Human Molecular Genetics, 2012, Vol. 21, No. 22 ***4957–4965***
doi:10.1093/hmg/dds340
Advance Access published on August 16, 2012

Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots

Yudong Li[1,2], Li Zhang[2], Robyn L. Ball[3], Xinle Liang[1], Jianrong Li[1], Zhenguo Lin[4] and Han Liang[2,*]

[1]Department of Bioengineering, School of Food Sciences and Biotechnology, Zhejiang Gongshang University, Hangzhou 310035, PR China [2]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA [3]Department of Statistics, Texas A&M University, College Station, TX 77843, USA [4]Department of Ecology and Evolution, The University of Chicago, Chicago, IL 60637, USA

Received May 12, 2012; Revised July 20, 2012; Accepted August 11, 2012

Somatic copy-number alterations (SCNAs) play a crucial role in the development of human cancer. However, it is not well understood what evolutionary mechanisms contribute to the global patterns of SCNAs in cancer genomes. Taking advantage of data recently available through The Cancer Genome Atlas, we performed a systematic analysis on genome-wide SCNA breakpoint data for eight cancer types. First, we observed a high degree of overall similarity among the SCNA breakpoint landscapes of different cancer types. Then, we compiled 19 genomic features and evaluated their effects on the observed SCNA patterns. We found that evolutionary indel and substitution rates between species (i.e. humans and chimpanzees) consistently show the strongest correlations with breakpoint frequency among all the surveyed features; whereas the effects of some features are quite cancer-type dependent. Focusing on SCNA breakpoint hotspots, we found that cancer-type-specific breakpoint hotspots and common hotspots show distinct patterns. Cancer-type-specific hotspots are enriched with known cancer genes but are poorly predicted from genomic features; whereas common hotspots show the opposite patterns. This contrast suggests that explaining high-frequency SCNAs in cancer may require different evolutionary models: positive selection driven by cancer genes, and non-adaptive evolution related to an intrinsically unstable genomic context. Our results not only present a systematic view of the effects of genetic factors on genome-wide SCNA patterns, but also provide deep insights into the evolutionary process of SCNAs in cancer.

INTRODUCTION

Cancer arises when cells grow in an unregulated fashion as a result of having acquired critical somatic changes. These changes include point mutations, insertions/deletions (indels), rearrangements (e.g. translocations and inversions) and somatic copy-number alterations (SCNAs, distinguished from germline copy-number variations, CNVs) (1). Among them, SCNAs often play a crucial role in the development of human cancers through the amplification of oncogenes or deletion of tumor suppressors (2,3). Recent advances in high-throughput genomic technologies such as array comparative genomic hybridization (4) and next-generation DNA sequencing (5) have enabled us to characterize genome-wide SCNAs in patient samples of cancer in a timely and cost-efficient manner (2,6). For example, one recent study presented a collection of high-resolution SCNA profiles from >3000 samples of 26 cancer types and systematically identified cancer SCNA hotspots consistently across different cancer types (2). Further computational analyses on the DNA breakpoints resulting from SCNAs revealed that the breakpoints are often clustered in some regions within the human genome (so-called breakpoint hotspots) and these hotspots are preferably associated with G-quadruplexes (G4) (7).

*To whom correspondence should be addressed at: Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA. Tel: +1 7137459815; Fax: +1 7135634242; Email: hliang1@mdanderson.org

Moreover, DNA three-dimensional structures (8) and DNA replication timing (9) have been suggested as the key genetic factors predicting the genome-wide SCNA patterns observed. While these efforts greatly advance our view on the mechanistic basis of SCNA occurrence, it remains not well understood what evolutionary mechanisms contribute to the global patterns of SCNAs in cancer genomes.

Previous SCNA genomic analyses have focused primarily on the data sets of a single cancer type or pooled data sets from different cancer types. Although these studies elucidate some important general trends, in-depth comparative analyses on SCNA data of different cancer types are crucial for understanding the evolutionary mechanisms underlying SCNA occurrence and identifying cancer-type-specific characteristics. The Cancer Genome Atlas (TCGA) project (10,11) is a valuable resource for this purpose, providing an unprecedented collection of patient samples that have been characterized for each cancer type in multiple dimensions using consistent platforms.

In this study, we performed a comparative analysis on SCNAs across eight different types of human cancer using TCGA data. We first demonstrated that the overall SCNA breakpoint landscapes among different types of cancer show high similarities; and then we evaluated the effects of various genomic features on the genome-wide SCNA patterns. Importantly, we found that cancer-type-specific SCNA breakpoint hotspots and common hotspots show distinct patterns: cancer-type-specific hotspots are enriched with known cancer genes; whereas common hotspots do not show such enrichment but can be relatively well predicted from genomic context. Our results not only present a systematic view of the genome-wide SCNA breakpoint patterns in human cancer, but also provide deep insights into the evolutionary mechanisms underlying SCNA breakpoint hotspots.

RESULTS

The overall similarity of SCNA breakpoint distributions across different types of cancer

TCGA employs several platforms to characterize genome-wide SCNAs, such as the Affymetrix SNP6.0 array (AffySNP6), Agilent 244K array and Illumine 1MDUO array (11). For this study, we focused on SCNA data generated from the AffySNP6 platform, which contains more than 1.8 million probes for detecting CNVs. We chose the AffySNP6 array because it offers (i) one of the highest probe densities, and (ii) large sample sizes for several different types of cancer. For comparative purposes, we chose eight types of cancer with the largest sample sizes (Table 1), each with at least 150 tumor samples (with paired normal tissue specimens from the same individual). For each cancer type, we first filtered the outlier samples, those with extremely large numbers of segments. Then, for each sample, we identified confident SCNA segments by considering the position, length and amplitude of change in normalized copy-number data (see details in Materials and Methods).

Table 1 shows the statistical summary of SCNA data used in our study. On average, we identified ~80 SCNAs (including both gain and loss) per sample, which is comparable to that observed in the literature (2). Among different cancer types, ovarian cancer has the highest segment number (125 per sample), consistent with the notion that ovarian cancer is a disease driven by SCNAs (11,12). To characterize the genome-wide SCNA patterns, we divided the human genome into 1 Mb non-overlapping blocks and calculated the number of SCNA breakpoints within each block, as in De and Michor (7). In order to validate the SCNA breakpoint data we obtained, we further examined 246 TCGA samples of ovarian cancer profiled with both AffySNP6 6.0 and Agilent 1M platforms. Using the same data-processing method, we found the breakpoint frequency for 1 Mb blocks obtained from these two platforms to be highly correlated ($Rs = 0.8$, $P < 2.2 \times 10^{-16}$). This result supports the robustness of our SCNA breakpoint data.

As shown in Figure 1, the distributions of SCNA breakpoint frequencies across different types of cancer are highly similar, and there are many breakpoint peaks shared across different types of cancer. Such a similarity also holds true when the breakpoint distributions were calculated separately based on the amplification and deletion SCNAs (Supplementary Material, Table S1). On average, Spearman's rank correlation coefficient of the genome-wide breakpoint distributions between pair-wise cancer types is 0.5, ranging from 0.32 to 0.64 (Supplementary Material, Fig. S1). At the chromosomal level, chromosomes 9 and 22 appear to show higher similarities than the others (Supplementary Material, Fig. S2). Each cancer type is also characterized by specific breakpoint peaks (red peaks in Fig. 1); we will investigate them in greater detail in subsequent sections.

The effect of various genomic features on SCNA breakpoint occurrence

To investigate how various genetic factors potentially affect the genome-wide SCNA breakpoint patterns, we compiled a comprehensive list of 19 genomic features for each 1 Mb block (Table 2). These features can be generally classified into four groups: (i) sequence features, (ii) DNA secondary structure motifs, (iii) evolutionary features, and (iv) functional features. Only some of these features have been examined by previous studies for their effects on SCNA breakpoints (7,9). Figure 2A shows the correlation coefficients between each feature and the breakpoint frequency in different types of cancer (Supplementary Material, Fig. S3 shows the distribution of these genomic features and the breakpoint frequency along the human genome). Across different features, we observed that two lung cancer types are clustered together and endometrial cancer is clustered with ovarian cancer, which is consistent with a common organ of origin. Across different cancer types, we found that some features, such as the indel rate between humans and chimpanzees or conserved elements, show consistently positive or negative correlations with the breakpoint frequency across the cancer types. In contrast, the correlations of the other features strongly depend on the cancer type. These results suggest that the effects of genetic factors on SCNA breakpoint occurrence are more complicated than previously appreciated (7,9). Furthermore, when breakpoints resulting from amplifications and deletions

Table 1. Summary of TCGA SCNA data analyzed

Cancer type	Sample #[a]	SCNA breakpoint # (amplification/ deletion)	Breakpoint # per sample
Breast invasive carcinoma (BRCA)	667	121 780 (60 758/ 61 022)	182
Glioblastoma multiforme (GBM)	441	64 820 (32 496/ 32 324)	146
Ovarian serous cystadenocarcinoma (OV)	500	125 595 (622 601/ 62 994)	251
Kidney renal clear cell carcinoma (KIRC)	459	48 907 (24 140/ 24 020)	106
Colon adenocarcinoma (COAD)	392	55 107 (27 577/ 27 530)	140
Uterine corpus endometrioid carcinoma (UCEC)	272	38 673 (19 334/ 19 339)	142
Lung squamous cell carcinoma (LUSC)	187	32 271 (16 116/ 16 155)	172
Lung adenocarcinoma (LUAD)	163	23 161 (11 591/ 11 570)	142

[a]Excluding hyper-segmented outlier.

were considered separately, the observed correlations between genomic features and breakpoint frequencies are quite similar (Supplementary Material, Table S2). In addition, the observed correlations were unlikely to be affected by the probe distribution of AffySNP6 in the human genome (details in Materials and Methods; Supplementary Material, Fig. S4). However, the correlation pattern was quite different based on those filtered outlier samples (Supplementary Material, Fig. S5).

Next, we examined to what extent the observed genome-wide pattern of breakpoints could be explained by these genomic features. Since the breakpoint distributions of different cancers in the human genome are quite similar, we performed a multivariable analysis on the pooled breakpoint data. Through a stepwise forward regression on the transformed breakpoint data, the indel rate shows the best predictive power, and an integrated model using the top four selected features (indel rate + exon + substitution rate + SINE) can collectively explain nearly 14% of the variation in the test data (Fig. 2B, Supplementary Material, Fig. S6), indicating that the intrinsic properties of the local genomic context play a considerable role in affecting SCNA occurrence. We also performed similar analyses for each cancer type (adjusted R^2 ranges from 0.068 to 0.167), and found that the substitution rate, exon density and indel rate were among the most frequently selected features across different types of cancer (Supplementary Material, Table S3).

Enrichment of known cancer genes in cancer-type-specific breakpoint hotspots

As shown in Figure 1, although different types of cancer have many common breakpoint peaks, each type of cancer also has its specific hotspots. To rigorously identify SCNA breakpoint hotspots, we generated a background breakpoint frequency distribution for each type of cancer, and identified the blocks with significantly enriched breakpoints as hotspots (see details in Materials and Methods). We obtained a total of 471 breakpoint hotspots across the eight cancer types with a mean of ~240 per cancer type (ranging from 211 to 267) (Supplementary Material, Table S4). In a comparison with a prior study, we found that >70% of the SCNA hotspots previously identified (2) overlapped with the breakpoint hotspots we identified in this study. To test the relationship between these hotspots and cancer genes, we obtained a set of 468 cancer genes defined by Futreal *et al*. (13) that have been widely used as a set of known cancer genes. We found that the hotspots we identified are significantly enriched with known cancer genes (Supplementary Material, Fig. S7).

Next, we examined how frequently the breakpoint hotspots were detected in the eight types of cancer and found that they exhibited a clear bimodal distribution (Fig. 3A). We defined the hotspots found in only one type of cancer as cancer-type-specific hotspots (highlighted in red in Fig. 1) and defined those detected in all eight cancers as common hotspots. To understand how known cancer genes are related to these two classes of breakpoint hotspots, we examined the overlap between these hotspot blocks and known cancer genes. Interestingly, we found that the number of cancer genes that overlapped with a cancer-type-specific hotspot was significantly higher than what we would expect by random chance (Fig. 3A, $P < 0.001$, Supplementary Material, Table S5), but that the number of cancer genes that overlapped with a common hotspot was lower and not statistically different (Fig. 3B, $P < 0.11$). Indeed, the ratio of cancer-type-specific hotspots that overlapped with a known cancer gene was significantly higher than that for the common hotspots (30.6 versus 12%, respectively, Fisher's exact test, $P < 1.5 \times 10^{-5}$). Consistently, the distance from a type-specific hotspot to its nearest cancer gene was much smaller than that distance for a common breakpoint hotspot (2799 versus 5288 kb, Wilcoxon rank test, $P < 1.5 \times 10^{-6}$). This same observation was made when centromere and telemere-related blocks were excluded from our analysis (Supplementary Material, Fig. S8a and b). Moreover, we obtained similar results when using more flexible definitions (definition of a cancer-type-specific hotspot as that detected in one or two cancer types; a common hotspot, detected in 7 or 8 cancer types) (Supplementary Material, Fig. S8c and d). Taken together, our results indicate that these two classes of hotspots indeed show very different patterns in terms of their co-localization with cancer genes.

Predictability of common cancer breakpoint hotspots from genomic features

To investigate how genomic context can affect the distribution of breakpoint hotspots in the human genome, we compared the predictability of cancer-type-specific hotspots and common hotspots using a machine learning approach. We built a random forest classifier to distinguish 150 common hotspots (or 183 cancer-type-specific hotspots) from 2409 non-hotspot blocks using 19 genomic features. Figure 4A shows the receiver operating characteristic (ROC) curves for the predictive models (10-fold cross-validation). The area under the ROC curve (AUC) is a standard measure for the predictive power

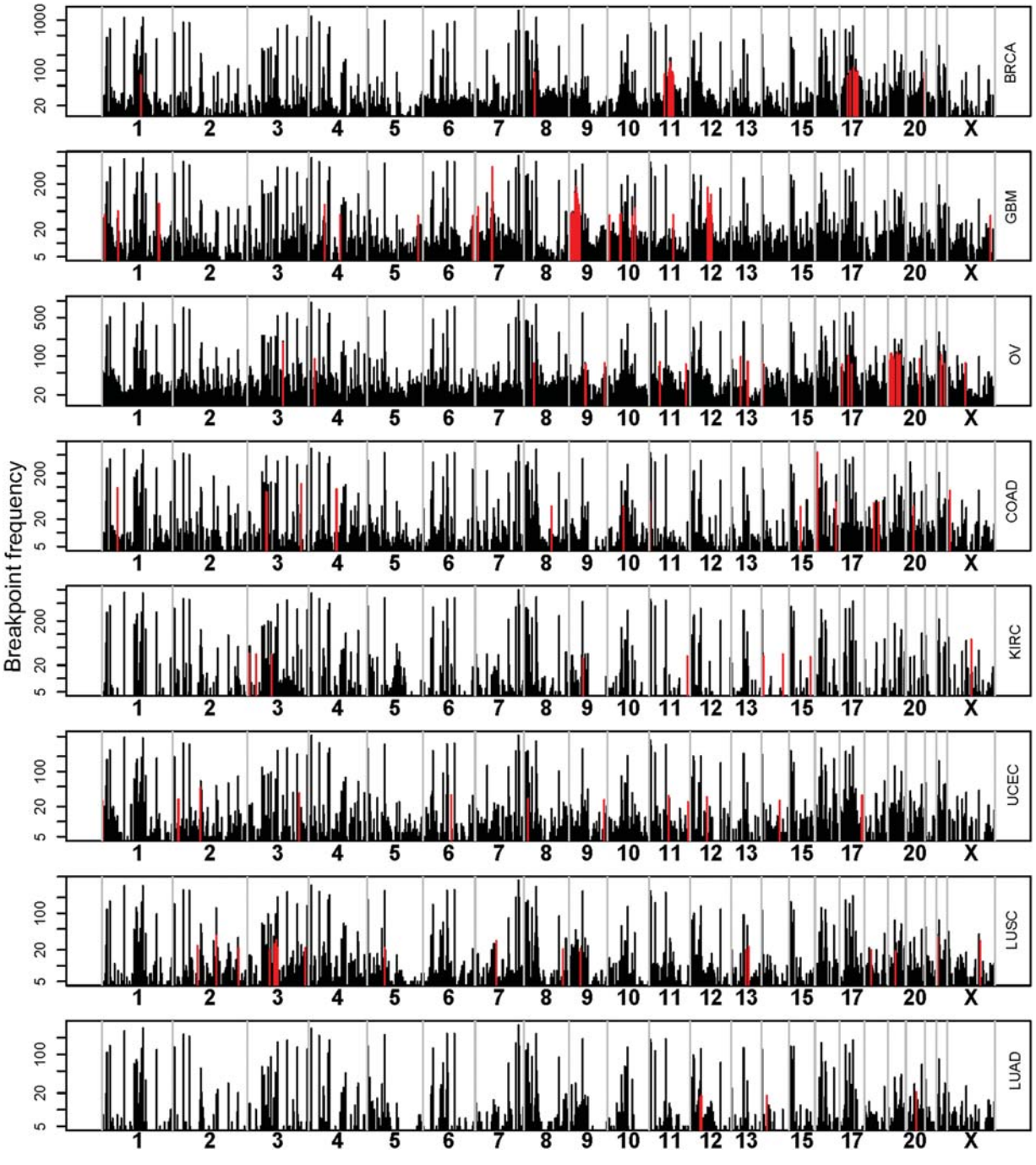

Figure 1. The distribution of SCNA breakpoint frequencies in eight types of human cancer. Cancer-type-specific breakpoint peaks (hotspots) are highlighted in red.

of different classifiers (the diagonal line in the ROC curve has an AUC value of 0.5, representing the power of a random guess). The AUC score for common hotspots is 0.754, whereas that for cancer-type-specific hotspots is only 0.592, indicating a much higher predictability for common hotspots. Furthermore, we obtained similar results when the two types of breakpoint hotspots were defined in a relaxed way (Supplementary Material, Fig. S9). Figure 4B shows a mean decrease in accuracy when a feature is permuted, which is commonly used to measure a feature's relative importance (14). Interestingly, the important features determined in the classifier for common hotspots, such as exon density, substitution rate and indel rate, were consistent with the results of the above multivariate regression analysis. These results suggest that common breakpoint hotspots strongly depend on the local genomic context.

Table 2. Genomic features investigated in this study

Features	Measurement (per Mb)	Sources
Sequence features		
GC content	GC nucleotide	'GC percent' track[a]
CpG islands	Coverage	'CpG island' track
SINEs (Alu, MIR)	Coverage	'RepeatMasker' track
LINEs (L1, L2, etc.)	Coverage	'RepeatMasker' track
Simple Repeats	Coverage	'Simple Repeat' track
Microsatellite	Coverage	'Microsatellite' track
Structure features		
PG4	Coverage	Quadruplex.org
Cruciform	Coverage	Non-B DB
Slipped motif	Coverage	Non-B DB
Z-DNA	Coverage	Non-B DB
Triplex	Coverage	Non-B DB
Evolutionary features		
Conserved elements	Coverage	'17-way most conserved' track
Substitution rate	Percentage	Human–Chimp alignment
Indel rate	Percentage	Human–Chimp–Macaque alignment
Recombination rate	Centimorgan (cM)	Myers *et al*. (23)
Functional features		
Fragile sites	Yes/no	'Gene' database of NCBI
Exon density[b]	Coverage	'UCSC genes' track
miRNA	Count	miRbase database
Replication time	Mean	Chen *et al*. (15)

LINE, long-interspersed repetitive elements; SINE, short-interspersed repetitive element; PG4, Potential guanine-quadruplex sequences.
[a]The track data in this table were downloaded from the UCSC Genome Browser, based on the human reference genome hg18.
[b]Exon base positions overlapping with multiple exons were counted once.

DISCUSSION

Taking advantage of recently available TCGA SCNA data, we systematically analyzed the effect of 19 genomic features on the SCNA breakpoint distribution in different cancer types. To the best of our knowledge, this is the most comprehensive analysis based on the largest SCNA data set of human cancer so far. Recent studies on pooled SCNA breakpoint data have shown that some genomic features we surveyed, such as repeat elements and G4 motifs, are associated with genome-wide SCNA patterns (7,9), but it remains unclear whether the same trend holds for every cancer type. We found that the correlations of these features depend on the cancer type, suggesting that the effect of these genetic factors on SCNA occurrence is more complicated than previously appreciated (7). The replication time data we used were from HeLa cells (15). It is known that replication time substantially varies among different tissues (16), which might contribute to the observed type-dependent correlations. Surprisingly, there is weak or no correlation between the breakpoint frequency and the presence of fragile sites in our analysis. This may be due to the low resolution of fragile-site data: fragile sites in our data set are usually long than 1 Mb, and many of them are longer than 10 Mb; and the data type is binary (0/1). Thus, the power to detect a meaningful correlation is limited. Importantly, some genomic features in our study, such as the indel and substitution rates between humans and chimpanzees, to our knowledge, have not been examined for their correlations with SCNAs previously. Interestingly, these genomic features characterized by evolutionary measures show the best and most consistent correlations with breakpoint frequency among all the features, indicating a parallel pattern between the copy-number gain or loss in the short-term evolution of tumor cells and that observed in long-term macro-evolution.

We observed that SCNA breakpoints in a single cancer type were often clustered into hotspots, which is consistent with previous findings (7). Like other genomic aberrations, the SCNAs observed in cancer genomes essentially result from the combination of two factors: the underlying SCNA mutation rate, which is largely determined by genomic context; and the fitness effect of an alteration, upon which nature selection acts. Our study reveals distinct patterns of two classes of breakpoint hotspots. Cancer-type-specific hotspots show a significant enrichment for known cancer genes, but are poorly predicted by genomic features. In contrast, common hotspots do not show this enrichment and can be well predicted from genomic features. These contrasting patterns may highlight two distinct evolutionary mechanisms underlying high-frequency SCNAs (Fig. 5). Some of these SCNAs are functionally driven: their occurrence is associated with a gain or loss of important cancer genes (oncogenes and tumor suppressors), through which tumor cells with such SCNAs acquire a selective growth advantage compared with other cells. Thus, their propagation in the tumor cell population is driven by positive selection (clonal selection) (17), and they become the dominant cell type. Since tumor cells are subjected to highly specialized micro-environments, such as anoxia, malnutrition and fluctuating hormonal influences (18), the fitness effect of the gain/loss of cancer genes depends on the tumor context and is often specific to the cancer type. Another group of high-frequency SCNAs occurs largely because they reside in intrinsically unstable genomic regions that are associated with a higher underlying SCNA mutation rate and can be characterized by various genomic features. Like many germline CNVs, their evolutionary process may be neutral (that is without positive selection) (19,20). Since the intrinsically unstable regions are independent of a specific tumor environment, they are usually shared across different cancer types. Of course, the above evolutionary models are, respectively, inferred from the group behaviors of the two classes of breakpoint hotspots, not necessarily holding true for every case in the corresponding class. Taken together, these results provide deep insights into the evolutionary mechanisms underlying SCNAs in human cancer, helping us to better define critical somatic change events in tumorigenesis and progression.

MATERIALS AND METHODS

SCNA breakpoint dataset

We obtained the copy-number segmentation profiling data (level 3) of the AffySNP6 platform from TCGA data portal (http://tcga-data.nci.nih.gov/tcga). We selected eight cancer types (BRCA, COAD, GBM, KIRC, LUSC, LUAD, OV and UCEC) (Table 1), each with more than 150 tumor samples in the data set (with paired normal samples from the same

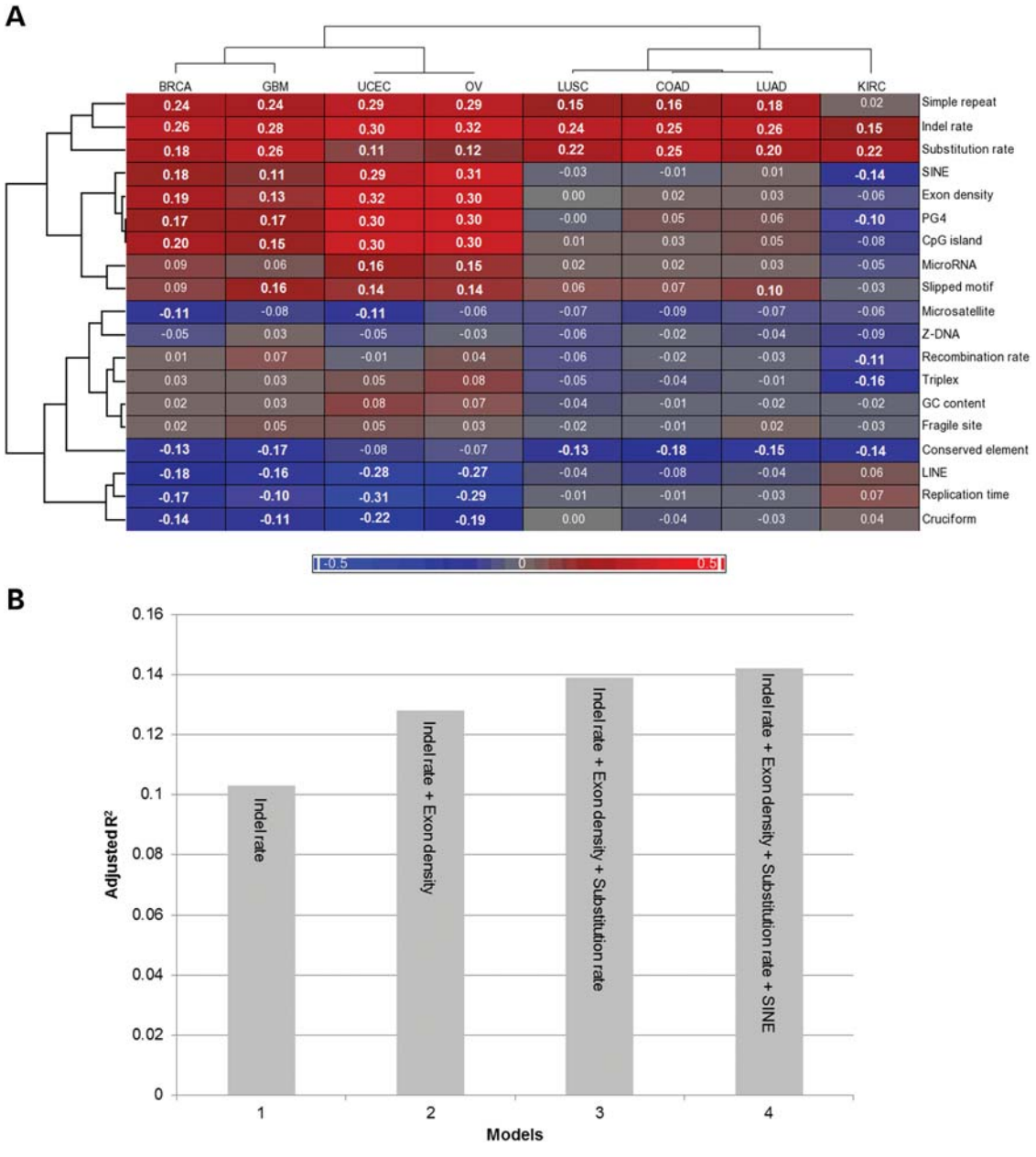

Figure 2. The effect of various genomic features on the genome-wide SCNA breakpoint pattern. (**A**) A heatmap of correlation coefficients between 19 genomic features and breakpoint frequencies in 8 cancer types. The heatmap is colored according to Spearman's rank correlation coefficient between a specific feature and a cancer type, significant correlations are shown in boldface ($|Rs| > 0.1$ and adjusted P-value $< 5 \times 10^{-5}$), and the clustering patterns are obtained with the average-linkage algorithm. (**B**) The predictive power of an integrative model with the top four selected features. The grey bars represent the adjusted R^2 in the test data.

individual). For each type of cancer, based on the raw segmentation-number distribution, we filtered the hyper-segmented samples that were defined as the top 5% of samples with the highest segmentation numbers (the top 10% of samples for LUSC because this cancer had an unusually long tail). To reduce false–positive SCNAs resulting from hypersegmentation, we further filtered segments in the remaining samples by using the following criteria: (i) an amplitude threshold at a neighboring copy-number difference of 0.3, and (ii) covering at least 25 probes. The remaining SCNAs were further used in the study. Finally, for each cancer type, we calculated the breakpoint numbers (each SCNA contributing two breakpoints) for 1 Mb non-overlapping blocks across the whole genome. Given the resolution (100–200 kb) of the AffySNP6 platform, using 1 Mb blocks as the analytic units can greatly reduce the uncertainty of breakpoint detection.

Genomic feature data collection

We obtained the following genomic features in the human genome (hg18) from the UCSC Genome Browser (21): GC content was downloaded as the percentage of G/C nucleotides

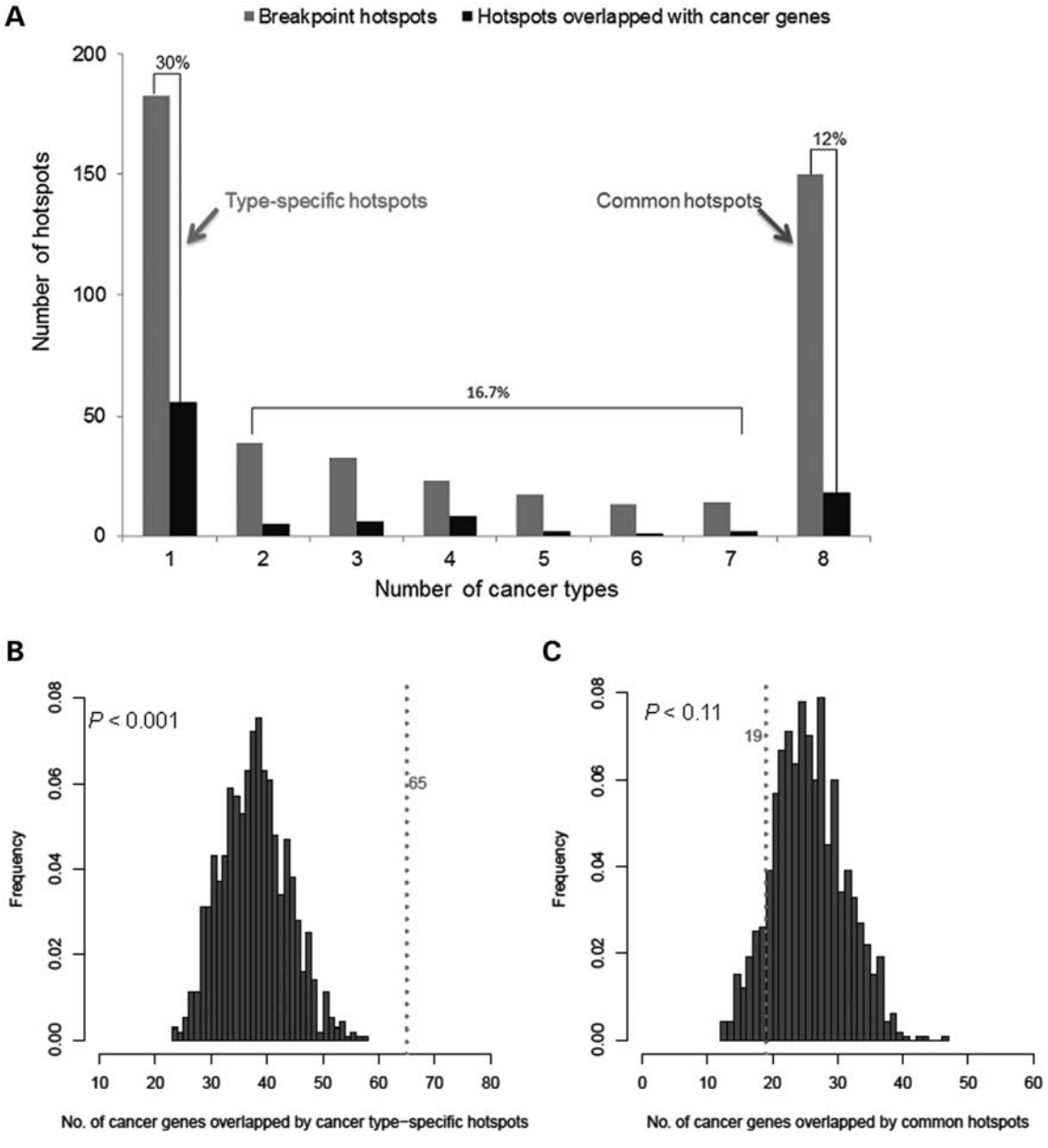

Figure 3. The enrichment of known cancer genes in cancer-type-specific breakpoint hotspots. (**A**) The distribution of breakpoint hotspot frequencies in eight cancer types. The fractions of breakpoint hotspots overlapping with cancer genes are shown for three groups (1, 2−7 and 8). (**B** and **C**) The statistical significance of the numbers of known cancer genes overlapped with cancer-type-specific breakpoint hotspots (B) and common hotspots (C). The background distributions were generated with 1000 random samplings, shown in blue bins. The dotted lines in the graphs represent the numbers of overlapping cancer genes observed from the real breakpoint hotspot data. The *y*-axis represents the frequency of the permutated hotspots overlapped with a specific number of cancer genes.

in 5 nt windows from the GC percent data (gc5base); CpG islands were obtained from the cpgIslandExt table; information about LINEs, SINEs, simple repeats and microsatellites was obtained from the RepeatMasker track, respectively; the coordinates of conserved elements were obtained from the '17-way most conserved' track (phastConsElements17way); exon information was obtained from the knownGene tracks (22) and recombination rates were obtained from the recombRate table (23). The nucleotide substitution (or indel) rate for a 1 Mb block was computed as the ratio of the total number of such substitutions (indels) to the total number of nucleotides in the human−chimpanzee alignment within the 1 Mb block using Galaxy tools (22).

We obtained potential PG4 motifs from the PG4 website (http://www.quadruplex.org) (24), in which PG4 motifs were predicted using Quadparser. We obtained the other four DNA secondary structure motifs (cruciform, slipped motif, Z-DNA, triplex) from the non-B DB database (25). We obtained the fragile-site information according to Durkin and Glover (26); fragile sites were at a megabase resolution. We obtained human miRNA genes from miRBase (release 18), which contains 1426 miRNAs (27). We obtained replication

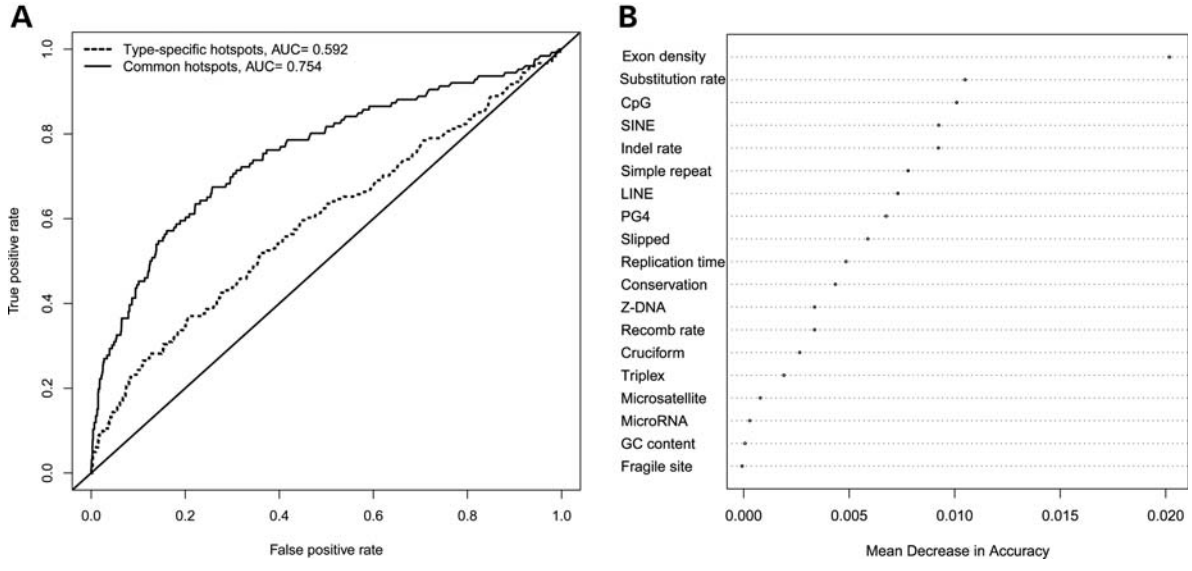

Figure 4. Predictability of cancer-type-specific breakpoint hotspots versus common hotspots from genomic features. (**A**) ROC–AUC curves of two random forest classifiers and (**B**) the mean decrease in accuracy values of genomic features in the classifier distinguishing common hotspots from non-hotspot breakpoint blocks.

time data (in HeLa cells) from Chen *et al*. (15). For all the features under survey, we calculated their counts or coverage for 1 Mb genomic blocks using in-house Perl scripts or Galaxy tools (28).

Correlation and multivariate regression analysis

To quantify the relationship between the breakpoint frequency and a genomic feature, which is not necessarily a linear relationship, we used Spearman's rank correlation coefficient, a non-parametric measure for a monotonic function. To study the effect of the AffySNP6 probe distribution on our results, we obtained the probe annotation file (snpArrayAffy6/snpArrayAffy6SV) from the UCSC Genome Browser, and analyzed the probe density of 1 Mb blocks. We repeated the correlation analyses between the breakpoint frequency and genomic features after removing the genomic blocks with extremely high or low probe density (mean $\pm$ 2 standard deviation) and observed very similar observations.

In the multivariate regression analysis, breakpoint frequency was Box–Cox transformed: $f(x) = x/(m + x)$, where m is the median value of x. With this transformation, $f(x)$ was approximately normally distributed. To effectively select the informative features, we divided the data set into a training data set (70%) and a test data set (30%). A step-wise forward regression was performed, in which the transformed breakpoint frequency was the response variable, and the 19 genomic features were potential explanatory variables. All statistical analyses were performed with the R package, version 2.13 (29).

Identification of breakpoint hotspots

We identified breakpoint hotspots according to the method of De and Michor (7). Briefly, we divided the cancer genomes into 1 Mb non-overlapping blocks and counted the number of SCNA breakpoints in each block. Then, we randomized the position of the breakpoints 10 000 times for each chromosome and generated a distribution of breakpoint densities for the first 1 Mb block. The genomic blocks that had a higher breakpoint frequency than that expected from the top 5% obtained from the simulation across the whole genome were identified as breakpoint hotspots. Lists of genes involved in cancer were downloaded from the census of human cancer genes from the Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP/Census/) (13). The UCSC liftOver program was used to convert human hg19 assembly coordinates for some data to the hg18 assembly coordinates. The coordinates of telomeres and centromeres were obtained from the Gap track of the UCSC Genome Browser.

Random forest classifiers

We used the randomForest package (30) in R and chose parameter values in accordance with Breiman's methodology (14). Although the default number of trees is 500, we chose to build 1000 trees (ntree = 1000) to obtain more robust results. Each tree was grown to its full depth (nodesize = 1) and was not pruned. At each node of each tree, a different random subset of the features was selected, and the Gini criterion was used to determine the feature in this subset that produced the best split of the data. The size of this subset (mtry) was the square root of the number of possible features (mtry = sqrt(p), where p is the number of features). Otherwise, the parameter values were left at their default values. To assess the predictive power, we performed 10-fold cross-validation: in each round, 90% of the data were used as the training data, and the remaining 10% was used as the test data.

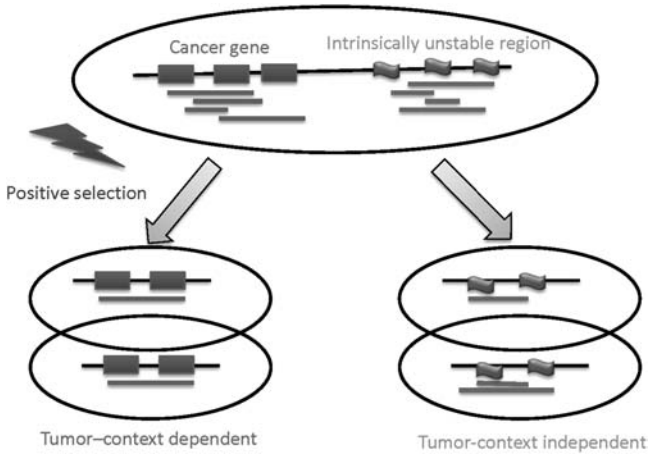

Figure 5. Illustration of two evolutionary mechanisms for high-frequency SCNAs. The overlapping SCNAs with cancer genes may be under positive selection, while common SCNAs may be due to their intrinsically genomic instability.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank LeeAnn Chastain for editorial assistance, and Siyuan Zheng for assistance in preparing the SCNA data. We also thank The Cancer Genome Atlas (TCGA) Research Network for providing SCNA data.

Conflict of Interest statement. None declared.

FUNDING

This study was supported by the National Institutes of Health (CA143883 to H.L., CA016672 to L.Z. and H.L.); UTMDACC – G.S. Hogan Gastrointestinal Research Fund (H.L.) and the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine to H.L.; and a grant from the National Natural Science Foundation of China (No. 31101339 to Y.L.).

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.S., Dobson, J., Urashima, M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
3. Albertson, D.G. (2006) Gene amplification in cancer. *Trends Genet.*, **22**, 447–455.
4. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
5. Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, **11**, 685–696.
6. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K. *et al.* (2010) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
7. De, S. and Michor, F. (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat. Struct. Mol. Biol.*, **18**, 950–955.
8. Fudenberg, G., Getz, G., Meyerson, M. and Mirny, L.A. (2011) High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotech.*, **29**, 1109–1113.
9. De, S. and Michor, F. (2011) DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.*, **29**, 1103–1108.
10. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
11. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
12. Kuo, K.T., Mao, T.L., Chen, X., Feng, Y., Nakayama, K., Wang, Y., Glas, R., Ma, M.J., Kurman, R.J., Shih Ie, M. *et al.* (2010) DNA copy numbers profiles in affinity-purified ovarian clear cell carcinoma. *Clin. Cancer Res.*, **16**, 1997–2008.
13. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
14. Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
15. Chen, C.-L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O. *et al.* (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.*, **20**, 447–457.
16. 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.
17. Nowell, P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–28.
18. Cahill, D.P., Kinzler, K.W., Vogelstein, B. and Lengauer, C. (1999) Genetic instability and darwinian selection in tumours. *Trends Cell Biol.*, **9**, M57–M60.
19. Nei, M., Niimura, Y. and Nozawa, M. (2008) The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat. Rev. Genet.*, **9**, 951–963.
20. Nozawa, M., Kawahara, Y. and Nei, M. (2007) Genomic drift and copy number variation of sensory receptor genes in humans. *Proc. Natl Acad. Sci. USA*, **104**, 20421–20426.
21. Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
22. Ananda, G., Chiaromonte, F. and Makova, K. (2011) A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol.*, **12**, R27.
23. Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310**, 321–324.
24. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
25. Cer, R.Z., Bruce, K.H., Mudunuri, U.S., Yi, M., Volfovsky, N., Luke, B.T., Bacolla, A., Collins, J.R. and Stephens, R.M. (2011) Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.*, **39**, D383–D391.
26. Durkin, S.G. and Glover, T.W. (2007) Chromosome fragile sites. *Annu. Rev. Genet.*, **41**, 169–192.
27. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
28. Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
29. Team, R.D.C. (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
30. Liaw, A. and Wiener, M. (2002) Classification and regression by random forest. *R News*, **2**, 18–22.