

Chapter 1

Site-Specific Self-Catalyzed DNA Depurination, the Basis of a Spontaneous Mutagenic Mechanism of Wide Evolutionary Significance

Jacques R. Fresco, Olga Amosova, Peter Wei, Juan R. Alvarez-Dominguez,
Damian Glumcher, and Rafael Torres

Abstract This chapter focuses on the nature of site-specific self-catalyzed DNA depurination as a spontaneous mechanism inherent in the chemical structure and dynamics of DNA that has contributed to evolutionary change. It describes the essential molecular features of the mechanism, the short consensus sequence elements that form the catalytic intermediate, the basics of the reactions that lead to the creation of apurinic sites, and the means by which those sites give rise to substitution and short deletion mutations. The consensus sequences are widely distributed in double-stranded genomes across the phyla at high frequency that increases up the phylogenetic tree. In the human genome, they constitute $>2 \times 10^6$ potential mutagenic sites, non-randomly scattered among very many genes, some containing multiple sites. Examples are presented of genes in which the mutations coincide with their self-depurination consensus sequences, the most striking being those in the β -globin gene that are responsible for six anemias and two β -thalassemias. Those of the olfactory receptor genes and the hypervariable regions of the immunoglobulin genes are shown to have utilized the mechanism to evolve their high degree of diversity and/or to develop their contemporaneous diversity for their present function.

1.1 Introduction

Spontaneous mutations are among the primary engines of evolutionary change. Until now, the major mode of their occurrence has been thought to be a consequence of errors in DNA replication, resulting in substitution and frameshift mutations. In reality, such mutational errors are not due to the enzymatic process having gone awry. Rather, substitution mutations are the consequences of intrinsic

J.R. Fresco • O. Amosova • P. Wei • J.R. Alvarez-Dominguez • D. Glumcher • R. Torres
Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
e-mail: jrfresco@princeton.edu

chemical equilibria associated with the tautomers of the bases and the steric isomers of the nucleotide precursors of DNA replication (Topal and Fresco 1976). Frame-shift mutations, on the other hand, arise as a consequence of template or growing strand DNA misfoldings resulting in sequences that misdirect the polymerase, which is possible because of the flexibility of the nucleic acid backbone (Fresco and Alberts 1960).

Recently, we discovered a very different source of spontaneous mutations that is due to a self-catalyzed, site-specific DNA depurination mechanism inherent in a DNA sequence element 14–16 base pairs long (Amosova et al. 2006). The mechanism is mediated by this stem-loop-forming sequence that is widely present in all the more than 100 double-stranded genomes we have examined from Archaea to Homo sapiens, occurring within many lower forms with a frequency of every ~5,000 base pairs for self-depurination of G-residues, and gradually rising up the phylogenetic tree to every ~3,000 base pairs for self-depurination of those pairs in man, i.e., $\sim 1.25 \times 10^6$ such sites.

Self-depurination has the potential to occur wherever the consensus sequences specific for self-catalysis of depurination of G or of A residues (see below and Fig. 1.1) are present, and so give rise to apurinic sites in the DNA backbone. Those sites can, in turn, result in point mutations as a consequence of their highly error-prone repair (Boiteux and Guillet 2004; Chakravarti et al. 2000; Korolev 2005; Simonelli et al. 2005). On the one hand, such repair can cause a substitution mutation. Alternatively, if the apurinic site gives rise to a backbone break by either

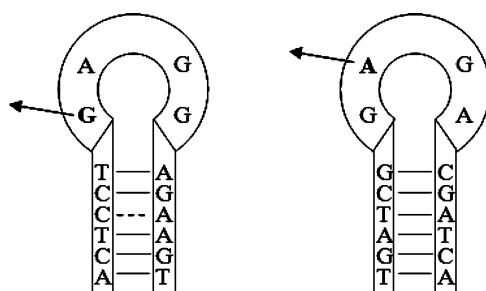


Fig. 1.1 Schematic diagrams of the stem-loop structure of the catalytic intermediates for site-specific self-catalyzed depurination in DNA of the indicated G-residue (*left*) and A-residue (*right*). In each case, the self-depurinating stem-loop is one side of a cruciform extruded from duplex DNA. The complementary inverted repeat sequence forms the other stem-loop of the cruciform that is, however, incapable of self-depurination. Certain sequence elements of the self-depurinating stem-loops are required for self-depurination to occur: the residues shown for each tetra-loop; particular complementary base pairs at the base of the loop, T•A or G•C on the left, and T•A, G•C or A•T on the right. These should be followed by any sequence of four or more additional complementary base pairs that can generally tolerate one mismatched pair (as shown on the left) or a single extrahelical residue. Thus, the specificity for self-depurination lies in the loop sequence and first base pair. The helical stem otherwise functions to maintain some strained loop structure required for the self-catalysis, and by its stability, to extend the lifetime of the intermediate and so enhance the kinetics of the self-catalysis

the action of the widely occurring enzyme apurinic endonuclease (Korolev 2005) or by way of the well-known spontaneous β -elimination reaction (Lhomme et al. 1999; Sugiyama et al. 1994), the strand break site results in a frayed end accessible to exonuclease attack, particularly if the first base pair after the break is A•T, and so leads to a short deletion mutation.

In this report, some essential features of the underlying self-depurination mechanism are first described. We then proceed to issues of biological relevance, including indications that the mechanism has played a role in the evolution of some biological phenomena. Because the consensus sequence for self-depurination of G-residues was discovered several years ago, and that for A-residues only very recently, our data is much more extensive for the former. Nevertheless, it will become apparent that both mechanisms are very similar, and appear to have played comparable mutagenic roles in several biological processes, including those of molecular evolution.

1.2 Essential Features of DNA Self-Catalyzed Depurination

Self-catalyzed depurination is a remarkably site-specific reaction that does not involve the direct participation of either a protein enzyme or any multivalent cation or cofactor, and can occur under essentially physiological conditions (Amosova et al. 2006). As such, it represents the first natural deoxyribozyme activity discovered. It is mediated by formation of two very similar stem-loop-forming consensus sequences, which we have found to be present at rather high frequency in every double-stranded DNA genome searched, from the lowest to the highest form.

Figure 1.1 shows schematically the stem-loop structure of two sequences that are highly site-specific for self-catalyzed depurination of a G-residue (left) and an A-residue (right). As the figure indicates, the self-depurination mechanism removes the 5'G-residue of the loop in the former case, and the A-residue toward the 5' end in the latter one. In either case, the initial product of the catalytic event is an apurinic site in the loop sequence. This chain backbone site is thereby labilized, carrying a potential for intracellular backbone cleavage either by the enzyme apurinic endonuclease, or else as a result of its susceptibility to spontaneous backbone cleavage by a β -elimination reaction that can occur at slightly alkaline intracellular pH. Such an apurinic site, susceptible to error-prone repair, is potentially highly mutagenic and can give rise to a substitution or a short deletion. It is this resultant mutagenic potential that confers on self-depurination a role in evolution, all the more so because the self-catalytic depurination rate we have measured in vitro (Amosova et al. 2006) occurs 10^4 – 10^5 \times faster than the background spontaneous depurination rate that has been estimated in vivo (Lindahl and Nyberg 1972).

DNA is typically double-stranded, whereas the catalytic intermediate for self-depurination is a single-stranded stem-loop. Hence, the self-depurination mechanism requires that the inverted repeat sequence harboring the self-depurinating loop

first extrude as a cruciform (half of which contains the single-stranded self-depurinating stem-loop), and that the cruciform have a sufficient lifetime for the reaction to occur. That such cruciform extrusion can take place has been demonstrated previously (Alvarez et al. 2002; Inagaki et al. 2009; Kim et al. 1998; Shlyakhtenko et al. 1998). We have recently performed experiments under physiological conditions *in vitro* with stem-loop-forming sequences for G-residue self-depurination embedded in supercoiled plasmids, in which such extrusion has been directly shown to be crucial for the self-catalytic depurination (Amosova et al. 2011b).

As Fig. 1.1 indicates, the essential features of the stem-loops are very similar, i.e., both have tetra-loops with different highly specific sequences: 5'G-A/T-G-G for depurination of the extreme 5' G-residue and 5'G-A-G-A for depurination of the second residue in from the 5' end, which is an A. Interestingly, both can form a homopurine base pair within the loop, G⁺•G and A⁺•A, in which the residue to be depurinated is protonated at N3 of the base (Lavelle and Fresco, in preparation). It is this base pair formation that likely explains why acid-catalyzed depurination can actually occur in the neutral pH environment of most cells.

The G-residue self-depurination activity exhibits very limited tolerance for loop sequence variation (Amosova et al. 2011a). For example, the three G-residues in the loop are replaceable only by hypoxanthine, a closely related purine analogue, with only modest reduction in the activity; and the A-residue in the G-self-depurinating sequence is replaceable only by a T-residue, in this case with activity enhancement. In contrast, there is total tolerance for variation in the complementary base pairs of the helical stem, except for the first one at the base of the loop. Even a single base pair mismatch or a single extrahelical base elsewhere in the stem can be tolerated. Apparently, the role of the stem is to stabilize and maintain the loop in some strained configuration favorable to glycosyl bond cleavage. Thus, the more stable the stem, as affected by length (Blake and Fresco 1973; Brahms et al. 1967), G•C content (Marmur and Doty 1959), base pair sequence (Ornstein and Fresco 1983), base pair mismatches (Lomant and Fresco 1973), and the presence of extrahelical bases (Lomant and Fresco 1973), the faster the rate of self-depurination. The nature of the first base pair is somewhat restricted, possibly because it orients the water molecule involved in the hydrolysis of the glycosyl bond of the residue to be depurinated.

1.3 Biological Relevance of the Self-Depurination Mechanism

The self-depurination reaction was discovered in the course of working with 29-residue long complementary deoxyoligonucleotide strands of the human β -globin gene that contained the sickle cell anemia mutation site. Whereas the duplex formed from those strands, as well as the noncoding strand by itself, were characteristically stable, the coding strand was not. Rather, under solvent, pH, and temperature that mimic physiological conditions, it rapidly self-fragmented not

randomly, but in a unique way. This fragmentation was found to arise from backbone cleavage that was ultimately traced to spontaneous β -elimination at the apurinic site caused by self-catalyzed depurination mediated by stem-loop formation (Amosova et al. 2006). The occurrence of such a reaction in a strand segment of a significant human gene, immediately upstream of the sickle cell mutation site, provided the impetus for trying to understand the significance of what was obviously a DNA self-catalyzed reaction.

With the finding that the stem base-paired sequence is tolerant of great variation, it was decided to determine the number of potential stem-loops for self-depurination of G-residues in the human genome, and their distribution across the phyla. The numbers proved to be surprisingly large, and indicative of substantial overrepresentation relative to random expectation, which was not the case for stem-loops of similar size with non-self-depurinating loop sequences. The overrepresentation was therefore taken to be indicative of some important bio-functionality of the stem-loops for self-depurination. The whole-genome search for self-depurinating stem-loops was complemented first by identifying all human genes containing those sequences, including the loci of their occurrence, i.e., exons, introns, control elements, untranslated regions, intergenic regions, etc. This was followed by analysis of G-residue consensus sequence occurrence within more than 100 individual genes and their degree of overrepresentation relative to random expectation (see Table 1.3 as an example). These searches, together with the uncovering of the stem-loop consensus sequence for self-depurination of A-residues by way of its overrepresentation in the human genome, led us to the indications of the role of self-depurinating sequences in evolution; and it is from this vantage point that we present the findings that follow.

1.4 Self-Depurination in the Human β -Globin Gene

This gene, which has 148 codons (in its message), contains no A-residue self-depurination site, and but one site for G-residue self-depurination, of which the first three residues of the loop correspond to codon 6. It is the second residue of the loop that is the site of the sickle cell anemia mutation. Figure 1.2 shows a plot of the number of independent variations (i.e., those in different haplotypes) per codon of this gene. The plot is based upon data obtained from two databases, HB Var (<http://globin.cse.psu.edu>), which includes all human hemoglobin variants and β -thalassaemia mutations reported in the literature over more than half a century, and the recently assembled Human Gene Mutation Database (HGMD) (<http://www.hgmd.cf.ac.uk>). As such, the plot represents a compendium of all unique β -globin alleles revealed in most of the human populations in the world. It is striking that the most prominent mutation site in the plot in Fig. 1.2, that at codon 6, corresponds to three of the four loop residues of that single self-depurinating consensus sequence in this gene. In this glutamic acid codon, residue #1, the self-depurinating G-residue, and residue #2, the site of the sickle cell mutation, are the sites of readily detectable

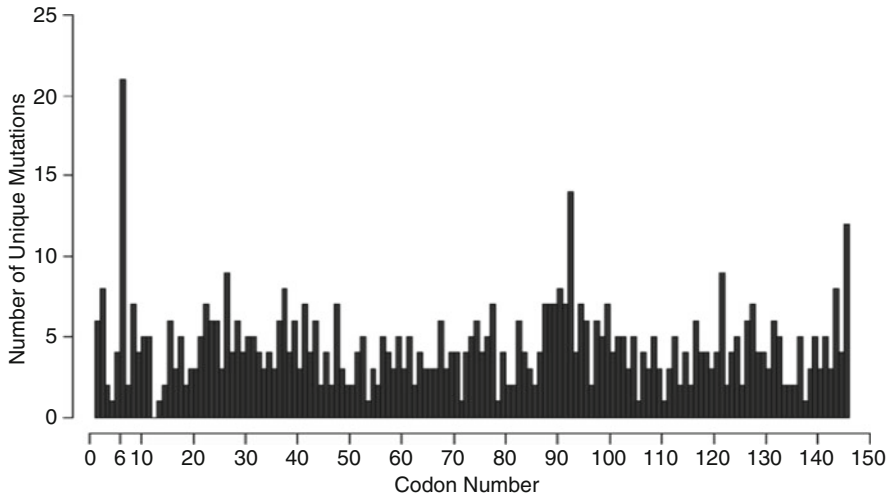


Fig. 1.2 Distribution of mutations appearing in independent haplotypes, i.e., unique mutations, among the codons of the β -globin gene. Codon 6, with the highest mutation frequency, is the only site in this gene capable of forming a stem-loop for self-depuration. Codon 6 and an additional 3'G-residue constitute the loop, which is preceded and followed by residues that form the stem shown in Fig. 1.1 (*left*). As can be deduced from Table 1.1, error-prone repair of the apurinic site resulting from self-depuration of the 5'G-residue of the loop must give rise to the observed substitution and deletion mutations, and thereby to the various anemias and β -thalassemias listed

Table 1.1 Coincidence of mutations reported in codon 6 of the β -globin gene with the first three loop residues of its only stem-loop-forming G-residue self-depuration consensus sequence^a

Type of human hemoglobin	Reference	Loop residues		
		#1	#2	#3
Hb A (wild-type)		G	A	G
Hb C	Siriboon et al. (1993)	A	A	G
Hb Machida	Harano et al. (1982)	C	A	G
Hb Grignoli	Grignoli et al. (2000)	T	A	G
Hb S (Sickle cell)	Engelke et al. (1988)	G	T	G
Hb Lavagna	Tanca et al. (2008)	G	G	G
Hb G Makassar	Sangkitporn et al. (2002)	G	C	G
β -Thalassemia	Kazazian et al. (1983)	G	D	G
β -Thalassemia	Dejong et al. (1968); Juricic et al. (1983)	D	D	D

^aA, G, C, T are standard base symbols; D indicates a deleted residue

substitutions (Table 1.1). Due to the partial degeneracy in the glutamic acid codon, a transition mutation in residue #3 would be silent, i.e., result in no amino acid change; but a transversion would encode aspartic acid, with little effect on hemoglobin function. Such a transversion should be detectable by the protein sequencing used to identify most of these mutations, and both types should be by DNA sequence analysis. So far, SNPs at that codon residue have not been reported. Codon 6 is also the site of a single base deletion at residue #2 and a deletion of

the entire codon. As noted earlier, this range of substitutions and short deletions at an apurinic site is just what is expected as a consequence of error-prone repair. All these are known to be inherited germline mutations that must have occurred over evolutionary time, and are responsible for different anemias and β -thalassemias. As such, the coincidence of these inherited mutations with the self-depurination consensus sequence at codon 6 in the β -globin gene provides strong support for the occurrence of the self-depurination mechanism in vivo, at least in germline cells. In this connection, it is worthy of mention that several haplotypes of the sickle cell mutation have been identified (Chebloune et al. 1988; Wailoo 1991), each likely representing an independent occurrence traceable to at least four different places in the Indian/Saudi Arabian subcontinent and in Africa (Kulozik et al. 1986; Lapoum eroulie et al. 1992; Pagnier et al. 1984; Schroeder et al. 1989). These occurrences speak to the mutagenicity associated with the error-prone repair of the apurinic sites created by the self-depurinating mechanism.

1.5 Coincidence of Substitution and Short Deletion Mutations with the Consensus Sequence for Self-Depurination of G-Residues in Some Human Genes

Having discovered the self-depurination mechanism at the site of codon 6 in the β -globin gene, and then found that nearly all detectable mutations anticipated for this mechanism have occurred over time at this site, it was of interest to see whether such mutations have occurred as well in other genes by this mechanism. With $\sim 1.25 \times 10^6$ potential G-residue self-depurinating sites in the human genome, and given their fairly regular distribution among exons, introns, control elements, etc., numbers of those sites should coincide with mutations detected in exons of other genes. A number of other such coincidences (see below) have indeed been found. At this early stage of human gene sequencing and the analysis of SNPs and short deletions, the finding of such coincidences in some cases can be viewed as more significant than their absence in others.

Table 1.2 provides details of nine such examples in five different genes. In each example, the entire consensus sequence is capable of forming a stem-loop catalytic intermediate. This sampling includes examples with the two loop sequences, 5'G-A-G-G and 5'G-T-G-G, each with T•A or G•C as the complementary pair at the base of the loop. Substitutions are seen to occur in the depurinated first loop residue, in the second loop residue, in the third, and in the fourth; there are three examples of deletion, one of the depurinated G-residue, another of the two consecutive G residues in the loop, and one of two stem residues.

As was apparent from the very limited tolerance for stem-loop sequence variation, it is reasonable to assume that any site in a genome which meets the sequence criteria for the self-depurinating mechanism is a potential site for mutation. That this is in fact the case is made even more convincing based upon the findings that follow.

Table 1.2 Coincidence of self-depuration consensus sequence and mutation sites in some human genes

Gene	Loop + first base-pair sequence ^a	Loop codons	Codon at mutation site	Mutation	Reference
TTR, transthyretin	T-GAG-GAG	61–62	61	GAG → AAG	Rosenzweig et al. (2007)
TTR, transthyretin	T-GAG-GAG	61–62	61	GAG → GGG	Shiomi et al. (1993)
SERPINC1, antithrombin	G-GAGG-C	42–43	42	G-GAG → TAG	Jochmans et al. (1998)
SERPINC1, antithrombin	T-GTGG-A	224–225	225	G-TGG → CGG	Lane et al. (1997)
BRCA1, breast cancer 1	T-GAGG-A	1,004–1,005	1,003	2 nt deletion of stem residues	Wagner et al. (1999)
BRCA1, breast cancer 1	T-GAGG-A	1,004–1,005	1,004	Loop G deleted	Goelen et al. (1999)
BRCA1, breast cancer 1	T-GAGG-A	1,219–1,220	1,219	GAG → GAC	Durocher et al. (1996)
TP53, tumor protein 53	T-GAGG-A	190–191	190	deletion of 2 loop Gs	Varley et al. (1997)
MAPT, tau protein	G-GAGG-C	271–272	272	GGC → GTC	Hutton et al. (1998)

^aThe G-residue in bold is the self-depuration site. In each case, the entire consensus sequence is capable of forming a stable stem-loop catalytic intermediate.

1.6 Stem-Loop-Forming Consensus Sequences for Site-Specific Self-Catalyzed Depurination of G-Residues Are Highly Overrepresented in Some Human Genes

Although genomes have not evolved as strictly random sequences relative to some defined base pair composition, it is not unreasonable to assume that sequences of the size of the consensus sequences for self-depuration (14 nt minimal length) should be present at frequencies that do not deviate too far from random unless they have some biofunctional role to play that led to their selection. Besides, the frequency of random sequence stem-loops can be used as a control or basis for comparison.

Table 1.3 gives “random probabilities” of the consensus sequence for each gene listed, and for comparison, the actual or “observed” number of consensus sequences in each. It will be noticed that the random probabilities are quite close to the “observed” number for non-depurinating stem-loops for each gene (the ratios of the “observed” to “calculated” are in the range of 0.7–2.1), in contrast to the wide variation of this ratio for the G-residue self-depurinating stem-loops (from 4 to 48). If the random probabilities are generally slightly less than the observed numbers, it is probably because these stem-loop-forming sequences represent a subclass of inverted repeats. These are known to be overrepresented in the genome (Cox and Mirkin 1997), which, as noted, is not a true random sequence. This overrepresentation of inverted repeats could also contribute to the number of “observed”

Table 1.3 Overrepresentation of the stem-loop-forming consensus sequence for site-specific, self-catalyzed depurination of G-residues in some human genes^a

Gene symbol	Total gene size (#bp) ^b	Self-depurinating consensus sequences			Non-self-depurinating stem-loop sequences		
		Observed ^b	Random expectation	Ratio of observed to random expectation	Observed ^b	Random expectation	Ratio of observed to random expectation
HRAS	3,309	0	0.08	0.0	81	38.05	2.1
CDKN1A	8,623	4	0.22	18.4	119	99.16	1.2
IGF2	9,653	1	0.24	4.1	132	111.01	1.2
CYP7A1	9,984	3	0.25	11.9	199	114.82	1.7
THBS1	16,389	3	0.41	7.3	238	188.47	1.3
TP53	19,179	22	0.48	45.5	156	220.56	0.7
CDKN2A	26,740	5	0.67	7.4	554	307.51	1.8
THRA	31,058	32	0.78	40.9	410	357.17	1.1
MDM2	32,235	31	0.81	38.2	600	370.70	1.6
HPRT1	40,440	28	1.02	27.5	705	465.06	1.5
ErbB2	40,523	5	1.02	4.9	699	466.01	1.5
KRAS	45,675	19	1.15	16.5	896	525.26	1.7
BARD1	81,090	32	2.04	15.7	1,530	932.54	1.6
BRCA1	81,155	99	2.05	48.4	879	933.28	0.9
BRCA2	84,193	50	2.12	23.6	965	968.22	1.0
APC	108,353	40	2.73	14.6	1,872	1246.06	1.5
ATM	146,268	68	3.69	18.4	2,639	1682.08	1.6

^aConsensus and non-self-depurinating stem-loops were identified based upon the presence and absence, respectively, of the consensus loop and first base pair (T•A or G•C for the self-depurinating stem-loops and any base pair for the non-self-depurinating one) plus a minimum of four base pairs within the adjacent five base pair stretch.

^bExons + introns + control elements

depurinating consensus sequences, and its effect should not be vastly different from what is observed for the non-depurinating stem-loops, i.e., it can contribute at most a factor of 2, not of 48. The fact that G-residue self-depurination sites throughout the genome occur far in excess of random expectation, by a factor of more than 5, suggests that they probably have some significant or essential biological role(s) that might be connected to the potential they create for mutation. Some critical insights in this regard followed upon the discovery of the consensus sequence for self-depurination of A-residues.

1.7 Discovery of the Consensus Sequence for Self-Depurination of A-Residues

It is the recognition of the overrepresentation of the consensus sequence for G-residue self-depurination, and the assumption that the consensus stem-loop sequence for similar self-depurination of A-residues might also be overrepresented

that led us to proceed with a search for it in the human genome. An important clue in that search was obtained from the work on the mechanism of the toxic protein ricin, which had been found to depurinate an adenine residue at a unique site, position 4,324, in 23S ribosomal RNA (Endo and Tsurugi 1987). The sequence at that site had been shown to form a stem-loop with the same number of loop and stem residues as the stem-loop for DNA self-depurination of G-residues, but with a different loop sequence, 5'G-A-G-A instead of 5'G-T/A-G-G (Amukele et al. 2005). Why then the requirement for a protein to depurinate the RNA target but not a DNA target? Our explanation was that the deoxyriboglycosyl bond, in being some three orders of magnitude more susceptible to acid-catalyzed hydrolysis than the riboglycosyl bond (Shabarova and Bogdanov 1994), did not require an enzyme catalyst. Based on this reasoning, we initiated a search in the human genome for stem-loop sequences with a 5'G-A-G-A loop and all four possible first base pairs. Once they were found, calculations were made to determine which if any were overrepresented. Interestingly, three of the four base pairs at the base of the loop were found to be similarly overrepresented significantly, which was taken to mean that they were likely self-catalytic for self-depurination of A-residues. This was confirmed in preliminary experiments.

More informative searches of the entire human genome followed to determine whether there are any genes with common functional annotation among the top 100 genes in which the A-residue self-depurinating consensus sequence is most highly overrepresented. This information was sought in order to learn whether there are any groups of genes that might have exploited the A-residue consensus sequence for self-depurination. It was indeed gratifying to find that there are, in fact, two groups of such genes, each with a common functional annotation, for which a built-in mutagenic mechanism to create sequence diversity is consistent with their function. The group with the greatest overrepresentation is the one encoding the hypervariable regions of immunoglobulins, while that with the second highest overrepresentation consists of the olfactory receptor genes.

1.8 Self-Depurination of A-Residues and Antibody Diversity

Why should the genes for the hypervariable regions of immunoglobulins encode sequences particularly disposed to undergoing mutation? In order for antibodies to perform their function, they must recognize a wide variety of antigens, including many that have not been encountered previously. Rather than the immune response mechanism having encoded what, in effect, would have to be so large a number of antibody genes in the germline as would be wasteful, biologically unproductive, and require a genome of excessive size, antibody diversity is generated by a capacity for targeted mutagenesis within a relatively small number of genes in the genome of certain somatic cells (Tonegawa 1983). From the diversity of somatic antibody genes so created by mutagenic mechanisms, clonal selection can enable production of a group of antibodies appropriate to any new antigenic challenge.

Two mechanisms were uncovered previously in (somatic) B cells to generate the requisite diversity of the different polypeptide components of the Y-shaped antibody structure (see Fig. 1.3). One such mechanism involves recombination of certain immunoglobulin gene sequence elements (Roth et al. 1992). A second mechanism, occurring during B cell division, involves somatic hypermutation that targets a different segment of antibody structure by way of enzymatic deamination of cytosine to uracil, thereby creating transition mutations (Neuberger et al. 2003). With our finding that immunoglobulin genes are the most significantly overrepresented with the A-residue self-depurination consensus sequence, and that they contain G self-depurination sites as well, it would appear that nature has selected this additional mutagenic mechanism to create antibody diversity.

In this connection, it is interesting that in the constant regions of immunoglobulin genes, the self-depurinating consensus sequence for A-residues is not highly overrepresented. This is a further indication of its selection for the purpose of creating the hypervariable regions. It is apparent then that these genes evolved so as to exploit the self-depurinating mechanism in their contemporary function.

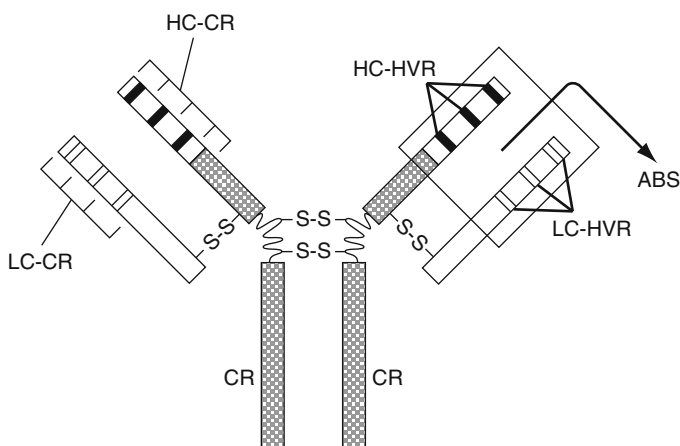


Fig. 1.3 A schematic diagram of the Y-shaped structure of a typical human antibody (protein) molecule showing the location of the constant regions (CR) that form the skeletal framework from which are extended various combinations of hypervariable heavy chain (HC-HVR) and light chain (LC-HVR) segments. Those hypervariable sequence segments, three in each heavy and each light chain, are interspersed by constant segments (HC-CR and LC-CR), and together form antigen-binding sites (ABS). Previously, those hypervariable segments had been found to attain great sequence diversity by a combination of genetic recombination and C-residue deamination. Now we have discovered that over the course of evolution, nature selected the self-depurination mechanism for this purpose as well, since the hypervariable regions are very heavily endowed with A-residue self-depurinating sites. Hence, the gene sequences for those sites provide another major mechanism to enable them to undergo mutation in order to create the extraordinary sequence diversity required to meet the challenges to the immune response

1.9 Self-Depurination of A-Residues in the Evolution of Olfactory Receptor Genes

As with the immunoglobulins, the function of the structurally similar olfactory receptors requires them to be able to recognize a large number of different odorific molecules. However, in contrast to the genes encoding the hypervariable regions of the immunoglobulins, the olfactory receptor genes are all already encoded in the genome, and there is no indication that their somatic DNA sequences undergo hyper mutation in their coding regions to any significant extent (Sharon et al. 1998, 1999). In effect, they are currently utilized with the diversity with which they evolved over time. Of the ~850 such human genes, of which some 55–60% are pseudogenes (Olender et al. 2004), a small fraction are singlets, but the majority are in clusters distributed among all chromosomes but #20 and Y. The gene clusters are likely to have arisen by repeated duplication of individual genes and clusters (Glusman et al. 1996), as many of the genes in tandem arrays are closely related (Niimura and Nei 2003). At the same time, during their evolution, they apparently exploited mutagenesis, at least by the self-depurination of A-residues to differentiate in function.

With such high overrepresentation of the self-depurination consensus sequence in olfactory receptor genes, it is reasonable to expect that over evolutionary time, the same mutagenic mechanism may have also resulted in the loss of the self-depurinating capacity in some of those genes. In that case, corresponding nonfunctional or pseudogenes would have accumulated that arose, e.g., as a consequence of substitution mutations in the loop residues of the catalytic intermediate. Evidence for this possibility was sought by searching these gene clusters for highly overrepresented pseudogenes whose consensus loop sequences had been mutated from 5' G-A-G-A to 5'G-T-G-A and 5' G-G-G-A, some of the gene products expected at those self-depurination sites. Such a mechanism could account, then, for the very high fraction of these clustered genes that were found to be pseudogenes (Rouquier et al. 1998a, b). In fact, pseudogenes with these very A → T and A → G mutations at the self-depurinating sites were found to be overrepresented. These observations do indicate, then, that over the course of evolution the same mutagenic mechanism has served to create mutations leading both to olfactory receptor diversity and to erasure of such genes.

1.10 Discussion

Our goal in this chapter has been to summarize that aspect of our knowledge of self-catalyzed DNA depurination that particularly relates to ways in which the mechanism played a role in the evolution of certain biological processes and some inherited diseases.

Thus, we have presented several types of observations that support the notion that the self-depurination mechanism is likely to have played a role in molecular

evolution. One relates to the distribution of the consensus sequence for self-depurination of G-residues. Not only is that consensus sequence found in every double-stranded genome we have examined from various Archaea species to *Homo sapiens*, but its frequency has been found to be very high, in the neighborhood of once every 3,000–5,000 base pairs. In mitochondrial genomes, which are also double-stranded, they are present in most species at a frequency similar to those in the genomes of lower forms. While they appear to be lacking in the mitochondrial genomes of some species, it must be kept in mind that we have not yet searched them for the consensus sequence for self-depurination of A-residues. Nor, in fact, have we yet searched any genomes other than that of *Homo sapiens* for the A-residue consensus sequence.

It is for this same reason that we are not yet able to interpret the absence of G-residue self-depurination sequences in half the single-stranded viral genomes recorded in the viral genome database. If the viral genomes with no G-residue self-depurination sites are in fact lacking in those for A-residues as well, that would provide important evidence that evolution selected against the self-depurination mechanism for these species. Such negative selection would arise because once an apurinic site resulted in backbone fragmentation of the single-stranded genome, the fragments might have no template complementary strand to enable their repair.

Another type of evidence supporting a role for self-depurination in molecular evolution relates to the highly error-prone repair of apurinic sites in all species in which it has been examined, from viruses (whose hosts carry the repair mechanism) and bacteria (Shearman and Loeb 1979) to lower and higher eukaryotes (Chakravarti et al. 2000). Moreover, in a number of cases, such sites have been specifically identified as mutational hot spots (Kunkel 1984).

We have additionally presented three examples of genes for which the evidence would seem to be compelling that self-depurination must have played a role in their evolution, in their loss of function or conversion to pseudogenes, and/or in their contemporary functioning.

In the case of the human olfactory receptor genes, the self-depurination consensus sequence, at least for A-residues, is clearly associated with the evolutionary development of their diversity; and just as these sites are significantly overrepresented in this group of genes, so are their mutation products overrepresented in their nonfunctional pseudogenes. Yet, in the human species, these olfactory receptor genes do not appear to be evolving now at any detectable rate. They are already all encoded in the genome, with no evidence of somatic cell DNA rearrangements or somatic mutations in the coding regions of those genes. So it is reasonable to conclude that the mechanism played a role in the course of the evolution of those genes to their current state of diversity, but that it does not play a significant role contemporaneously.

The case of the genes encoding the hypervariable regions of the immunoglobulins would appear to be somewhat different. Here we have genes for which mechanisms for variation are essential elements for their contemporary function. The consensus sequences for both G- and A-residue self-depurination are again highly overrepresented in these genes. The consequent mutagenic mechanism may

have played a role in their evolution, something which our analysis cannot say with certainty at the present time. What is much more apparent, however, is that it is one of the several mechanisms by which antibody diversity is contemporaneously achieved in response to confrontation with an antigen.

Finally, we have presented the case of the unique occurrence of six different anemias due to substitutions and two β -thalassemias due to short deletions, all within codon 6 of the human β -globin gene. That codon is the only self-depurination consensus sequence site for G-residues in the entire gene, and there are none for A-residues. While these mutations are all explicable as a consequence of the self-depurination mechanism, there are no indications that they readily occur as somatic mutations (and even if they do, they are unlikely to be detected, since a majority of hemoglobin-producing cells would still have the wild-type DNA sequence). Consequently, they appear as germline mutations, evolutionarily retained, no doubt, as a consequence of the resistance to malaria that at least the sickle cell anemia mutation confers (Friedman 1978), and possibly some of the others as well.

Taken together, the foregoing examples serve as a strong indication that the self-depurination mechanism, coupled to the process of error-prone repair, has played an evolutionary role. But why has a mechanism capable of causing DNA damage been selected for and even concentrated in genomes in the course of evolution? Perhaps the advantages accruing from the availability of a mechanism inherent in the sequence and structural dynamics of DNA to help create gene sequence diversity far outweighs the negative consequences of genetic instability, at least where the products of gene sequence diversification have positive value. It is interesting in this connection that we have found that some genes appear to be favored by the presence of either G-residue or of A-residue self-depurinating consensus sequences; this might possibly indicate different evolutionary origins for such genes. Moreover, genes that are highly conserved, e.g., those encoding HOX, histone core, ribosomal proteins, contain very few if any such consensus sequences. It appears then, that these two consensus sequences have sometimes been selected against, but other times selected for, which would indicate that self-depurination is a significant mutagenic mechanism that has been taken note of and sometimes harnessed by evolution.

Acknowledgments This research was supported in part by a grant from the U.S. Army Research Office (W911NF-07-1-0152) and by fellowships to D. Glumcher from Dickinson College and Princeton University, to R. Torres from the Howard Hughes Medical Institute, to J. Alvarez-Dominguez and P. Wei from NIH grant P50 GM071508.

References

- Alvarez D, Novac O, Callejo M, Ruiz M, Price G, Zannis-Hadjopoulos M (2002) 14-3-3 sigma is a cruciform DNA binding protein and associates in vivo with origins of DNA replication. *J Cell Biochem* 87:194–207
- Amosova O, Coulter R, Fresco JR (2006) Self-catalyzed site-specific depurination of guanine residues within gene sequences. *Proc Natl Acad Sci USA* 103:4392–4397

- Amosova O, Smith A, Fresco J (2011a) The consensus sequence for self-catalyzed, site-specific G-residue depurination in DNA. *J Biol Chem*, in press
- Amosova O, Kumar V, Deutsch A, Fresco J (2011b) Self-catalyzed, site-specific depurination of G-residues mediated by cruciform extrusion in closed circular DNA plasmids. *J Biol Chem*, in press
- Amukele TK, Roday S, Schramm VL (2005) Ricin A-chain activity on stem-loop and unstructured DNA substrates. *Biochemistry* 44:4416–4425
- Blake RD, Fresco JR (1973) Polynucleotides. XI. Thermodynamics of $(A)_N \cdot 2(U)_\infty$ from the dependence of T_{mN} on oligomer length. *Biopolymers* 12:775–786
- Boiteux S, Guillet M (2004) Abasic sites in DNA: repair and biological consequences in *Saccharomyces cerevisiae*. *DNA Repair (Amst)* 3:1–12
- Brahms J, Maurizot JC, Michelson AM (1967) Conformation and thermodynamic properties of oligocytidylic acids. *J Mol Biol* 25:465–480
- Chakravarti D, Mailander PC, Cavalieri EL, Rogan EC (2000) Evidence that error-prone DNA repair converts dibenzo[a, l]pyrene-induced depurinating lesions into mutations: formation, clonal proliferation and regression of initiated cells carrying H-ras oncogene mutations in early preneoplasia. *Mutat Res* 456:17–32
- Chebloune Y, Pagnier J, Trabuchet G, Faure C, Verdier G, Labie D, Nigon V (1988) Structural analysis of the 5' flanking region of the beta-globin gene in African sickle cell anemia patients: further evidence for three origins of the sickle cell mutation in Africa. *Proc Natl Acad Sci USA* 85:4431–4435
- Cox R, Mirkin SM (1997) Characteristic enrichment of DNA repeats in different genomes. *Proc Natl Acad Sci USA* 94:5237–5242
- Dejong WWW, Went LN, Bernini LF (1968) Haemoglobin leiden - deletion of beta6 or 7 glutamic acid. *Nature* 220:788–790
- Durocher F, ShattuckEidens D, McClure M, Labrie F, Skolnick MH, Goldgar DE, Simard J (1996) Comparison of BRCA1 polymorphisms, rare sequence variants and/or missense mutations in unaffected and breast/ovarian cancer populations. *Hum Mol Genet* 5:835–842
- Endo Y, Tsurugi K (1987) RNA N-glycosidase activity of ricin A-chain. Mechanism of action of the toxic lectin ricin on eukaryotic ribosomes. *J Biol Chem* 262:8128–8130
- Engelke DR, Hoener PA, Collins FS (1988) Direct sequencing of enzymatically amplified human genomic DNA. *Proc Natl Acad Sci USA* 85:544–548
- Fresco JR, Alberts BM (1960) The accommodation of noncomplementary bases in helical polyribonucleotides and deoxyribonucleic acids. *Proc Natl Acad Sci USA* 46:311–321
- Friedman M (1978) Erythrocytic mechanism of sickle cell resistance to malaria. *Proc Natl Acad Sci USA* 75:1994–1997
- Glusman G, Clifton S, Roe B, Lancet D (1996) Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity. *Genomics* 37(2):147–160
- Goelen G, Teugels E, Bonduelle M, Neyns B, De Greve J (1999) High frequency of BRCA1/2 germline mutations in 42 Belgian families with a small number of symptomatic subjects. *J Med Genet* 36:304–308
- Grignoli CR, Carvalho MH, Kimura EM, Sonati MF, Arruda VR, Saad ST, Costa FF (2000) Beta0-thalassemia resulting from a novel mutation: beta66/u- > stop codon. *Eur J Haematol* 64:137–138
- Harano T, Harano K, Ueda S, Shibata S, Imai K, Seki M (1982) Hemoglobin Machida [beta-6 (A3) Glu-JGln], a new abnormal hemoglobin discovered in a Japanese family - structure, function and biosynthesis. *Hemoglobin* 6:531–535
- Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, Chakraverty S, Isaacs A, Grover A, Hackett J, Adamson J, Lincoln S, Dickson D, Davies P, Petersen RC, Stevens M, de Graaff E, Wauters E, van Baren J, Hillebrand M, Joesse M, Kwon JM, Nowotny P, Che LK, Norton J, Morris JC, Reed LA, Trojanowski J, Basun H, Lannfelt L, Neystat M, Fahn S, Dark F, Tannenberg T, Dodd PR, Hayward N, Kwok JBJ, Schofield PR, Andreadis A, Snowden J, Craufurd D, Neary D, Owen F, Oostra BA, Hardy J,

- Goate A, van Swieten J, Mann D, Lynch T, Heutink P (1998) Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* 393:702–705
- Inagaki H, Ohye T, Kogo H, Kato T, Bolor H, Taniguchi M, Shaikh TH, Emanuel BS, Kurahashi H (2009) Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res* 19:191–198
- Jochmans K, Lissens W, Seneca S, Capel P, Chatelain B, Meeus P, Osselaer JC, Peerlinck K, Seghers J, Slacmouder M, Stibbe J, van de Loo J, Vermeylen J, Liebaers I, De Waele M (1998) The molecular basis of antithrombin deficiency in Belgian and Dutch families. *Thromb Haemost* 80:376–381
- Juricic D, Ruzdic I, Beer Z, Efremov GD, Casey R, Lehmann H (1983) Hemoglobin Leiden [beta 6 or 7 (A3 or A4) Glu leads to O] in a Yugoslavian woman arisen by a new mutation. *Hemoglobin* 7:271–277
- Kazazian HH, Orkin SH, Boehm CD, Sexton JP, Antonarakis SE (1983) Beta-Thalassemia due to a deletion of the nucleotide which is substituted in the beta-S-globin gene. *Am J Hum Genet* 35:1028–1033
- Kim E, Peng H, Esparza F, Maltchenko S, Stachowiak M (1998) Cruciform-extruding regulatory element controls cell-specific activity of the tyrosine hydroxylase gene promoter. *Nucleic Acids Res* 26:1793–1800
- Korolev VG (2005) Base excision repair: AP endonucleases and DNA polymerases. *Russ J Genet* 41:1063–1070
- Kulozik A, Wainscoat J, Serjeant G, Kar B, Al-Awamy B, Essan G, Falusi A, Haque S, Hilali A, Kate S et al (1986) Geographical survey of beta S-globin gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. *Am J Hum Genet* 39:239–244
- Kunkel TA (1984) Mutational specificity of depurination. *Proc Natl Acad Sci USA* 81:1494–1498
- Lane DA, Bayston T, Olds RJ, Fitches AC, Cooper DN, Millar DS, Jochmans K, Perry DJ, Okajima K, Thein SL, Emmerich J (1997) Antithrombin mutation database: 2nd (1997) update. For the Plasma Coagulation Inhibitors Subcommittee of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. *Thromb Haemost* 77:197–211
- Lapoum eroulie C, Dunda O, Ducrocq R, Trabuchet G, Mony-Lob e M, Bodo J, Carnevale P, Labie D, Elion J, Krishnamoorthy R (1992) A novel sickle cell mutation of yet another origin in Africa: the Cameroon type. *Hum Genet* 89:333–337
- Lhomme J, Constant JF, Demeunynck M (1999) Abasic DNA structure, reactivity, and recognition. *Biopolymers* 52:65–83
- Lindahl T, Nyberg B (1972) Rate of depurination of native deoxyribonucleic acid. *Biochemistry* 11:3610–3618
- Lomant AJ, Fresco JR (1973) Polynucleotides. 13. Stoichiometric and thermodynamic studies of polynucleotide helices with non-complementary residues. *Biopolymers* 12:1889–1903
- Marmur J, Doty P (1959) Heterogeneity in deoxyribonucleic acids. I. Dependence on composition of the configurational stability of deoxyribonucleic acids. *Nature* 183:1427–1429
- Neuberger M, Harris R, Di Noia J, Petersen-Mahrt S (2003) Immunity through DNA deamination. *Trends Biochem Sci* 28(6):305–312
- Niimura Y, Nei M (2003) Evolution of olfactory receptor genes in the human genome. *Proc Natl Acad Sci USA* 100(21):12235–12240
- Olender T, Feldmesser E, Atarot T, Eisenstein M, Lancet D (2004) The olfactory receptor universe—from whole genome analysis to structure and evolution. *Genet Mol Res* 3:545–553
- Ornstein RL, Fresco JR (1983) Correlation of Tm and sequence of DNA duplexes with ΔH computed by an improved empirical potential method. *Biopolymers* 22:1979–2000
- Pagnier J, Mears JG, Dunda-Belkhdja O, Schaefer-Rego KE, Beldjord C, Nagel RL, Labie D (1984) Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc Natl Acad Sci* 81:1771–1773
- Rosenzweig M, Skinner M, Prokava T, Th eberge R, Costello C, Drachman BM, Connors LH (2007) A new transthyretin variant (Glu61Gly) associated with cardiomyopathy. *Amyloid* 14:65–71

- Roth DB, Nakajima PB, Menetski JP, Bosma MJ, Gellert M (1992) V(D)J recombination in mouse thymocytes: double-strand breaks near T cell receptor delta rearrangement signals. *Cell* 69:41–53
- Rouquier S, Friedman C, Delettre C, van den Engh G, Blancher A, Crouau-Roy B, Trask B, Giorgi D (1998a) A gene recently inactivated in human defines a new olfactory receptor family in mammals. *Hum Mol Genet* 7(9):1337–1345
- Rouquier S, Taviaux S, Trask B, Brand-Arpon V, van den Engh G, Demaille J, Giorgi D (1998b) Distribution of olfactory receptor genes in the human genome. *Nat Genet* 18(3):243–250
- Sangkitporn S, Rerkamnuaychoke B, Mitrakul C, Sutivigit Y (2002) Hb G Makassar (beta 6:Glu-Ala) in a Thai family. *J Med Assoc Thai* 85:577–582
- Schroeder W, Powars D, Kay L, Chan L, Huynh V, Shelton J, JR S (1989) Beta-cluster haplotypes, alpha-gene status, and hematological data from SS, SC, and S-beta-thalassemia patients in southern California. *Hemoglobin* 13:325–353
- Shabarova Z, Bogdanov A (1994) *Advanced organic chemistry of nucleic acids*. VCH (UK) Ltd, Cambridge, England, p 60
- Sharon D, Glusman G, Pilpel Y, Horn-Saban S, Lancet D (1998) Genome dynamics, evolution, and protein modeling in the olfactory receptor gene superfamily. *Olfaction Taste* Xii: 182–193
- Sharon D, Glusman G, Pilpel Y, Khen M, Gruetzner F, Haaf T, Lancet D (1999) Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. *Genomics* 61(1):24–36
- Shearman CW, Loeb LA (1979) Effects of depurination on the fidelity of DNA synthesis. *J Mol Biol* 128:197–218
- Shiomi K, Nakazato M, Matsukura S, Ohnishi A, Hatanaka H, Tsuji S, Murai Y, Kojima M, Kangawa K, Matsuo H (1993) A basic transthyretin variant (Glu(61)-JLys) causes familial amyloidotic polyneuropathy - protein and DNA-sequencing and PCR-induced mutation restriction analysis. *Biochem Biophys Res Commun* 194:1090–1096
- Shlyakhtenko LS, Potaman VN, Sinden RR, Lyubchenko YL (1998) Structure and dynamics of supercoil-stabilized DNA cruciforms. *J Mol Biol* 280:61–72
- Simonelli V, Narciso L, Dogliotti E, Fortini P (2005) Base excision repair intermediates are mutagenic in mammalian cells. *Nucleic Acids Res* 33:4404–4411
- Siriboon W, Srisomsap C, Winichagoon P, Fucharoen S, Svasti J (1993) Identification of Hb C [beta 6(A3)Glu- > Lys] in a Thai male. *Hemoglobin* 17:419–425
- Sugiyama H, Fujiwara T, Ura A, Tashiro T, Yamamoto K, Kawanishi S, Saito I (1994) Chemistry of thermal degradation of abasic sites in DNA. Mechanistic investigation on thermal DNA strand cleavage of alkylated DNA. *Chem Res Toxicol* 7:673–683
- Tanca D, Devoto G, Deiana F, Luciano B, Lisi A, Ivaldi G (2008) Characterization of a new hemoglobin variant. *Ital J Lab Med* 4:142
- Tonegawa S (1983) Somatic generation of antibody diversity. *Nature* 302:575–581
- Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. *Nature* 263:285–289
- Varley JM, Evans DGR, Birch JM (1997) Li-Fraumeni syndrome - a molecular and clinical review. *Br J Cancer* 76:1–14
- Wagner T, Stoppa-Lyonnet D, Fleischmann E, Muhr D, Pagès S, Sandberg T, Caux V, Moeslinger R, Langbauer G, Borg A, Oefner P (1999) Denaturing high-performance liquid chromatography detects reliably BRCA1 and BRCA2 mutations. *Genomics* 62:369–376
- Wailoo K (1991) “A disease sui generis”: the origins of sickle cell anemia and the emergence of modern clinical research, 1904–1924. *Bull Hist Med* 65:185–208