Human Molecular Genetics, 2013, Vol. 22, No. 13 **2642–2651**
doi:10.1093/hmg/ddt113
Advance Access published on March 7, 2013

Increased genome instability in human DNA segments with self-chains: homology-induced structural variations via replicative mechanisms

Weichen Zhou[1,†], Feng Zhang[1,†,*], Xiaoli Chen[2,3], Yiping Shen[2,4,5], James R. Lupski[6,7,8] and Li Jin[1]

[1]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China, [2]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA, [3]Beijing Municipal Key Laboratory of Child Development and Nutriomics, Capital Institute of Pediatrics, Beijing 100020, China, [4]Department of Laboratory Medicine, Boston Children's Hospital, Boston, MA 02115, USA, [5]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA, [6]Department of Molecular and Human Genetics and [7]Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA [8]Texas Children's Hospital, Houston, TX 77030, USA

Received January 17, 2013; Revised and Accepted March 1, 2013

Environmental factors including ionizing radiation and chemical agents have been known to be able to induce DNA rearrangements and cause genomic structural variations (SVs); however, the roles of intrinsic characteristics of the human genome, such as regional genome architecture, in SV formation and the potential mechanisms underlying genomic instability remain to be further elucidated. Recently, locus-specific observations showed that 'self-chain' (SC), a group of short low-copy repeats (LCRs) in the human genome, can induce autism-associated SV mutations of the *MECP2* and *NRXN1* genes. In this study, we conducted a genome-wide analysis to investigate SCs and their potential roles in genomic SV formation. Utilizing a vast amount of human SV data, we observed a significant biased distribution of human germline SV breakpoints to SC regions. Notably, the breakpoint distribution pattern is different between SV types across deletion, duplication, inversion and insertion. Our observations were coincident with a mechanism of SC-induced DNA replicative errors, whereas SC may sporadically be used as substrates of nonallelic homologous recombination (NAHR). This contention was further supported by our consistent findings in somatic SV mutations of cancer genomes, suggesting a general mechanism of SC-induced genome instability in human germ and somatic cells.

INTRODUCTION

Genomic rearrangements can lead to various structural variations (SVs) in the human genome, including deletions, duplications, insertions, inversions and complex rearrangements (1–3). SVs represent a significant source of the human genome variations. SVs contribute substantially to inter-individual variations (4); they can segregate across generations via germline transmission, whereas other SV mutations occur *de novo* and may result in sporadic diseases and genomic disorders in offspring (5). In addition to germline mutations, somatic SVs in the human genome are also prevalent; such events are thought to play an important role in cancer development (6,7). Perhaps an extreme example of such structural changes of the human genome is the phenomenon of 'chromothripsis' observed in cancers (8) and 'chromoanasynthesis' seen in developmental disorders (9).

SV mutations do not occur at a uniform rate throughout the human genome, but arise more frequently in regions with genomic instability, i.e. mutational hotspots (10,11). Environmental factors that contribute to genome instability could act consistently across whole genomes, whereas the intrinsic

*To whom correspondence should be addressed. Tel: +86 2165643301; Fax: +86 2155664885; Email: zhangfeng@fudan.edu.cn
†The first two authors have contributed equally to this study and they should be regarded as joint First Authors.

characteristics of local genomic segments may be more important in underlying SV hotspots and regional instability in the human genome. Notably, low-copy repeats (LCRs, usually 10–400 kb in length and ≥97% in identity) represent a classic region-specific genome architecture that has been known to induce SV-associated genomic disorders (5) and other human diseases (12,13). Previous studies also showed that segmental duplications (SDs, an extended group of LCRs with a size of ≥1 kb) (14) play an important role in recurrent genome rearrangements and genome evolution (10,11,13). However, SDs account for a minority of the investigated SV hotspots (11); therefore, other intrinsic characteristics of the human genome and molecular mechanisms underlying SV instability remain to be elucidated.

In a recent study on the intragenic deletion of *NRXN1* in autism, an association of human self-chains (SC; human chained self-alignments archived in UCSC Genome Browser) (15,16), a set of short LCRs in the human genome, with genomic deletion instability in the *NRXN1* gene, was identified (17). Interestingly, another recent study identified an SV mutation mediated by inverted SC pairs in the autism-associated *MECP2* gene (18). However, the genome-wide contribution of SCs in human genomic structural mutations was not well understood.

To investigate whether the SC-mediated genomic rearrangements are only locus-specific in sporadic cases or reflect a general mechanism of SV formation, we conducted a genome-wide analysis based on four sets of SV data from both human populations and cancer genomes. Our findings document significantly biased distributions of SV breakpoints to adjacent SC pairs and suggest a general mechanism of homology-induced genome instability via DNA replicative errors and nonallelic homologous recombination (NAHR) (Fig. 1) in both human germ and somatic cells.

RESULTS

Self-chains as a novel group of short low-copy repeats in the human genome

SCs were previously mapped by alignment of the human genome with itself, using *BLASTZ*, a gap scoring system that allows long gaps (15,16). The SCs are short in length; most of the SCs investigated in this study range from 150 bp to 1 kb in size (Supplementary Material, Fig. S1). Furthermore, different from the transposon-derived high-copy repeats such as *Alu* elements (19), SCs only have a limited number of matched alignments in the human genome. The distribution of the genomic regions with adjacent SC pairs is shown in Supplementary Material, Figures S2 and S3. Thus, SCs represent a distinct type of short LCRs in the human genome.

Adjacent SC pairs and their possible involvement in structural variation formation

In previous observations of SC-induced SV mutations at the *MECP2* (18) and *NRXN1* (17) loci, frequent microhomologies were found at breakpoints, suggesting a possible involvement of DNA replicative mechanisms in these SC-induced events (20). Accordingly, adjacent SC pairs are likely to facilitate formation of DNA secondary structures during replication and cause replication fork stalling (Fig. 1) (21), which can be a prelude to SV mutations via subsequent template switching to resume replication (22,23). DNA replication fork U-turns may also occur frequently at inverted repeat structures (24).

In this study, we investigated SC pairs located within 30 kb intervals that correspond to the sizes of mammalian replicons (25); therefore, these adjacent SCs can possibly mediate DNA secondary structures within a replicon or replication factory (26), and render susceptibility to SVs via DNA replication errors. A preliminary ‘clean-up’ of SC alignments was performed before analyses (see Materials and Methods). In total, we elucidated 26 624 SC regions (SCRs) using our definitional criteria (Supplementary Material, Fig. S2); these include 18 434 +SCRs (only having direct SC repeats), 7794 −SCRs (only having inverted SC repeats), and 396 complex SCRs (having both direct and inverted SC repeats). The difference in number between +SCRs and −SCRs is due to the over-representation of direct SC repeats compared with inverted ones, which is potentially shaped during human genome evolution. Such an evolutionary phenomenon has been observed to explain for the biased distribution of direct and inverted *Alu* repeats in the human genome (27).

Biased breakpoint distribution of germline SVs to SCRs

The SCRs were used to investigate the breakpoint distribution of SVs in the human genome (Fig. 2). Given that adjacent paired SCs could induce genomic structural mutations, the SV breakpoint densities were anticipated to increase when approaching SCRs, i.e. more breakpoints in proximity to SCRs than in those genomic intervals far from SCRs (Supplementary Material, Fig. S4).

The first set of SVs analyzed in this study was generated in previous studies using high-resolution comparative genomic hybridization (CGH) microarrays or single nucleotide polymorphism (SNP) genotyping microarrays or next-generation sequencing (NGS) read depth in human populations (see Materials and Methods), including 31 043 SVs (18 526 copy number losses, i.e. deletions; 11 555 gains including duplications and insertions; 962 SVs with gains and losses). These breakpoint data have a level of genome resolution of kilo-base pairs (kb). Intriguingly, we observed that the density of SV breakpoints increased when narrowing the sizes of SCR-flanking regions for investigation (Fig. 3A). A significant difference in breakpoint density between the regions flanking SCRs and those flanking control regions was observed when the flanking regions were 10 kb or less (Fig. 3A). Thus, our preliminary analysis suggested that the paired SCs in this study were associated with SV breakpoints and thus likely to be involved in SV formation in the human genome.

The second dataset (see Materials and Methods) included 38 250 SVs (30 973 deletions, 2177 duplications, 672 inversions, and 4428 insertions) that were identified in human populations using NGS split-read and/or assembly methods in previous studies including the 1000 Genome Project (Pilot 2). When investigating these SV breakpoints at nucleotide resolution, we observed a similar biased breakpoint distribution to SCRs (Fig. 3B). In addition, the size of SCR-flanking

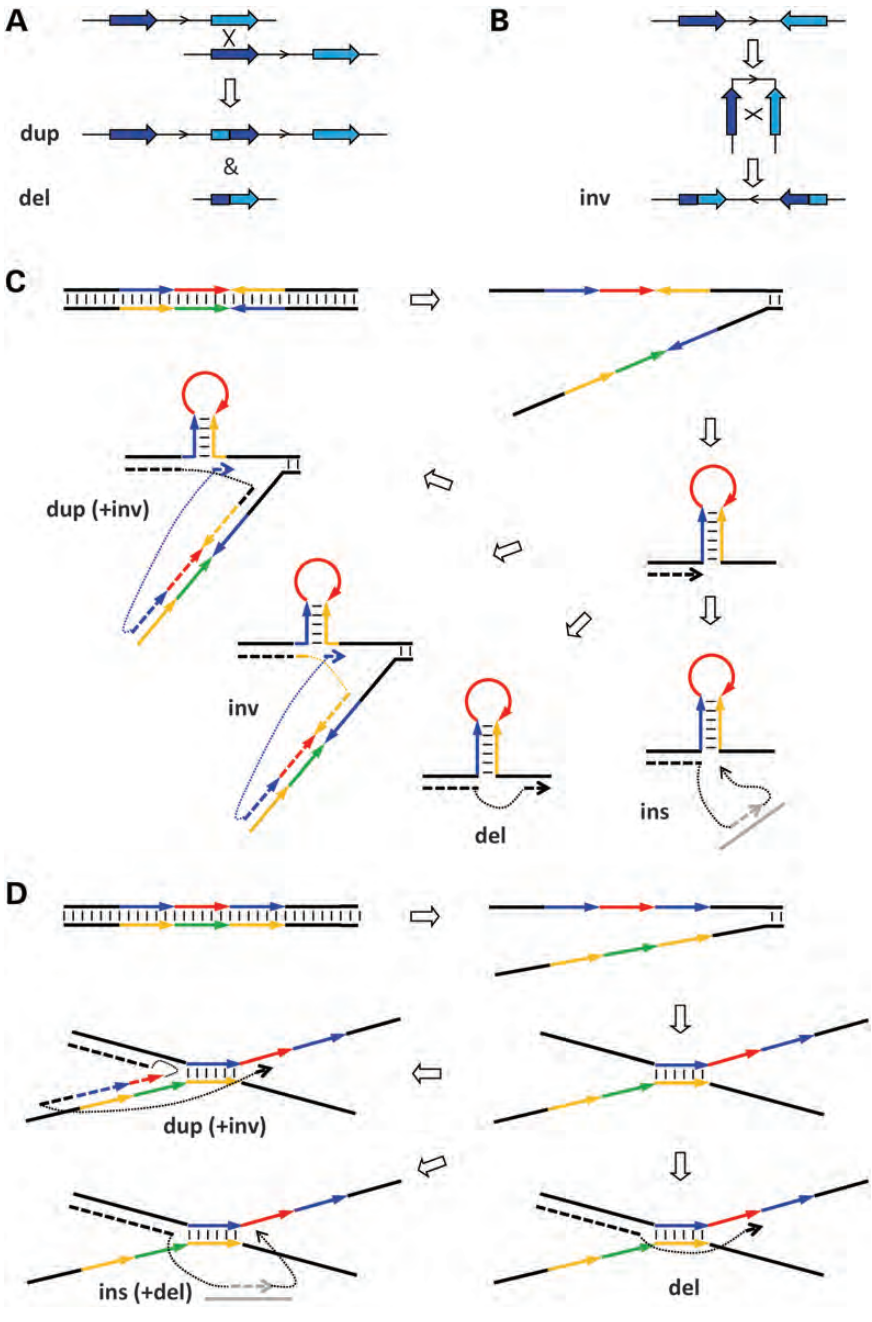

Figure 1. Genomic rearrangements and homology-driven mechanisms of SV formation in the human genome. (**A**) NAHR between repeats (dark and light blue arrowed bars) in direct orientation can cause reciprocal duplications (dup) and deletions (del). The cross depicts a DNA recombination event. (**B**) The recombination between inverted repeats can lead to inversions (inv). The likely involvement of inverted (**C**) and direct repeats (**D**) in DNA replicative mechanisms and SV mutations. The thick lines and arrows depict single DNA strands and short thin lines represent the annealing between Watson–Crick base pairs (blue-yellow or red-green). During DNA replication, adjacent short repeats could lead to secondary structures and consequently cause replication fork stalling. The newly synthesized DNA strands are shown by dashed lines. Based on the replicative mechanisms (20), DNA template switching can occur to resume replication and generate SVs as well. SV types and template switching patterns: deletions, switching forward; duplication, switching to the opposite strand and backward; insertion (ins), switching to a template in another replicon (shown by a grey line) and backward; inversion, switching between leading and lagging strands via homology and/or microhomology. The dotted lines represent the DNA template switching events.

intervals with a significantly increased breakpoint density was narrowed from 10 kb in the first dataset to 5 kb in the second one, which is possibly due to the improvement in analysis resolution of SV breakpoints. Our consistent observations in the above two sets of SV data suggested that SV incidence increases in proximity to SCRs, supporting the contention that adjacent SC pairs can indeed induce genomic instability as evidenced by SVs.

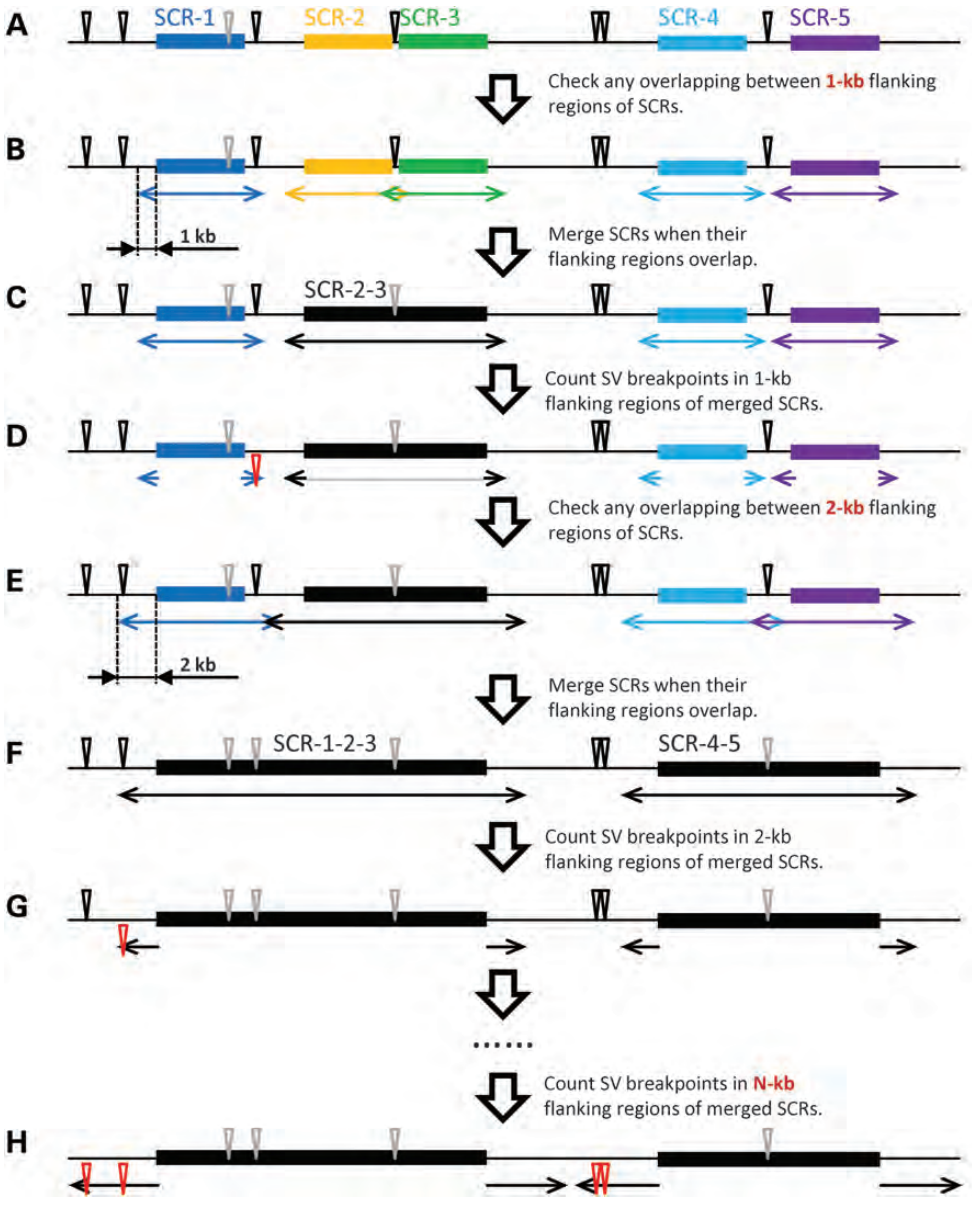

Figure 2. Counting SV breakpoints in the flanking regions of SCRs. The example, in which the starting size of 1 kb and the increment of 1 kb were adopted for SCR-flanking regions, was illustrated. (**A**) Five non-overlapping SCRs (SCR-1 to SCR-5) are shown. Triangles depict SV breakpoints; black depicts breakpoints outside SCRs; grey depicts breakpoints in SCRs. (**B**) Before counting SV breakpoints, we check whether any two of the SCR-flanking regions overlap. Here, we show the example that the 1 kb flanking regions of SCR-2 and SCR-3 overlap each other. (**C**) To avoid counting the breakpoints between SCR-2 and SCR-3 twice, we merge these two SCRs into a new SCR (SCR-2−3). (**D**) The numbers of the SV breakpoints in 1 kb SCR-flanking regions are counted (shown by red triangles). (**E**) Then, the size of SCR-flanking regions for investigation is increased to 2 kb. Check whether any two of 2 kb flanking regions overlap. (**F**) Merge any two SCRs whether their flanking regions overlap each other. (**G**) The numbers of the SV breakpoints in 2 kb SCR-flanking regions are counted. (**H**) Increase the size of SCR-flanking regions for investigation again and repeat the steps E-G till the size of SCR-flanking regions reaches N kb.

Different breakpoint distribution patterns between SV sub-types

The second dataset based on NGS split-read and/or assembly approaches (see Materials and Methods) resolved SVs into various sub-types, including deletions, duplications, inversions and insertions. Therefore, SV breakpoint distributions were further investigated based on SV sub-types.

For deletions, significantly increased breakpoint densities were found in the regions flanking SCRs (Fig. 4A and B). The biased distribution is more significant for +SCRs than for −SCRs, possibly and partially due to the difference in sample size of deletion breakpoints flanking +SCRs and −SCRs, respectively.

For duplications, increased breakpoint densities were consistently found for both +SCRs and −SCRs when the size of their flanking regions was narrowed to 4 to 5 kb or less (Fig. 4C and D).

For inversions, the biased breakpoint distribution is dominantly associated with the −SCRs, whereas the increase in breakpoint density in the regions flanking +SCRs is not significant (Fig. 4E and F). Notably, this finding is consistent

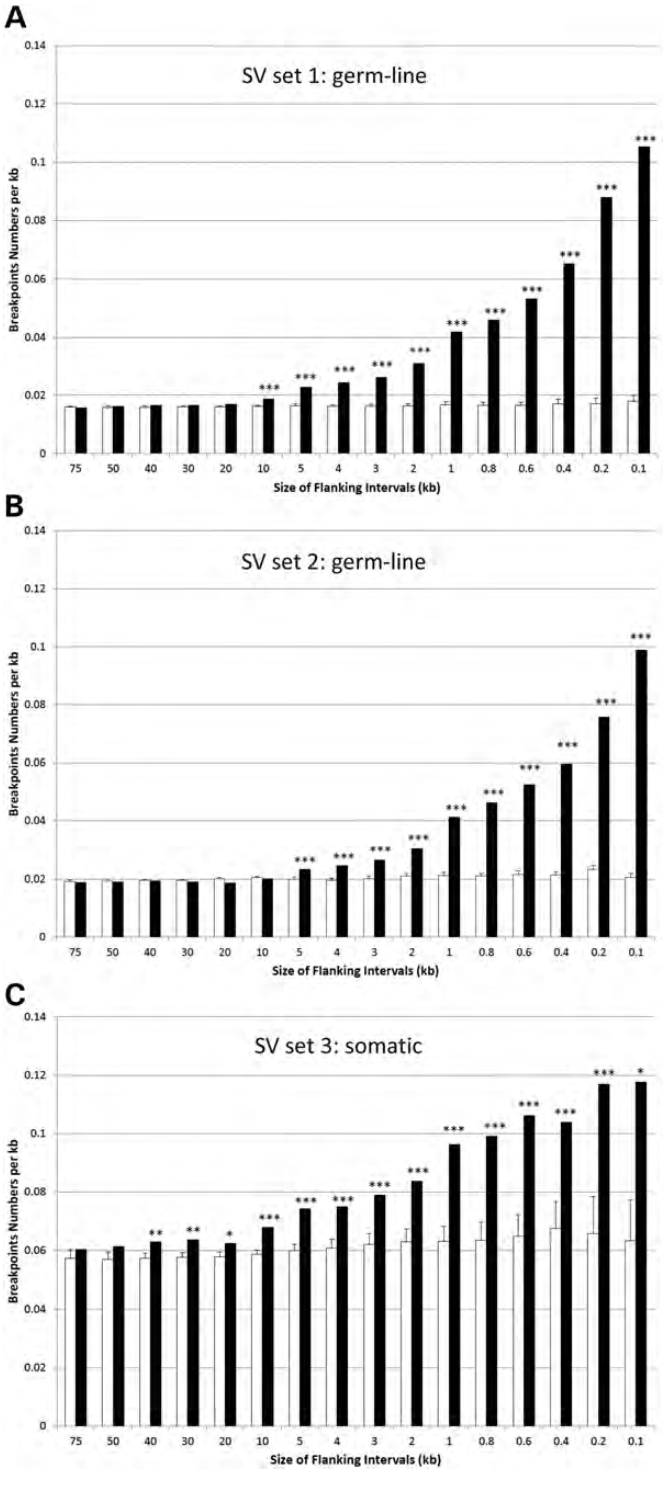

Figure 3. Both germline SVs in human populations and somatic SVs in cancer genomes have a significant biased breakpoint distribution to SCRs. (**A**) Breakpoint distribution of the SVs resolved by microarray and/or NGS read-depth methods in human populations. (**B**) Breakpoint distribution of the SVs resolved by NGS split-read and/or assembly methods in human populations. (**C**) Breakpoint distribution of the SVs resolved by microarray methods in cancer genomes. *X*-axis, size of SCR-flanking regions. From left to right, the narrowing-down of SCR-flanking regions. *Y*-axis, SV breakpoint density (number per kb). Black columns, SCRs; open columns, simulated control regions. The significant differences in breakpoint density between the flanking regions of SCRs and those of control regions are indicated by asterisks (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

with the rearrangement models based on DNA replicative mechanisms (Fig. 1). When inverted SCs mediate secondary structures and cause replication fork stalling, templates can switch to the opposite strand within a replicon. Although microhomology of a few base pairs in length may be able to help DNA template switch back and resume replication (20), both direct and inverted SC homologies that are much longer than microhomologies exist on the two opposite strands in a −SCR, which may facilitate template switching and cause inversions (Fig. 1C). Such structures may potentially also facilitate a fork reversal of the replisome. However, only two direct or two inverted SC repeats exist on each strand in a +SCR, which cannot be used as a long homology to drive template switching between strands (Fig. 1D).

Importantly, and in contrast to our above findings in deletions/duplications/inversions, our observations on insertions showed no obvious bias in breakpoint distribution in the regions flanking either +SCRs or −SCRs (Supplementary Material, Fig. S5). Notably, these results for insertions were anticipated. The insertion breakpoint coordinates actually showed the replicated segments (illustrated by the grey solid lines in Fig. 1C and 1D), which were not the unstable regions inducing replication fork stalling but the targets for DNA template switching. Therefore, the coordinate data of insertion breakpoints may not reflect the origins of genomic instability.

Biased breakpoint distribution of somatic SVs to SCRs

The SVs identified in human populations represent germline mutations, which can result from meiotic events during gametogenesis and/or mitotic cycles of germ stem cells. Since our observations suggested the involvement of DNA replication errors in SC-induced SV mutations, we hypothesized that a biased distribution of SV breakpoints to SCRs might also be applicable to mitotically derived somatic SVs, which are prevalent in cancer genomes (6). Therefore, we investigated the third dataset including 108 882 somatic copy number alterations (44 345 deletion-associated losses and 64 537 duplication/insertion-associated gains) derived from glioblastoma multiforme, lung squamous cell carcinoma and ovarian serous cystadenocarcinoma of the Cancer Genome Atlas (see Materials and Methods).

Intriguingly, a significant biased distribution of somatic SV breakpoints to SCRs was observed (Fig. 3C), which was consistent with our findings in germline SVs in human populations. Notably, after sub-categorizing somatic SVs into deletions (losses) and duplications/insertions (gains) and sub-dividing SCRs into ±SCRs, the SV breakpoints still show biased distributions to SCRs (Fig. 5).

In aggregate, our observations on somatic SVs in cancer genomes are consistent with the finding in germline SVs of human populations, potentially suggesting a general involvement of adjacent SC homologies in inducing genomic instability.

DISCUSSION

Regional genome architecture is important in underlying SV hotspots and genome instability. Observations in model organisms suggest that the non-B DNA structure can induce

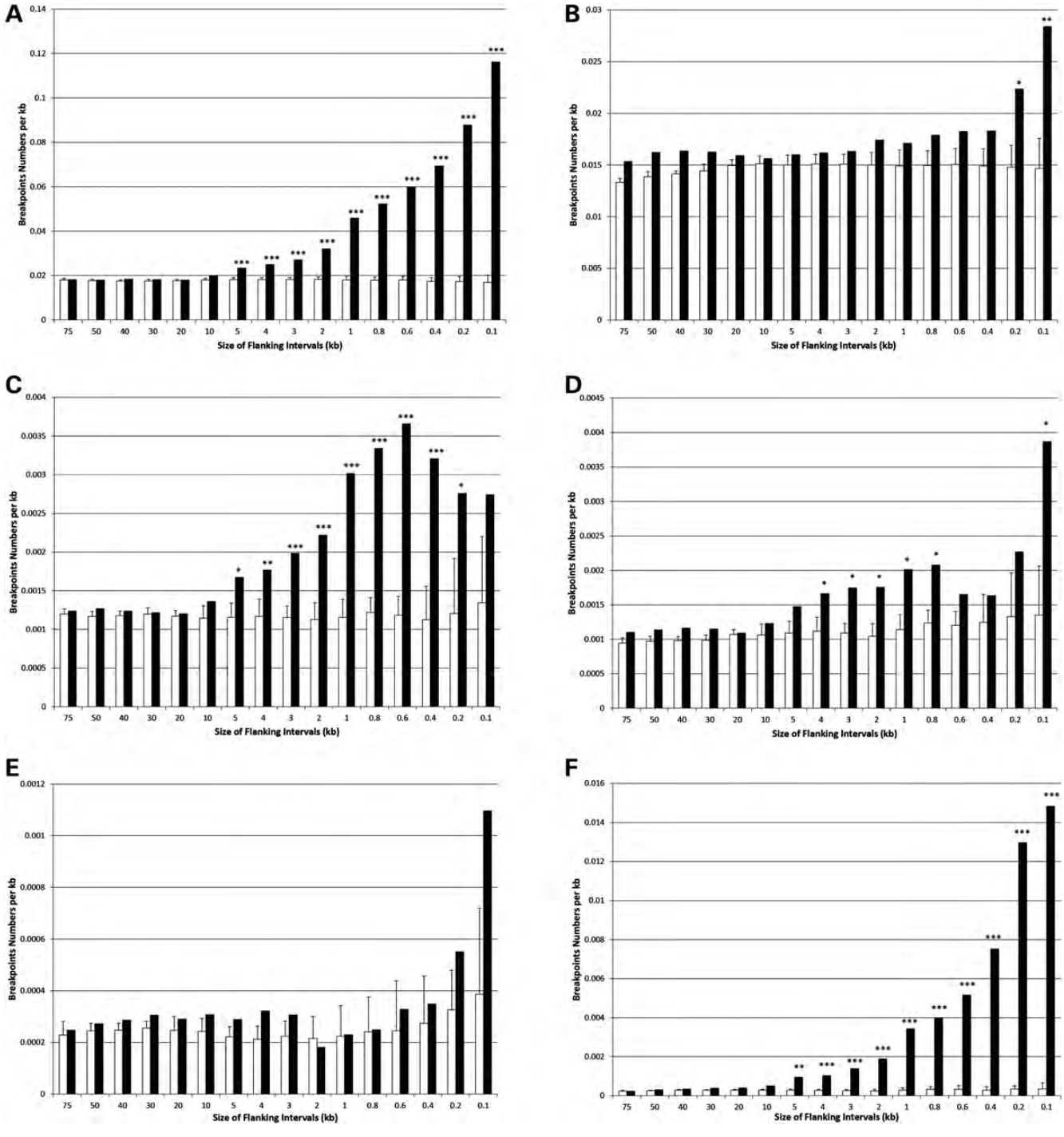

Figure 4. The correlations of breakpoint distributions between SV types and SC orientations. Based on the germline SVs identified by NGS split-read and/or assembly methods, deletion breakpoints have a biased distribution to both +SCRs (**A**) and −SCRs (**B**). Duplication breakpoints also have a biased distribution to both +SCRs (**C**) and −SCRs (**D**). Distribution of inversion breakpoints in flanking regions of +SCRs (**E**) and −SCRs (**F**). The significant differences in breakpoint density between the regions flanking SCRs (black columns) and those flanking control regions (open columns) are indicated by asterisks ($^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$).

genomic instability (28,29). In addition, NAHR between paired long repeats plays an important role in SV hotspots in the human genome (11), potentially by facilitating an ectopic synapsis allowing an ectopic recombination or NAHR to occur (30). In this current study, we investigated a novel group of short LCRs (i.e. SCs) in the human genome; these are distinct from classic LCRs/SDs in repeat length (Supplementary Material, Fig. S1) and distinguishable from transposon-derived short repeats in repeat number. Significant biased distributions of SV breakpoints to the regions with SCs

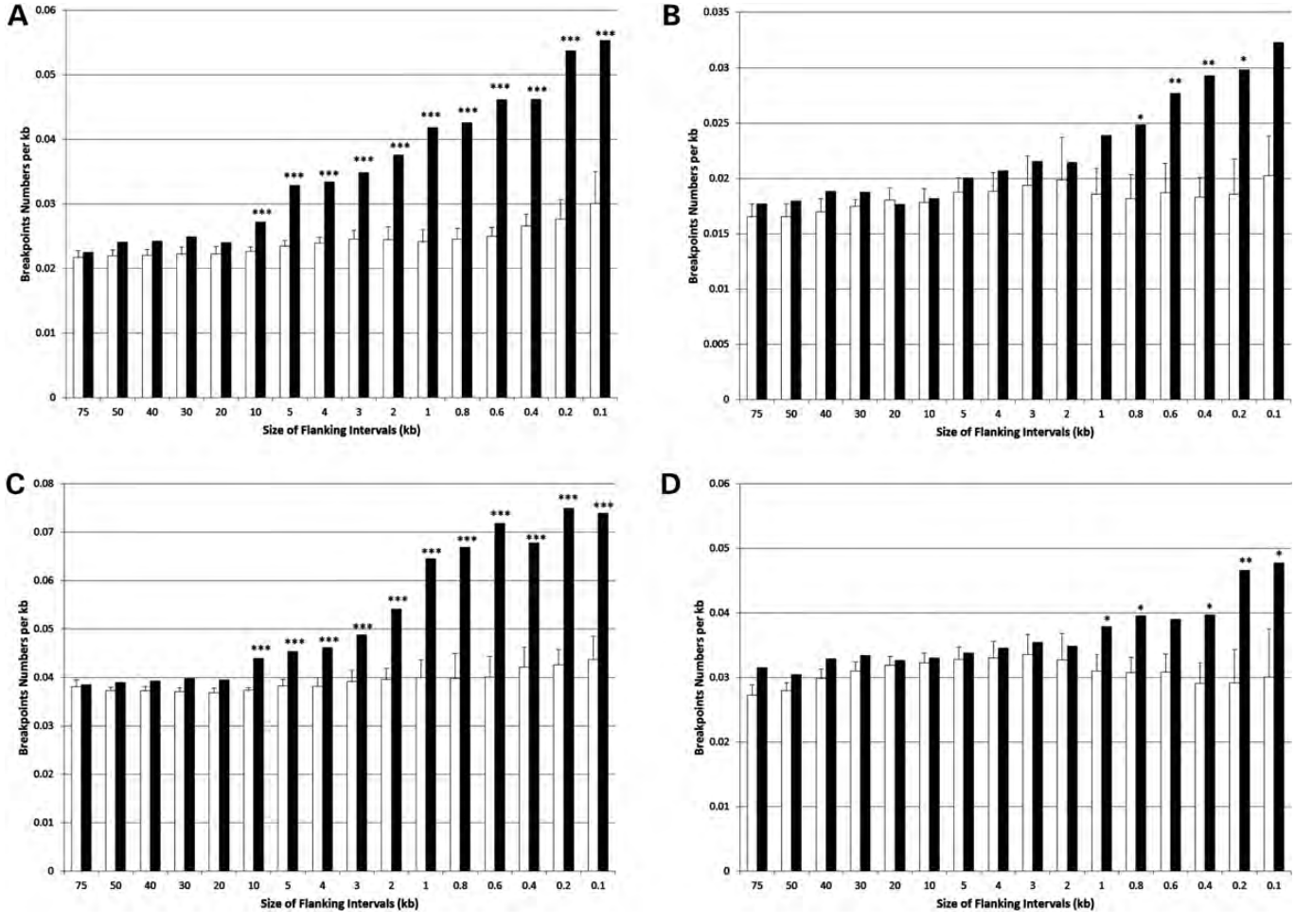

Figure 5. Biased breakpoint distributions of the somatic SVs identified by microarray methods in cancer genomes. The breakpoints of deletions (i.e. copy number losses) in the regions flanking +SCRs (**A**) and −SCRs (**B**). The breakpoints of duplications and insertions (i.e. copy number gains) in the regions flanking +SCRs (**C**) and −SCRs (**D**). The significant differences in breakpoint density between the regions flanking SCRs (black columns) and those flanking control regions (open columns) are indicated by asterisks ($^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$).

were consistently observed based on three aforementioned SV datasets derived from both germline and somatic SV events using various SV investigative methods, suggesting a general mutational mechanism. However, SCs can act as recombination substrates for NAHR; although the repeat length is correlated with the frequency of events. Thus, the longer LCRs may be the ones more frequently involved in NAHR, particularly to those ectopic recombinations occurring during meiosis that are potentially driven by an ectopic synapsis versus those occurring during mitosis. In addition, SCs are also likely to induce DNA replication errors (Fig. 1). Therefore, the contributions of NAHR and replicative mechanisms in SC-induced mutational events require further investigations. Based on the mutational models (Fig. 1), breakpoint mapping may be informative; the NAHR breakpoints must be located within repeats, whereas the breakpoints of DNA replicative mechanisms may be mapped inside or outside repeats.

We noted three issues in the first and third sets of SV data (see Materials and Methods) derived using microarrays and NGS read-depth: (i) both duplications and insertions may be grouped into copy number gains only; (ii) since their investigative methods can only detect alterations in the copy number (31), balanced SVs such as inversions were not included; (iii) since these methods map SVs in fine resolution but cannot resolve SV breakpoints to the nucleotide level, these SV breakpoints may not precisely reflect the breakpoint positions. Therefore, the second and fourth datasets (the sequencing data of germline and somatic SVs, respectively; see Materials and Methods) can reveal more accurate and more informative distributions of SV breakpoints.

The sequenced SV breakpoints available in the second and fourth datasets were further mapped (Supplementary Material, Fig. S6). The SVs with one or both breakpoints within SCs and/or SC-flanking regions are likely to be affected by SCs. Given that the length of flanking regions was 1 kb, the details of SV breakpoint mapping are shown in Table 1. In 38 250 germline SVs of human populations, 5865 (15.3%) and 1257 (3.3%) have at least one breakpoint in SCs and/or flanking regions of direct and inverted SC pairs, respectively. Similarly, 162 (6.2%) and 72 (2.7%) out of 2629 somatic SVs from cancer genomes were related to direct and inverted SC pairs, respectively.

Table 1. The SVs with sequenced breakpoints located in SCs or its 1 kb flanking regions

SV Type	Number	SCR-associated SVs Number/% in all SVs		SC1−SC2/%	Remaining/%
Germline SVs with sequenced breakpoints					
All	38 250	−SCR	1257/3.3	76/6.0	1181/94.0
		+SCR	5865/15.3	1952/33.3	3913/66.7
Deletion	30 973	−SCR	983/3.2	39/4.0	944/96.0
		+SCR	5436/17.6	1933/35.6	3503/64.4
Duplication	2177	−SCR	58/2.7	1/1.7	57/98.3
		+SCR	197/9.1	17/8.6	180/91.4
Inversion	672	−SCR	155/23.1	36/23.2	119/76.8
		+SCR	83/12.4	1/1.2	82/98.8
Insertion	4428	−SCR	61/1.4	0/0.0	61/100.0
		+SCR	149/3.4	1/0.7	148/99.3
Somatic SVs with sequenced breakpoints					
All	2629	−SCR	72/2.7	0/0.0	72/100.0
		+SCR	162/6.2	1/0.6	161/99.4
Deletion	581	−SCR	20/3.4	0/0.0	20/100.0
		+SCR	33/5.7	1/3.0	32/97.0
Duplication	794	−SCR	20/2.5	0/0.0	20/100.0
		+SCR	45/5.7	0/0.0	45/100.0
Inversion	421	−SCR	13/3.1	0/0.0	13/100.0
		+SCR	36/8.6	0/0.0	36/100.0
Insertion	833	−SCR	19/2.3	0/0.0	19/100.0
		+SCR	48/5.8	0/0.0	48/100.0

We further characterized these SC-related SVs based on their patterns of breakpoint mapping and found that 33.3% of + SC-related germline SVs and 6.0% of −SC-related ones showed the SC1−SC2 pattern (Supplementary Material, Fig. S6), consistent with the NAHR mechanism (Table 1). Notably, the rearrangement types of these SC1−SC2 SVs were associated with SC repeat orientations. Based on our observations on the germline SVs, more SC1−SC2 deletions and duplications were associated with direct SC repeats rather than inverted ones; 35.6 versus 4.0% ($P = 1.3 \times 10^{-113}$, Fisher's exact test) for deletions, and 8.6 versus 1.7% ($P = 0.08$) for duplications (Table 1). In contrast, the inversions have a reverse distribution pattern; 1.2% of SC1−SC2 inversions were associated with direct SC repeats, while 23.2% associated with inverted repeats ($P = 9.1 \times 10^{-7}$). No obvious distribution bias was observed for SC1−SC2 insertions. These observations were similar to the findings in NAHR-mediated SVs that the NAHR events between direct SDs can cause deletions and duplications, and instead NAHR between inverted SDs can only lead to inversions (12). Therefore, it is possible for SC repeats to cause SV mutations via the NAHR mechanism.

NAHR is more frequent in germline than in somatic events (32). Consistently, only one (0.6%) of the SC-related somatic SVs was identified to have an SC1−SC2 pattern (Table 1). However, we note that a minimum of 200−300 bp of uninterrupted homology, also known as a minimal efficient processing segment, seems to be required for NAHR to occur (12); but many of the paired SCs investigated in this study do not share such a long stretch of identical sequence for NAHR. Therefore, other mechanism(s) may also be involved in SC-induced SV mutations.

We showed that our hypothesized SC-induced replicative mechanism can also generate SC1−SC2 SVs via replication template switching driven by paired SCs sharing hundreds of base pairs of homology instead of by classic microhomologies of only a few base pairs in length (Fig. 1) (20). Notably, in addition to the SC1−SC2 SVs, the majority of SC-related SVs have breakpoints outside SCs (Table 1), which is not consistent with NAHR, but further supports the replicative mechanisms (Fig. 1C and D).

In the aggregate, our observations on the germline SVs in human populations and the somatic SVs in cancer genomes reveal that adjacent SC pairs are associated with SV mutations and instability in the human genome, potentially by DNA replication errors that occur via SC-mediated secondary structures and DNA replication template switching or fork reversal. Alternatively, such repeats may facilitate NAHR between SCs in some instances. These SCs represent a type of genome architecture distinct from known long LCRs/SDs and transposon-derived high-copy repeats in the human genome, and may be an important architectural feature of the human genome that underlies regional susceptibility to genomic instability in human germ and somatic cells.

MATERIALS AND METHODS

Adjacent self-chain pairs in the human genome

The SCs (15,16) and other data, including centromeres, SDs, and sequencing gaps in the human genome, were obtained from the UCSC Genome Browser website (http://genome.ucsc.edu/; genome assembly hg18). The SCs include plus (+, in direct orientation) and minus (−, in inverted orientation) pairs. The segment of any paired SCs in the same chromosome and their spacing gap was defined as an SC segment (SCS) (see Supplementary Material, Fig. S2). The self-aligned inverted SCs were regarded as an SCS. The paired SCs located in different chromosomes and those in the same chromosome but having long spacing intervals (SCS size >30 kb in this study) were filtered out. In addition, any SCS overlapping with the human genome gaps (33), centromeres, or SDs (14) was further filtered out. To accurately count SV breakpoints in the regions flanking SCs, overlapping SCSs were further merged into single SCRs (see Fig. 2 and Supplementary Material, Fig. S2), whereas any non-overlapping SCS was treated as an SCR.

Whole-genome random control regions

We randomly generated control regions of 30 kb in length in the human genome. The number of control regions corresponds to that of SCRs during our analyses. In addition, any control region overlapping with the human genome gaps (33), centromeres or SDs (14) was filtered out, the same as the data pre-processing that was done with SCRs. The strategies of merging control regions and counting breakpoints were also the same as those for SCRs (Fig. 2 and Supplementary Material, Fig. S2). Control simulations were repeated 10 times. Eventually, we performed one-way analysis of variance to test the significance of breakpoint density difference in the flanking regions between SCRs and control regions.

Germline structural variations and breakpoints in human populations

Many germline SVs previously identified in human populations have been archived in Database of Genomic Variants (http://projects.tcag.ca/variation/). A portion of these data had fine breakpoint resolution. These data can be categorized into two sets based on their original investigative methods. The first dataset (SV set 1) includes the SVs that were detected by high-resolution CGH or SNP microarrays or NGS read-depth methods (34–39).

The second dataset (SV set 2) includes the SVs that were detected by NGS split-read and/or assembly methods (34,38,40–44). The SV sequencing data of two additional studies, including Kidd et al. (45) and the 1000 Genome Project (Pilot 2; 59 unrelated individuals from YRI, 60 from CEU, 30 from CHB, and 30 from JPT) (4), were included in the second dataset.

Somatic structural variations and breakpoints in cancer genomes

The third dataset (SV set 3) was derived from the somatic SVs that were identified in patients with tumors by using high-resolution microarrays. The interpreted SV data of the Cancer Genome Atlas (TCGA, Level 3) were obtained. To identify somatic SVs, only the tumor data with matched normal samples were analyzed by excluding germline SVs that should be shared by tumors and matched normal tissues. The SV calls showing mean signal values of >0.3 were taken as copy number gains (including duplications and insertions), whereas ≤ 0.5 for copy number losses (i.e. deletions). Based on the genomic resolutions of various investigative microarrays, the split SV calls were merged if the gap between gains or between losses is <10 kb (CGH-1M), 20 kb (CGH-415K) or 30 kb (CGH-244K). Then, the SVs shared by tumors and matched normal tissues (50% overlapping), which potentially represent the germline SVs, were filtered out. The following SVs were also excluded in the further analyses: the large SVs of >5 Mb that could be caused by chromosomal abnormality, and the small SVs (<10 kb for CGH-1M, <20 kb for CGH-415K, and <30 kb for CGH-244K; or being called by less than five investigative probes for all microarray formats) that were beyond microarray resolutions and might be false-positive calls.

Besides the above microarray-derived SVs in tumors, many somatic SVs have also been recently identified by cancer genome sequencing. Therefore, the fourth dataset (SV set 4) was obtained from 24 breast cancers (46), 13 metastatic pancreatic cancers (47) and seven prostate cancers (48).

SV breakpoint counting and density calculation

As a surrogate measure for genomic instability, we investigated the densities of SV breakpoints. To avoid counting a specific SV breakpoint twice or more times by different SCRs and overestimating SV breakpoint density, any potential overlapping between the flanking regions of different SCRs were examined first (Fig. 2). Given the starting size of SCR-flanking regions for investigation as S_0, the S_0-kb regions flanking all the SCRs were examined for overlapping. If any two S_0-kb SCR-flanking regions overlap, their corresponding SCRs were merged into a new SCR. Then, numbers of the SV breakpoints (N_{BR}) located in all the S_0-kb SCR-flanking regions were counted, and breakpoint densities (per kb) were $N_{BR}/(N_{SCR} \times S_0 \times 2)$, while N_{SCR} is the number of non-overlapping SCRs. In the next round of counting, the size of investigated SCR-flanking regions was increased from S_0 to S_1 (i.e. S_n to S_{n+1}) kb. The procedure of SCR merging, breakpoint counting and density calculation were repeated (Fig. 2). The following sizes of SCR-flanking regions were applied in this study: 0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 4, 5, 10, 20, 30, 40, 50 and 75 kb.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank C.M.B. Carvalho, W. Fu, P. Liu, P. Stankiewicz, K. Xu and Y. Wang for their critical reviews.

Conflict of Interest statement. J.R.L. is a consultant for Athena Diagnostics, has stock ownership in 23 and Me and Ion Torrent Systems and is a coinventor on multiple United States and European patents for DNA diagnostics.

FUNDING

This work was supported by National Basic Research Program of China (2012CB944600 and 2011CBA00401), National S&T Major Special Project (2011ZX09102-010-01), National Natural Science Foundation of China (81222014, 31171210 and 31000552), Shanghai Pujiang Program (10PJ1400300), Shu Guang Project (12SG08), and National Institute of Neurological Disorders and Stroke, NIH (R01NS058529).

REFERENCES

1. Stankiewicz, P. and Lupski, J.R. (2010) Structural variation in the human genome and its role in disease. *Annu. Rev. Med.*, **61**, 437–455.
2. Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.
3. Liu, P., Carvalho, C.M., Hastings, P. and Lupski, J.R. (2012) Mechanisms for recurrent and complex human genomic rearrangements. *Curr. Opin. Genet. Dev.*, **22**, 211–220.
4. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
5. Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.*, **14**, 417–422.
6. Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
7. Tanaka, H. and Yao, M.C. (2009) Palindromic gene amplification—an evolutionarily conserved role for DNA inverted repeats in the genome. *Nat. Rev. Cancer*, **9**, 216–224.
8. Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.
9. Liu, P., Erez, A., Nagamani, S.C., Dhar, S.U., Kolodziejska, K.E., Dharmadhikari, A.V., Cooper, M.L., Wiszniewska, J., Zhang, F., Withers,

M.A. *et al.* (2011) Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, **146**, 889–903.
10. Itsara, A., Cooper, G.M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R.M., Myers, R.M., Ridker, P.M., Chasman, D.I. *et al.* (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.*, **84**, 148–161.
11. Fu, W., Zhang, F., Wang, Y., Gu, X. and Jin, L. (2010) Identification of copy number variation hotspots in human populations. *Am. J. Hum. Genet.*, **87**, 494–504.
12. Stankiewicz, P. and Lupski, J.R. (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet.*, **18**, 74–82.
13. Dittwald, P., Gambin, T., Gonzaga-Jauregui, C., Carvalho, C.M., Lupski, J.R., Stankiewicz, P. and Gambin, A. (2013) Inverted low-copy repeats and genome instability—a genome-wide analysis. *Hum. Mutat.*, **34**, 210–220.
14. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W. and Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
15. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
16. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
17. Chen, X., Shen, Y., Zhang, F., Chiang, C., Pillalamarri, V., Blumenthal, I., Talkowski, M., Wu, B.L. and Gusella, J.F. (2013) Molecular analysis of a deletion hotspot in the *NRXN1* region reveals the involvement of short inverted repeats in deletion CNVs. *Am. J. Hum. Genet.*, **92**, 375–386.
18. Carvalho, C.M., Ramocki, M.B., Pehlivan, D., Franco, L.M., Gonzaga-Jauregui, C., Fang, P., McCall, A., Pivnick, E.K., Hines-Dowell, S., Seaver, L.H. *et al.* (2011) Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.*, **43**, 1074–1081.
19. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
20. 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.
21. Voineagu, I., Narayanan, V., Lobachev, K.S. and Mirkin, S.M. (2008) Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc. Natl Acad. Sci. USA*, **105**, 9936–9941.
22. Lee, J.A., Carvalho, C.M. and Lupski, J.R. (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, **131**, 1235–1247.
23. Hastings, P.J., Ira, G. and Lupski, J.R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.*, **5**, e1000327.
24. Mizuno, K., Miyabe, I., Schalbetter, S.A., Carr, A.M. and Murray, J.M. (2013) Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature*, **493**, 246–249.
25. Berezney, R., Dubey, D.D. and Huberman, J.A. (2000) Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma*, **108**, 471–484.
26. Kitamura, E., Blow, J.J. and Tanaka, T.U. (2006) Live-cell imaging reveals replication of individual replicons in eukaryotic replication factories. *Cell*, **125**, 1297–1308.
27. Stenger, J.E., Lobachev, K.S., Gordenin, D., Darden, T.A., Jurka, J. and Resnick, M.A. (2001) Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. *Genome Res.*, **11**, 12–27.
28. Bacolla, A., Jaworski, A., Larson, J.E., Jakupciak, J.P., Chuzhanova, N., Abeysinghe, S.S., O'Connell, C.D., Cooper, D.N. and Wells, R.D. (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl Acad. Sci. USA*, **101**, 14162–14167.
29. 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.
30. Liu, P., Lacaria, M., Zhang, F., Withers, M., Hastings, P.J. and Lupski, J.R. (2011) Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *Am. J. Hum. Genet.*, **89**, 580–588.
31. Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.
32. Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S. and Hurles, M.E. (2008) Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. *Nat. Genet.*, **40**, 90–95.
33. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
34. Kim, J.I., Ju, Y.S., Park, H., Kim, S., Lee, S., Yi, J.H., Mudge, J., Miller, N.A., Hong, D., Bell, C.J. *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011–1015.
35. Matsuzaki, H., Wang, P.H., Hu, J., Rava, R. and Fu, G.K. (2009) High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. *Genome Biol.*, **10**, R125.
36. 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.
37. Ju, Y.S., Hong, D., Kim, S., Park, S.S., Lee, S., Park, H., Kim, J.I. and Seo, J.S. (2010) Reference-unbiased copy number variant analysis using CGH microarrays. *Nucleic Acids Res.*, **38**, e190.
38. Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.*, **11**, R52.
39. Park, H., Kim, J.I., Ju, Y.S., Gokcumen, O., Mills, R.E., Kim, S., Lee, S., Suh, D., Hong, D., Kang, H.P. *et al.* (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.*, **42**, 400–405.
40. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
41. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L. and Bignell, H.R. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
42. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
43. Ahn, S.M., Kim, T.H., Lee, S., Kim, D., Ghang, H., Kim, D.S., Kim, B.C., Kim, S.Y., Kim, W.Y., Kim, C. *et al.* (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.*, **19**, 1622–1629.
44. McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.
45. Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K. and Eichler, E.E. (2010) A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, **143**, 837–847.
46. Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.
47. Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.L. *et al.* (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, **467**, 1109–1113.
48. Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214–220.