Non-B DNA conformations as determinants of mutagenesis and human disease

ARTICLE *in* MOLECULAR CARCINOGENESIS · APRIL 2009

Impact Factor: 4.81 · DOI: 10.1002/mc.20507 · Source: PubMed

2 AUTHORS:

Albino Bacolla
University of Texas at Austin
75 PUBLICATIONS **2,375** CITATIONS
SEE PROFILE

Robert D Wells
Institute of Biosciences and Technology Texas A …
291 PUBLICATIONS **14,898** CITATIONS
SEE PROFILE

MOLECULAR CARCINOGENESIS 48:273–285 (2009)

IN PERSPECTIVE

Non-B DNA Conformations as Determinants of Mutagenesis and Human Disease

Albino Bacolla and Robert D. Wells*

Center for Genome Research, Institute of Biosciences and Technology, Texas A&M University System Health Science Center, Texas Medical Center, Houston, Texas

Repetitive DNA motifs may fold into non-B DNA structures, including cruciforms/hairpins, triplexes, slipped conformations, quadruplexes, and left-handed Z-DNA, thereby representing chromosomal targets for DNA repair, recombination, and aberrant DNA synthesis leading to repeat expansion or genomic rearrangements associated with neurodegenerative and genomic disorders. Hairpins and quadruplexes also determined the relative abundances of simple sequence repeats (SSR) in vertebrate genomes, whereas strong base stacking has permitted the expansion of purine·pyrimidine-rich SSR during evolutionary time. SSR are enriched in regulatory and cancer-related gene classes, where they have been actively recruited to participate in both gene and protein functions. SSR polymorphic alleles in the population are associated with cancer susceptibility, including within genes that appear to share regulatory circuits involving reactive oxygen species.

Key words: DNA structure; DNA repair; cruciforms; microsatellites; ROS signaling

INTRODUCTION

Chromosomal DNA is the critical target for the complex, multistep process of carcinogenesis [1]. The integrity of the DNA molecule is of paramount importance for the survival of the host cell and for the maintenance of the species. This principle holds for all types of animal, plant, insect, and microbial species. All cells and most viruses utilize DNA as the carrier of the genetic language. DNA is constantly subjected to a variety of insults from numerous sources including radiation and chemical damage such as carcinogens. Whereas a large amount of effort has been expended to understand these processes, much less work has been devoted to the relatively recent discovery that a feature of the DNA molecule itself, namely its three-dimensional conformation, can also be mutagenic [2,3].

Genetic instabilities are involved in the etiology of cancers [4] as well as hereditary neurological diseases [5]. The molecular mechanisms involved in genetic instabilities, specifically replication, recombination, repair, and transcription, have been investigated extensively [5–8]; these instabilities are a form of mutagenesis. The concept of non-B DNA conformations as an inextricable component of the instabilities of repeating tri-, tetra-, and pentanucleotide sequences in neurological disease genes (such as fragile X syndrome, myotonic dystrophy types 1 and 2, and Friedreich ataxia) laid the foundation for this principle in other disease states [5–7,9].

Herein, we shall review features of the non-B DNA conformations and their role in defining the double-strand breakpoints responsible for gross deletions and genomic rearrangements (translocations, deletions, inversions, insertions, etc.) that serve as the molecular basis for a family of genomic disorders [10]. DNA repair and recombination mediate these processes. Incisive investigations have demonstrated that the non-B DNA conformations, not the primary sequence per se in the orthodox right-handed Watson–Crick structure, are responsible. Second, we shall review the discovery that folded-back hairpin conformations in repeated tetranucleotide and trinucleotide tracts in chromosomal DNA are responsible for the relative abundance, or even absence, of certain sequence motifs. The thermodynamic stabilities of the folded-back conformations determine the propensity of a repeating tract motif to survive through evolutionary time. Third, we present new data on human cancer susceptibility and DNA repeat polymorphism. Only a modest attempt is made herein to thoroughly

Abbreviations: DSB, double-strand break; SSR, simple sequence repeats; EGFR, epidermal growth factor receptor; ROS, reactive oxygen species.

*Correspondence to: Center for Genome Research, Institute of Biosciences and Technology, Texas A&M University System Health Science Center, Texas Medical Center, 2121 W. Holcombe Blvd., Houston, TX 77030.

Received 28 August 2008; Revised 29 October 2008; Accepted 29 October 2008

DOI 10.1002/mc.20507

Published online in Wiley InterScience (www.interscience.wiley.com)

reference all important and relevant studies because this article is one in a series of related topics.

BACKGROUND AND OVERVIEW

Non-B DNA Structures

Tracts of chromosomal DNA are structurally polymorphic. A large number of simple DNA repeat sequences can exist in at least two conformations. All sequences adopt the orthodox right-handed B-form, probably for the majority of the time, with Watson–Crick A·T and G·C bp (dot separates the DNA complementary bases or strands). However, at least 10 non-B DNA conformations (Figure 1) are formed, perhaps transiently, at specific sequence motifs as a function of negative supercoil density, generated in part by transcription, protein binding, and other factors [reviewed in [2,5,6,11]]. Studies over the past 35 yr have revealed the presence of cruciforms, triplexes, slipped (hairpin) structures, tetraplexes, and left-handed Z-DNA at different repeating sequence features in a variety of systems. These higher energy states are facilitated at specific repeating tracts by the free energy derived from negative supercoiling and other factors. In addition to the five non-B DNA structures shown in Figure 1, nodule DNA, flexible and writhed DNA, bent DNA, and sticky DNA have been characterized and investigated with respect to their biological functions [reviewed in [2,5,6,11]]. All of these conformations have bond angles that are contorted from the orthodox right-handed B conformation and/or unpaired nucleotides; these features render the non-B conformations vulnerable to potential attack by certain chemical and/or enzymatic agents. A comprehensive review [12] describes the features of non-B DNA conformations.

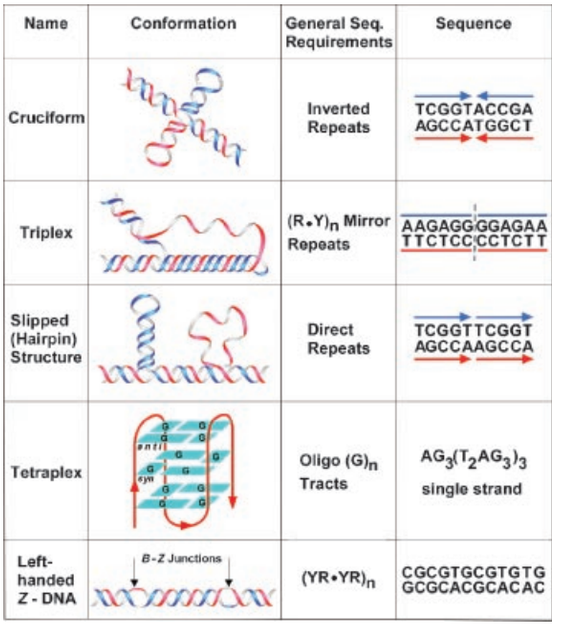

Figure 1. Non-B DNA conformations involved in genetic instabilities and chromosomal rearrangements. Features of these conformations and sequences have been described previously [2,6]. One of the duplex DNA strands is shown in blue whereas the complementary strand is shown in red. Reproduced with permission from Ref. [21].

Non-B DNA Structures and Hereditary Neurological Diseases

Studies on non-B DNA structures took a giant leap forward in the early 1990s when a number of the researchers in this area turned their attention to the mechanisms of genetic instabilities related to hereditary neurological diseases. At that time, it was discovered by the human genetics community that expansions (genetic instabilities) of repeating tri- and tetranucleotide sequences were associated with a number (~20) of hereditary neurological diseases [5–7]. Some of these diseases are myotonic dystrophy type 1 (DM1), myotonic dystrophy type 2 (DM2), fragile X syndrome, and Friedreich ataxia (FRDA). These diseases are associated with the massive expansions of repeat sequences (CTG·CAG, CCTG·CAGG, CGG·CCG, and GAA·TTC, respectively), which are able to form hairpin (CTG·CAG, CCTG·CAGG, CGG·CCG, and GAA·TTC), quadruplex (CGG·CCG) and triplex/sticky (GAA·TTC) structures, in the non-protein coding regions of their respective genes (type 1 diseases, as opposed to type 2 diseases, in which much smaller expansions occur in the protein-coding portions of genes). A large body of studies has demonstrated [5–7,13,14] that these instabilities occur by the molecular processes of DNA replication, recombination, and repair, probably acting in concert; the slippage of the two repetitive DNA complementary strands relative to each other is an integral component of the molecular mechanism. Slipped strand DNA and hairpin loops have been invoked in virtually all models that account for the expansion and deletion mechanisms. Furthermore, DNA unwinding elements, tetraplexes, triplexes, and sticky DNA are also important components in certain instability mechanisms related to individual diseases [13,14]. The replication mechanisms are influenced by pausing of replication forks, orientation of the repeat strands, location of the repeat sequences relative to the replication origins, and the flap endonuclease [13]. In addition, methyl-directed mismatch repair, nucleotide excision repair, and repair of damage caused by mutagens are also significant. Genetic recombination and double-strand break (DSB) investigations in model systems [13,15] have provided important new information on the expansion mechanisms.

Genomic Rearrangements, DNA Structure, and Human Genomic Disorders

Recent investigations [2,3,11,16–19] have shown that breakpoints of gross rearrangements, such as

deletions and translocations, coincide with non-B DNA conformations. These results are dramatic because they tie together the roles of non-B DNA structures with important biological functions and human disease. Also, they serve as a springboard for further investigations into bioinformatics related to these discoveries. Breakpoints had been recognized for many years to be important in the etiology of various translocations but these discoveries introduced the concept of non-B DNA conformations as signals for the specific breakpoints. These works have revealed that slipped structures, cruciforms, triplexes, tetraplexes, and left-handed Z-DNA (formed by runs of CG·CG dinucleotide repeats and to a lesser extent by TG·CA repeats; Figure 1) are formed in chromosomes and elicit far-reaching genetic consequences via recombination–repair. Repeating sequences, probably in their non-B conformations, cause gross genomic rearrangements (translocations, deletions, insertions, inversions, and duplications). These rearrangements are the genetic molecular basis for numerous human genomic diseases including polycystic kidney disease, adrenoleukodystrophy, follicular lymphomas, and spermatogenic failure [20]. A large number (at least 70) of diseases may fall in this category [2,16–18,21,22]. The reader is encouraged to consult these reviews for further details on these diseases.

Overarching Hypothesis

DNA polymorphisms are an important molecular component of hereditary neurological diseases as well as a variety of genomic disorders. Certain non-B DNA conformations (Figure 1 and other structures) are formed transiently in chromosomes at specific loci. These non-B DNA conformations are in equilibrium with the orthodox right-handed B-form. When the sequences are in a non-B conformation, at small direct or inverted repeat homologies, DNA repair systems recognize the altered base pairs at non-B DNA structures, which are at or near the breakpoints. The helical distortions associated with these altered base pairs are believed to be cleaved, thereby giving rise to DSBs and subsequently triggering repair–recombination events [18,23–25]. These DSBs may reside either on the same chromosome or on two distinct chromosomes. After the DNA is joined and healed by the recombination–repair systems, the mutagenic events that include gross deletions, translocations, inversions, insertions, and duplications are observed. Subsequently, through numerous steps in pathology these genetic events ultimately give rise to genomic disorders [10]. Recent bioinformatics analyses demonstrate that long homopurine·homopyrimidine sequences (in which mirror repeats may form triplexes, Figure 1) are characteristics of genes expressed in brain and the pseudoautosomal region [26], whereas quadruplex-forming motifs (formed by four runs of G_{2-4} each separated by 1–7 intervening bases, Figure 1) are enriched at regions of gene transcription initiation [27,28]. Thus, these sequences have important biological functions.

Is the DNA Conformation the Provocateur or the Sequence?

A substantial uncertainty in the history of investigations on the biological roles of non-B DNA structures concerns the sequence versus conformation issue. Specifically, is the provoking agent the sequence in the orthodox right-handed B-conformation per se or the non-B DNA conformation that can be adopted by this sequence? A recent investigation [29] developed an ingenious family of three experiments to rigorously reveal that the non-B DNA conformations are the culprit regarding mutagenesis, not the sequences per se. Long repeating tracts of CTG·CAG, CCTG·CAGG, and GAA·TTC were studied in *E. coli* and three types of mammalian fibroblast-like cells by genetic and biochemical studies including: the in vivo modulation of global negative supercoil density using topoisomerase mutants in *E. coli*; the in vivo cleavage of hairpin loops which are an obligate consequence of DNA slipped strand structures, cruciforms, and intramolecular triplexes; by the inactivation of the SbcC protein which removes hairpin loops that arise as intermediates in replication and transcription processes; and by genetic instability studies with plasmids containing long repeating sequence inserts that do, or do not, adopt non-B structures in vitro.

Prior investigations [3,16,19,30–37] were also conducted to assess the role of non-B DNA conformations in genetic instability. In these studies, sequences with different abilities to form alternative DNA structures and their orientation relative to the replication origin were investigated. In addition, chemical and genetic means to alter negative supercoiling were employed and patient samples were analyzed. These composite data clearly indicate that the non-B DNA conformations are critical for the mutagenesis mechanisms, not the sequence per se in its orthodox right-handed structure.

NON-B DNA STRUCTURES AND THE GENOMIC REPRESENTATION OF SIMPLE SEQUENCE REPEATS

Background

Simple sequence repeats (SSR) are perfect or near perfect tandem repeats of a particular *k*-mer ($k = 1–13$, also called microsatellites) and are abundant in vertebrate genomes. They comprise ~3% of the human genome [38]. Their principal characteristic is the high frequency with which the number of repeating units at a given chromosomal location varies among individuals and/or ethnic groups, giving rise to polymorphic alleles. The most common

mechanism believed to be responsible for such length changes is DNA polymerase slippage during replication; however, other mechanisms such as recombination may contribute. A considerable number of SSR have been found to be an integral part of gene and protein function, and their polymorphic nature has been associated with phenotypic variation [39] and susceptibility to disease [40]. For any given *k*-mer (i.e., 4-mer or tetranucleotide repeats) different sequence combinations display highly variable representation and length distribution in vertebrate genomes. For example, of the 33 unique sequence combinations that compose the family of tetranucleotide repeats, six sequences are present in >1000 copies at lengths ≥ 8 units (32 bp) whereas seven others are not present at all in the human genome [41]. (In the context of this review ''genetically unstable'' sequences indicate their rare occurrence genome-wide, whereas ''genetically stable'' indicates their abundance.) Also, SSR comprising only purines·pyrimidines (R•Y), such as AAAG•CTTT, AAGG•CCTT, and AAG•CTT, which are capable of forming triplex DNA, display a strong bias towards longer length distributions than the other sequences [42]. Interestingly, this latter behavior is manifested not only when genomic sequences are compared among vertebrate species and therefore analyzed in the context of evolutionary time scales, but also when the extent of microsatellite length variations is assessed on the comparatively fewer somatic cell divisions in cancer cells. Indeed, length alterations are often observed at selected tetranucleotide repeats (enhanced microsatellite alterations at selected tetranucleotide repeats or EMAST), such as AAAG·CTTT, AAGG•CCTT, and AGAT•ATCT, in specific types of cancers of the respiratory tract, skin, and bladder [43,44].

These observations underlie the notion that mutation rates, during DNA replication and recombination/repair, are determined by specific features intrinsic to the DNA sequence composition. In other words, the kinetic parameters for the chemical reactions carried out by the enzyme complexes involved depend upon the sequence composition of the substrate DNA. Thus, the varying sequence representation and length distributions of SSR may reflect the types of interactions between those intrinsic physical properties of DNA and the processing enzymes, thereby either increasing genome-wide instability (rare repeats) or, alternatively, conferring genome-wide stability (abundant/long repeats).

Genome-Wide Instability and Non-B DNA Structures

A recent analysis of tetranucleotide repeat-containing tracts in nine vertebrate genomes (human, chimpanzee, mouse, rat, dog, cow, chicken, fugu, and zebrafish) indicated that the frequencies at which each of the 33 possible sequence combination occurred were similar in all genomes, implying that repeat abundance was a function of sequence composition rather than a species-specific phenomenon [41]. Hence, the dependence of genome-wide SSR instability upon potentially intrinsic DNA features appeared to extend to all species examined.

Molecular modeling of the single-stranded sequences that composed the family of 4-mer SSR suggested that repeat abundance would correlate inversely with the thermal stability of folded-back hairpin intrastrand structures. Subsequent temperature-dependent absorption spectroscopy (melting) and circular dichroism studies established an inverse relationship between the temperatures at which each molecule would fold back upon itself to form either hairpins or quadruplex structures, and the average frequencies at which the corresponding repeat occurred in the nine vertebrate genomes. The correlation also remained significant after the fractions of CG·CG-containing tetranucleotide repeats were corrected for their low frequencies, as if methylation-induced transitions had not occurred during evolutionary time.

Similar studies on the family of 3-mer SSR (triplet repeats), which comprises 10 sequence combinations, also yielded similar conclusions. Hence, the varying thermodynamic stability of intramolecular folded-back structures adopted by each of the SSR sequences within a specific *k*-mer family appears to be a key determinant governing genome-wide SSR instability. Because SSR undergo dynamic changes as a result of replication slippage and recombination, resulting in both expansions and contractions, the depletion of certain long sequences within vertebrate genomes appears therefore to originate from their propensity to fold into increasingly (with length) stable folded-back structures, such as hairpins and quadruplexes, which are then selectively targeted for repair (Figure 2, left). By contrast, sequences devoid of hairpin-forming potential are maintained genome-wide during evolutionary time and may generate length-polymorphic alleles as a result of loop-induced mispairing during replication (replication slippage; Figure 2, right). Nevertheless, rare (unstable) SSR, such as CTG·CAG and CGG·CCG, which may form strong DNA secondary structures (see above), have been under strong positive selection at specific genomic loci (see later section), and they also exhibit length polymorphisms as a result of loop/hairpin-mispairing and replication slippage [reviewed in [6,45]]. However, contrary to the more abundant (stable) SSR, the expanded rare SSR display strong association with neuropathological conditions due to their effects in promoting gene silencing, mRNA aggregation, and protein unfolding [reviewed in [5,45,46]]. Hence, as elaborated later, length polymorphisms within genes have become both a source of susceptibility to disease and, at the same time, a means for fine-tuning

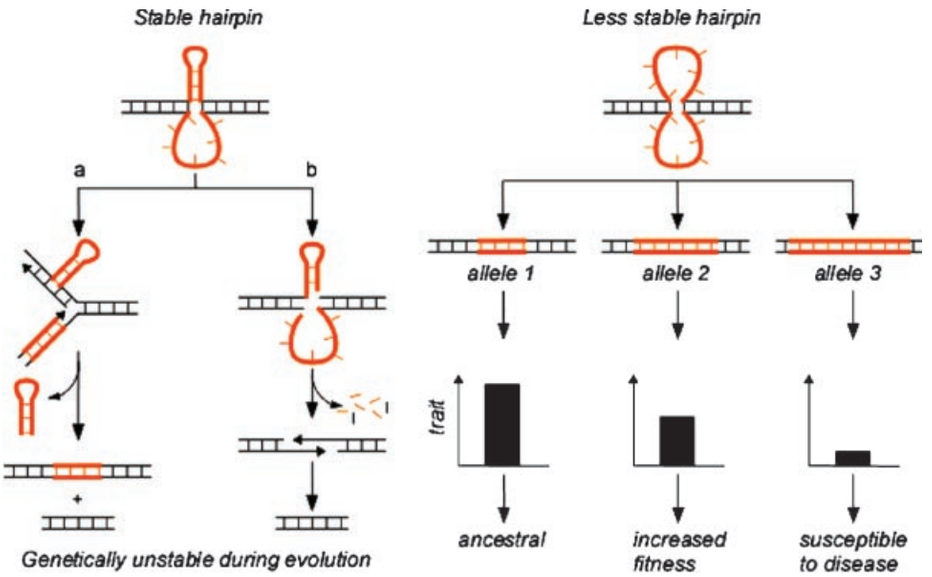

Figure 2. Model for relationships among repeat abundance, DNA structure, repeat polymorphism, and disease susceptibility. If a repeating sequence can form a stable hairpin, this self base-paired tract may be bypassed by DNA replication (left side, a) or induce DSBs (left side, b), which then trigger deletions resulting, over evolutionary time, in the loss of the underlying duplex DNA. Alternatively, if the sequence forms a less stable hairpin, it may generate longer repeat containing alleles as a consequence of DNA slippage. These longer alleles will go on to generate further length polymorphisms (right side, alleles 1–3). Functional repeat polymorphisms within human genes may be associated with variable phenotypic traits (filled bars), such as blood pressure, heart rate, and muscular tension [reviewed in [41]]. Such traits have the potential to confer either increased fitness (allele 2) or allele-specific susceptibility to complex diseases (allele 3) or repeat expansion disorders [41]. Orange, SSR. Note that overall repeat lengths are not drawn to scale and that the choice of decreased gene expression with increasing repeat length is arbitrary. Reprinted with permission from Ref. [41].

gene regulation/function, thereby increasing population fitness [41].

Long R·Y-Rich Sequences and Base Stacking

Analyses of the length distributions for SSR comprising the triplet and tetranucleotide repeats indicated that, whereas the sequence(s) exhibiting the most extended distributions may vary among species, they nevertheless involved exclusively R·Y asymmetric motifs, such as AAAG·CTTT, AAGG·CCTT, GGAT·ATCC, AGAT·ATCT, and AAG·CTT repeats [41,42]. Although the exact mechanisms underlying this behavior remain unclear, melting studies on synthetic single-stranded oligonucleotides, as well as theoretical calculations on free energies of base stacking, revealed a significant and direct correlation between the maximal lengths attained by the families of triplet and tetranucleotide repeats in the human genome and the strength of base stacking along the purine-rich strands [41]. Hence, as for the case of genome-wide SSR instability described above, an intrinsic DNA property appears to play a key role in determining the extent to which lengthy tandemly repeated sequences are tolerated within vertebrate genomes. Little is known about the exact roles of base stacking interactions on enzymatic reactions, including DNA repair. Nevertheless, current data suggest that base stacking may slow the kinetic rates for some of the reactions by inhibiting the accessibility of the target base(s) to the enzymes' active site [47]. Because both stacked bases and enzyme active sites possess a hydrophobic environment, it is possible that extrusion of a target base away from the double helix and within the enzymatic pocket would involve intermediate hydrated states, whose activation energies would be proportional to the stacking free energies. If this were the case, then the rates of enzymatic reactions would be expected to decrease with increasing free energies of base stacking.

Positive Selection for Unstable SSR

SSR which have the potential of forming stable DNA secondary structures are rare in vertebrate genomes. Nevertheless, based on the relationships between DNA structure stabilities and genome-wide numbers of repeats, certain unstable sequences, such as the CTG·CAG triplet repeat, were found to occur more frequently than expected, suggesting that selection might have prevented their loss [41]. Analyses of the distributions of tri- and tetranucleotide repeats in intergenic and intragenic (coding DNA) regions in the human genome showed that >90% of repeat tracts were located in intergenic regions. However, the proportion of tracts in DNA coding regions were the highest (~30%) for the CTG·CAG and CGG·CCG trinucleotide repeats (≥30 bp). In addition, among tetranucleotide repeats most CG·CG-containing sequences were overrepresented (>10%) in coding regions at length ≥12 bp

(no, or very few, tracts ≥32 bp existed for such sequences). Because the repeats that were over-represented in coding DNA comprised sequences with strong propensities to fold into DNA secondary structures (hairpins and quadruplexes), their high numbers were suggestive of selective pressures acting to preserve their activity within gene's coding regions. Other analyses on the frequencies of triplet repeats within coding and non-coding sequences in 12 Drosophila species also revealed the strong over-representation of CTG·CAG repeats in coding DNA [48]. In summary, selective pressures appear to have played critical roles in maintaining certain intrinsically unstable SSR in the DNA coding regions of eukaryotic genomes.

Homo-Amino Acid Runs and Coding SSR

Within the past 6 yr several analyses were conducted on proteomic databases to address the question of the fractions of proteins containing homo-amino acid runs. In eukaryotes, between 13% (worm) and 27% (fly) (~20% in humans) of predicted proteins were found to contain at least one run of ≥5 identical amino acid residues [49]. In prokaryotes, these fractions were much smaller (~9%) [50], consistent with the higher proportions of SSR in the eukaryotic genomes than in prokaryotes [51].

Considering the distribution of SSR within genes (5′UTR, ORF, and 3′UTR), the fraction of human genes harboring SSR was 8.6% (2315 entries) of the ~27 000 currently annotated genes [41]. (The lower fraction of SSR within genes relative to that of homo-amino acid runs is not unexpected because homo-amino acid runs are generally encoded by mixed repeat sequences, such as tracts comprising CAG and CAA repeats encoding Q_n.) In addition, different types of SSR were found to localize within different gene regions, such that dinucleotide repeats were present almost exclusively (~90%) within the 3′UTR, whereas triplet repeats were mostly represented (60%) within the ORF. Therefore, dinucleotide repeats may have been recruited to be involved preferentially in gene expression regulation and/or mRNA stability/transport, whereas trinucleotide repeats would have served protein functions (Figure 2).

Functional Gene Categories

Enrichment analyses using either proteins containing homo-amino acid runs or SSR-containing genes, both in the human and other eukaryotic species, were concordant in revealing strong over-representation in cellular activities associated with transcription, regulation of cellular processes, signaling pathways such as WNT and MAPK and regulation of development, particularly in the nervous system and at synapses [41,49,50,52,53]. By contrast, homo-amino acid runs were comparatively rare in the orthologous prokaryotic genes, suggesting an association with the increased complexity of eukaryotic processes, particularly those related to intracellular signaling and cell–cell communication [50].

Analyses of the types of amino acids encoded by SSR indicated that polar amino acid residues predominate over hydrophobic residues and, within the polar amino acids, acidic residues occur more frequently than basic residues [49,50]. Glutamine (Q) was generally found to be the most abundant, followed by glutamic acid (E), alanine (A), serine (S), glycine (G), leucine (L), and proline (P), although their ranking varied, depending on the species. By contrast, aromatic [phenylalanine (F), tyrosine (Y), and tryptophan (W)] and other hydrophobic amino acid residues such as isoleucine (I), were rarely found [41,49,50,52]. In addition to encoding preferred homo-amino acid runs, the location of specific SSR codons varied within translated exons. Specifically, polyL, polyA, polyG, and polyS-encoding repeats were strongly overprepresented in the first exon, implying that some of these amino acid runs localized close to the N-terminus may be part of signal peptides [41,49,50]. On the other hand, SSR encoding polyQ and polyE runs were found preferentially in internal exons. Remarkably, CTG·CAG triplet repeats would encode preferentially polyQ runs in the ''CAG reading frame'' within the first exon and polyL runs in the ''CTG reading frame'' in internal exons [41]. Finally, polyE runs were found to be associated with proteins involved in DNA binding, whereas polyL and polyA runs were characteristic of transmembrane receptors and transcription factors, respectively [52]. These composite observations support the conclusion that specific SSR were recruited within genes of eukaryotic species (Figure 2) to endow their products with novel or enhanced functions relative to their prokaryotic protein counterparts.

Homo-Amino Acid Runs and Protein Disorder

The number of trinucleotide repeats genome-wide is greater in rodents than in the human species; however, their fraction within DNA coding regions is higher in the human lineage than in rodents, implying the action of purifying selection along the human lineage [52]. As stated, selection appears to have been particularly strong for the most unstable sequences, CTG·CAG and CGG·CCG [41]. Comparative studies indicate that these repeats, particularly CTG·CAG, have undergone strong bias towards expansion in genes such as *TBP*, *MEF2A*, and *SCA2* [41,54] along the human lineage, where they are found to encode polyQ runs. In fact, whereas both the CAG and CAA codons may be used to encode polyQ runs, CAG was found to be the predominant codon employed. In instances where conserved polyQ tracts encoded by both CAA and CAG repeats were traced back to eukaryotic species that arose earlier during evolution, such as the TATA-

box binding protein, it was the CAG repeat that displayed selective expansion [41]. These observations suggest that, during evolutionary time, genome-wide SSR instability was exploited as a means to modulate protein structure, perhaps to explore new domains of functional significance (Figure 2).

The group of proteins containing homo-amino acid runs consists of intracellular constituents that form large multipeptide complexes, such as ribosomes, the DNA replication apparatus, and the transcription and translation machineries, all of which involve extensive protein–protein and protein–DNA/RNA interactions. Experimental and theoretical analyses support the notion that homo-amino acids constitute disordered regions within the subunits of such macromolecular complexes, which are critical for protein–protein and/or protein–DNA interactions. These observations suggest a model for the role of homo-amino acids in protein function whereby in the disordered state, protein domains would provide a means for low affinity interactions aimed at identifying cognate partners. However, upon binding to specific target sites, high affinity interactions would be established through domain dynamics, which would afford induced fit between the interacting protein surfaces. Hence, disordered domains may act as modular interfaces for the interaction with multiple partners. Alternatively, disordered regions may function as hinges to permit protein domains to reach out and engage binding with distant recipient domains. Therefore, homo-amino acids would serve as an easy-to-fine-tune tool to engage in protein–protein and/or protein–DNA interactions, without perturbing neighboring constrained structural elements, essential for catalysis [55].

REPEATING DNA AS AN INTRINSIC DETERMINANT OF CANCER

Genome-Wide Analyses

The current census of genes whose mutations, mostly translocations, have been associated with cancer lists a total of 378 entries as of July 2008 (www.sanger.ac.uk), 1.4% of the currently ~27 000 annotated genes. A proteomic analysis of these genes (http://david.abcc.ncifcrf.gov) indicates an exceedingly strong enrichment (P-value 9.4×10^{-51}) for functions related to the regulation of cellular processes, similar to the functions noted for the SSR-containing genes (see above). Indeed, 95 of the 378 genes (25%) also contain at least one SSR within the coding DNA [[41] and Author's table at http://www.ibt.tamhsc.edu/labs/cgr/documents/table-cancer-micro.xls], a significantly high proportion ($P < 0.001$). These results extend previous analyses conducted on triplet repeat-containing coding DNA [56], as well as the distribution of homo-amino acid runs within proteins [49], and indicate that genes whose recurrent mutations lead to cancer generally contain higher fractions of repetitive DNA than genome average. Additional investigations showed that the classes of cancer-related and triplet repeat-containing primary transcripts are longer than average [56], a property which also characterizes regulatory genes highly expressed in the brain and containing long intronic R·Y tracts [26]. Hence, repetitive DNA, such as SSR and long R·Y tracts, are more likely to be found in genes that are longer than the genome-wide average [26].

Genomic Rearrangements and Repetitive DNA

As elaborated earlier, a frequent mechanism of gross chromosomal rearrangements consists in the fusion of DNA broken ends that share either limited or no sequence homology by the non-homologous end joining (NHEJ) pathway. The DNA sequence features found in the vicinity of such breakpoints have been analyzed both through genetic and computational approaches in an attempt to identify recurrent motifs potentially involved in these mutational processes [16,17,57]. Non-B DNA-forming sequences, such as inverted and direct repeats, were found to be significantly overrepresented close to, or directly abutting, breakpoints, supporting the hypothesis that DNA secondary structures might have triggered the initial breaks and/or served a function during the subsequent repair reaction. In support of this hypothesis, linker-mediated PCR analyses of mutations, such as large deletions and complex rearrangements, triggered by an H-DNA (triplex DNA)-forming sequence at the *c-myc* locus in COS-7 cells, indicated the formation of DSB near the H-DNA-forming sequence [16]. Because the overall frequencies of these and other mutations in plasmids transfected in COS-7 were ~20-fold higher in the presence of *c-myc* H-DNA than in the presence of nt substitutions inhibiting the H-DNA structures, a role for non-B DNA in inducing DSB leading to high mutation rates was revealed [16]. These results are significant given the involvement of *c-myc* gene translocations in tumors such as the Burkitt's lymphoma [16]. Indeed, several non-B DNA conformations, including hairpin/cruciforms, Z-DNA, and H-DNA have been shown to act as naturally occurring mutagens through their interactions with cellular pathways, including replication, repair, and transcription [3,19].

Particularly intriguing was the observation that sequence complexity (whose value correlates inversely with the propensity of forming DNA secondary structures) was lower at the sites of chromosomal fusions after the rearrangements took place than in the original sequences, suggesting that during the interaction between the two broken ends and their fusion, a DNA secondary structure was formed to

mediate, favor, or assist the repair process. Hence, these genome-wide investigations support the notion that repetitive DNA may trigger occasional chromosomal breaks, which are then repaired resulting in gross rearrangements. Cancer-related genes would be at particularly high risks of incurring DSB because of the extended genomic intervals they occupy and the high fraction of repetitive DNA they harbor.

SSR Polymorphisms and Cancer Susceptibility

A large number of studies support the notion that microsatellite length polymorphism in non-coding DNA may confer either susceptibility to, or protection from, complex diseases, such as diabetes, obesity, hyperandrogeny, osteoporosis, arthritis, schizophrenia, and cancer [reviewed in [41] and Figure 2]. The mechanisms underlying SSR-dependent disease susceptibility are distinct from the gross chromosomal rearrangements outlined above, which may disrupt gene function, and consist instead of subtle variations in protein levels, which are specific to certain individuals or to specific ethnic groups. As an example, herein we review cancer susceptibility associated with SSR polymorphisms in the *ERBB1* gene, encoding the epidermal growth factor receptor (EGFR).

EGFR participates in at least three signaling pathways: first, the phosphatidylinositol 3-kinase-(PIP3K)-dependent phosphorylation of AKT and suppression of apoptosis through BAD; second, SOS-1/GRB2 activation of RAS and subsequent cell cycle progression through the ERK and JNK1 cascade and, third, phospholipase-C-gamma1 phosphorylation and actin reorganization via the PIP2 pathway [58]. In the heart, EGFR also leads to cardiac hypertrophy through the MAPK pathway and reactive oxygen species (ROS), by activating RAS and RHO. EGFR is activated by a variety of ligands, including EGF, heparin-binding EGF-like growth factor (HB-EGF), TGF alpha, betacellulin, amphiregulin, epiregulin, epigen, neuregulins and trefoil factor family, or TFF [reviewed in [59]]. The receptor undergoes dimerization upon ligand binding and transactivation, following cross-phosphorylation of the intracellular tyrosine kinase domain. EGFR is frequently overexpressed in many malignancies, including non-small cell lung and head and neck cancers, and overexpression is generally associated with poor prognosis [58,60,61]. Transcription of the *ERBB1* gene starts at multiple sites within a G+C-rich region that lacks consensus TATA- and CAAT-box sequences, and it is regulated by three enhancers, two upstream of exon 1 and the third in intron 1. A variable number of CA dinucleotide repeats (usually referred to as CA-SSR1), ranging from 14 to 21 units, are located upstream of enhancer 3. Transcriptional activity decreases with increasing number of CA repeats, such that up to 10-fold differences were noted in vitro between the shortest and the longest repeat lengths. Allelic frequencies of the CA-SSR1 repeat vary among populations, with the shorter alleles occurring more frequently in Europeans than in East Asians. Hence, the levels of EGFR are on average lower in the latter population than in the former, both in normal and in cancer tissues.

Allele polymorphisms associated with high EGFR expression are a risk factor for cancer [58,60–66]. In a case–control study involving lung cancer patients, individuals with CA-SSR1 length sums of $\leq$36 repeats were significantly more represented among cases than controls, whereas longer allele length sums ($>$36 repeats) were more frequent among controls [62]. In a separate study, among patients suffering from non-small cell lung cancer, subjects with allele CA-repeat sums $\leq$35 had a median overall survival of 29.2 mo, whereas subjects with a CA-allele sums $\geq$35 had a median overall survival of 41.0 mo [63]. Studies conducted on East Asian populations indicated that EGFR overexpression in cancer cells, such as in lung cancer, results from a combination of CA allele length and other mutations within the *ERBB1* gene. For example, in Japanese patients with lung cancer, shorter alleles were significantly associated with in-frame (small 9–24 bp) deletions in exon 19 but not with single amino acid substitutions in exons 18 or 21 [64]. Intrinsic mRNA expression in non-cancerous tissues and doubling time measurements in human bronchial epithelial (HBE) cells in these individuals were higher when short CA alleles and exon 19 mutations were found to be present in combination [64]. An additional form of *ERBB1* mutation, also associated with increased gene expression, consists of selective amplifications of intron-1 sequences, including the CA-SSR1 and flanking sequences [58,67]. A comparative study, whereby polymorphisms and *ERBB1* amplification were investigated in non-small cell lung cancer patients from various ethnic populations, found that in individuals of Japanese descent *ERBB1* amplification occurred significantly more frequently than in patients of Caucasian origins [60]. Because the CA-SRR1 length distributions are generally longer (and therefore *ERBB1* gene expression is lower) in the former population than in the latter [58,60], these composite data strongly support the conclusion that cells exploit *ERBB1* mutations to achieve EGFR overexpression, thereby supporting cancer development [60]. However, whether or not long CA-SRR1 repeats are directly involved in triggering additional *ERBB1* gene mutations remains to be determined.

The types of mutations that lead to EGFR overexpression also appear to determine the extent of responsiveness to anticancer drugs, such as the tyrosine kinase inhibitors, gefitinib and erlotinib. Specifically, response rates were found to be higher in patients harboring short CA-SRR1 lengths than in patients with longer repeat lengths [65,66].

Cancer Susceptibility and ROS Signaling

In addition to *ERBB1*, microsatellite polymorphisms in the non-coding portions of genes associated with susceptibility to cancer have been reported between a CA repeat of *IGF1* (encoding the insulin-like growth factor I, IGF-1) and non-polyposis colorectal cancer and breast cancer [68–70], a TA repeat in *ESR1* (encoding the estrogen receptor alpha) and prostate [71] and breast cancers [72,73], a GT repeat in *HMOX1* (encoding heme oxigenase 1, HO-1) and lung adenocarcinoma [74,75], gastric adenocarcinoma and lymphovascular tumor invasion [75] and postmenopausal breast cancer [76]. All these polymorphisms are present in the promoter regions. In addition, a GT polymorphism in intron 13 of the *HIF1A* gene (encoding hypoxia inducible factor-1 alpha, HIF-1 alpha) with increased protein expression has been reported in non-small cell lung cancer cells [77].

Analyses (http://cgap.nci.nih.gov) of the pathways in which the above genes are involved revealed the existence of relationships among the EGFR, IGF-1, HO-1, and HIF-1 alpha proteins, their genes and the roles of ROS in signal transduction. ROS species are often found to be elevated in tumors. Specifically, HO-1 is known to catalyze the NADPH and P450 reductase-dependent oxidation of heme to biliverdin IX alpha, free iron, and CO [78,79]. CO stimulates the rapid production of ROS, that is, superoxide anion ($O_2^{\cdot -}$), hydrogen peroxide (H_2O_2), and hydroxyl radicals (OH^-) thereby triggering pleiotropic effects, including the modulation of guanylyl cyclase, MAPK, the PIP3K/AKT pathway, STATs signaling, PPAR-gamma and HIF-1 alpha activation [80]. In addition, EGFR directly impinges on *HMOX1* (HO-1) transactivation through the JAK/STATs pathway, thereby enhancing cell growth, survival, and differentiation, which are all steps involved in malignant transformation. The *ERBB1* (EGFR) gene itself is upregulated by ROS through IFG-1 [81]. Indeed, IFG-1 increases ROS production and, in complex with its receptor (IGFR, which is activated by stress, membrane depolarization, integrins, cytokine receptors, etc.) transactivates *ERBB1* through Src, whose phosphorylation is also ROS-dependent [81]. Finally, ROS amplifies the EGFR-dependent activation of the antiapoptotic PIP3K/AKT pathway [81,82]. Hence, positive feed-back loops have been established among these genes, which share SSR- and ROS-dependent regulatory mechanisms. Other disease predispositions, such as inflammation, asthma, Alzheimer's disease, cystic fibrosis, allograph rejection, and restenosis have been noted, in addition to cancer, for ROS-dependent and microsatellite-containing genes, including *IL10* [83–85], *CCL3* (macrophage inflammatory protein-1 alpha, MIP-1 alpha) [86], *NOS1* (nitric oxide synthase 1) [87], *SLC11A1* (solute carrier MA1 protein) [88], *MMP9* (matrix metalloproteinase 9) [53,89,90], *NOX1* (gp91phox subunit of NADPH oxidase) [91], and *GCLC* (catalytic subunit of gamma-glutamylcysteine ligase) [92]. Hence, it appears that a large regulatory network exists in the cell which links microsatellite-dependent gene expression regulation (Figure 2) to ROS-dependent signaling. These interactions predict the composite effect of repeat polymorphisms in multiple ROS pathway-related genes on cancer susceptibility, as indeed recently found for selected genes [76].

Repeating DNA Motifs as Effectors of Transcriptional Regulation

DNA amplification at specific loci leading to increased gene expression is a hallmark of cancer and is generally associated with advanced disease [93]. A critical step in DNA amplification is the occurrence of DSB followed by the formation of large newly synthesized regions with inverted symmetry. Small inverted repeats play a crucial role in this process by determining the boundaries of DNA amplification. Current models suggest that following DSB formation, a 5′-to-3′resection exposes single stranded ends which fold-back upon themselves (intrastrand annealing) into small hairpin structures at sites of inverted repeat symmetry. These structures would be crucial in priming DNA synthesis, thereby providing nucleation sites for copying extensive chromosomal regions. The process may then reiterate to yield gene amplification [93].

Small inverted repeats, which may form cruciforms, and other non-B DNA-forming structures are also often found at rearrangement breakpoints [17], suggesting that they may provide a substrate for the initial DSB, as elaborated earlier. Myc binding sites have also been observed in association with amplified regions [94], suggesting that Myc overexpression may favor gene amplification by relaxing the restrictions upon aberrant DNA synthesis.

Recently, a group of primate-specific transposable elements (Made1) comprising miniature inverted repeats (MITE) was found to encode a family of microRNA genes (hsa-mir-548) active in downregulating genes related to cell proliferation, mitosis, and apoptosis [95], all processes associated with cancer. Genes with putative hsa-mir-548 target sites were found to be downregulated in cancer cells, particularly in colorectal cancer [95], when compared to normal tissue, suggesting MITE overexpression in cancer. Therefore, inverted repeats are intimately associated with cancer in at least three ways: first, by providing a substrate for DSB in their non-B cruciform conformation; second, by nucleating DNA synthesis leading to gene amplification, and third, by encoding cell-proliferation-specific regulatory miRNAs. Finally, sites in which stable DNA secondary structures are predicted to form in the non-templating strand during RNA synthesis

have also been associated with high mutation rates during somatic hypermutation in the course of antibody production [96]. This mechanism may also contribute to increasing DSB formation within transcribed units.

By contrast, the mechanisms by which polymorphic microsatellite repeats alter gene transcription remain poorly characterized. One notable exception is the ETS gene family of transcription factors, which are frequently dysregulated by chromosomal translocations in cancer. The most recurrent translocation in Ewing's sarcoma encodes ESW/FLI, an oncogenic transcription factor which, upon translocation, acquires high affinity binding properties for tandem GGAA repeats [97]. Commonly present in promoter regions, GGAA repeats then function as necessary and sufficient response elements for oncogenic transformation. In addition, the strength of ESW/FLI transactivation increases with the number of GGAA repeats. Therefore, it is conceivable that length polymorphisms in key genes may contribute to cancer susceptibility and/or development. A second study, unrelated to cancer, showed an oscillatory behavior in the pattern of gene expression levels as a function of TA repeat numbers in the *NOX1* gene promoter [91]. Repeat numbers equaling integral helical turns of DNA induced transcription maximally, whereas repeat numbers providing half helical turns were least effective. These results strongly support an effect of DNA looping upon gene transcription. Hence, repeat polymorphism (Figure 2) is likely to modulate gene transcription through a variety of mechanisms.

PROSPECTS FOR THE FUTURE

Non-B DNA structures have profound effects on several biological processes, including the regulation of gene function, evolution, and human disease. Despite tremendous progress, this field is just on the threshold of revealing its complexities. These developments have generated a number of important questions to be addressed in future studies, including:

- The establishment of experimental models aimed at elucidating the exact molecular mechanisms of selected repeat expansion in neurological disorders.
- The identification of the enzymatic complex(es), their in vitro reconstitution and kinetic characterization of the process(es) of DSB repair that give rise to chromosomal rearrangements.
- The roles of SSR in gene regulation and their origins within selected gene classes.
- The mechanisms by which SSR-encoded homoamino acid runs participate in protein structure and function.
- The relationships between SSR and the length-dependence regulation of gene expression.
- The execution of association studies involving a multigene approach to susceptibility to disease by SSR- and other polymorphism-containing genes.
- Furthering the relationships of SSR-containing genes along common signal transduction pathways, such as ROS.

Studies at the interfaces of biochemistry, genomics, bioinformatics, and human genetics will be most informative. Also, future therapeutic strategies may be explored by modulating the stability of the non-B DNA structures, thereby influencing the frequencies of the genetic events described above.

ACKNOWLEDGMENTS

We thank the NIH (ES11347), the Robert A. Welch Foundation, Friedreich's Ataxia Research Alliance, and Seek a Miracle for support. We also express appreciation to J.E. Larson for 40 yr of research on non-B DNA structures. All materials from the R.D. Wells laboratory were transferred to Dr. Sergei Mirkin (Tufts University, Department of Biology, Barnum 101B, 163 Packard Ave., Medford, MA 02155. E-mail: sergei.mirkin@tufts.edu). The authors convey their apologies to colleagues whose fine work could not be cited due to a journal restriction on the number of references and the length of the review.

REFERENCES

1. Heidelberger C. Chemical carcinogenesis. Annu Rev Biochem 1975;44:79–121.
2. Wells RD. Non-B DNA conformations, mutagenesis and disease. Trends Biochem Sci 2007;32:271–278.
3. Wang G, Vasquez KM. Non-B DNA structure-induced genetic instability. Mutat Res 2006;598:103–119.
4. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396:643–649.
5. Wells RD, Ashizawa T, editors. Genetic instabilities and neurological diseases. 2nd edition. San Diego: Elsevier/Academic Press; 2006. 766 p.
6. Wells RD, Dere R, Hebert ML, Napierala M, Son LS. Advances in mechanisms of genetic instability related to hereditary neurological diseases. Nucleic Acids Res 2005;33:3785–3798.
7. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: Mechanisms and common principles. Nat Rev Genet 2005;6:743–755.
8. Lin Y, Hubert L, Jr., Wilson JH. Transcription destabilizes triplet repeats. Mol Carcinog this issue.
9. Duker NJ. Chromosome breakage syndromes and cancer. Am J Med Genet 2002;115:125–129.
10. Lupski JR, Stankiewicz P, editors. Genomic disorders: The genomic basis of disease. Totowa, NJ: Humana Press; 2006. 426 p.
11. Wells RD. *Reflections*: Discovery of the role of non-B DNA structures in mutagenesis and human diseases. J Biol Chem (in press).
12. Sinden RR. DNA structure and function. San Diego, California: Academic Press; 1994.
13. Bowater RP, Wells RD. The intrinsically unstable life of DNA triplet repeats associated with human hereditary disorders. Progr Nucleic Acid Res Mol Biol 2000;66:159–202.

14. Wells RD. DNA triplexes and Friedreich ataxia. FASEB J 2008;22:1625–1634.
15. Jakupciak JP, Wells RD. Genetic instabilities of triplet repeat sequences by recombination. Life 2000;50:355–359.
16. Wang G, Vasquez KM. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. Proc Natl Acad Sci USA 2004;101:13448–13453.
17. Bacolla A, Jaworski A, Larson JE, et al. Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci USA 2004;101:14162–14167.
18. Raghavan SC, Swanson PC, Wu X, Hsieh CL, Lieber MR. A non-B-DNA structure at the *Bcl*-2 major breakpoint region is cleaved by the RAG complex. Nature 2004;428:88–93.
19. Wang G, Christensen LA, Vasquez KM. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. Proc Natl Acad Sci USA 2006;103:2677–2682.
20. Bacolla A, Wojciechowska M, Kosmider B, Larson JE, Wells RD. The involvement of non-B DNA structures in gross chromosomal rearrangements. DNA Repair 2006;5:1161–1170.
21. Bacolla A, Wells RD. Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem 2004;279:47411–47414.
22. Bacolla A, Wells RD. Non-B DNA and chromosomal rearrangements. In: Lupski JR, Stankiewicz P, editors. Genomic disorders: The genomic basis of disease. Totowa, NJ: Humana Press; 2006. p 89–99.
23. Sharan SK, Kuznetsov SG. Resolving *RAD51C* function in late stages of homologous recombination. Cell Div 2007;2:15.
24. Inagaki K, Ma C, Storm TA, Kay MA, Nakai H. The role of DNA-PKcs and artemis in opening viral DNA hairpin termini in various tissues in mice. J Virol 2007;81:11304–11321.
25. Ma Y, Schwarz K, Lieber MR. The artemis: DNA-PKcs endonuclease cleaves DNA loops, flaps, and gaps. DNA Repair 2005;4:845–851.
26. Bacolla A, Collins JR, Gold B, et al. Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. Nucleic Acids Res 2006;34:2663–2675.
27. Du Z, Zhao Y, Li N. Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. Genome Res 2008;18:233–241.
28. Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res 2007;35:406–413.
29. Wojciechowska M, Napierala M, Larson JE, Wells RD. Non-B DNA conformations formed by long repeating tracts of DM1, DM2 and FRDA genes, not the sequences per se, promote mutagenesis in flanking regions. J Biol Chem 2006;281:24531–24543.
30. Samadashwily GM, Raca G, Mirkin SM. Trinucleotide repeats affect DNA replication *in vivo*. Nat Genet 1997;17:298–304.
31. Raghavan SC, Chastain P, Lee JS, et al. Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation. J Biol Chem 2005;280:22749–22760.
32. Bacolla A, Jaworski A, Connors TD, Wells RD. *PKD1* unusual DNA conformations are recognized by nucleotide excision repair. J Biol Chem 2001;276:18597–18604.
33. Freudenreich CH, Stavenhagen JB, Zakian VA. Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol Cell Biol 1997;17:2090–2098.
34. Gotter AL, Shaikh TH, Budarf ML, Rhodes CH, Emanuel BS. A palindrome-mediated mechanism distinguishes translocations involving LCR-B of chromosome 22q11.2. Hum Mol Genet 2004;13:103–115.
35. Kato T, Inagaki H, Yamada K, et al. Genetic variation affects *de novo* translocation frequency. Science (New York, NY) 2006;311:971.
36. Rooney SM, Moore PD. Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. Proc Natl Acad Sci USA 1995;92:2141–2144.
37. Mulvihill DJ, Nichol Edamura K, Hagerman KA, Pearson CE, Wang YH. Effect of CAT or AGG interruptions and CpG methylation on nucleosome assembly upon trinucleotide repeats on spinocerebellar ataxia, type 1 and fragile X syndrome. J Biol Chem 2005;280:4498–4503.
38. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
39. Fondon JW, III, Garner HR. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA 2004;101:18058–18063.
40. Riancho JA, Zarrabeitia MT, Valero C, et al. Aromatase gene and osteoporosis: Relationship of ten polymorphic loci with bone mineral density. Bone 2005;36:917–925.
41. Bacolla A, Larson JE, Collins JR, et al. Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res 2008;18:1545–1553.
42. Clark RM, Bhaskar SS, Miyahara M, Dalgliesh GL, Bidichandani SI. Expansion of GAA trinucleotide repeats in mammals. Genomics 2006;87:57–67.
43. Xu L, Chow J, Bonacum J, et al. Microsatellite instability at AAAG repeat sequences in respiratory tract cancers. Int J Cancer 2001;91:200–204.
44. Danaee H, Nelson HH, Karagas MR, et al. Microsatellite instability at tetranucleotide repeats in skin and bladder cancer. Oncogene 2002;21:4894–4899.
45. Mirkin SM. Expandable DNA repeats and human disease. Nature 2007;447:932–940.
46. Orr HT, Zoghbi HY. Trinucleotide repeat disorders. Annu Rev Neurosci 2007;30:575–621.
47. Yang W. Poor base stacking at DNA lesions may initiate recognition by many repair proteins. DNA Repair 2006;5:654–666.
48. Huntley MA, Clark AG. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. Mol Biol Evol 2007;24:2598–2609.
49. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci USA 2002;99:333–338.
50. Faux NG, Bottomley SP, Lesk AM, et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 2005;15:537–551.
51. Subramanian S, Madgula VM, George R, et al. MRD: A microsatellite repeats database for prokaryotic and eukaryotic genomes. Genome Biol 2002;3: Preprint 0011.
52. Alba MM, Guigo R. Comparative analysis of amino acid repeats in rodents and humans. Genome Res 2004;14:549–554.
53. Zalba G, Fortuno A, Orbe J, et al. Phagocytic NADPH oxidase-dependent superoxide production stimulates matrix metalloproteinase-9: Implications for human atherosclerosis. Arterioscler Thromb Vasc Biol 2007;27:587–593.
54. Yu F, Sabeti PC, Hardenbol P, et al. Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. PLoS Genet 2005;1:e41.
55. Mularoni L, Veitia RA, Alba MM. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics 2007;89:316–325.
56. Haberman Y, Amariglio N, Rechavi G, Eisenberg E. Trinucleotide repeats are prevalent among cancer-related genes. Trends Genet 2008;24:14–18.
57. Chuzhanova N, Abeysinghe SS, Krawczak M, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. Hum Mutat 2003;22:245–251.
58. Brandt B, Meyer-Staeckling S, Schmidt H, Agelopoulos K, Buerger H. Mechanisms of egfr gene transcription modulation: Relationship to cancer risk and therapy response. Clin Cancer Res 2006;12:7252–7260.

59. Holgate ST, Lackie P, Wilson S, Roche W, Davies D. Bronchial epithelium as a key regulator of airway allergen sensitization and remodeling in asthma. Am J Respir Crit Care Med 2000;162:S113–S117.
60. Nomura M, Shigematsu H, Li L, et al. Polymorphisms, mutations, and amplification of the EGFR gene in non-small cell lung cancers. PLoS Med 2007;4:125.
61. Etienne-Grimaldi MC, Pereira S, Magne N, et al. Analysis of the dinucleotide repeat polymorphism in the epidermal growth factor receptor (EGFR) gene in head and neck cancer patients. Ann Oncol 2005;16:934–941.
62. Zhang W, Weissfeld JL, Romkes M, Land SR, Grandis JR, Siegfried JM. Association of the EGFR intron 1 CA repeat length with lung cancer risk. Mol Carcinog 2007;46:372–380.
63. Dubey S, Stephenson P, Levy DE, et al. EGFR dinucleotide repeat polymorphism as a prognostic indicator in non-small cell lung cancer. J Thorac Oncol 2006;1:406–412.
64. Sueoka-Aragane N, Imai K, Komiya K, et al. Exon 19 of EGFR mutation in relation to the CA-repeat polymorphism in intron 1. Cancer Sci 2008;99:1180–1187.
65. Ichihara S, Toyooka S, Fujiwara Y, et al. The impact of epidermal growth factor receptor gene status on gefitinib-treated Japanese patients with non-small-cell lung cancer. Int J Cancer 2007;120:1239–1247.
66. Han SW, Jeon YK, Lee KH, et al. Intron 1 CA dinucleotide repeat polymorphism and mutations of epidermal growth factor receptor and gefitinib responsiveness in non-small-cell lung cancer. Pharmacogenet Genomics 2007;17:313–319.
67. Buerger H, Packeisen J, Boecker A, et al. Allelic length of a CA dinucleotide repeat in the egfr gene correlates with the frequency of amplifications of this sequence—First results of an inter-ethnic breast cancer study. J Pathol 2004;203:545–550.
68. Reeves SG, Rich D, Meldrum CJ, et al. IGF1 is a modifier of disease risk in hereditary non-polyposis colorectal cancer. Int J Cancer 2008;123:1339–1343.
69. Cleveland RJ, Gammon MD, Edmiston SN, et al. IGF1 CA repeat polymorphisms, lifestyle factors and breast cancer risk in the Long Island Breast Cancer Study Project. Carcinogenesis 2006;27:758–765.
70. Zecevic M, Amos CI, Gu X, et al. IGF1 gene polymorphism and risk for hereditary nonpolyposis colorectal cancer. J Natl Cancer Inst 2006;98:139–143.
71. McIntyre MH, Kantoff PW, Stampfer MJ, et al. Prostate cancer risk and ESR1 TA, ESR2 CA repeat polymorphisms. Cancer Epidemiol Biomarkers Prev 2007;16:2233–2236.
72. Boyapati SM, Shu XO, Ruan ZX, et al. Polymorphisms in ER-alpha gene interact with estrogen receptor status in breast cancer survival. Clin Cancer Res 2005;11:1093–1098.
73. Cai Q, Gao YT, Wen W, et al. Association of breast cancer risk with a GT dinucleotide repeat polymorphism upstream of the estrogen receptor-alpha gene. Cancer Res 2003;63:5727–5730.
74. Kikuchi A, Yamaya M, Suzuki S, et al. Association of susceptibility to the development of lung adenocarcinoma with the heme oxygenase-1 gene promoter polymorphism. Hum Genet 2005;116:354–360.
75. Lo SS, Lin SC, Wu CW, et al. Heme oxygenase-1 gene promoter polymorphism is associated with risk of gastric adenocarcinoma and lymphovascular tumor invasion. Ann Surg Oncol 2007;14:2250–2256.
76. Hong CC, Ambrosone CB, Ahn J, et al. Genetic variability in iron-related oxidative stress pathways (Nrf2, NQ01, NOS3, and HO-1), iron intake, and risk of postmenopausal breast cancer. Cancer Epidemiol Biomarkers Prev 2007;16:1784–1794.
77. Koukourakis MI, Papazoglou D, Giatromanolaki A, et al. C2028T polymorphism in exon 12 and dinucleotide repeat polymorphism in intron 13 of the HIF-1alpha gene define HIF-1alpha protein expression in non-small cell lung cancer. Lung Cancer (Amsterdam, Netherlands) 2006;53:257–262.
78. Wang J, Evans JP, Ogura H, La Mar GN, Ortiz de Montellano PR. Alteration of the regiospecificity of human heme oxygenase-1 by unseating of the heme but not disruption of the distal hydrogen bonding network. Biochemistry 2006;45:61–73.
79. Liu Y, Ortiz de Montellano PR. Reaction intermediates and single turnover rate constants for the oxidation of heme by human heme oxygenase-1. J Biol Chem 2000;275:5297–5307.
80. Bilban M, Haschemi A, Wegiel B, Chin BY, Wagner O, Otterbein LE. Heme oxygenase and carbon monoxide initiate homeostatic signaling. J Mol Med 2008;86:267–279.
81. Meng D, Shi X, Jiang BH, Fang J. Insulin-like growth factor-I (IGF-I) induces epidermal growth factor receptor transactivation and cell proliferation through reactive oxygen species. Free Radic Biol Med 2007;42:1651–1660.
82. Giannoni E, Buricchi F, Grimaldi G, et al. Redox regulation of anoikis: Reactive oxygen species as essential mediators of cell survival. Cell Death Differ 2008;15:867–878.
83. Astermark J, Oldenburg J, Pavlova A, Berntorp E, Lefvert AK. Polymorphisms in the IL10 but not in the IL1beta and IL4 genes are associated with inhibitor development in patients with hemophilia A. Blood 2006;107:3167–3172.
84. Cochery-Nouvellon E, Vitry F, Cornillet-Lefebvre P, Hezard N, Gillot L, Nguyen P. Interleukin-10 promoter polymorphism and venous thrombosis: A case-control study. Thromb Haemost 2006;96:24–28.
85. Murray PJ. Understanding and exploiting the endogenous interleukin-10/STAT3-mediated anti-inflammatory response. Curr Opin Pharmacol 2006;6:379–386.
86. Li K, Dai D, Yao L, et al. Association between the macrophage inflammatory protein-l alpha gene polymorphism and Alzheimer's disease in the Chinese population. Neurosci Lett 2008;433:125–128.
87. Texereau J, Marullo S, Hubert D, et al. Nitric oxide synthase 1 as a potential modifier gene of decline in lung function in patients with cystic fibrosis. Thorax 2004;59:156–158.
88. Bayele HK, Peyssonnaux C, Giatromanolaki A, et al. HIF-1 regulates heritable variation and allele expression phenotypes of the macrophage immune response gene SLC11A1 from a Z-DNA forming microsatellite. Blood 2007;110:3039–3048.
89. Kodama D, Saito M, Matsumoto W, et al. Longer dinucleotide repeat polymorphism in matrix metalloproteinase-9 (MMP-9) gene promoter which correlates with higher HTLV-I Tax mediated transcriptional activity influences the risk of HTLV-I associated myelopathy/tropical spastic paraparesis (HAM/TSP). J Neuroimmunol 2004;156:188–194.
90. Shin MH, Moon YJ, Seo JE, Lee Y, Kim KH, Chung JH. Reactive oxygen species produced by NADPH oxidase, xanthine oxidase, and mitochondrial electron transport system mediate heat shock-induced MMP-1 and MMP-9 expression. Free Radic Biol Med 2008;44:635–645.
91. Uhlemann AC, Szlezak NA, Vonthein R, et al. DNA phasing by TA dinucleotide microsatellite length determines in vitro and in vivo expression of the gp91phox subunit of NADPH oxidase and mediates protection against severe malaria. J Infect Dis 2004;189:2227–2234.
92. Nichenametla SN, Ellison I, Calcagnotto A, Lazarus P, Muscat JE, Richie JP, Jr. Functional significance of the GAG trinucleotide-repeat polymorphism in the gene for the catalytic subunit of gamma-glutamylcysteine ligase. Free Radic Biol Med 2008;45:645–650.
93. Tanaka H, Cao Y, Bergstrom DA, Kooperberg C, Tapscott SJ, Yao MC. Intrastrand annealing leads to the formation of a large DNA palindrome and determines the boundaries of

genomic amplification in human cancer. Mol Cell Biol 2007; 27:1993–2002.
94. Neiman PE, Elsaesser K, Loring G, Kimmel R. Myc oncogene-induced genomic instability: DNA palindromes in bursal lymphomagenesis. PLoS Genet 2008;4:e1000132.
95. Piriyapongsa J, Jordan IK. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE 2007;2:e203.
96. Wright BE, Schmidt KH, Davis N, Hunt AT, Minnick MF, II. Correlations between secondary structure stability and mutation frequency during somatic hypermutation. Mol Immunol 2008;45:3600–3608.
97. Gangwal K, Sankar S, Hollenhorst PC, et al. Microsatellites as EWS/FLI response elements in Ewing's sarcoma. Proc Natl Acad Sci USA 2008;105:10149–10154.