.0001$) with gene content within the affected chromosomal bands. This association also remains highly significant in separate analyses of recurrent and nonrecurrent chromosome abnormalities as well as of specific subtypes of cancer ($P < 0.0001$ for all comparisons). In contrast, the impact of band length was negligible. The breakpoint distribution is thus not stochastic—gene-rich regions are preferentially affected. Several genomic features relating to transcription, replication, and chromatin organization have been found to enhance chromosome breakage frequencies; this indicates that gene-rich regions may be more break-prone. The salient finding in the present study is that a substantial fraction of all structural chromosome abnormalities, not only those specifically associated with certain tumor types, may affect genes that are pathogenetically important. If this interpretation is correct, then the prevailing view that the great majority of cancer chromosome aberrations is cytogenetic noise can be seriously questioned.

KEYWORDS

breakpoint distribution, cancer chromosome abnormalities, cancer chromosome breakpoints, cytogenetic noise, gene-rich regions

1 | INTRODUCTION

Most cancer cells have acquired clonal cytogenetic aberrations.[1] An increasing number of characteristic aberrations, in particular balanced changes such as translocations and inversions, are with remarkable specificity associated with distinct morphological and clinical disease characteristics and some are even pathognomonic for certain tumor entities. Thus, identifying recurrent aberrations is an important tool in the clinical management of cancer patients to help establish a correct diagnosis, to predict prognosis, and to select the most appropriate treatment.[2,3] Furthermore, cancer cytogenetics has provided invaluable information on pathogenetically important genes located in the breakpoints of structural chromosome aberrations. Practically all balanced chromosome abnormalities in cancer that have been characterized at the molecular level have been shown to exert their effect through one of two mechanisms: deregulation, usually resulting in the overexpression of a seemingly normal gene in one of the breakpoints, or creation of a chimeric gene through the fusion of parts of two genes, one in each breakpoint.[4] Gene fusions are, however, not only formed by balanced rearrangements; there are several examples of gene fusions caused by the juxtaposition of two genes as a consequence of deletions or amplifications.[4,5]

Irrespective of their mode of origin, there is ample evidence that gene fusions represent early driver mutations.[4,6]

However, it should be emphasized that only a few neoplasms carry molecularly characterized specific chromosome changes. In fact, most human tumors display apparently unique, usually complex, structural, and/or numerical karyotypic changes and many malignant epithelial tumors harbor cytogenetically unrelated clones.[1] This heterogeneity, complexity, and confusing variety of cytogenetic abnormalities in most cancer cases have led to the view that only a few recurrent primary or secondary abnormalities are pathogenetically important, whereas most cytogenetic changes in fact represent unimportant "noise"—they are not the cause but rather the consequence of the neoplastic state, that is, they merely reflect the chromosomal instability that characterizes cancer cells.[7,8]

Ascertainment of the distribution of the breakpoints in structural chromosome abnormalities in cancer cells in relation to various characteristics of the chromosome bands involved provides a possibility indirectly to assess their importance. If larger bands are affected by aberrations more often, this would no doubt favor the assumption that they are just random chance events of no significance. If, however, the breakpoints are related to other genomic features, such as gene content, this would support the view that they indeed reflect important mechanisms in tumor development. We here present evidence that the distribution of the breakpoints in recurrent as well as nonrecurrent abnormalities, in all tumor entities investigated, is significantly associated with gene content, rather than with band length, within the affected chromosomal bands. This finding may have important ramifications for our understanding of the role played by chromosome aberrations in cancer development.

2 | MATERIALS AND METHODS

2.1 | Cases and chromosomal changes ascertained

All clonal structural chromosome abnormalities reported in the literature were extracted from the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.[1] When queried on 21 May 2018, it contained 68 487 neoplastic disorders with at least one clonal numerical and/or structural chromosome change. Furthermore, 7223 cytogenetically abnormal unpublished cases from our laboratory were also ascertained, making a total of 75 710 cases available for inclusion. All structural abnormalities, balanced as well as unbalanced, in which the breakpoints had been localized to specific chromosome bands were considered. Hence, all descriptions of aberrations indicating breakpoint uncertainty according to the ISCN nomenclature,[9] that is, denoting alternative interpretations (or), containing a question mark (?), having a breakpoint interval indicated as an approximate sign (~), or lacking band specifications were excluded. Y chromosome breakpoints were also disregarded. Furthermore, to obtain an as unbiased general view as possible of the breakpoints involved, that is, to avoid bias in the existing data due to overrepresentation in the literature of characteristic aberrations reported in specific tumor entities and selective reporting of recurrent tumor-associated abnormalities, identical aberrations were counted only once within each tumor entity. For example, t(9;22)(q34;q11) was counted once in chronic myeloid leukemia, once in acute lymphoblastic leukemia (ALL), and once in acute myeloid leukemia (AML). After applying the above-mentioned exclusion criteria, 22 344 aberrations in 49 626 cases remained for analyses.

2.2 | Parameters ascertained

The breakpoint distribution of all 22 344 aberrations in each of the 317 bands of the standard human karyotype (excluding the Y chromosome) was compared with the gene content (number of genes in a band) and band length (number of nucleotides); these data were retrieved from Ensembl genome database assembly GRCh38 (http://www.ensembl.org/index.html). In addition, gene occupancy, that is, the accumulated length occupied by gene sequences within each band, and band staining properties (G-band positive vs G-band negative) were also included in the analyses.

The total number of genes in the 22 autosomes and the X chromosome was 27 268. The range of number of genes per band was 2 to 897. Twenty-two bands, primarily the short arms of the acrocentric chromosomes 13–15, 21, and 22, contained no genes. Band lengths varied between 1.00 and 28.44 Mb and gene occupancy ranged from 0.01 to 18.3 Mb. Of the 280 informative bands, 155 were G-band negative (light) and 125 G-band positive (dark); the heterochromatic bands in chromosomes 1, 3–6, 9, 11–16, 19, 21, 22, and X, and the satellite stalks in chromosomes 13–15, 21, and 22 were excluded when comparing light and dark G-bands.

2.3 | Statistical analyses

Linear regression analyses were used to study associations between breakpoint localizations in relation to gene content, band length, gene occupancy, and G-band positivity/negativity. The coefficient of determination (R^2) was calculated for each model. To compare the predictive value of band length and gene content, the increase in R^2 (denoted R^2 diff) when adding one of the variables to a univariable model including the other variable, was calculated. The following aberration groups were analyzed separately in the total material of benign and malignant tumors: (1) all aberrations, (2) all recurrent abnormalities (defined as at least two identical aberrations within a particular tumor entity), (3) all sole anomalies, (4) all aberrations found only once within any tumor entity (designated "non-recurrent aberrations"), (5) all translocations, and (6) all recurrent translocations.

The following specific tumor entities had a sufficient number of abnormalities to allow analyses of breakpoint localizations of all aberrations in relation to gene content and band length: AML, ALL, malignant lymphomas excluding Hodgkin lymphoma (ML), malignant bone and soft tissue tumors (MBST), and malignant epithelial tumors (MET).

All analyses were performed in SAS 9.4 (SAS Institute Inc., Cary, NC) and a P-value lower than 0.05 was considered significant.

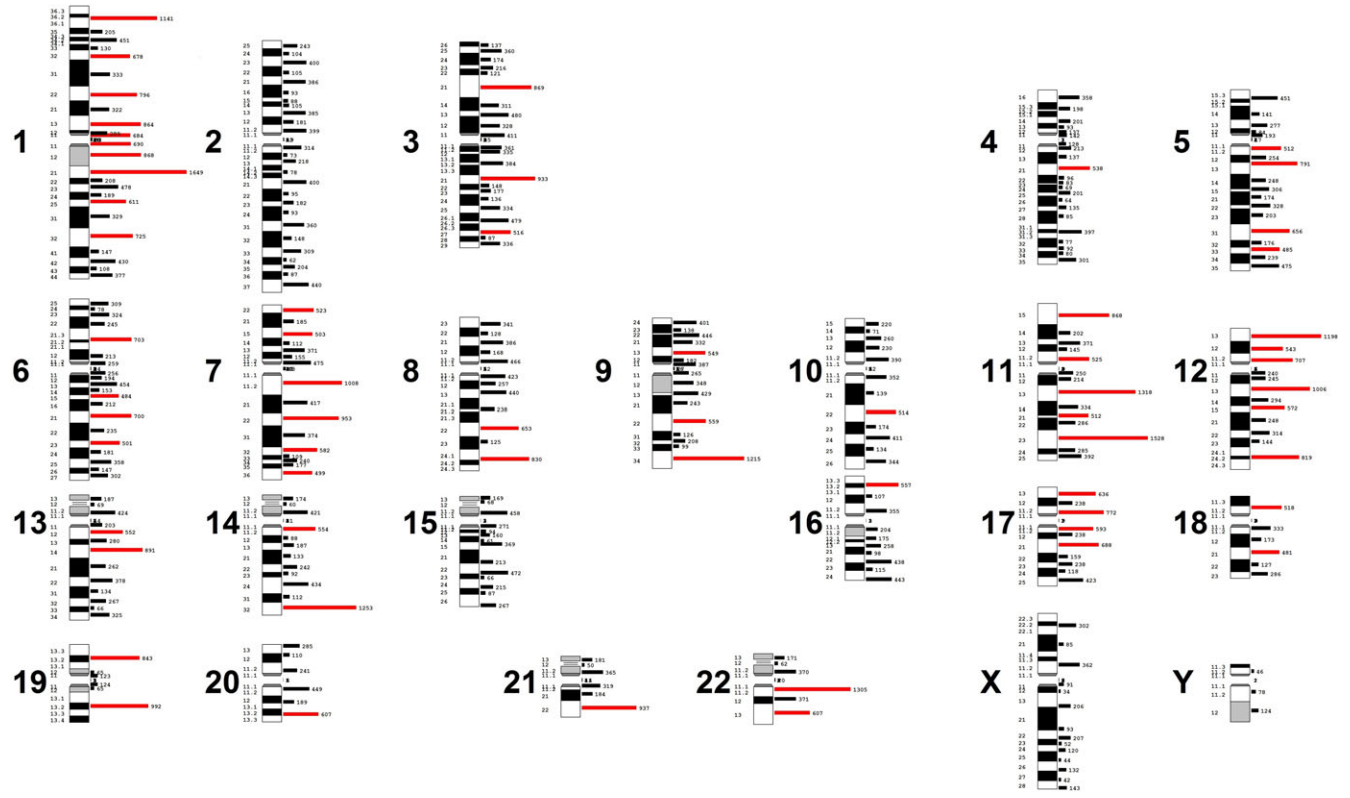

FIGURE 1 Breakpoint distribution of all 22 344 structural chromosome aberrations in all 49 626 informative benign or malignant neoplasms. For each band, the number of breakpoints involved in all aberrations is shown as a bar to the right of each chromosome. The bars with the highest numbers of breakpoints, defined as the 90th percentile, are presented in red. Wherever sub-bands are indicated in the ideogram, the bars show the number of breakpoints at the corresponding band level. The figure may be enlarged for better readability in the online version

3 | RESULTS

Supporting Information Table S1 lists all 317 chromosome bands ascertained together with information on their staining characteristics, gene content, length, and gene occupancy as well as their numbers of breakpoints in all aberrations, all recurrent aberrations, all aberrations found as the sole anomaly, all nonrecurrent aberrations, all translocations, all recurrent translocations, and their numbers of breakpoints in all aberrations in AML, ALL, ML, MBST, and MET. Figure 1 shows the breakpoint localizations in the total material—the

TABLE 1 Univariable and multivariable regression analyses of associations between number of breakpoints in relation to gene content and band length in the aberration types ascertained

Model[a]	Gene content		Band length		Multivariable full model[b]		
	P-value	R^2	*P*-value	R^2	R^2	R^2 diff	
						Length (genes)	Genes (length)
All aberrations							
Univariable	<0.0001	37.43	<0.0001	10.16			
Multivariable	<0.0001		0.3146		37.64	0.20	27.48
Recurrent aberrations							
Univariable	<0.0001	31.43	<0.0001	8.80			
Multivariable	<0.0001		0.4493		31.56	0.13	22.75
Sole aberrations							
Univariable	<0.0001	27.23	<0.0001	6.94			
Multivariable	<0.0001		0.3141		27.46	0.23	20.52
Nonrecurrent aberrations							
Univariable	<0.0001	33.03	<0.0001	9.33			
Multivariable	<0.0001		0.4520		33.15	0.12	23.82

[a] Univariable includes only one of the variables as predictor.
[b] The R^2 diff shows how much R^2 increases when the variable outside the parentheses is added to a model with the variable inside the parentheses.

TABLE 2 Univariable and multivariable regression analyses of associations between number of breakpoints in relation to gene content and band length in the tumor types analyzed separately

Model[a]	Gene content		Band length		Multivariable full model[b]		
	P-value	R^2	*P*-value	R^2	R^2	R^2 diff	
						Length (genes)	Genes (length)
Acute myeloid leukemia							
Univariable	<0.0001	24.83	<0.0001	7.73			
Multivariable	<0.0001		0.7771		24.84	0.02	17.12
Acute lymphoblastic leukemia							
Univariable	<0.0001	20.41	<0.0001	5.77			
Multivariable	<0.0001		0.5907		20.48	0.07	14.71
Malignant lymphomas							
Univariable	<0.0001	26.78	<0.0001	8.85			
Multivariable	<0.0001		0.9395		26.78	0.00	17.93
Malignant bone and soft tissue tumors							
Univariable	<0.0001	36.88	<0.0001	7.23			
Multivariable	<0.0001		0.0213		37.94	1.06	30.71
Malignant epithelial tumors							
Univariable	<0.0001	25.53	<0.0001	5.13			
Multivariable	<0.0001		0.0915		26.21	0.67	21.07

[a] Univariable includes only one of the variables as predictor.
[b] The R^2 diff shows how much R^2 increases when the variable outside the parentheses is added to a model with the variable inside the parentheses.

distribution is nonrandom with light G-bands being more frequently involved (details presented below). Univariable and multivariable regression analyses of the associations between breakpoints and gene content and band length for all aberrations (n = 22 344), all recurrent aberrations (n = 3089), all sole aberrations (n = 5776), and all nonrecurrent aberrations (n = 13 945) are presented in Table 1. Table 2 shows the associations between number of breakpoints in relation to gene content and band length for all aberrations in AML (n = 6234), ALL (n = 3816), ML (n = 6282), MBST (n = 3858), and MET (n = 5361). Separate analyses for all translocations (n = 16 055) and all recurrent translocations (n = 1660) are given in Table 3; the remaining major aberration types (additions, deletions, duplications, and inversions) amounted to only 3624 in total and could not be analyzed separately.

As seen in Tables 1–3, there were highly significant associations between the breakpoint distribution and gene content/band length at univariable analyses in the entire cohort as well as in all different subsets (P < 0.0001). At multivariable analyses, however, only gene content remained statistically significant (P < 0.0001 for all pairwise comparisons), whereas band length, with the exception of MBST (P = 0.0213), was not significant in any comparison (P = 0.0633-0.9395). The R^2 differences clearly showed that for all aberrations, irrespective of whether sole, recurrent, or nonrecurrent, in the total material as well as in the specific tumor types analyzed separately, gene content provided the best explanation of the breakpoint distribution of cancer-associated chromosome abnormalities. The contribution of band length was negligible (R^2 differences = 0.00-1.06). Figure 2 illustrates the strong association between the breakpoint distribution of all aberrations in all tumor types in relation to gene content. Gene occupancy did not provide any relevant additional information in this context (data not shown), presumably because of its strong association with both gene content and band length (Pearson correlations

TABLE 3 Univariable and multivariable regression analyses of associations between number of breakpoints in relation to gene content and band length in all translocations and in all recurrent translocations

Model[a]	Gene content		Band length		Multivariable full model[b]		
	P-value	R^2	*P*-value	R^2	R^2	R^2 diff	
						Length (genes)	Genes (length)
All translocations							
Univariable	<0.0001	40.11	<0.0001	9.11			
Multivariable	<0.0001		0.0633		40.77	0.66	31.66
Recurrent translocations							
Univariable	<0.0001	27.90	<0.0001	6.22			
Multivariable	<0.0001		0.1425		28.39	0.49	22.17

[a] Univariable includes only one of the variables as predictor.
[b] The R^2 diff shows how much R^2 increases when the variable outside the parentheses is added to a model with the variable inside the parentheses.

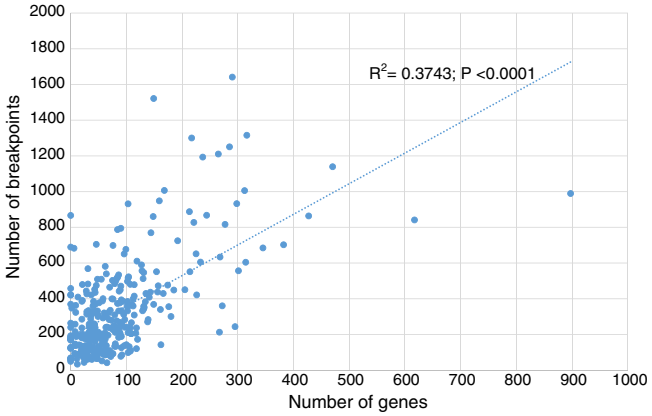

FIGURE 2 Regression analysis of the total material, comprising 22 344 structural chromosome abnormalities in 49 626 informative cases, shows a highly significant association between gene content and the number of breakpoints in the 317 bands [Color figure can be viewed at wileyonlinelibrary.com]

0.7716 and 0.8060, respectively). Table 4 shows the analyses of the breakpoint distributions in light and dark G-bands. The total band lengths of light and dark bands were almost equal (light = 1.49×10^9 bp; dark = 1.36×10^9 bp). Multivariable analyses revealed that gene content and band length contributed equally to the breakpoint distribution in dark G-bands, whereas gene content but not band length was significant for light G-bands ($P < 0.0001$), which contained 2.5 times as many genes and 3.5 times as many breakpoints as dark G-bands (Figure 1).

Thus, of all the parameters analyzed (gene content, band length, gene occupancy, and staining properties of the chromosome bands), gene content was the prime determinant of the breakpoint distribution of tumor-associated structural chromosome aberrations.

4 | DISCUSSION

Cytogenetic analyses of neoplastic tissues have advantages and disadvantages. The obvious strength is that the analysis is based on individual cells, thus providing cell-specific information on heterogeneous cell populations. The most important weakness is that the localization of breakpoints is sometimes imprecise; it may be difficult to decide whether a break has occurred in a particular or a neighboring band. However, we believe that this shortcoming is insignificant in the present study based on more than 22 000 different aberrations identified by different banding techniques in numerous studies over five decades by scientists all over the world. It is unreasonable to believe that so many investigators would systematically have favored particular breakpoint localizations, thereby introducing a serious bias of the total material.

Since the 1970s, chromosome abnormalities in cancer cells have been subdivided into primary and secondary.[10] Primary aberrations, such as t(9;22)(q34;q11) in chronic myeloid leukemia, are frequently found as the sole karyotypic abnormalities and are often specifically associated with particular tumor types. Such aberrations will, by virtue of their tumorigenic importance, be present already in the earliest disease phases. In contrast, secondary abnormalities, which represent the overwhelming majority of all chromosome changes in cancer,[1] develop in cells that already carry a primary abnormality. Although sometimes nonrandom, they typically lack the specificity seen for primary changes.[2] In later disease stages, secondary aberrations may be so numerous as to completely dominate the karyotype. This diverse and confusing chromosome variability has led to the impression that most secondary aberrations represent incidental phenomena, consequently often designated cytogenetic noise. However, we here show that in all tumor types there were strong associations between breakpoint localizations and gene content, but only a negligible contribution of band length, for all kinds of aberrations (sole, recurrent, and nonrecurrent), encompassing primary as well as secondary abnormalities. This clearly indicates that the breakpoint distribution is not stochastic and that gene-rich regions are preferentially affected.

The clustering of breakpoints in gene-rich regions may be due to several genomic features. Numerous studies have indicated relationships between chromosome breakage in normal and cancer cells and chromatin composition and function, such as GC content, transcriptional activity, replication timing, and chromatin accessibility.[11–15] These features are more or less interdependent and the correlations with breakage resulting in structural chromosomal aberrations are complex and still insufficiently investigated. However, they no doubt point in the direction that gene-rich regions may be more breakprone. Our data provide compelling support for the assumption that a

TABLE 4 Univariable and multivariable regression analyses of associations between number of breakpoints in relation to gene content and band length in light and dark G-bands

Model[a]	Gene content		Band length		Multivariable full model[b]		
	P-value	R^2	*P*-value	R^2	R^2	R^2 diff	
						Length (genes)	Genes (length)
Light G-bands (n = 155 with a total of 19 245 genes)							
Univariable	<0.0001	35.92	<0.0001	25.59			
Multivariable	<0.0001		0.1295		36.89	0.96	11.30
Dark G-bands (n = 125 with a total of 7804 genes)							
Univariable	<0.0001	13.18	<0.0001	14.85			
Multivariable	0.0130		0.0035		19.06	5.87	4.21

[a] Univariable includes only one of the variables as predictor.
[b] The R^2 diff shows how much R^2 increases when the variable outside the parentheses is added to a model with the variable inside the parentheses.

substantial fraction of structural chromosome abnormalities, not only those specifically associated with certain tumor types, affects genes that may be of importance for the carcinogenic process. Our results thus indicate that many more genes than previously appreciated may be involved in this process and that the prevailing view that the great majority of cancer chromosome aberrations is noise can be seriously questioned. It may very well be, as our results suggest, that the breakpoints of most chromosome abnormalities, be they primary or secondary, play pathogenetically important roles in tumor development.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

ORCID

Felix Mitelman https://orcid.org/0000-0002-3757-7212

REFERENCES

1. Mitelman F, Johansson B, Mertens F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer http://cgap.nci.nih.gov/Chromosomes/Mitelman. Accessed May 21, 2018.
2. Heim S, Mitelman F, eds. *Cancer Cytogenetics. Chromosomal and Molecular Genetic Aberrations of Tumor Cells*. 4th ed. Oxford: Wiley-Blackwell; 2015.
3. Rowley JD, Le Beau MM, Rabbitts TH, eds. *Chromosomal Translocations and Genome Rearrangements in Cancer*. Switzerland: Springer; 2015.
4. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer*. 2015;15:371-381.
5. Mertens F, Antonescu CR, Mitelman F. Gene fusions in soft tissue tumors: recurrent and overlapping pathogenetic themes. *Genes Chromosomes Cancer*. 2016;55:291-310.
6. Greaves M. Leukemia 'firsts' in cancer research and treatment. *Nat Rev Cancer*. 2016;16:163-172.
7. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. *Nature*. 1998;396:642-649.
8. Bakhoum SF, Cantley LC. The multifaceted role of chromosomal instability in cancer and its microenvironment. *Cell*. 2018;174:1347-1360.
9. McGowan-Jordan J, Simons A, Schmid M, eds. *ISCN: An International System for Human Cytogenomic Nomenclature*. S. Karger: Basel; 2016.
10. Mitelman F, Levan G. Clustering of aberrations to specific chromosomes in human neoplasms. III. Incidence and geographic distribution of chromosome aberrations in 856 cases. *Hereditas*. 1978;89:207-232.
11. Drier Y, Lawrence MS, Carter SL, et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res*. 2013;23:228-235.
12. Morganella S, Alexandrov LB, Glodzik D, et al. The topography of mutational processes in breast cancer genomes. *Nat Commun*. 2016;7:11383.
13. Glover TW, Wilson TE, Arit MF. Fragile sites in cancer: more than meets the eye. *Nat Rev Cancer*. 2017;17:489-501.
14. Anderson ND, de Borja R, Young MD, et al. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science*. 2018;361:eaam8419.
15. Corces MR, Granja JM, Shams S, et al. The chromatin accessibility landscape of primary human cancers. *Science*. 2018;362:eaav1898.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Mitelman F, Johansson B, Mertens F, Schyman T, Mandahl N. Cancer chromosome breakpoints cluster in gene-rich genomic regions. *Genes Chromosomes Cancer*. 2019;58:149–154. https://doi.org/10.1002/gcc.22713