Human Molecular Genetics, 2016, Vol. 25, No. 5 1019–1030

doi: 10.1093/hmg/ddv623
Advance Access Publication Date: 4 January 2016
Association Studies Article

ASSOCIATION STUDIES ARTICLE

Genomic determinants of somatic copy number alterations across human cancers

Yanping Zhang[1,†], Hongen Xu[1,†] and Dmitrij Frishman[1,2,3,*]

[1]Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany, [2]Helmholtz Center Munich – German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany and [3]St Petersburg State Polytechnic University, St Petersburg 195251, Russia

*To whom correspondence should be addressed at: Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany. Tel: +49 8161712134; Fax: +49 8161712186; Email: d.frishman@wzw.tum.de

Abstract

Somatic copy number alterations (SCNAs) play an important role in carcinogenesis. However, the impact of genomic architecture on the global patterns of SCNAs in cancer genomes remains elusive. In this work, we conducted multiple linear regression (MLR) analyses of the pooled SCNA data from The Cancer Genome Atlas (TCGA) Pan-Cancer project. We performed MLR analyses for 11 individual cancer types and three different kinds of SCNAs—amplifications and deletions, telomere-bound and interstitial SCNAs and local SCNAs. Our MLR model explains >30% of the pooled SCNA breakpoint variation, with the explanatory power ranging from 13 to 32% for different cancer types and SCNA types. In addition to confirming previously identified features [e.g. long interspersed element-1 (L1) and short interspersed nuclear elements], we also identified several novel informative features, including distance to telomere, distance to centromere and low-complexity repeats. The results of the MLR analyses were additionally confirmed on an independent SCNA data set obtained from the catalogue of somatic mutations in cancer database. Using a rare-event logistic regression model and an extremely randomized tree classifier, we revealed that genomic features are informative for defining common SCNA breakpoint hotspots. Our findings shed light on the molecular mechanisms of SCNA generation in cancer.

Introduction

Cancer is fundamentally a disease characterized by a diversity of somatic alterations (1). Recently developed technologies, such as single-nucleotide polymorphism (SNP) arrays and next-generation DNA sequencing have created unprecedented opportunities for studying different classes of mutations, including single-base substitutions, small indels, genomic rearrangements, and somatic copy number alterations (SCNAs) (1–3). The landscape of SCNAs has been charted across different types of cancer, with recurrent SCNAs often pointing at novel oncogenes and tumor suppressor genes (2,4,5). Although SCNAs affect a sizeable fraction of the genome and are functionally important in carcinogenesis, their generation mechanisms are not yet fully understood.

Previous analyses of SCNA data have provided insights into the mechanisms shaping SCNA occurrence (2,5–7). SCNA breakpoints are not uniformly distributed in the genome, but rather tend to be spatially clustered in breakpoint hotspots (6). For instance, G-quadruplex sequences (G4s) are enriched in the vicinity of SCNA breakpoints, suggesting the contribution of genomic properties to SCNA formation (6). A recent comparative analysis has identified two types of SCNA breakpoint hotspots—cancer-type-specific SCNA breakpoint hotspots, which are enriched in known cancer genes, and common hotspots (CHSs). The latter can be relatively well predicted from genomic context by a multiple linear regression (MLR) model (8). However, the model presented in (8) explains only a small part of the SCNA breakpoint variance [with the top four features—indel rate, exon density,

[†]Y.Z. and H.X. contributed equally to this work.

Received: November 1, 2015. **Revised:** December 15, 2015. **Accepted:** December 21, 2015

substitution rate and short interspersed nuclear element (SINE) coverage—being collectively responsible for 14% of the variation]. A model considering a much wider spectrum of genomic properties would be expected to better illuminate how different genomic features contribute to the global patterns of SCNAs in cancer genomes.

Many endogenous factors (such as non-B DNA conformations and repetitive sequences) can cause double-strand breaks (DSBs). Subsequent erroneous DNA repairs will result in copy number alterations (6,9,10). Indeed, genome-wide mapping of DSBs has shown that DSB regions are enriched in genomic regions frequently rearranged in cancers (11). Under certain circumstances, DNA can assemble into non-B conformations at specific sequence motifs including A-phased repeats, G-quadruplex, Z-DNA, inverted repeats, mirror repeats and direct repeats (12). The resulting DNA secondary structures have been implicated in the formation of structural alterations including copy number variations (CNVs), inversions and translocations, such as G-quadruplexes (6), Z-DNA (13), cruciforms formed by inverted repeats (14) and triplexes (also known as H-DNA) formed by mirror repeats (15). Transposable elements are dispersed at high-copy numbers throughout the human genome, and non-allelic homologous recombination between different copies of transposable elements can result in CNVs. For example, homologous recombination of non-allelic copies of L1 and human endogenous retroviral elements leads to the formation of CNVs (16,17). Moreover, a 13-mer CCNCCNTNNCCNC motif was found to associate with recombination hotspots in humans and was clustered in common mitochondrial deletion hotspots (18). Recently, Zhou *et al.* (19) have revealed a significant enrichment of human germline and somatic structural variant breakpoints in self-chain (SC) regions, a group of low-copy repeats <1 kb. Besides the effects of local genomic context on CNV formation, TCGA Pan-Cancer analysis has suggested different mechanisms for telomere-bound SCNAs and those SCNAs that are interstitial to chromosomes, highlighting the importance of the chromosome structure (e.g. telomeres and centromeres) (5).

In this study, we selected genomic features, which have been proposed to affect SCNAs across the human genome, of which DSBs, SCs, recombination motifs, and distance to telomeres and centromeres have not been investigated in previous studies. We also include the histone marker H3K9me3, which accounts for >40% of mutation rate variation in cancer cells (20). We built MLR and logistic regression (LR) models to explore the intrinsic basis of observed SCNA patterns. These statistical methods have been successful in contrasting common fragile sites and non-fragile sites (21) and investigating the effects of diverse sequence features on integration sites of DNA transposons (22).

The overview of our study is presented in Figure 1. Taking advantage of SCNAs data from the TCGA Pan-Cancer project and collected genomic features, we first selected predictors (genomic features) to reduce multicollinearity and identified common SCNA breakpoint hotspots and non-hotspots (NHSs) across Pan-Cancer types. We then built MLR models to investigate whether and how different genomic features contribute to the genome-wide patterns of SCNA breakpoints. We also applied LR and extremely randomized tree classifier to contrast between common SCNA breakpoint hotspots and NHSs. Our MLR models can explain >30% of SCNA breakpoint variation. The power of the models remain stable when one considers separately different SCNA types (amplifications and deletions), SCNA types of possible different generation mechanisms (telomere-bound SCNAs and interstitial SCNAs), and SCNAs from different cancer types. We also demonstrate that these genomic features are informative for telling apart common SCNA breakpoint hotspots and NHSs by logistic models and extremely randomized tree classifiers. This suggests that common breakpoint hotspots strongly depend on the local genomic context.

Results

Identification of SCNA breakpoint hotspots

In this work, we analyzed data on 404 488 SCNA breakpoints (5) in 11 cancer types (Table 1). To characterize the genome-wide patterns of SCNA occurrence, we divided the human genome into 1 Mb non-overlapping windows, after removing gaps, and calculated the density of SCNA breakpoints within each window.

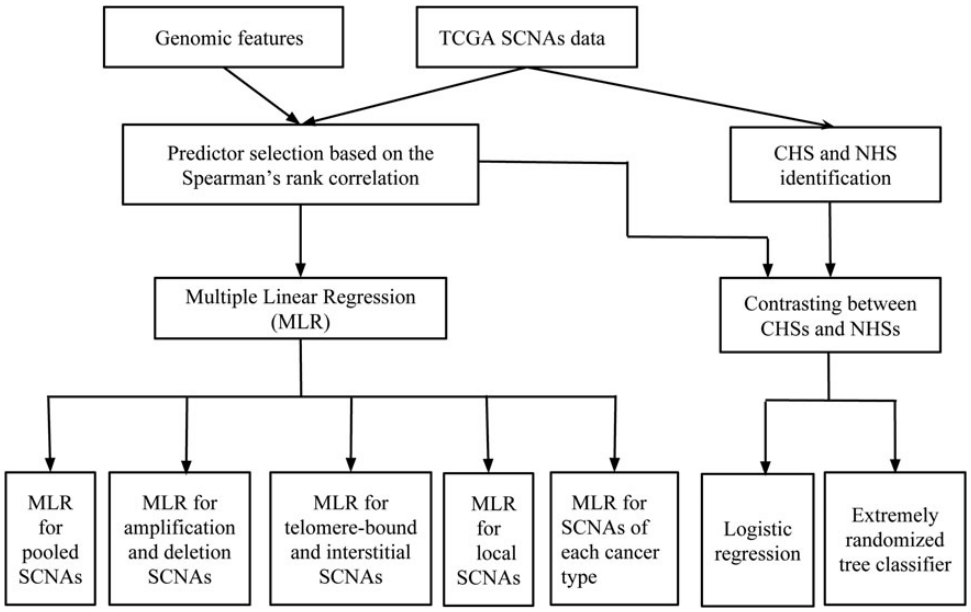

Figure 1. An overview of the study design. TCGA: the cancer genome atlas; SCNA, somatic copy number alterations; CHS, common hotspots; NHS, non-hotspots.

Table 1. Summary of SCNA data from TCGA Pan-Cancer project

Cancer type	Abbr.	Sample size	SCNA breakpoints	Breakpoint Amplification			Deletion		
				Interstitial Local	Telomere-bound Chr. level	Telomere-bound Local	Interstitial Local	Telomere-bound Chr. level	Telomere-bound Local
Bladder urothelialc carcinoma	BLCA	90	13 344	4562	802	1172	3900	1326	1582
Breast invasive carcinoma	BRCA	745	99 574	42 268	2624	8792	25 414	8610	11 866
Colon adenocarcinoma	COAD	349	21 650	4222	2318	2004	6672	3966	2468
Glioblastoma multiforme	GBM	485	28 462	10 162	2078	1074	10 234	2556	2358
Head and neck squamous cell carcinoma	HNSC	270	24 272	6990	1130	3068	5586	3320	4178
Kidney renal clear cell carcinoma	KIRC	373	9040	1818	1024	860	1756	2230	1352
Lung adenocarcinoma	LUAD	292	34 952	12 080	1890	3430	8006	4882	4664
Lung squamous cell carcinoma	LUSC	261	34 400	10 828	1106	3998	7992	4628	5848
Ovarian serous cystadenocarcinoma	OV	457	92 216	41 238	2762	10 720	19 200	7176	11 120
Rectum adenocarcinoma	READ	147	12 358	2620	1114	1090	3694	2328	1512
Uterine corpus endometrial carcinoma	UCEC	376	34 220	18 014	1196	2726	6570	2132	3582
Total		3845	404 488	154 802	18 044	38 934	99 024	43 154	50 530

Abbr., abbreviation; Chr., chromosome.

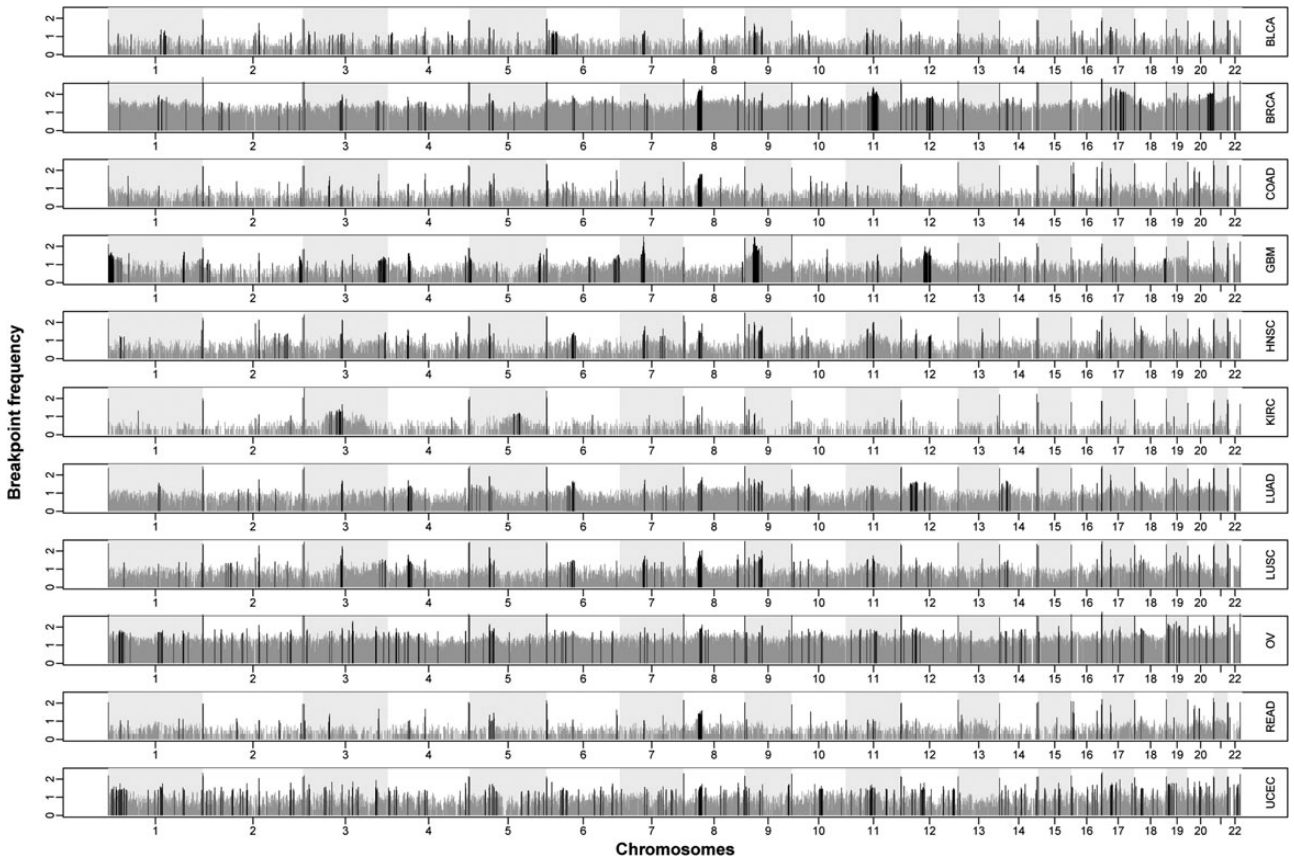

Figure 2. The distribution of SCNA breakpoint frequencies in 11 cancer types—BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, READ and UCEC (see Table 1 for full names), calculated as $=\log_{10}$ (the number of SCNA breakpoints in each block +1). Breakpoint hotspots in each cancer type are colored in black.

Based on the randomization procedure described in the *Materials and Methods* section, we identified 81–331 breakpoint hotspots in individual cancers [false discovery rate (FDR) corrected $P < 0.05$]. As seen in Figure 2 different types of cancer often share breakpoint hotspots, but also have their specific hotspots. Based on the definitions in the *Materials and Methods* section, we identified 29 CHSs, 1824 NHSs and 685 non-common hotspots (NCHSs).

Human genomic features

To identify potential correlates of SCNA breakpoint patterns, we compiled a set of diverse genomic features, of which some, including non-B DNA sequences, and transposable elements, were previously investigated for their effects on SCNA breakpoints (8), while several other features, such as distance to centromere and DSBs, are used for this purpose in this work for

the first time. In total, we examined 29 features that can be generally categorized into six groups: non-B DNA conformations; DNA sequence; gene regulation and expression; evolutionary features; chromosome structures and functional features (Table 2). Following Fungtammasan *et al.* (21) and Campos-Sánchez *et al.* (22), we used hierarchical clustering with Spearman's rank correlation to remove some strongly correlated features (Supplementary Material, Fig. S1). Finally, 25 features were used for subsequent regression analyses.

Impact of genomic features on the frequencies of SCNA breakpoints

We examined to what extent the observed genome-wide patterns of breakpoints could be explained by genomic features. Following an approach similar to the one described in (21,22), the density of SCNA breakpoints (response) calculated in each 1 Mb window was represented as a function of the 25 genomic features (predictors) measured in the same 1 Mb window. The resulting MLR model accounted for 32.03% of the variation in the breakpoint density and contained 11 significant predictors (Table 3). The predictor with the strongest positive effect in the model is direct repeat coverage (10.35%). Other predictors with a significant positive effect are L1 coverage, low-complexity repeat coverage, SINE count, conserved DNA element count, CpG island coverage and inverted repeat coverage with the relative contribution to variance explained (RCVE) ranging from 0.89 to 2.06% (Table 3; Fig. 3). The predictors with the strongest negative effect are distance to telomere (29.15%) and distance to centromere (14.55%). Less significant predictors with a negative effect are mirror repeat count (6.68%), Z-DNA repeat coverage (1.14%) and simple repeat coverage (0.98%).

We repeated the same analysis replacing some of the predictors with highly correlated predictors. For example, A-phased repeat coverage was replaced with G4 count or recombination motif and we observed slight changes in both the RCVE of predictors and R^2 of models. Most of genomic features remained significant in these alternative models (Supplementary Material, Tables S1–S4).

We next applied MLR for breakpoints of two SCNA types—amplifications and deletions—separately. The MLR model explained 29.52% (amplifications) and 27.88% (deletions) of response variance. Notably, the predictors and the sign of their effect revealed by these two MLR models are similar to those of pooled SCNA breakpoints (Supplementary Material, Tables S5 and S6), although some differences were apparent. For instance, Z-DNA repeat coverage, which had negative effect when both types of breakpoints were considered, disappeared in the MLR model for amplification breakpoints. Likewise, inverted repeat coverage lost its positive effect in the MLR model for deletion breakpoints.

Distance to telomere is a predictor with the strongest negative effect for both pooled SCNA breakpoints and the breakpoints corresponding to the two individual SCNA types—amplifications and deletions (Table 3 and Supplementary Material, Tables S5 and S6). In order to remove the confounding effect of this parameter, we next divided SCNAs into two categories: telomere-bound SCNAs, with one boundary located in the telomere and interstitial SCNAs, with both boundaries interstitial to the chromosome (5). MLR models accounted for 31.90 and 20.24% of the variation for telomere-bound SCNAs and interstitial SCNAs, respectively. Significant predictors of telomere-bound and interstitial SCNAs

Table 2. Genomic features used in the regression analyses

Category	Predictor	Measure	Source
DNA conformation	A-phased repeats	Coverage	Non-B DB version 2
	Mirror repeats	Count	Non-B DB version 2
	Direct repeats	Coverage	Non-B DB version 2
	Inverted repeats	Coverage	Non-B DB version 2
	Z-DNA	Coverage	Non-B DB version 2
	G4	$\log_{10}$ (count)	Non-B DB version 2
DNA sequence	Microsatellites	Coverage	UCSC Genome Browser
	SINEs	$\log_{10}$ (count)	UCSC Genome Browser
	L1	Coverage	UCSC Genome Browser
	L2	Coverage	UCSC Genome Browser
	LTR retrotransposons	Coverage	UCSC Genome Browser
	DNA transposons	Coverage	UCSC Genome Browser
	Low-complexity repeats	Coverage	UCSC Genome Browser
	Double-strand breaks	Coverage	Tchurikov *et al.* (2013)
	Self-chain segments	Coverage	This work
	GC content	Coverage	This work
	Simple repeats	Coverage	UCSC Genome Browser
Expression and gene regulation	H3K9me3	Count	Barski *et al.* (2007)
	CpG islands	Coverage	UCSC Genome Browser
Chromosome structure	Distance to centromere	$\log_{10}$ (distance in bp)	This work
	Distance to telomere	$\log_{10}$ (distance in bp)	This work
Evolutionary features	Recombination motif	Coverage	This work
	Conserved DNA elements	Count	Siepel *et al.* (2005)
	Indel rate	Coverage	Human-Chimp alignment
	Substitution rate	Coverage	Human-Chimp alignment
Functional features	Replication timing	Sum	Hansen *et al.* (2010)
	Exon	Coverage	UCSC Genome Browser
	miRNA genes	Coverage	miRbase database
	Fragile sites	Yes/no	Fungtammasan *et al.* (2012)

Table 3. The MLR model for pooled SCNA breakpoints

Predictor	Standardized coefficient	Variance inflation factor	P-value	Relative contribution, %	Five-fold relative contribution, %
Distance to centromere	−0.2428	1.265	4.24E−38	14.55	19.76
Conserved element count	0.1132	3.382	1.88E−04	1.18	1.07
CpG island coverage	0.0722	1.133	3.88E−05	1.43	1.11
Direct repeat coverage	0.425	5.433	7.69E−28	10.35	11.97
Inverted repeat coverage	0.0976	3.330	1.17E−03	0.89	0.51
L1 coverage	0.1361	3.677	1.66E−05	1.57	1.67
Low-complexity repeat coverage	0.1424	3.069	8.34E−07	2.06	2.78
Mirror repeat count	−0.3028	4.284	1.12E−18	6.68	7.70
SINE count	0.2231	3.762	4.84E−06	1.77	1.87
Distance to telomere	−0.4194	1.883	2.81E−72	29.15	32.21
Z-DNA coverage	−0.1083	3.146	2.46E−04	1.14	Not significant
Simple repeat coverage	−0.0874	2.434	6.67E−04	0.98	1.12
Adjusted R^2					31.36
Five-fold adjusted R^2					25.31

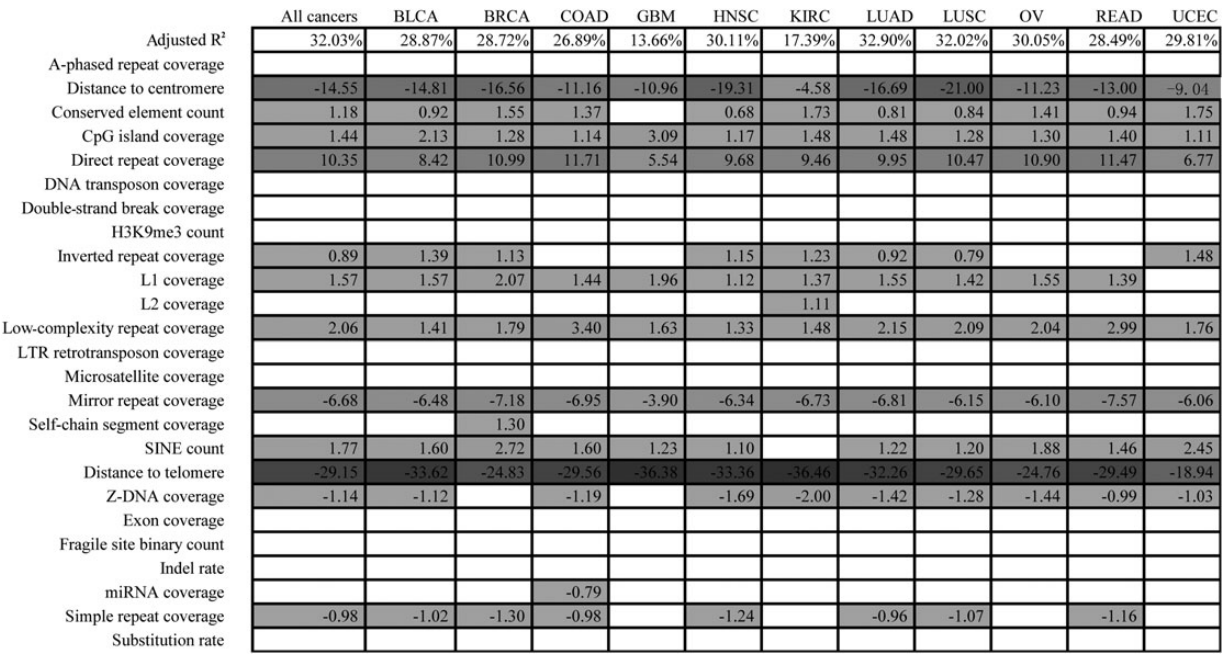

Figure 3. The effect of genomic features in MLR models. The intensity of grey color is proportional to the RCVE in each model. Predictors in white color are not significant. See Table 1 for full names of cancer types.

are listed in Supplementary Material, Tables S7 and S8. Distance to telomere is a dominant predictor for telomere-bound SCNAs (relative contribution of 29.97%), while for interstitial SCNAs the most significant predictor is distance to centromere (relative contribution of 45.91%). Distance to centromere and SINEs are also significant for both SCNA types. However, the relative contribution of distance to centromere is substantially reduced for the telomere-bound SCNAs compared with interstitial SCNAs. Moreover, the other significant predictors for telomere-bound SCNAs are quite different from the significant predictors for the interstitial SCNAs.

By definition, the breakpoints of chromosome-level SCNAs are fixed at telomeres. We, therefore, excluded chromosome-level SCNAs from all the pooled SCNAs before conducting MLR analyses. We found that the model could explain 30.36% of the variation and included 10 significant predictors (Supplementary Material, Table S9). Notably, the predictors and their effect are similar to those of pooled SCNAs.

We also performed similar analyses for each cancer type and found the adjusted R^2 of models to be >26% for all cancer types except for glioblastoma multiforme (13.66%) and kidney renal clear cell carcinoma (17.39%). Similar to the MLR model of the pooled SCNA breakpoints, we identified direct repeat coverage, L1 coverage, low-complexity repeat coverage and SINE count as significant positive predictors for almost all cancer types (Fig. 3). The distance to telomere, distance to centromere and mirror repeat count remained significant negative predictors for each cancer type (Fig. 3).

We also conducted 5-fold cross validation for all the MLR models. While the MLR model trained over the pooled breakpoint data set yielded an adjusted R^2 of 32.03%, the R^2 of the 5-fold MLR built from the pooled breakpoint data set was 25.31% (Table 3).

Moreover, the significant predictors and their effects identified in 5-fold MLR are similar to those of MLR (Table 3). The 5-fold MLR results for the other MLR models are provided in Supplementary Material, Tables S1–S9 and Figure S2. The consistency between the MLR model and 5-fold MLR model indicates that the MLR regression model demonstrates good predictive ability and generalizes well on validation data sets.

We also assessed the generalization ability of our MLR model on an independent data set obtained from the catalogue of somatic mutations in cancer (COSMIC) database (see the Materials and Methods section). On this data set the MLR model and the 5-fold MLR model accounted for 41.16 and 36.99% of breakpoint variation, respectively (Supplementary Material, Table S10). The most significant predictors, e.g. distance to telomere, mirror repeats and distance to centromere identified in the MLR model for pooled breakpoints from TCGA are also found to be significant in the MLR model on the independent data set. However, predictors, including exon coverage, H3K9me3 count, long terminal repeat (LTR) retrotransposon coverage and indel rate, gained significance in this data set. Exon coverage and indel rate are among the top four features in the model presented in (8).

Contrasting between CHSs and NHSs by LR

We investigated how genomic context affects the distribution of common breakpoint hotspots in cancer genomes. To this end, we built a standard LR model using 25 features. The final standard LR model had a pseudo-R^2 51.83% and comprised two highly significant genomic features: distance to telomere (individual contribution 20.70%) and direct repeat coverage (individual contribution 5.16%).

However, the standard LR model may suffer from small-sample bias and class imbalance. In this work, the sample size of CHSs is small (sample size: 29) and sample sizes for NHSs and CHSs are imbalanced (1824 versus 29). For this reason, besides standard LR, we performed the rare-events logistic regression (RELR). The estimates of a RELR model are corrected for class imbalance. Moreover, to eliminate the possible small-sample bias, we increased the number of common cancer hotspots by a sliding process, in which we divided the human genome into 1 Mb overlapping windows with a step size of 100 kb. Following the hotspot identification procedure described in the Materials and Methods section, we identified 231 CHSs. The RELR model has a pseudo R^2 51.83% and contains 12 significant predictors (Table 4; Fig. 4). The strongest feature discriminating CHSs and NHSs was distance to telomere (individual contribution 20.70%). This was a negative predictor, indicating that CHSs tend to be positioned closely to telomere. Direct repeat coverage is the strongest significant positive predictor (with the individual contribution is 5.16%), which implies that CHSs are located preferably in a genomic context that is enriched in direct repeats. We also performed RELR to contrast between NCHSs and NHSs as well as between NCHSs and CHSs. We found that genomic features cannot discriminate between them (data not shown).

Interestingly, the important features determined by the model, such as distance to telomere, direct repeat coverage, distance to centromere and L1 coverage, were also identified to have significant effects on SCNA breakpoint in the MLR models.

Table 4. RELR for contrasting CHSs with NHSs

Predictor	Standardized coefficient	P-value	Relative contribution, %
Conserved elements count	5.0288	5.18E−04	1.01
CpG island coverage	1.8248	1.04E−06	1.14
Direct repeats coverage	11.2571	2.16E−11	5.16
DNA coverage	−5.2514	3.82E−05	2.02
L1 coverage	8.2532	1.87E−09	2.95
L2 coverage	−4.8572	2.02E−05	1.61
Low-complexity repeats coverage	3.7462	1.56E−04	1.08
Mirror repeat count	−2.7408	5.41E−03	0.67
SINE count	10.5131	6.26E−08	2.50
Distance to telomere	−44.2594	4.50E−27	20.70
Z-DNA coverage	−4.0246	1.16E−05	1.61
Simple repeat coverage	−6.7009	9.29E−04	1.02
Explained deviance			51.83

Extremely randomized tree classifier for telling apart CHSs and NHSs

We applied the extremely randomized tree classifier to distinguish CHSs and NHSs using the same 25 features. For the CHSs, this classifier reaches the area under the receiver operating characteristic (ROC) curve (AUC) of 0.96 (Fig. 5A). The important features determined by the classifier for CHSs are distance to telomere, indel rate and direct repeats (Fig. 5B), which is generally consistent with the predictors identified in the RELR model. These results suggest that the positions of common breakpoint hotspots can be reasonable well predicted from local genomic properties.

Discussion

Using a MLR model trained on 19 genomic properties, a previous study revealed top four genomic features, including indel rate, exon density, substitution rate and SINE coverage, contributing to SCNA breakpoint formation (8). Taking advantage of the TCGA Pan-Cancer SCNA data, we considered a wider range of genomic features than in (8) and performed prescreening of features to reduce the effect of multicollinearity. Our MLR model is more than two times more powerful than that in (8) (32% of breakpoint variance explained versus 14%) and maintains its strong performance upon 5-fold cross validation. By including six novel genomic features, our models revealed two novel predictors—distance to telomere and distance to centromere—which made the strongest contribution to our model (relative contribution of 29.15 and 10.35% to MLR model for pooled SCNA breakpoints). The inclusion of these two features may explain the superiority of our model compared with that described in (8). Notably, out of the top four features reported in (8) SINE coverage ranked sixth in predictive importance in our model, while the other three features—indel rate, exon density and substitution rate—were not among the significant predictors in our model (rank below 13th, see Supplementary Material, Table S11). When applying the same model to an independent data set, exon density and indel rate have some predictive power and rank second and last, respectively (Supplementary Material, Table S10). We, thus, encountered some discrepancies between the results obtained on the TCGA data and the independent COSMIC data set. However, we found that distance to telomere, distance to centromere, CpG island coverage and mirror repeat count affect SCNA formation in both data sets, and the general consistency of the results obtained on these two datasets emphasizes the reliability of our findings. The power of the models was upheld for different SCNA types (amplifications and deletions),

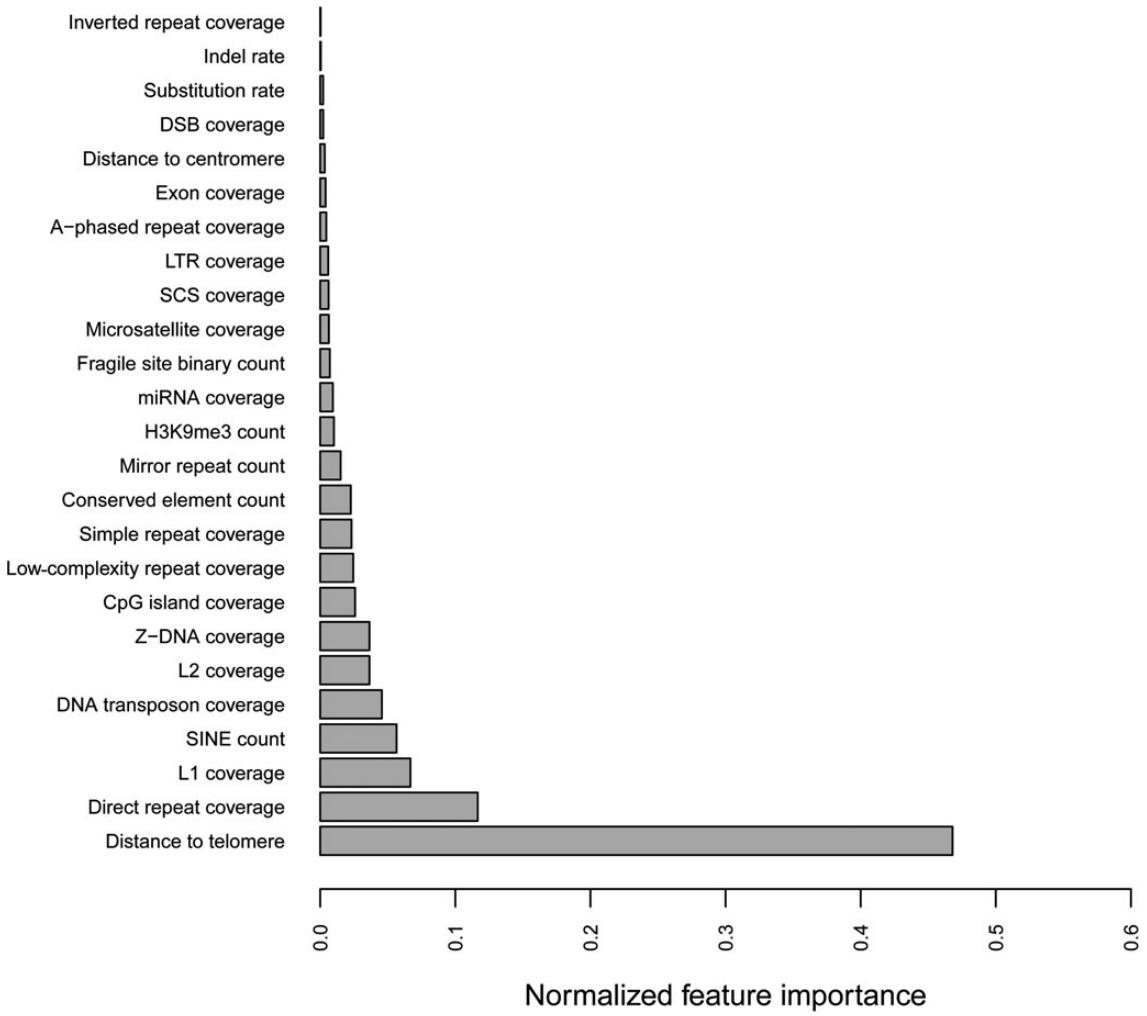

Figure 4. The normalized relative contribution of predictors in terms of distinguishing CHSs and NHSs for the RELR model.

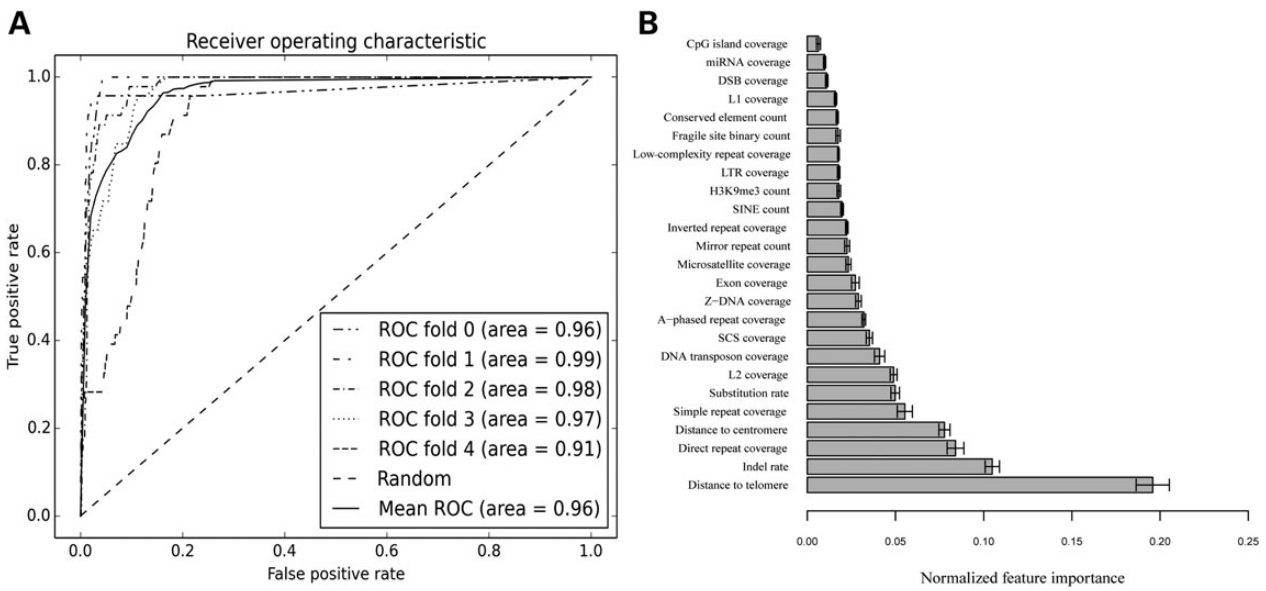

Figure 5. Distinguishing CHSs from NHSs from genomic features. (**A**) ROC–AUC curves of the extremely randomized forests. (**B**) The normalized relative contribution of predictors in terms of distinguishing CHSs and NHSs.

for SCNAs generated by distinct mechanisms (telomere-bound SCNAs and interstitial SCNAs) and for SCNAs from different cancer types. The TCGA Pan-Cancer analysis has revealed two types of SCNAs: interstitial SCNAs and telomere-bound ones (5). The frequency of interstitial SCNAs is inversely correlated with their lengths (2,5), while the telomere-bound ones tend to follow a uniform length distribution (5), which reflects distinct mechanisms underlying their formation. Indeed, in our study distance to centromere contributes strongly to the MLR model for interstitial SCNAs, while distance to centromere has a much smaller role than distance to telomere and direct repeat coverage in the MLR model for telomere-bound SCNAs. According to the MLR model the breakpoints of interstitial SCNAs are overrepresented close to centromeres, which is consistent with the previous observations (5,23,24). Frequent breakages near centromeres may lead to their dysfunction and further cause chromosomal instability (25), which is a hallmark of diverse cancers (26). The prevalence of telomere-bound SCNAs in cancers may relate to telomere dysfunction (27), and those breakpoints of telomere-bound SCNAs that are not located in telomeres were speculated to occur at regions with DSBs (5). Our MLR models for telomere-bound SCNAs favor this hypothesis and demonstrate frequent occurence of DSBs in regions enriched in direct repeats. Direct repeats have been documented previously to cause hairpins and to overlap with chromosome regions undergoing somatic rearrangements (28). The high-prediction power of direct repeats in every cancer type suggests their significant common role in shaping the distribution of SCNA breakpoints.

We also demonstrate that mirror repeat count, L1 coverage, SINE count, low-complexity repeat coverage and several other features have important albeit smaller roles in our MLR models. SINEs and L1 have been extensively studied for their roles in non-allelic homologous recombination, which leads to deletions, duplications and inversions (16,29). The significant positive effect of low-complexity repeats for all cancer types is in line with the fact that they are usually AT-rich and prone to causing the replication fork to pause or stall (30) and thus induce breaks. Moreover, AT-rich repeats constitute unstable regions of the genome, conferring susceptibility to rearrangements (31). These results suggest a general mechanism of genome instability induced by genomic context.

Using the same 25 genomic features to contrast CHSs and NHSs of SCNA breakpoints, we applied extremely tree classifiers to train the model and obtained a more powerful model compared with that in (8) (AUC: 0.96 versus 0.75). RELR and extremely tree classifiers both revealed distance to telomere and direct repeat coverage as being particularly potent in distinguishing CHSs and NHSs of SCNA breakpoints. The consistency of the results obtained by rare-event logistic models and extremely tree classifiers corroborates the robustness of our conclusions. It is noteworthy that indel rate is an important predictor in extremely randomized tree classifiers, but not in rare-event logistic models. The strong contrast between CHSs and NHSs for SCNA breakpoints in terms of the distance to telomere and direct repeat coverage indicates that CHSs strongly depend on the local genomic context. Given that only few known cancer genes are located in common breakpoint hotspot regions (2,8), Li *et al.* (8) hypothesized that the high frequency of SCNAs in these CHSs across cancer types is largely due to regionally higher mutation rate (8). The regions with intrinsically higher mutation rate are independent of tumor type (or tissue origin) and are usually shared across different caner types. Since the regions enriched in direct repeats and/or those close to telomeres are susceptible to mutations, our models comply with this hypothesis.

Materials and Methods

SCNAs data

The first SCNA data published in (5) were kindly provided by Travis I. Zack and Rameen Beroukhim (Dana-Farber Cancer Institute, USA). SCNAs were obtained by mapping the signal intensities from the Affymetrix Genome-Wide Human SNP Array 6.0 in each cancer sample upon removing the probes in regions of recurrent germline CNVs identified from normal tissue samples. The data were provided as files with 105 890 and 96 354 individual SCNAs corresponding to amplifications and deletions. For each individual SCNA the files contain its chromosomal coordinates (chromosome number as well as start and end positions), TCGA barcode (sample identity), amplitude of copy number change and other information. We grouped SNCAs from the same cancer type based on the Pan-Cancer project sample information from synapse.org (syn1710466). Both boundaries of each SCNA were defined as breakpoints with a precision of ~1 kb (the median inter-marker distance for Affymetrix Genome-Wide Human SNP Array 6.0 is <700 bases). In total, we obtained 404 488 SCNA breakpoints from 4943 samples across 11 cancer types, of which 211 780 and 192 708 breakpoints correspond to amplifications and deletions, respectively (Table 1). We also subdivided all SCNAs into two categories: telomere-bound SCNAs, with at least one boundary situated on a telomere, and interstitial SCNAs, with both boundaries interstitial to the chromosome. Specifically, for each chromosome we defined those SCNAs started at the left-most position or ended at the right-most position of the chromosome as telomere-bound SCNAs (see Fig. 6). All the remaining SCNAs were considered to be interstitial. We further subdivided SCNAs into local and chromosome-level ones. Chromosome-level SCNAs were defined as those having the left boundary at the left-most position and the right boundary at the right-most position in the given chromosome, while all other SCNAs were considered local (Fig. 6). By definition, all chromosome-level SCNAs are also telomere-bound, and all interstitial SCNAs are also local SCNAs. The second data set was from the COSMIC database (version 73) (32), and we retrieved 699 492 SCNAs generated by studies other than TCGA (COSMIC study identifiers: 328, 382, 538, 585, 586, 589 and 650).

Data collection on genomic features

A total of 29 genomic features were considered as potential predictors of the SCNA patterns (Table 2). Their genomic coordinates were either obtained from public databases and published studies or identified in this study. All coordinates correspond to the human genome assembly hg19 and, where necessary, the University of California, Santa Cruz (UCSC) liftOver tool was used to convert the hg18 coordinates to hg19 (33).

Chromosomal coordinates of the following genomic features were downloaded from the UCSC Genome Browser (33): probes of the Affymetrix Genome-Wide Human SNP Array 6.0 (retrieved from the SNP/CNV Arrays track); LTR retrotransposons, L1, L2, SINE, DNA transposons and low-complexity repeats (retrieved from the RepeatMasker track); telomeres, centromeres and genome assembly gaps (retrieved from the Gap track); microsatellites; simple repeats; CpG islands; exons and SCs. The latter elements are essentially pairs of short (up to 1 kb) low-copy repeats either in direct (+) or inverted (–) orientation (19). Following

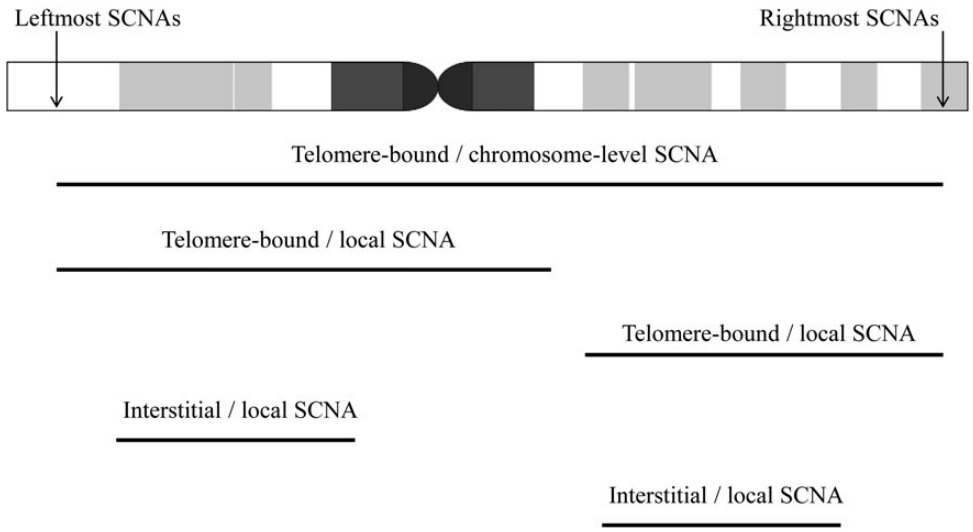

Figure 6. Schematic illustration of SCNA categories considered in this work.

(19) we only considered self-chain segments (SCSs) consisting of paired SCs located on the same chromosome as well as their spacing gaps with the total lengths of up to 30 kb. Furthermore, we removed any SCSs overlapping with gaps in the human genome assembly (including centromeres, telomeres, heterochromatin regions etc.) and segmental duplications.

Non-B DNA motifs (A-phased repeats, direct repeats, inverted repeats, mirror repeats, G-quardruplexes (G4) and Z-DNA) were downloaded from the non-B DB version 2 (12). We used the data set of conserved DNA elements in vertebrates published by Siepel *et al.* (34). Regions containing DSBs were downloaded from Tchurikov *et al.* (35). Genomic coordinates for each histone modification marker H3K9me3 in $CD4^+$ T cells were obtained from the study of Barski *et al.* (36). Replication timing (RT) data for the lymphoblastoid cell line GM06990 were obtained from Hansen *et al.* (37). For each 1 kb window of the genome sequence, we obtained percent-normalized tag density values for the six phases of the cell cycle (denoted G1b, S1, S2, S3, S4 and G2). As suggested by the authors, a weighted average of the data based on the progression of each cell cycle was utilized, and RT was defined by the following formula:

$$\begin{aligned} RT = {} & (0.917 \times G1b) + (0.75 \times S1) + (0.583 \times S2) \\ & + (0.417 \times S3) + (0.25 \times S4) + (0 \times G2). \end{aligned}$$

Higher RT values correspond to earlier replication events. The percentage of G/C nucleotides (GC coverage) for specific genomic regions was calculated using the *nuc* utitlity, which is part of BEDTools (38). The genome-wide distribution of the 13-mer CCNCCNTNNCCNC motifs related to recombination hotspots was obtained by *FUZZNUC* searches [as implemented in the European Molecular Biology Open Software Suite package (39)]. We obtained the coordinates for fragile sites and miRNA genes from a previous study (21) and miRbase (40), respectively. The rates of nucleotide substitutions and indels were calculated based on human–chimpanzee alignments as described in (8).

Data transformation and prescreening of SCNA predictors

Genomic features described above were considered as potentially affecting the patterns of SCNA occurrence across the genome. We partitioned the human genome into non-overlapping 1 Mb windows, after excluding gaps in the genome assembly. The features were measured as counts (number of copies in a window), coverage (fraction of a window occupied by the feature), distance in base pairs to a telomere or a centromere or sum (specifically, the sum of the RT values of 1 kb fragments in a 1 Mb window) (Table 2). All features were evaluated for normality, and if necessary transformed by the logarithm function to approximate it (Table 2). In order to improve the efficiency of model selection for the subsequent regression analyses (see below) and reduce the influence of multicollinearity, we performed the same filtering process for the genomic features as in (21,22). We used hierarchical clustering to identify clusters of features based on Spearman's rank correlation coefficient using a threshold of 0.8. From each such cluster, we selected one representative feature, thus ensuring relatively low linear dependencies.

Identification of CHSs and NHSs for breakpoints across cancer types

Breakpoint hotspots, i.e. genomic regions in which breakpoints are significantly enriched, were identified according to the method described in (6,8,41). We split the human genome into non-overlapping 1 Mb windows and excluded from consideration windows with extremely low Affymetrix Genome-Wide Human SNP Array 6.0 probe density (below three standard deviations from the mean). The number of breakpoints for each cancer type was counted in each 1 Mb window. The same procedure was applied to SCNA breakpoint positions randomized 1000 times in order to generate the null distribution expected by chance. Randomization and counting of breakpoints were performed using BEDTools (38). We assumed a normal distribution for the randomly generated samples and computed *P*-values from the parameterized normal cumulative density function. The windows with FDR-corrected $P < 0.05$ were defined as breakpoint hotspots. We defined the 1 Mb breakpoint hotspots shared in all 11 cancer types as CHSs and the 1 Mb windows which are not identified as breakpoint hotspot in any cancer type as NHSs. The remaining 1 Mb breakpoint hotspots were defined as NCHSs, including hotspots found in only one cancer type and hotspots identified in some, but not all cancer types.

MLR analysis

MLR models an approximately continuous response on the predictors. MLR builds the linear relationship between the predictors and the response. All surveyed genomic features measured in 1 Mb segments were used as potential predictors of SCNA occurrence across the human genome. The density of SCNA breakpoints in every 1 Mb window was determined both for all cancer types pooled together and for each cancer type individually. In addition, in each window we also calculated the breakpoint density of copy number amplifications and deletions, as well as telomere-bound and interstitial SCNAs. Further, for each window, we also computed the SCNA breakpoint densities after excluding chromosome-level SCNAs with both boundaries located approximately at telomeres. These densities were used as response variables for MLR.

To diagnose multicollinearity of each predictor, variance inflation factors (VIFs) were calculated to avoid problems caused by the instability of the coefficients. R^2 was used to capture the explanatory power of the MLR model. For the MLR model, the RCVE of each predictor was defined as:

$$\mathrm{RCVE} = 1 - R^2_{\mathrm{reduced}}/R^2_{\mathrm{full}}$$

where R^2_{full} and R^2_{reduced} denote the residual sum of squares of the full model (including all of the tested predictors) and the reduced model without the predictor of interest, respectively. Moreover, we tested the robustness of the MLR model by substituting some of the predictors with other highly correlated features. We performed k-fold cross validation (42) of the MLR model by randomly dividing the data into k-folds of the same size, using $k - 1$ folds of the data as a training dataset, and testing the model on the remaining fold. The results from each fold test are combined to produce a single estimate, which we call k-fold MLR. The mean of the k-fold adjusted R^2 for the model and k-fold RCVE for each predictor are denoted as k-fold adjusted R^2 and k-fold RCVE, respectively.

All statistical analyses were performed in the R environment (43). The MASS (44) and Car (45) packages were used to generate the common diagonostic plots (e.g. residual plots, Q–Q plots) and the QuantPsyc (46) package was used to calculate the standardized coefficient of predictors (with the signs of plus or minus denoting the positive or negative effect that predictors have on the response). The DAAG (47) package was used to perform k-fold cross validation. RCVEs were represented graphically in heatmaps. Predictors with FDR-corrected $P < 0.05$ are considered to be significant.

Distinguishing between CHSs and NHSs by LR

LR was used to distinguish between CHSs (binary response 1) and NHSs (binary response 0) using the same predictors as in the MLR model. To eliminate the possible small-sample size bias, we increased the number of CHSs by applying a sliding procedure. Specifically, we divided the human genome into sliding windows of 1 Mb in length with a step size of 100 kb. We also applied RELR (48) to reduce the sample imbalance bias. The RELR analysis was performed with the help of the statistical software Zelig (http://gking.harvard.edu/zelig) (49) using the same predictors as in the LR model. We used pseudo R^2 to capture the explanatory power of the LR and RELR models. The relative contribution of each predictor for both models (RCVE) was calculated by the formula:

$$\mathrm{RCVE} = [(D_0 - D) - (D_0 - D_{(-p)})]/(D_0 - D),$$

where D_0 and D are the null deviance and residual deviance of the model, respectively, and $D_{(-p)}$ is the deviance of the resulting model after removing the predictor of interest.

Distinguishing between CHSs and NHSs by an extremely randomized tree classifier

A classification decision tree (50) is an input–output model represented by a tree structure. As a single-decision tree usually suffers from high variance, ensembles of decision trees have been proposed to circumvent this problem. In this work, we applied the extremely randomized tree classifier to distinguish between CHSs and NHSs using the same features as in the MLR and LR models. The extremely randomized tree classifier is implemented in Scikit-Learn, a collection of Python modules of common machine learning algorithms (http://scikit-learn.org) (51). We chose to build 500 trees to obtain robust results, growing each tree to its full depth. To balance the input data classes, sample weights were passed to the classifier. The predictive performance of the classifier was assessed by AUC obtained on the data set by 5-fold cross-validation: in each validation around 80% of the data were used as the training data and the remaining 20% were used as the test data. The final AUC values were computed by averaging AUCs over the 5-folds. Feature importance in extremely randomized tree classifiers was assessed based on the mean decrease impurity importance, which gets computed and normalized in Scikit-Learn by default.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We thank Travis I. Zack for kindly providing SCNA data and for his critical comments, Yudong Li and Subhajyoti De for their suggestions on randomization of SCNA breakpoints, Weichen Zhou for his help in preparing the data for SCSs, Feng Zhang for helpful comments on our manuscript and Norbert Krautenbacher for his advice on statistics.

Conflict of Interest statement. None declared.

Funding

Y.Z. and H.X. gratefully acknowledge the financial support of the China Scholarship Council.

References

1. Stratton, M.R. (2011) Exploring the genomes of cancer cells: progress and promise. *Science*, **331**, 1553–1558.
2. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J., Dobson, J., Urashima, M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
3. Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
4. Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
5. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.Z., Wala, J.,

Mermel, C.H. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, **45**, 1134–1140.
6. De, S. and Michor, F. (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat. Struct. Mol. Biol.*, **18**, 950–955.
7. Fudenberg, G., Getz, G., Meyerson, M. and Mirny, L.A. (2011) High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.*, **29**, 1109–1113.
8. Li, Y., Zhang, L., Ball, R.L., Liang, X., Li, J., Lin, Z. and Liang, H. (2012) Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots. *Hum. Mol. Genet.*, **21**, 4957–4965.
9. Gu, W., Zhang, F. and Lupski, J.R. (2008) Mechanisms for human genomic rearrangements. *Pathogenetics*, **1**, 4.
10. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. and Ira, G. (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.
11. Crosetto, N., Mitra, A., Silva, M.J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K. *et al.* (2013) Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods*, **10**, 361–365.
12. Cer, R.Z., Donohue, D.E., Mudunuri, U.S., Temiz, N.A., Loss, M. A., Starner, N.J., Halusa, G.N., Volfovsky, N., Yi, M., Luke, B.T. *et al.* (2013) Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.*, **41**, D94–D100.
13. Wang, G., Christensen, L.A. and Vasquez, K.M. (2006) Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl Acad. Sci. USA*, **103**, 2677–2682.
14. Inagaki, H., Ohye, T., Kogo, H., Kato, T., Bolor, H., Taniguchi, M., Shaikh, T.H., Emanuel, B.S. and Kurahashi, H. (2009) Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res.*, **19**, 191–198.
15. Wang, G. and Vasquez, K.M. (2004) Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl Acad. Sci. USA*, **101**, 13448–13453.
16. Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L. and Batzer, M.A. (2008) L1 recombination-associated deletions generate human genomic variation. *Proc. Natl Acad. Sci. USA*, **105**, 19366–19371.
17. Campbell, I.M., Gambin, T., Dittwald, P., Beck, C.R., Shuvarikov, A., Hixson, P., Patel, A., Gambin, A., Shaw, C.A., Rosenfeld, J.A. *et al.* (2014) Human endogenous retroviral elements promote genome instability via nonallelic homologous recombination. *BMC Biol.*, **12**, 74.
18. Myers, S., Freeman, C., Auton, A., Donnelly, P. and McVean, G. (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.*, **40**, 1124–1129.
19. Zhou, W., Zhang, F., Chen, X., Shen, Y., Lupski, J.R. and Jin, L. (2013) Increased genome instability in human DNA segments with self-chains: homology-induced structural variations via replicative mechanisms. *Hum. Mol. Genet.*, **22**, 2642–2651.
20. Schuster-Bockler, B. and Lehner, B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.
21. Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K.A. and Makova, K.D. (2012) A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome Res.*, **22**, 993–1005.
22. Campos-Sánchez, R., Kapusta, A., Feschotte, C., Chiaromonte, F. and Makova, K.D. (2014) Genomic landscape of human, bat and ex vivo DNA transposon integrations. *Mol. Biol. Evol.*, **31**, 1816–1832.
23. Alsop, A.E., Teschendorff, A.E. and Edwards, P.A. (2006) Distribution of breakpoints on chromosome 18 in breast, colorectal, and pancreatic carcinoma cell lines. *Cancer Genet. Cytogenet.*, **164**, 97–109.
24. Nguyen, D.Q., Webber, C. and Ponting, C.P. (2006) Bias of selection on human copy-number variants. *PLoS Genet.*, **2**, e20.
25. Manning, A.L., Longworth, M.S. and Dyson, N.J. (2010) Loss of pRB causes centromere dysfunction and chromosomal instability. *Genes Dev.*, **24**, 1364–1376.
26. Negrini, S., Gorgoulis, V.G. and Halazonetis, T.D. (2010) Genomic instability—an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.*, **11**, 220–228.
27. Artandi, S.E., Chang, S., Lee, S.L., Alson, S., Gottlieb, G.J., Chin, L. and DePinho, R.A. (2000) Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature*, **406**, 641–645.
28. Zhao, J., Bacolla, A., Wang, G. and Vasquez, K.M. (2010) Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.*, **67**, 43–62.
29. Konkel, M.K. and Batzer, M.A. (2010) A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Semin. Cancer Biol.*, **20**, 211–221.
30. Freudenreich, C.H. (2007) Chromosome fragility: molecular mechanisms and cellular consequences. *Front Biosci.*, **12**, 4911–4924.
31. Edelmann, L., Spiteri, E., McCain, N., Goldberg, R., Pandita, R. K., Duong, S., Fox, J., Blumenthal, D., Lalani, S.R., Shaffer, L.G. *et al.* (1999) A common breakpoint on 11q23 in carriers of the constitutional t(11;22) translocation. *Am. J. Hum. Genet.*, **65**, 1608–1616.
32. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
33. Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
34. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
35. Tchurikov, N.A., Kretova, O.V., Fedoseeva, D.M., Sosin, D.V., Grachev, S.A., Serebraykova, M.V., Romanenko, S.A., Vorobieva, N.V. and Kravatsky, Y.V. (2013) DNA double-strand breaks coupled with PARP1 and HNRNPA2B1 binding sites flank coordinately expressed domains in human chromosomes. *PLoS Genet.*, **9**, e1003429.
36. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
37. Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M. and Stamatoyannopoulos, J.A. (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA*, **107**, 139–144.

38. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
39. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
40. Kozomara, A. and Griffiths-Jones, S. (2011) miRbase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
41. De, S., Pedersen, B.S. and Kechris, K. (2014) The dilemma of choosing the ideal permutation strategy while estimating statistical significance of genome-wide enrichment. *Brief. Bioinform.*, **15**, 919–928.
42. Olson, D.L. and Delen, D. (2008) *Advanced Data Mining Techniques*, Springer, Berlin, Germany.
43. R Development Core Team. (2011) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
44. Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*, Springer, New York.
45. Fox, J. and Weisberg, S. (2011) *An {R} Companion to Applied Regression*, Sage, Thousand Oaks, CA.
46. Fletcher, T.D. (2012) *QuantPsyc: Quantitative Psychology Tools*, http://CRAN.R-project.org/package=QuantPsyc (date last accessed, March 30, 2015).
47. Maindonald, J.H. and Braun, W.J. (2010) *Data Analysis and Graphics Using R*, Cambridge University Press, Cambridge, UK.
48. King, G. and Zeng, L. (2001) Logistic regression in rare events data. *Polit. Anal.*, **9**, 137–163.
49. Imai, K., King, G. and Lau, O. (2008) Toward a common framework for statistical analysis and development. *J. Comput. Graph. Stat.*, **17**, 1–22.
50. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth, Belmont, CA.
51. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.