

Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification

Hisashi Tanaka¹, Donald A Bergstrom², Meng-Chao Yao^{1,3} & Stephen J Tapscott⁴

Breakage-fusion-bridge cycles contribute to chromosome instability and generate large DNA palindromes that facilitate gene amplification in human cancers. The prevalence of large DNA palindromes in cancer is not known. Here, by using a new microarray-based approach called genome-wide analysis of palindrome formation, we show that palindromes occur frequently and are widespread in human cancers. Individual tumors seem to have a nonrandom distribution of palindromes in their genomes, and a subset of palindromic loci is associated with gene amplification. This indicates that the location of palindromes in the cancer genome can serve as a structural platform that supports subsequent gene amplification. Genome-wide analysis of palindrome formation is a new approach to identify structural chromosome aberrations associated with cancer.

Amplification of large genomic regions is common in human cancers and contributes to tumor progression^{1–3}. The incidence and distribution of amplified loci has been well-documented in many tumors, but little is known about the mechanisms that determine whether a particular locus is susceptible to amplification. The initial formation of a large inverted repeat, or palindrome, is a rate-limiting step in gene amplification, and the molecular mechanisms that mediate palindrome formation are conserved among eukaryotes^{4–6}. Because breakage-fusion-bridge (BFB) cycles cause gene amplification in model systems^{7–10} and palindrome formation is a structural component of BFB cycles, palindrome formation might mark regions susceptible to subsequent gene amplification.

Using a new microarray-based approach for genome-wide analysis of palindrome formation (GAPF), we found that palindromes occur frequently in human cancers and seem to cluster at specific loci in the genome. Most palindrome-containing regions are not associated with gene amplification, but they are more likely to be amplified than are random loci. These data indicate that the formation of palindromes

broadly alters the cancer genome and provides a structural platform for subsequent gene amplification.

RESULTS GAPF

We developed a new method to obtain a genome-wide assessment of palindrome formation based on the efficient intrastrand base pairing in large palindromic sequences^{11,12}. Palindromic sequences can rapidly anneal intramolecularly to form ‘snap-back’ DNA under conditions that do not favor intermolecular annealing. We demonstrated this property by heat denaturation and rapid cooling of genomic DNA from D79IR-8 Sce2 cells containing two or three copies of a large DNA palindrome of the transgene *DHFR* (Fig. 1a)⁴. Southern-blot analysis showed that, in the presence of NaCl, renaturation occurred through intrastrand base-pairing, because the restriction fragments migrated at one-half the predicted size: the 10-kb *KpnI* fragment became 5 kb and the 22-kb *XbaI* fragment became 11 kb.

To determine whether this snap-back property could be used to enrich for large palindromic sequences in total genomic DNA, we digested genomic DNA from D79IR-8 Sce2 cells with *SallI*, denatured it, then rapidly renatured it with 100 mM NaCl and then digested it with the single-strand-specific nuclease S1 (Fig. 1b). Snap-back DNA formed from palindromes is double-stranded and should be resistant to S1, whereas the remainder of the genomic DNA is single stranded and should be sensitive to S1. We digested S1-resistant double-stranded DNA with *MspI* or *TaqI*, amplified it by ligation-mediated PCR (LM-PCR) using linker-specific primers and analyzed it by Southern blotting with either a probe in the inverted repeat (probe 1) or a probe in an adjacent nonpalindromic fragment (probe 2). We detected a signal only with probe 1 in snap-back DNA, indicating that this method enriches palindromic sequences.

We carried out a dilution experiment to test the ability of this technique to identify palindromes present in only a subpopulation of cells, as might occur in a tumor with a heterogeneous population of

¹Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. ²Department of Laboratory Medicine, University of Washington, Seattle, Washington, USA. ³Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan. ⁴Division of Human Biology, Fred Hutchinson Cancer Research Center, Mailstop C3-168, 1100 Fairview Avenue North, Seattle, Washington 98109-1024, USA. Correspondence should be addressed to S.J.T. (stapscot@fhcrc.org) or M.-C.Y. (mcyao@imb.sinica.edu.tw).

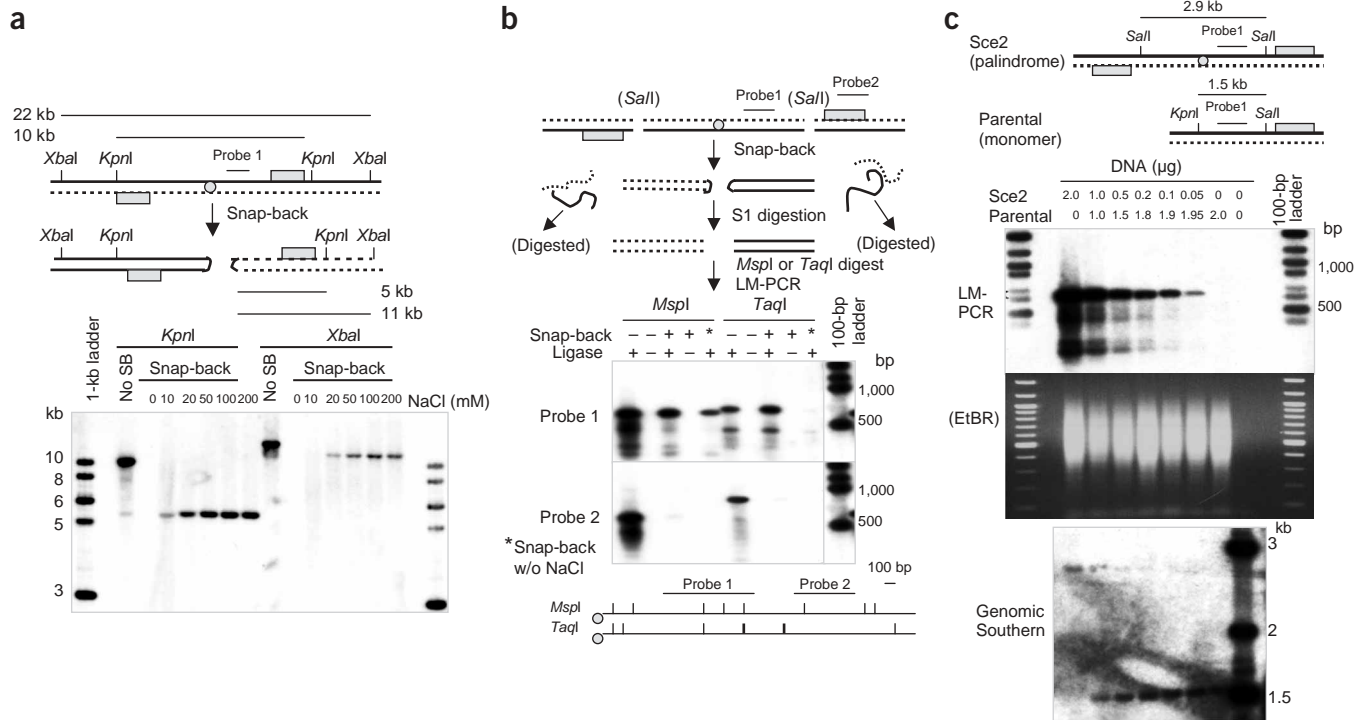


Figure 1 Snap-back DNA and S1 digestion enriches for palindromic DNA. **(a)** NaCl-dependent formation of snap-back DNA. *KpnI* or *XbaI* digestion of DNA and Southern blotting showed efficient intrastrand hybridization of the duplicated region in the presence of NaCl. Solid lines and dotted lines represent complementary strands of single-stranded DNA. Probe 1 (used for hybridization) is indicated. The gray circle represents the center of the palindrome, and the gray rectangle represents the *DHFR* coding sequence. no SB, non-snap-back DNA. **(b)** The same genomic DNA as in **a** was digested with *Sall*, processed for snap-back DNA purification using *MspI* and *TaqI* and analyzed by Southern blotting with a probe for a fragment that contained an inverted repeat (probe 1) or a probe to an adjacent region that did not contain an inverted repeat (probe 2). The experimental procedure is schematically shown on the top, and restriction sites, along with locations of probe 1 and probe 2, are shown on the bottom. **(c)** Detection of a somatic palindrome after dilution with nonpalindromic DNA. *Sall*-digested genomic DNA from D79-8-Sce2 cells and parental D79-8 cells⁴ was mixed in a variety of ratios. Two micrograms of mixed DNA were subjected to LM-PCR amplification of snap-back DNA for Southern-blot analysis (upper panel), and another 2 μg were digested with *KpnI* and analyzed by genomic Southern blotting (lower panel) with probe 1. the *KpnI* site is present only in the *DHFR* transgene in parental cells and not in the *DHFR* palindrome in D79-8-Sce2 cells⁴.

genetically altered cells. We serially diluted genomic DNA from D79IR-8 Sce2 cells with DNA from the parental D79IR-8 cells⁴ that contain a single nonpalindromic copy of the transgene. We analyzed the DNA mixes directly or subjected them to formation of snap-back DNA, *MspI* digestion, amplification by LM-PCR and Southern-blot analysis (Fig. 1c). Using probe 1, we detected signal from the palindrome even in a 1:40 dilution, indicating that this technique can detect a palindrome that is present in only a small population of cells.

Using this technique, distribution of somatic palindromes throughout the genome can be assessed using DNA array hybridization (GAPF). Initially, we used genomic DNA from primary cultures of human fibroblasts derived from skin biopsies (HDF1 and HDF3) or from a foreskin sample (HFF2) from different individuals. We assumed that DNA palindromes are formed as a consequence of genetic instability and that normal fibroblasts would not have many differences between them. We digested snap-back DNA from each of the fibroblasts with S1 nuclease and restriction enzymes (*MspI*, *TaqI* or *MseI*), ligated it to a linker specific for each enzyme and amplified it by PCR with Cy5-labeled linker-specific primers (Fig. 2). For the common standard competitor DNA, we used genomic DNA from similarly processed HFF2 fibroblasts but without denaturation (non-snap-back DNA) and amplified it using Cy3-labeled linker-specific

primers. We competitively hybridized Cy5-labeled snap-back DNA from HFF2, HDF1 or HDF3 fibroblasts against Cy3-labeled non-snap-back DNA from HFF2 fibroblasts on spotted arrays containing 17,569 clones from the sequence-verified ResGen Human UniGene (31,105) Clone Set¹³, generating comparable GAPF profiles of fibroblasts from each individual. For DNA from each type of fibroblast, the procedure was done in triplicate to produce three independent preparations of LM-PCR-amplified DNA for hybridization. We calculated Storey's q value¹⁴, an estimate of the false discovery rate, for each gene in each comparison between fibroblasts to control for multiple testing errors. At a threshold of $q < 0.1$, no feature was significantly different between any two of the normal fibroblast samples.

Palindrome prevalence in human cancers

To determine whether GAPF can detect palindromes in cancer cells, we initially used the Colo320DM human colon cancer cell line. *MYC* amplification in this cell line is organized as large inverted repeats¹². We labeled snap-back DNA from Colo320DM cells with Cy5 and cohybridized it with Cy3-labeled non-snap-back DNA from HFF2 fibroblasts. Experiments were done in triplicate. We compared the GAPF profile with a 'common baseline' GAPF profile consisting of two triplicate data sets of snap-back DNA from HDF1 and HDF3 fibroblasts (Fig. 3a). The signal of 150 genes (GAPF-positive genes)

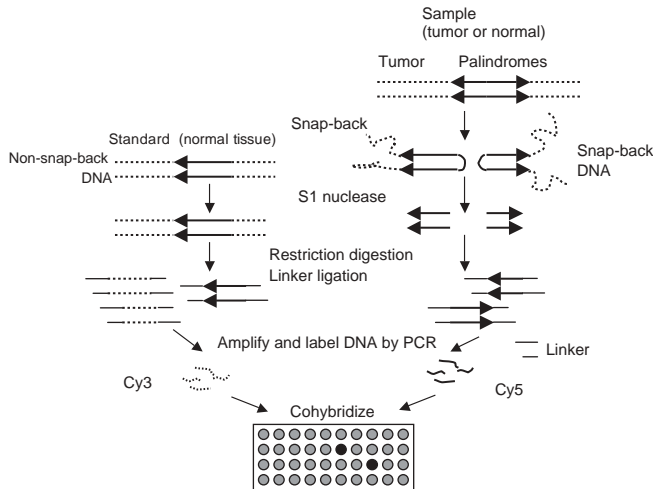


Figure 2 GAPF. To enrich for palindromic regions, DNA is denatured and rapidly renatured to create snap-back DNA, which contains double-stranded snap-back DNA near palindromic centers and largely single-stranded DNA in nonpalindromic regions. Snap-back DNA is treated with the single-strand-specific nuclease S1; digested with *MspI*, *TaqI* or *MseI*; ligated to specific linkers; and then amplified by PCR with Cy5-labeled linker-specific primers. A standard DNA for competitive hybridization was made from normal human fibroblast DNA using the same procedure on non-denatured (non-snap-back) DNA labeled with Cy3. Labeled DNAs are cohybridized onto a human spotted cDNA microarray¹³.

was significantly higher ($q < 0.05$ and relative change > 2) in the Colo320DM sample than in the common baseline samples (Supplementary Table 1 online). Validating the technique, *MYC* was GAPF-positive on the array, as confirmed using the same Southern-blot-based approach that was originally used to show its palindromic structure¹⁵ (Supplementary Fig. 1 online). This result was not simply due to the high degree of *MYC* amplification, because digestion with *EcoRI* before formation of snap-back DNA and digestion with S1 eliminated the signal probably because the *EcoRI* fragment does not contain the center of the palindrome.

For a genome-wide analysis of GAPF-positive loci, we placed the data from each gene on the array onto a physical map and grouped adjacent genes in one of two ways: (i) into 521 cytogenetic bands that ranged in size from 1 to 132 genes, with an average of 18 genes per cytogenetic band, or (ii) into 588 regions of 5-Mb bins, with an average of 22 genes per bin. Because large palindromes might extend over more than one gene, grouping genes into contiguous regions provides more statistical power by increasing the number of measurements for each region. There were no significant differences ($q < 0.05$ and relative change > 1) between the common baseline profile and the GAPF profile of HFF2 fibroblasts, or between the common baseline profile and a profile from an independent preparation of snap-back DNA from HDF3 fibroblasts, indicating that neither normal individual variance nor interexperimental variance was detected at this level of stringency. In contrast, the GAPF profile of Colo320DM cells showed 81 GAPF-positive cytogenetic bands and 60 GAPF-positive bins (Fig. 3a,b). As expected, the cytogenetic band (8q24.1) containing *MYC* was GAPF-positive (Fig. 4a and Supplementary Table 2 online).

Cytogenetic band 1q21 was also GAPF-positive (Fig. 4b and Supplementary Table 2 online). Placing the array features on the physical map of 1q21 showed that this band contains four

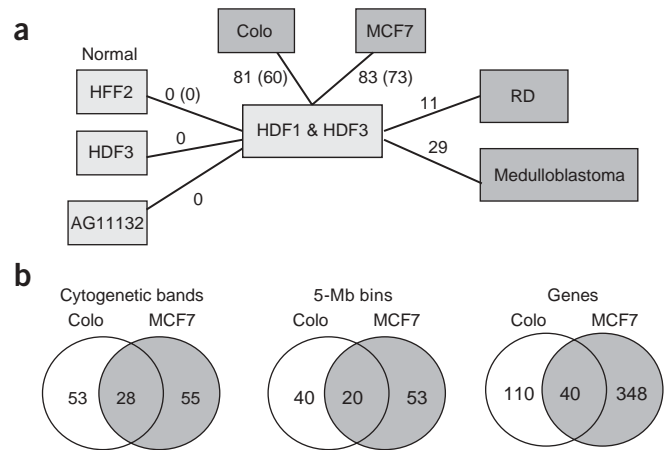


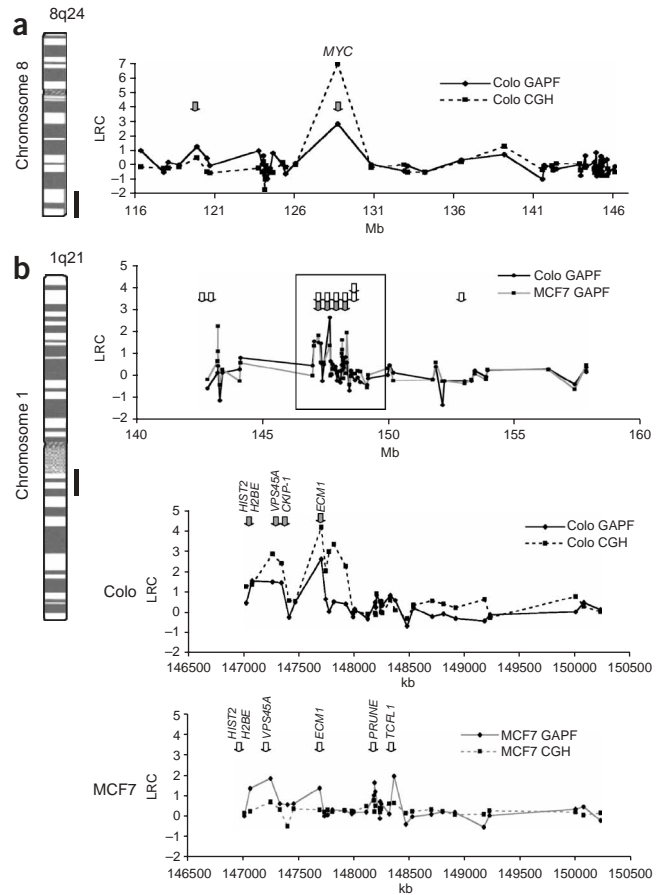
Figure 3 GAPF-positive features are not randomly distributed in cancer cells. (a) GAPF profiles from HFF2 fibroblasts (three independent snap-back samples hybridized on microarray, $n = 3$), AG11132 cells ($n = 3$), HDF3 fibroblasts (prepared independently from the HDF3 sample included in the common baseline, $n = 5$), Colo320DM cells (Colo; $n = 3$), MCF7 cells ($n = 3$), RD cells ($n = 3$) and five independent primary medulloblastoma samples were compared with a common baseline profile consisting of two triplicate data sets of snap-back DNA from HDF1 and HDF3 fibroblasts. The numbers of GAPF-positive cytogenetic bands in each cell line relative to the common baseline are given. For Colo320DM cells, MCF7 cells and HFF2 fibroblasts, the number of GAPF-positive 5-Mb bins is shown in parentheses. A list of cytogenetic bands and 5-Mb bins is given in Supplementary Table 2 online. Because the analyses based on grouping features by cytogenetic bands versus by 5-Mb bins yielded the same results, only analyses based on grouping by cytogenetic bands was done on the RD and medulloblastoma samples. (b) Significant overlap of GAPF-positive features between Colo320DM (Colo) and MCF7 cells for a total of 521 cytogenetic bands ($P = 4 \times 10^{-6}$), 588 5-Mb bins ($P = 3 \times 10^{-6}$) or 11,871 individual genes ($P < 1 \times 10^{-99}$). Statistical analysis was based on the hypergeometric distribution.

individual GAPF-positive genes (*HIST2H2BE*, *VPS45A*, *CKIP-1* and *ECM1*) that closely cluster within 600 kb (Fig. 4b and Supplementary Table 3 online). Of these four genes, *ECM1* had the highest relative change and the lowest q value (0.003) and was therefore the most likely to be near the center of the palindrome. Southern-blot analysis using a probe within *ECM1* showed that restriction fragments from snap-back DNA migrated at one-half of the fragment size in non-snap-back DNA (Fig. 5a), and additional restriction mapping based on the genomic sequence placed the center of a palindrome near the 5' end of *ECM1* in Colo320DM cells (Fig. 5b and Supplementary Fig. 2 online).

Colo320DM and MCF7 share common regions of palindromes

To determine whether GAPF sites have similar distribution and frequency in different cancer cells, we obtained GAPF profiles for a breast cancer cell line (MCF7), a normal breast epithelial cell line (AG11132)¹⁶ and an embryonal rhabdomyosarcoma cell line (RD)¹⁵. No cytogenetic bands were GAPF-positive in the comparison of AG11132 cells with the normal HDF fibroblast baseline, whereas 83 cytogenetic bands and 73 bins were GAPF-positive in MCF7 cells relative to HDF fibroblasts (Fig. 3a), including cytogenetic band 1q21 ($q = 0.0056$; Supplementary Table 2 online). Overall, there was a significant overlap of GAPF-positive cytogenetic bands and bins between Colo320DM and MCF7 cells (28 bands, $P = 3 \times 10^{-6}$; 20 bins, $P = 4 \times 10^{-6}$; Fig. 3b and Supplementary Table 2 online),

Figure 4 Physical maps of GAPF-positive cytogenetic bands. Genes from each cytogenetic band and the surrounding region are plotted, and the average log₂ relative changes (LRCs; *n* = 3) of GAPF and array CGH for Colo320DM (Colo) or MCF7 cells relative to HDF fibroblasts are shown. Arrows indicate GAPF-positive genes (*q* < 0.05 and a log₂ relative change > 1) in Colo320DM (gray) or MCF7 (white) cells. (a) *MYC* is amplified as a DNA palindrome in Colo320DM cells. GAPF (solid line) and CGH profiles (dotted line) of 32-Mb regions at 8q24 are shown. (b) Clustering of GAPF-positive genes in a 600-kb region at 1q21 in Colo320DM and MCF7 cells. Upper panel: GAPF profiles of the 18-Mb region at 1q21 from both Colo320DM and MCF7 cells are shown. GAPF-positive genes are clustered (highlighted in rectangle). Lower panel: Detailed GAPF and array-CGH profiles of the 4-Mb region highlighted in the rectangle in the panel above from Colo320DM and MCF7 cells. (The GAPF-negative features interspersed between GAPF-positive features, such as between *PRUNE* and *TCFL1* in the MCF7 sample or between *CSNK1A1* and *ECM1* in the Colo320DM sample, might represent areas difficult to amplify or hybridize by GAPF for technical reasons or could indicate that the cell population has two different palindromic centers within this region.) Additional data are given in **Supplementary Table 3** online.



indicating that these epithelial tumor cell lines have common sites of palindrome formation. There was also a significant overlap of GAPF-positive genes between Colo320DM (150 genes) and MCF7 cells (388 genes; 40 genes in common, *P* < 1 × 10⁻⁹⁹, **Fig. 3b** and **Supplementary Table 1** online). At 1q21, the GAPF-positive genes in MCF7 cells are clustered in a 1.3-Mb region and include three genes that are also GAPF-positive in Colo320DM cells (*HIST2H2BE*, *VPS45A* and *ECM1*; **Fig. 4b** and **Supplementary Table 3** online). The GAPF profile of the RD cell line identified 11 GAPF-positive cytogenetic bands (**Fig. 3a** and **Supplementary Table 2** online), which had no significant overlap with those of Colo320DM (*P* = 0.294) or MCF7 (*P* = 0.296) cells. These results suggest that GAPF patterns are not random but vary among different tumor cell lines.

A subset of GAPF-positive genes is amplified

Palindrome formation is an initial step of gene amplification^{4,17}. To determine whether palindromes are associated with increases in gene copy number, we carried out array comparative genomic hybridization (CGH) for Colo320DM and MCF7 cells using the same spotted cDNA arrays that we used for the GAPF study. Most GAPF-positive genes were not associated with an increase in gene copy number, either because CGH was not sensitive enough to detect it or because the cell population was heterogeneous and only a subset of cells carried the palindrome. But GAPF-positive genes were more likely than other loci to be amplified (**Table 1**), indicating that a subset of GAPF-positive loci is selected for amplification. For example, at 1q21, subsets of the GAPF-positive genes were amplified in Colo320DM but not MCF7 cells (**Fig. 4b** and **Supplementary Table 3** online). At 7q35, where a common fragile site (*FRA7I*) has been implicated in the palindromic amplification of the *PIP* oncogene in breast cancer cells¹⁸, three genes (*ZNF282*, *REP1N1* and *KCNH2*) within a 1.8-Mb region were GAPF-positive in Colo320DM cells but showed no differences in copy number by CGH (**Supplementary Table 3** online). These differences indicate that GAPF and array CGH measure different features in cancer cells: GAPF measures a structural feature (palindrome) and CGH measures the average copy number of genes in cell populations. The fact that GAPF-positive genes are more likely than random loci to be amplified is consistent with our demonstration that palindrome formation is a rate-limiting step in gene amplification.

Palindromes in primary medulloblastomas

Colo320DM, MCF7 and RD cell lines are derived from primary tumors, and the widespread palindrome formation identified by GAPF might be the result of multiple passages in culture. To examine palindrome formation in primary tumors, we carried out GAPF analysis on DNA isolated from five independent primary medulloblastomas, the most common central nervous system malignancy of childhood. We detected palindrome formation at 29 cytogenetic bands in the primary human medulloblastomas (**Fig. 3a** and **Supplementary Table 2** online). These GAPF-positive bands include loci that are commonly amplified in medulloblastoma tissues, including 6q (6q12 and 6q14), 4q (4q24 and 4q25) and 7q (7q21.1, 7q22.1 and 7q31)¹⁹. Other GAPF-positive bands, such as 1p34.2, 5p15.2, 5p15.3 and 13q34, include loci that are highly amplified in a subset of medulloblastomas²⁰, again suggestive of a link between gene amplification and palindrome formation.

In contrast to the similarity between the GAPF profiles of Colo320DM and MCF7 cells, there was no significant overlap of GAPF-positive cytogenetic bands between medulloblastomas and Colo320DM cells (seven cytogenetic bands, *P* = 0.099) or between medulloblastomas and MCF7 cells (seven bands, *P* = 0.092); however, significant overlap was evident between medulloblastomas and RD cells (three bands, *P* = 0.017; **Supplementary Table 2** online). Therefore, the widespread occurrence of palindromes is not simply a tissue culture artifact but occurs in primary tumors as well. In addition, these results indicate that the distribution of palindromes is different in two types of pediatric tumor (medulloblastomas and rhabdomyosarcomas) than in the colon and breast cancer cell lines that we analyzed.

Figure 5 Palindromic amplification of *ECM1* in Colo320DM cells. (a) Snap-back DNA analysis of amplified *ECM1* by Southern blotting indicates a palindromic structure. Southern-blot analysis of snap-back DNA (SB) from Colo320DM (Colo) cells shows an *EcoRI* fragment that is one-half the size (6-kb) of that of non-snap-back DNA (12 kb; schematically shown in lower panel), indicative of a palindromic amplification of *ECM1*. The 12-kb palindromic *EcoRI* fragment comigrates with the 12.9-kb nonrearranged *EcoRI* fragment (**Supplementary Fig. 2** online). The human myogenin probe was serially hybridized to the same blot as a loading control. (b) Restriction mapping of snap-back DNA identifies the center of the palindrome. The restriction map of snap-back DNA (SB) for *AflII* (A), *EcoRI* (E), *EcoRV* (EV), *HindIII* (H) and *PstI* (P) shown is based on the human genome sequence. The gray rectangle represents the region of *ECM1*. *EcoRI* and *AflII* sites are 1.3 kb and 2.8 kb, respectively, telomeric to the *EcoRV* site. The 7.5-kb *AflII* fragment, 6-kb *EcoRI* fragment and 4.7-kb *EcoRV* fragment from snap-back DNA map the center of the palindrome at the 5' end of *ECM1*. The non-snap-back *AflII*, *EcoRI* and *EcoRV* fragments (~15, 12 and 9.4 kb, respectively) are present in Colo320DM cells (**Supplementary**

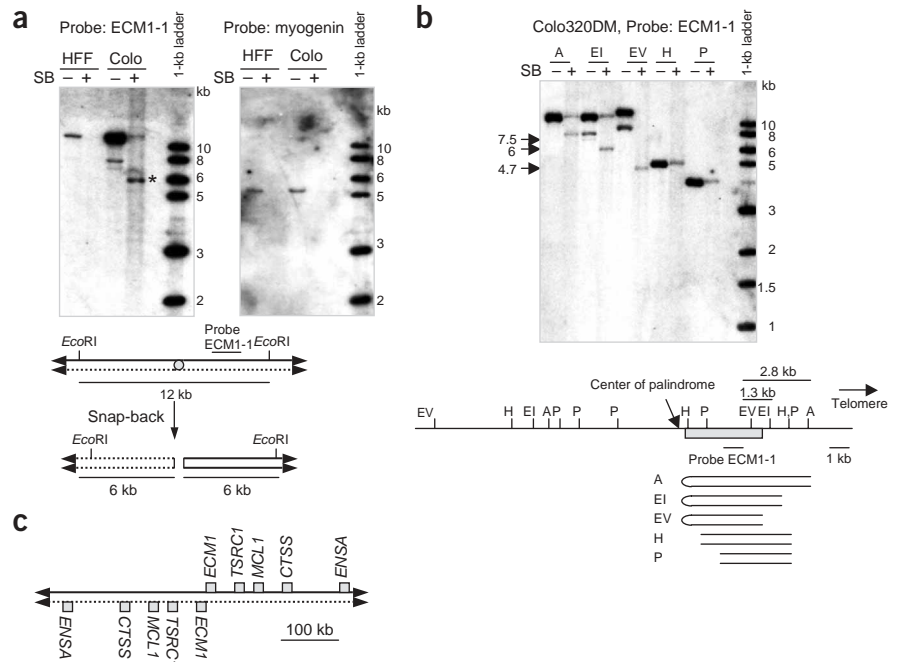


Fig. 2 online). (c) Inferred structure of the palindromic amplification at 1q21 in Colo320DM cells. Array CGH identifies amplification at this locus covering five genes in a region spanning at least 220 kb (**Fig. 4b** and **Supplementary Table 3** online), but only *ECM1* is GAPF-positive. Southern-blot analysis indicates that *ECM1* resides near the center of a palindromic region. The other four genes are not GAPF-positive, probably owing to their distance from the center of the palindrome and the size of the DNA fragments extracted from Colo320DM cells.

DISCUSSION

These results show that palindromes occur frequently and may have characteristic patterns of distribution throughout the genome in different cancer cells. Unlike conventional array-CGH analysis, which measures the average gene dosage in cell populations, GAPF provides a qualitative measurement of a structural chromosomal aberration (palindromes) that has previously been examined only by cytogenetic studies. Detailed mapping shows that the palindromes tend to cluster at specific regions, some of which undergo gene amplification. On the basis of the data currently available, we cannot determine whether the palindromes we identified in cancer cells were somatically generated secondary to a cancer-associated genomic

instability or were germline rearrangements, and we will need to analyze matched samples of normal and tumor tissue to distinguish between these possibilities. Because we have not yet identified similar widespread palindrome formation in noncancer cells (three different fibroblast cell lines and one breast epithelial cell line), we will discuss our results with the assumption that the palindrome formation we identified here is a cancer-associated event.

The BFB cycle has been well-studied cytogenetically as a common mechanism for gene amplification⁷⁻¹⁰. The BFB cycle is triggered by recombinogenic free ends, followed by fusion of broken sister chromatids, which generates a dicentric chromosome^{21,22}. This dicentric chromosome segregates into different daughter cells and results in chromosome breaks at mitosis. At the end of each cycle, extra copies of chromosomal regions accumulate at the chromosome end in a large DNA palindrome. Many amplifications of drug-resistance genes in rodent cells are associated with large DNA palindromes^{7-10,17}, and an initial palindromic duplication of *DHFR* triggers the BFB cycle and subsequent gene amplification⁴. In addition, the BFB cycle is a key mechanism for generating chromosome aberrations in cancer cells²³⁻²⁶. Telomere dysfunction leads to dicentric chromosomes through end-to-end fusions and triggers BFB cycles^{23,27}. Chromosome end-to-end fusions induced by spontaneous or ionizing radiation are seen in cells that have cancer-predisposing mutations, such as a deficiency in the DNA damage checkpoint and repair function²⁸⁻³¹. Thus, failure to maintain genetic integrity often leads to BFB cycles that can cause palindromes in the cancer genome.

ECM1 is a significantly GAPF-positive gene in Colo320DM cells, and Southern-blot analysis indicates that the center of a palindrome is located close to the 5' end of *ECM1* (**Fig. 5** and **Supplementary Fig. 2** online). Four genes telomeric to *ECM1* (*TSRC1*, *MCL1*, *CTSS* and *ENSA*) are also amplified in this region but are not GAPF-positive

Table 1 A subset of Colo320DM GAPF-positive genes is amplified

CGH relative change (log ₂) ^a	Copy-number change ^b	Observed ^c	Expected ^d
$x < -0.42$	≤ -1N	20	24
$-0.42 < x < 0.32$	ON	66	94
$0.32 < x < 0.81$	+1N	35	28
$x > 0.81$	≥ +2N	29	5

Frequency of the observed change in copy number of GAPF-positive genes in Colo320DM cells relative to HDF fibroblasts compared with the expected frequency based on the distribution of array CGH ratios across all genes. ^aLog-transformed relative change ratio of the CGH values in Colo320DM cells versus HDF fibroblasts. ^bN refers to the copy number in Colo320DM cells relative to HDF fibroblasts. -1N is the loss of one allele; ON is no change; +1N is the gain of 1 allele; +2N is the gain of two alleles. ^cThe CGH ratio for the 150 GAPF-positive genes. Based on the relative change value derived from the CGH study, each GAPF-positive gene was assigned to a copy-number change category. If there was no association between GAPF-positive genes and gene amplification, this column would be expected to approximate a normal distribution. The bold numbers deviate significantly ($P < 0.001$, χ^2 analysis) from the expected normal distribution based on post hoc analysis. ^dThe expected number of GAPF-positive genes in each category assuming no association with gene amplification and calculated according to the normal distribution of the 11,974 features from the CGH array data.



(Fig. 4b and Supplementary Table 3 online). This is consistent with a head-to-head organization of amplicons in this region (Fig. 5c) and confirms that GAPF was able to detect the center of a palindrome in an amplified region. The entire amplicon is complex, however, and the presence of restriction fragments of the predicted size for a nonrearranged allele (Supplementary Fig. 2 online) implies that additional palindromic centers or other forms of rearrangements, such as a tandem array amplification, exist^{32,33}.

The clustering of somatic palindromes within a subregion of a cytogenetic band could be due to clustering of chromosome breakage sites in the genome, as chromosome breakage facilitates palindrome formation⁴. The clastogenic drugs colomycin and actinomycin D induce initial chromosome breaks at loci bracketing the palindromic amplification of the gene *AMPD2* (ref. 10), and aphidicolin-induced common fragile sites are involved in oncogene amplification in human cancer cells^{18,34}, suggesting that specific sites of chromosome breaks make a region susceptible to amplification. The distinct distributions of GAPF sites in different cancers suggest that whatever mechanism induces DSB, cell-specific factors may alter the site of occurrence. Additional GAPF studies on higher resolution arrays might identify specific chromosomal regions or sequences that are susceptible to DNA breaks and palindrome formation.

In addition to the requirement for a double-strand break (DSB), complex genomic architecture might determine where palindromes can form. In the simple eukaryotes *Tetrahymena*^{5,35,36}, *Schizosaccharomyces pombe*³⁷, *Saccharomyces cerevisiae*³⁸ and *Leishmania*³⁹, palindrome formation is mediated by short inverted repeats that naturally exist in the genome. In *S. cerevisiae*, exogenous short inverted repeats inserted in the chromosome can induce chromosome breaks and palindrome formation⁴⁰. We showed directly that in mammalian cells, short inverted repeats can mediate palindrome formation following an adjacent DSB⁴. Inverted repeats are common in the human genome and are often involved in disease-related DNA rearrangements^{41,42}. Further studies might determine whether naturally existing short inverted repeats facilitate the widespread palindrome formation that we have characterized in cancer cells.

Several of the GAPF-positive genes that we identified are associated with translocations in some tumor types, such as T-cell leukemia/lymphoma 1A (*TCL1A*)⁴³, synovial sarcoma X breakpoint 4 (*SSX4*)⁴⁴ and myeloid leukemia factor 1 (*MLF1*)⁴⁵ (Supplementary Table 1 online). Therefore, DSBs at these genes might be resolved as either a palindrome or a translocation, with different consequences for the progression of the tumor. In RD cells, 2q35 was GAPF-positive, and the gene *PAX3* in this region was enriched by GAPF, although it did not meet the preset statistical criteria to be independently called GAPF-positive as a single gene. Alveolar rhabdomyosarcomas are characterized by a t(2;13)(q35;q14) translocation that fuses the genes *PAX3* and *FOXO1A* on chromosome 13, whereas embryonal rhabdomyosarcomas do not carry this translocation⁴⁶; however, the association of this region with a somatic palindrome formation in an embryonal rhabdomyosarcoma (RD cells) implies that *PAX3* also resides in a region susceptible to DSBs and suggests that the alternative resolutions of a DSB might determine the subtype of rhabdomyosarcoma generated.

For medulloblastomas, the fact that five independent primary tumors have common loci of somatic palindrome formation indicates that there is a shared mechanism of palindrome formation and that tumor-specific properties determine their genomic location. The palindromic regions contain genes that might contribute to tumor progression. For example, *SKP2* at 5p13 encodes a subunit of the ubiquitin ligase complex that regulates entry into S phase⁴⁷; *FZD1* at 7q21.1 encodes a receptor for Wnt signaling pathway that is often

dysregulated in medulloblastomas⁴⁸; and *TERT*, telomere reverse transcriptase, at 5p15.3 is often amplified in medulloblastomas⁴⁹.

Many of the GAPF-positive genes are not associated with an increase in gene copy number as measured by array CGH. Most somatic palindromes probably have not undergone further amplification, which indicates that palindrome formation is an initial and fundamental step in cancer formation. Because the palindromes we are detecting may be present in only a subpopulation of cells (see Fig. 1c), however, the absence of gene amplification in CGH analysis might be partly due to an heterogeneous cell population. Therefore, the genomic distribution of palindromes would provide a structural platform for subsequent gene amplification in the cancer cell. In this model, different tumor types might have a common set of palindromes that mark regions susceptible to amplification through BFB cycles, and selective advantage would determine whether the amplified loci are retained. For example, GAPF-positive genes cluster at 1q21 in both Colo320DM and MCF7 cells, but copy number is increased only in Colo320DM cells (Fig. 4), suggesting that a gene in this region gives selective advantage to Colo320DM but not MCF7 cells. Our observation that the genomic distribution of palindromes differs among cancer cells further suggests that specific mechanisms lead to the locus-specific formation of a palindrome, which might determine the susceptibility of that locus for subsequent gene amplification.

We developed a new and robust microarray-based approach to identify the genome-wide distribution of somatic palindrome formations in cancer cells. Our initial studies used a cDNA spotted array with ~18,000 features. Because our technique might not detect palindromes distant from the point of symmetry, use of a representative cDNA array might not show the full frequency or extent of genomic palindromes. Future studies with more saturating array technologies have the potential to map accurately the points of palindrome symmetry and possibly to identify the sequences that determine palindrome location. Delineation of tumor-specific sites of palindrome formation would provide a new opportunity to develop assays for early detection and residual disease and might have predictive value in tumor progression.

METHODS

Cell lines and cancer tissues. D791R-8 and D791R-8-Sce2 cells were previously described⁴. We obtained Colo320DM and RD cells obtained from American Type Culture Collection. MCF7 and AG11132 (ref. 16) cells were gifts from K. Swisshelm (University of Washington). We obtained skin biopsy-derived HDF1 and HDF3 fibroblasts from T. Bird (University of Washington) and human foreskin HFF2 fibroblasts from D. Miller (Fred Hutchinson Cancer Research Center) as anonymous cell lines. DNA samples, stripped of identifying information, from five primary medulloblastomas were provided by J. Olson (Fred Hutchinson Cancer Research Center). All samples were obtained after review and approval for use of anonymous human DNA samples and human cell lines by the Institutional Review Board of the Fred Hutchinson Cancer Research Center.

Microarray analysis. We extracted high-molecular-weight genomic DNA as described previously⁴. To make snap-back DNA, we boiled 2 µg of genomic DNA in 50 µl of water with 100 mM NaCl for 7 min and then placed it on ice to cool quickly. We added 6 µl of S1 nuclease buffer, 4 µl of 3 M NaCl and 100 U of S1 nuclease (Invitrogen) to the DNA and incubated it at 37 °C for 1 h. We inactivated S1 nuclease by adding 10 mM EDTA and carrying out phenol-chloroform extraction. We precipitated DNA by ethanol, dissolved it in water and digested it with 40 U of *MspI*, *TaqI* or *MseI* for 16 h. We precipitated DNA, dissolved in 21 µl of water and ligated it to *MspI*-, *TaqI*- or *MseI*-specific linker by adding 5 µl of 20 mM linker, 3 µl of T4 DNA ligase buffer and 400 U of T4 DNA ligase at 16 °C for 16 h. We precipitated DNA, dissolved it in 200 µl of Tris-EDTA and then applied it onto Microcon YM-50 to remove excess linker. We recovered DNA in 20 µl of water. For each cell line or tumor tissues, we

prepared templates with three different linkers. For PCR, we mixed 2 μ l of DNA, 0.5 μ l of Faststart Taq DNA polymerase (Roche), 2.5 μ l of 2 mM dNTP, 5 μ l of 10 \times PCR buffer with MgCl₂ (for Faststart Taq DNA polymerase, Roche) and 2 μ M of a Cy3- or Cy5-labeled linker-specific primer with water to a total volume of 50 μ l. For *Msp*I-linker ligated DNA, we also included 5 μ l of GC-rich solution (for Faststart Taq DNA polymerase, Roche). We carried out PCR at 96 °C for 6 min followed by 30 cycles of 96 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s on a GeneAmp PCR system 9700 (Perkin-Elmer). We mixed PCR reactions for the same template from three different linker-specific primers and purified them using PCR purification Kit (QIAGEN). The procedure was done in triplicate to produce three independent preparations of LM-PCR-amplified DNA for hybridization for Colo320DM, MCF7 and RD cells. We added 100 μ g of human Cot-1 DNA, 20 μ g of polyA/dT and 100 μ g of yeast tRNA for hybridizing 18k human cDNA array¹³. Cot-1 DNA suppresses hybridization of repetitive sequences (Supplementary Fig. 3 online). For primary medulloblastomas, we processed each tumor sample individually and compared the GAPF profiles from the five independent samples with the GAPF profile for the HDF fibroblasts. To prepare template DNA for array-CGH analysis, we digested genomic DNA with *Msp*I, *Taq*I or *Mse*I and ligated it with a linker specific for each restriction enzyme. We amplified three independent preparations of template DNA by either Cy3- or Cy5-labeled linker-specific primer. Cohybridization either of Cy3-labeled cancer (Colo320DM or MCF7) DNA with Cy5-labeled normal (HFF2) DNA or of Cy5-labeled cancer DNA with Cy3-labeled normal DNA was done in triplicate. Oligonucleotides were synthesized by QIAGEN Genomics. DNA sequence for primers and linkers are available on request.

Southern blotting. We carried out Southern blotting as described⁴. We digested 2 μ g of high-molecular-weight human genomic DNA with restriction enzymes, separated it on 0.8% agarose gels and blotted it to nylon membranes. We prepared snap-back DNA by boiling 2 μ g of genomic DNA in 50 μ l of water with 100 mM NaCl for 7 min and then immediately placing it on ice to cool. We precipitated DNA using ethanol and digested it with restriction enzyme. We used 2.5 kb of Molecular Ruler (BIO-RAD), a 1-kb DNA ladder and a 100-bp DNA ladder (New England Biolabs) as size markers. To make a probe for Southern-blot analysis, we amplified human genomic DNA by PCR and cloned a fragment by TOPO TA Cloning Kit (Invitrogen). Oligo primer sequences are available on request.

Statistical analysis. We normalized array data using the GeneSpring Analysis Package, version 6.2 (Silicon Genetics), by Lowess normalization (an intensity-dependent algorithm) and then transformed them into logarithmic space, base 2. We annotated data by cytogenetic band or by UniGene cluster using National Center for Biotechnology Information databases current as of February 2004. We carried out Welch's *t*-test for each cytogenetic band or UniGene cluster comparing replicate data sets. We used Storey's *q* value¹⁴ to control for multiple testing error and transformed each *P* value to a *q* value, which is an estimate of the false discovery rate.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J. Delrow and the Fred Hutchinson Cancer Research Center Genomics facility for advice on the array studies and P. Nieman, M. Groudine, B. Trask, C. Kemp, J. Radich, J. Roberts and members of the laboratory of M.-C.Y. for comments on the manuscripts. This work was supported by National Institutes of Health grants from General Medicine (M.-C.Y.) and National Institute of Arthritis, Musculoskeletal and Skin Diseases (S.J.T.) and by the Avon Breast Cancer Crusade Opportunity Fund (to H.T.). H.T. is a recipient of Interdisciplinary Dual Mentor Fellowship of Fred Hutchinson Cancer Research Center (1999 and 2001).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 11 November 2004; accepted 3 January 2005
Published online at <http://www.nature.com/naturegenetics/>

1. Tlsty, T.D. Genomic instability and its role in neoplasia. *Curr. Top. Microbiol. Immunol.* **221**, 37–46 (1997).

2. Lengauer, C., Kinzler, K.W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
3. Albertson, D.G., Collins, C., McCormick, F. & Gray, J.W. Chromosome aberrations in solid tumors. *Nat. Genet.* **34**, 369–376 (2003).
4. Tanaka, H., Tapscott, S.J., Trask, B.J. & Yao, M.C. Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells. *Proc. Natl. Acad. Sci. USA* **99**, 8772–8777 (2002).
5. Yasuda, L.F. & Yao, M.C. Short inverted repeats at a free end signal large palindromic DNA formation in *Tetrahymena*. *Cell* **67**, 505–516 (1991).
6. Butler, D.K., Yasuda, L.E. & Yao, M.C. Induction of large DNA palindrome formation in yeast: implications for gene amplification and genome stability in eukaryotes. *Cell* **87**, 1115–1122 (1996).
7. Toledo, F., Le Roscouet, D., Buttin, G. & Debatisse, M. Co-amplified markers alternate in megabase long chromosomal inverted repeats and cluster independently in interphase nuclei at early steps of mammalian gene amplification. *EMBO J.* **11**, 2665–2673 (1992).
8. Smith, K.A., Stark, M.B., Gorman, P.A. & Stark, G.R. Fusions near telomeres occur very early in the amplification of CAD genes in Syrian hamster cells. *Proc. Natl. Acad. Sci. USA* **89**, 5427–5431 (1992).
9. Ma, C., Martin, S., Trask, B. & Hamlin, J.L. Sister chromatid fusion initiates amplification of the dihydrofolate reductase gene in Chinese hamster cells. *Genes Dev.* **7**, 605–620 (1993).
10. Coquelle, A., Pipiras, E., Toledo, F., Buttin, G. & Debatisse, M. Expression of fragile sites triggers intrachromosomal mammalian gene amplification and sets boundaries to early amplicons. *Cell* **89**, 215–225 (1997).
11. Ish-Horowitz, D. & Pinchin, S.M. Genomic organization of the 87A7 and 87C1 heat-induced loci of *Drosophila melanogaster*. *J. Mol. Biol.* **142**, 231–245 (1980).
12. Ford, M. & Fried, M. Large inverted duplications are associated with gene amplification. *Cell* **45**, 425–430 (1986).
13. Graf, L., Iwata, M. & Torok-Storb, B. Gene expression profiling of the functionally distinct human bone marrow stromal cell lines HS-5 and HS-27a. *Blood* **100**, 1509–1511 (2002).
14. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
15. Bernasconi, M., Remppis, A., Fredericks, W.J., Rauscher, F.J. 3rd & Schafer, B.W. Induction of apoptosis in rhabdomyosarcoma cells through down-regulation of PAX proteins. *Proc. Natl. Acad. Sci. USA* **93**, 13164–13169 (1996).
16. Swisshelm, K. *et al.* SEMP1, a senescence-associated cDNA isolated from human mammary epithelial cells, is a member of an epithelial membrane protein superfamily. *Gene* **226**, 285–295 (1999).
17. Ruiz, J.C. & Wahl, G.M. Formation of an inverted duplication can be an initial step in gene amplification. *Mol. Cell. Biol.* **8**, 4302–4313 (1988).
18. Ciullo, M. *et al.* Initiation of the breakage-fusion-bridge mechanism through common fragile site activation in human breast cancer cells: the model of PIP gene duplication from a break at FRA71. *Hum. Mol. Genet.* **11**, 2887–2894 (2002).
19. Shlomit, R. *et al.* Gains and losses of DNA sequences in childhood brain tumors analyzed by comparative genomic hybridization. *Cancer Genet. Cytogenet.* **121**, 67–72 (2000).
20. Michiels, E.M. *et al.* Genetic alterations in childhood medulloblastoma analyzed by comparative genomic hybridization. *J. Pediatr. Hematol. Oncol.* **24**, 205–210 (2002).
21. McClintock, B. Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* **16**, 13–47 (1951).
22. Smogorzewska, A., Karlseder, J., Holtgreve-Grez, H., Jauch, A. & de Lange, T. DNA ligase IV-dependent NHEJ of deprotected mammalian telomeres in G1 and G2. *Curr. Biol.* **12**, 1635–1644 (2002).
23. Maser, R.S. & DePinto, R.A. Connecting chromosomes, crisis, and cancer. *Science* **297**, 565–569 (2002).
24. Gisselsson, D. *et al.* Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proc. Natl. Acad. Sci. USA* **97**, 5357–5362 (2000).
25. Gisselsson, D. *et al.* Abnormal nuclear shape in solid tumors reflects mitotic instability. *Am. J. Pathol.* **158**, 199–206 (2001).
26. Rudolph, K.L., Millard, M., Bosenberg, M.W. & DePinto, R.A. Telomere dysfunction and evolution of intestinal carcinoma in mice and humans. *Nat. Genet.* **28**, 155–159 (2001).
27. DePinto, R.A. The age of cancer. *Nature* **408**, 248–254 (2000).
28. Metcalfe, J.A. *et al.* Accelerated telomere shortening in ataxia telangiectasia. *Nat. Genet.* **13**, 350–353 (1996).
29. Bailey, S.M. *et al.* DNA double-strand break repair proteins are required to cap the ends of mammalian chromosomes. *Proc. Natl. Acad. Sci. USA* **96**, 14899–14904 (1999).
30. Ferguson, D.O. *et al.* The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. *Proc. Natl. Acad. Sci. USA* **97**, 6630–6633 (2000).
31. Hsu, H.L. *et al.* Ku acts in a unique way at the mammalian telomere to prevent end joining. *Genes Dev.* **14**, 2807–2812 (2000).
32. Ma, C., Looney, J.E., Leu, T.H. & Hamlin, J.L. Organization and genesis of dihydrofolate reductase amplicons in the genome of a methotrexate-resistant Chinese hamster ovary cell line. *Mol. Cell. Biol.* **8**, 2316–2327 (1988).
33. Kuwahara, Y. *et al.* Alternative mechanisms of gene amplification in human cancers. *Genes Chromosomes Cancer* **41**, 125–132 (2004).
34. Hellman, A. *et al.* A role for common fragile site induction in amplification of human oncogenes. *Cancer Cell* **1**, 89–97 (2002).



35. Yao, M.C., Yao, C.H. & Monks, B. The controlling sequence for site-specific chromosome breakage in *Tetrahymena*. *Cell* **63**, 763–772 (1990).
36. Butler, D.K., Yasuda, L.E. & Yao, M.C. An intramolecular recombination mechanism for the formation of the rRNA gene palindrome of *Tetrahymena thermophila*. *Mol. Cell. Biol.* **15**, 7117–7126 (1995).
37. Albrecht, E.B., Hunyady, A.B., Stark, G.R. & Patterson, T.E. Mechanisms of *sod2* gene amplification in *Schizosaccharomyces pombe*. *Mol. Biol. Cell* **11**, 873–886 (2000).
38. Maringele, L. & Lydall, D. Telomerase- and recombination-independent immortalization of budding yeast. *Genes Dev.* **18**, 2663–2675 (2004).
39. Grondin, K., Roy, G. & Ouellette, M. Formation of extrachromosomal circular amplicons with direct or inverted duplications in drug-resistant *Leishmania tarentolae*. *Mol. Cell. Biol.* **16**, 3587–3595 (1996).
40. Lobachev, K.S., Gordenin, D.A. & Resnick, M.A. The *mre11* complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. *Cell* **108**, 183–193 (2002).
41. Inoue, K. & Lupski, J.R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
42. Barbouti, A. *et al.* The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic low-copy repeats. *Am. J. Hum. Genet.* **74**, 1–10 (2004).
43. Hecht, F., Morgan, R., Hecht, B.K. & Smith, S.D. Common region on chromosome 14 in T-cell leukemia and lymphoma. *Science* **226**, 1445–1447 (1984).
44. Skytting, B. *et al.* A novel fusion gene, SYT-SSX4, in synovial sarcoma. *J. Natl. Cancer Inst.* **91**, 974–975 (1999).
45. YonedaKato, N. *et al.* The t(3;5)(q25.1;q34) of myelodysplastic syndrome and acute myeloid leukemia produces a novel fusion gene NPM-MLF1. *Oncogene* **12**, 265–275 (1996).
46. Anderson, J., Gordon, A., Pritchard-Jones, K. & Shipley, J. Genes, chromosomes, and rhabdomyosarcoma. *Genes Chromosomes Cancer* **26**, 275–285 (1999).
47. Carrano, A.C., Eytan, E., Hershko, A. & Pagano, M. SKP2 is required for ubiquitin-mediated degradation of the CDK inhibitor *p27*. *Nat. Cell Biol.* **1**, 193–199 (1999).
48. Yokota, N. *et al.* Role of Wnt pathway in medulloblastoma oncogenesis. *Int. J. Cancer* **101**, 198–201 (2002).
49. Fan, X. *et al.* hTERT gene amplification and increased mRNA expression in central nervous system embryonal tumors. *Am. J. Pathol.* **162**, 1763–1769 (2003).