The Dissertation Committee for Keith Albert Friedman Certifies that this is the approved version of the following dissertation:

Evaluation of Genome Designs for Oxidation Resistance: Guanine Minimization and Scavenger Guanine

Committee:

Adam Heller, Supervisor

George Georgiou

Nicholas A. Peppas

Christine E. Schmidt

Chaim N. Yarnitzky

Evaluation of Genome Designs for Oxidation Resistance: Guanine Minimization and Scavenger Guanine

by

Keith Albert Friedman, B.S., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2003

Acknowledgements

I am grateful beyond words for the patience and support of my family during my years in graduate school. Myself, I would not have put up with myself.

Adam Heller provided such significant ideas, light but careful mentoring and a wonderful environment. He really is as good as everyone says.

Teachers and fellow students helped me grow professionally. Especially, Chaim Yarnitzky from the Technion, who spent six months in my lab and taught by his example more than books can contain.

Evaluation of Genome Designs for Oxidation Resistance: Guanine Minimization and Scavenger Guanine

Publication No._____

Keith Albert Friedman, Ph. D

The University of Texas at Austin, 2003

Supervisor: Professor Adam Heller

The genome's environment contains strong oxidizers, some of which selectively attack guanine, the most readily oxidized nucleotide. The ranking of guanine oxidation rates is central G in GGG (G**G**G) ≥ 5' G in GG (**G**G) > isolated or 3' G. Vulnerability to selective oxidants puts mutation pressure on guanine. This is apparent in the differences between observed levels of G**G**G and levels predicted by probability from total G. G**G**G is below probability predictions in the genomes of *D. melanogaster*, *C. elegans*, *A. thaliana*, *S. cerevisiae* and *S. pombe*. G**G**G is statistically under-represented in *H. sapiens* exons, but over-represented in *H. sapiens* introns and intergenic domains. It is not under-represented in *E. cuniculi*. It is over-represented *P. falciparum* chromosomes 2 and 3, but this organism's total G levels are extremely low. **G**G generally is not under-represented in these genomes.

Beyond enzymatic elimination of the oxidizing agents and their precursors, and excision and repair of oxidative lesions, we propose that genomes are built to mitigate damage to essential domains. Resistance to oxidation could be enhanced by making genomes more "noble" by reducing the fractions of total G, **G**G and particularly G**G**G. Alternately, if the duplex conducts electron vacancies (holes) over ~100 bp, oxidation could

be shifted from essential domains to sacrificially oxidizable G**G**G and **G**G in nonessential domains. The distribution of G**G**G and **G**G in exons, introns and intergenic domains of eight model genomes suggests ennoblement in six, protection by sacrificial anodes in one, and no guanine-based protection in one (*E. cuniculi*). G**G**G triads are excluded or are statistically underrepresented in exons and short splicing-controlling introns of *D. melanogaster, C. elegans, A. thaliana, S. cerevisiae, S. pombe* and *P. falciparum* chromosomes 2 and 3. The introns of *H. sapiens*, which are about twenty times longer than those of the other organisms, are rich in sacrificially oxidizable G**G**G triads that are 50-100 bp from the exons. Their frequency correlates with the presence of protection-requiring G**G**G triads in the exons.

Table of Contents

Chapter 1: The Hypothesis of Cathodic Protection of Genes

INTRODUCTION

In the 2000 Spiers Memorial Lecture, "On the Hypothesis of Cathodic Protection of Genes" (1), Adam Heller said that an electrochemist looking at DNA sees a corrosion problem that engineers routinely solve with cathodic protection. Researchers on remote oxidation in DNA, especially, Barton et al. (2, 3), Giese et al. (4, 5), Kawanishi et al. (6), Schuster et al. (7), Thorp et al. (8), had accumulated evidence that suggested to them the plausibility of this hypothesis. Heller's lecture synthesized these results, defined and developed the hypothesis, and asked: are genomes organized to mitigate oxidative damage? This question was the impetus for this dissertation, so this introduction is based on that lecture, sometimes word-for-word.

DISCUSSION

Hypothesis: Cathodic Protection of Genes. The hypothesis of cathodic protection of genes states that some organisms have evolved so that sacrificial G-rich DNA sequences are preferentially oxidized to cathodically protect genes and other essential domains against oxidative damage. The sacrificially oxidized domains may be exclusively protective, or they may have essential functions, the transient loss of which is tolerated within the time required for damage recognition, excision and repair (9, 10). Consequently, loss of cathodic protection contributes to aging and death of non-proliferating cells, increased likelihood of cell mutation, and the particular sensitivity of some cancer cells to oxidative damage.

Cathodic Protection of Metals. When two electron conductors are in an oxidizing environment, immersed in the same electrolytic solution, and electrically connected, the more noble one is cathodically protected against oxidative damage and the less noble one is sacrificially oxidized. For example, zinc cathodically protects steel against oxida-

tion, being itself sacrificially oxidized ($Fe^{2+} + Zn \rightarrow Fe + Zn^{2+}$). The two metals are conductors, and the standard potentials for anodic oxidation of iron ($Fe \rightarrow Fe^{2+} + 2e^-$) and zinc ($Zn \rightarrow Zn^{2+} + 2e^-$) are –0.44 V and –0.76 V, respectively (11). (All potentials are relative to the normal hydrogen electrode.) Cathodic protection of steel hulls of ships by zinc, now widely used, was introduced by Humphrey Davy, who was assisted by Michael Faraday. For cathodic protection of genes, chromosomal DNA must conduct electrons and/or holes, and essential components of genes must be electrochemically noble relative to less essential domains that are sacrificially oxidized.

Electrical Conduction in DNA. The occurrence of a faradaic reaction proves transport of electrons or holes in a film. Although passage of an electrical current through a film can result from transport of ions, electrons or holes, faradaic reactions do not take place unless electrons or holes are transported. Hartwich et al. studied the occurrence of faradaic reactions on electrodes coated with calf-thymus DNA and found that thin films of randomly-oriented double-stranded DNA and thiol-terminated 50 Å thick monolayers of single-stranded DNA do not conduct electrons or holes (12). However, they found that double-stranded DNA monolayers, in which the duplexes are aligned in parallel and tilted by about 30° vs. the surface normal, conduct electrons or holes. In their experiments, $Fe(CN)_6^{3-/4-}$, which is electrostatically excluded from the film, and PQQ, which is covalently bound to the solution side of the monolayer (Figure 1-1), are electro-oxidized/reduced. The rate constants, calculated from cyclic voltammetry (Figure 1-2) (13, 14), are nil with single stranded DNA, and 2×10^{-3} cm s^{-1} for $Fe(CN)_6^{3-/4-}$ and 1.5 ± 0.2 s^{-1} for PQQ with double-stranded DNA. The rates decrease when two base pairs are mismatched in the 12 bp duplex.

Okahata and co-workers showed that steady-state DC conductivity increases more than a thousand-fold when DNA aggregates are aligned along the helices' axes (23). The steady-state current that they measured could only be electron or hole transport, because mobile ions are exhausted in less than one minute in their experiments

These demonstrations of DNA conduction in aqueous environments are consistent with conduction in solid DNA. Fink and Schönenberger studied ~600 nm long DNA ropes, and found that their one-dimensional conductance is comparable to that of degenerate semiconductors and only ~100 times smaller than that of iron and their current-voltage relationship is ohmic (15). Porath et al. found that ~10 nm long poly-GC duplexes, individually and in small clusters, conduct electrons or holes in air and *in vacuo* at both ambient and cryogenic temperatures (16).

These experiments demonstrate that DNA can conduct electrons and/or holes in the laboratory, not that it does so *in vivo*. Solid, aligned DNA aggregates do not represent flexible, randomly-oriented DNA strands in chromatin generally found in cells (17). Special conditions are required to form liquid crystalline DNA *in vivo* (18-20); intriguingly, this occurs in *E. coli* under oxidative stress (21). Demonstration of DNA conduction *in vivo* and *in vitro* under biological conditions comes from experiments showing remote oxidation of G, described below.

Electrochemical Series of DNA Bases. Faraggi and co-workers determined that the one-electron oxidation potentials of DNA bases of at pH 7 are 1.04 for G; 1.29 for T; 1.32 for A; and 1.44 V for C, where all potential have ± 0.02 V uncertainty (22). These values, particularly that G is the most reducing of the four nucleotides, were confirmed by Oliveira-Brett et al. (23) and Tomschik et al. (24). Hutter and Clark (25) and Sugiyama and Saito (26) calculated that the ionization potential of the G:C base pair is downshifted relative to that of the non-hybridized G base. Not only is G the most reducing base, its catalytic one-electron oxidation kinetics in poly-G sequences are particularly rapid (27).

Remote Oxidation of Guanine. Migration of holes to oxidize remote poly-G sequences in dissolved oligonucleotides has been reported by Barton et al. (3, 28-31), Schuster et al. (32-39), Giese, Michel-Beyerle and co-workers (40-44), Fukui and Tanaka (45) and Saito et al. (26, 46, 47). Their photochemical, spectroscopic and theoretical

studies established that guanines in GG and poly-G sequences are selectively oxidized upon oxidative attack at any position in ≤50 bp oligonucleotides. This distance substantially exceeds the maximum electron transfer distance in proteins, ≤20 Å (48), equivalent to ≤6 bp. These photochemical studies also show that a G oxidized to 8-oxo-7,8-dihydro-2′-deoxyguanosine stops hole transport (33); remote oxidation can extend from double-stranded to single-stranded DNA (39); and G:C base pairs retard carrier transport of far less than A:T base pairs (40-44).

Guanine-rich Domains as Anodes. A corrosion chemist envisioning a logical scheme for cathodic protection of genomes might idealize genes or chromosomes as conducting rods composed of a body with similar mole fractions of A:T and G:C and an end piece that is particularly rich in G (Figure 1-3). In this scheme, the G-rich end piece is the sacrificial anode that cathodically protects the body of the DNA rod against oxidation, as a zinc-rich end piece protects a brass rod. Genomic DNA has G-rich domains flanking genes and chromosomes that are less essential than genes, at least transiently, but it lacks the conductivity to use them as sacrificial anodes. Vertebrate housekeeping genes often have a CpG island at their 5' end (49), and the G of CpG dinucleotides is readily oxidized (50). Telomers at chromosome ends are G-rich in their repeat sequence, GGGTTA in humans (51), and in their single-stranded overhang (52, 53). Neither telomeres nor CpG islands are transcribed. In dissolved DNA, electron transfer across ~50 base pairs, which is much shorter than the distances required for the cathodic protection of genes many thousands of bases long (see also (2, 42)). Conduction over such distances probably is restricted to solid DNA arrays in which the molecules are aligned in parallel.

The search for G-rich domains that could function as sacrificial anodes within the constraints of DNA conduction was the impetus for this research..

Genome Maintenance in an Oxidizing Environment. Cellular DNA is exposed to strong oxidizers such as NO, H_2O_2, singlet oxygen, and $^{\bullet}OH$ radicals. Although buildup of NO and H_2O_2 usually is avoided, their transient concentration can exceed 1

nM (54), which means $>10^9$ oxidant molecules per genome copy. Although catalases abound in tissues, some of the continuously generated H_2O_2 reacts with oxidizable transition metal ions, such as Fe^{2+} or Cu^+, to produce $^{\bullet}OH$ radicals via the Fenton reaction. As a result, the DNA in each human cell undergoes $\sim 10^4$ oxidative attacks each day (55, 56) and requires continual excision and replacement of damaged segments (9, 10).

DNA conductivity could enable cathodic protection to reduce the likelihood of gene alteration by these and other oxidants. In an insulator, attack by an oxidizer results in a local chemical change at the attack site (Figure 1-4). In a conductor, the reaction mainly occurs at the most reducing site within the diffusion range of the hole injected by the oxidizer (Figure 1-3). In conducting DNA, the hole generally diffuses to and reacts at a remote, G-rich domain, where it forms a remote radical, often by releasing a proton. If the remote reaction site is less essential than the original attack site, then it acts as a sacrificial anode that provides cathodic protection.

Cathodic Protection and Cancer. If the hypothesis is valid, then the extent of cathodic protection provided by G-rich domains and the accumulated oxidative damage in them help define the likelihood of mutation under oxidative stress which is related to cancer (56-58). Some carcinogens are metabolized to DNA-binding oxidation catalysts. For example, carcinogenic aromatic hydrocarbons are hydroxylated to form DNA-binding ortho-diphenols that react rapidly and efficiently with molecular oxygen to form quinones that are rapidly reduced by NADH or NADPH. Thus, the carcinogen-derived phenols/quinones catalyze the reaction $NADH + O_2 + H^+ \rightarrow NAD^+ + H_2O_2$. Reaction of H_2O_2 with a reduced transition metal ion produces $^{\bullet}OH$ radicals in the Fenton reaction. Upon oxidation a G is converted to 8-oxo-7,8-dihydro-2′-deoxyguanosine which, unlike G itself, hybridizes not only with C, but also with A. This results in the attacked G:C base pair mutating to T:A after two round of replication.

Oxidation is a cause and a treatment for cancer. Theories that connect oxidation and cancer (58), including the cathodic protection hypothesis, resolve this paradox as fol-

lows. Some cells and genes are particularly susceptible to oxidative mutation, in some cases leading to cancer. Genes with poor cathodic protection, due to weak or exhausted sacrificial anodes, putatively are especially susceptible. In tumors, these same cells and genes still are especially sensitive to oxidants, such as those produced by ionizing radiation or chemotherapeutic drugs like cisplatin or doxorubicin, which are oxidation catalysts (59, 60). The more likely it is that a gene will mutate cancerously under oxidative stress, the greater will be its susceptibility to further mutation, leading to cell death.

REFERENCES

1. Heller, A. (2000) *Faraday Discuss.* **116,** 1-13.
2. Boon, E. M. & Barton, J. K. (2002) *Curr. Opin. Structural Biology* **12,** 320-329.
3. Hall, D. B., Holmlin, R. E. & Barton, J. K. (1996) *Nature* **382,** 731-735.
4. Giese, B. (2000) *Chemistry in Britain* **36,** 44-46.
5. Giese, B. (2002) *Annu. Rev. Biochemistry* **71,** 51-70.
6. Oikawa, S., Tada-Oikawa, S. & Kawanishi, S. (2001) *Biochemistry* **40,** 4763-4768.
7. Kanvah, S. & Schuster, G. B. (2002) *J. Am. Chem. Soc.* **124,** 11286-11287.
8. Szalai, V. A., Singer, M. J. & Thorp, H. H. (2002) *J. Am. Chem. Soc.* **124,** 1625-1631.
9. Croteau, D. L. & Bohr, V. A. (1997) *J. Biol. Chem.* **272,** 25409-25412.
10. Bohr, V. A. & Anson, R. M. (1995) *Mutat. Res.* **338,** 25-34.
11. Bard, A. J. & Faulkner, L. R. (2001) *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York).
12. Hartwich, G., Caruana, D. J., de Lumley-Woodyear, T., et al. (1999) *J. Am. Chem. Soc.* **121,** 10803-10812.
13. Laviron, E. (1979) *J. Electroanal. Chem.* **101,** 19-28.
14. Nicholson, R. S. (1965) *Anal. Chem.* **37,** 1351-5.
15. Fink, H.-W. & Schonenberger, C. (1999) *Nature* **398,** 407-410.
16. Porath, D., Bezryadin, A., De Vries, S., et al. (2000) *Nature* **403,** 635-638.

17. Zlatanova, J., Leuba, S. H. & Van Holde, K. (1998) *Biophys. J.* **74,** 2554-2566.

18. Leforestier, A. & Livolant, F. (1993) *Biophys. J.* **65,** 56-*72.

19. Livolant, F. (1991) *Physica A* **176,** 117-37.

20. Strey, H. H., Podgornik, R., Rau, D. C., et al. (1998) *Curr. Opin. Structural Biology* **8,** 309-313.

21. Wolf, S. G., Frenkiel, D., Arad, T., et al. (1999) *Nature* **400,** 83-85.

22. Faraggi, M., Broitman, F., Trent, J. B., et al. (1996) *J. Phys. Chem.* **100,** 14751-14761.

23. Oliveira-Brett, A. M., Vivan, M., Fernandes, I. R., et al. (2002) *Talanta* **56,** 959-970.

24. Tomschik, M., Jelen, F., Havran, L., et al. (1999) *J. Electroanal. Chem.* **476,** 71-80.

25. Hutter, M. & Clark, T. (1996) *J. Am. Chem. Soc.* **118,** 7574-7577.

26. Sugiyama, H. & Saito, I. (1996) *J. Am. Chem. Soc.* **118,** 7063-7068.

27. Sistare, M. F., Codden, S. J., Heimlich, G., et al. (2000) *J. Am. Chem. Soc.* **122,** 4742-4749.

28. Murphy, C. J., Arkin, M. R., Jenkins, Y., et al. (1993) *Science* **262,** 1025-9.

29. Murphy, C. J., Arkin, M. R., Ghatlia, N. D., et al. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 5315-19.

30. Arkin, M. R., Stemp, E. D. A., Pulver, S. C., et al. (1997) *Chem. Biol.* **4,** 389-400.

31. Nunez, M. E., Hall, D. B. & Barton, J. K. (1999) *Chem. Biol.* **6,** 85-97.

32. Gasper, S. M. & Schuster, G. B. (1997) *J. Am. Chem. Soc.* **119,** 12762-12771.

33. Henderson, P. T., Armitage, B. & Schuster, G. B. (1998) *Biochemistry* **37,** 2991-3000.

34. Gasper, S. M., Armitage, B., Shui, X., et al. (1998) *J. Am. Chem. Soc.* **120,** 12402-12409.

35. Henderson, P. T., Jones, D., Hampikian, G., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 8353-8358.

36. Ly, D., Sanii, L. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 9400-9410.

37. Sartor, V., Henderson, P. T. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 11027-11033.

38. Kan, Y. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 11607-11614.

39. Kan, Y. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 10857-10864.

40. Giese, B., Wessely, S., Spormann, M., et al. (1999) *Angew. Chem., Int. Ed.* **38,** 996-998.

41. Meggers, E., Michel-Beyerle, M. E. & Giese, B. (1998) *J. Am. Chem. Soc.* **120,** 12950-12955.

42. Bixon, M., Giese, B., Wessely, S., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11713-11716.

43. Meggers, E., Kusch, D., Spichty, M., et al. (1998) *Angew. Chem., Int. Ed.* **37,** 460-462.

44. Meggers, E. & Giese, B. (1999) *Nucleosides Nucleotides* **18,** 1317-1318.

45. Fukui, K. & Tanaka, K. (1998) *Angew. Chem., Int. Ed.* **37,** 158-161.

46. Saito, I., Takayama, M., Sugiyama, H., et al. (1995) *J. Am. Chem. Soc.* **117,** 6406-7.

47. Saito, I., Nakamura, T., Nakatani, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 12686-12687.

48. Moser, C. C., Keske, J. M., Warncke, K., et al. (1992) *Nature* **355,** 796-802.

49. Cross, S. H. & Bird, A. P. (1995) *Curr. Opin. Genet. Dev.* **5,** 309-14.

50. Pothukuchy, A., Mano, N., Salazar, M., et al. (2002) *submitted*.

51. Alberts, B., Johnson, A., Lewis, J., et al. (2002) *Molecular Biology of the Cell* (Garland Science, New York).

52. Hardin, C. C., Henderson, E., Watson, T., et al. (1991) *Biochemistry* **30,** 4460-72.

53. Sundquist, W. I. & Klug, A. (1989) *Nature* **342,** 825-9.

54. Amatore, C., Arbault, S., Bruce, D., et al. (2000) *Faraday Discuss.* **116,** 319-333.

55. Helbock, H. J., Beckman, K. B., Shigenaga, M. K., et al. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 288-293.

56. Setlow, R. B. (2001) *Mutat. Res.* **477,** 1-6.

57. Ambrosone, C. B. (2000) *Antioxidants & Redox Signaling* **2,** 903-917.

58. Jackson, A. L. & Loeb, L. A. (2001) *Mutat. Res.* **477,** 7-21.

59. Yokomizo, A., Ono, M., Nanri, H., et al. (1995) *Cancer Research* **55,** 4293-6.

60. Conklin, K. A. (2000) *Nutrition and Cancer* **37,** 1-18.

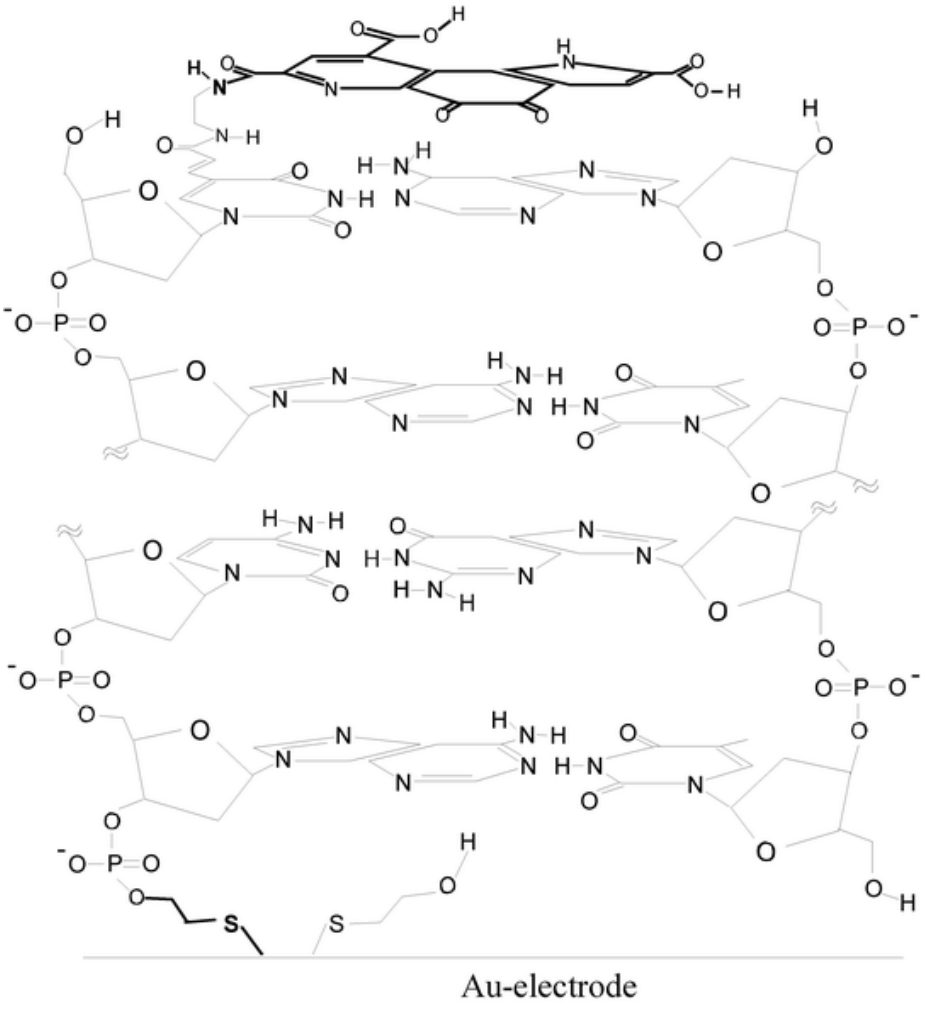


Figure 1-1. Schematic drawing of Au–S– $(CH_2)_2$–ds-oligo–NH–PQQ, the PQQ-bound, 12-base duplex oligonucleotide (3'-ACGAAGG CTGAT-5') on gold. The PQQ redox function is attached to C5 of the 5'-thymine via a C5–CH_2–CH=CH–CO–NH–CH_2–CH_2–NH_2 spacer arm. The length of the unit is ~49 ± 2 Å.

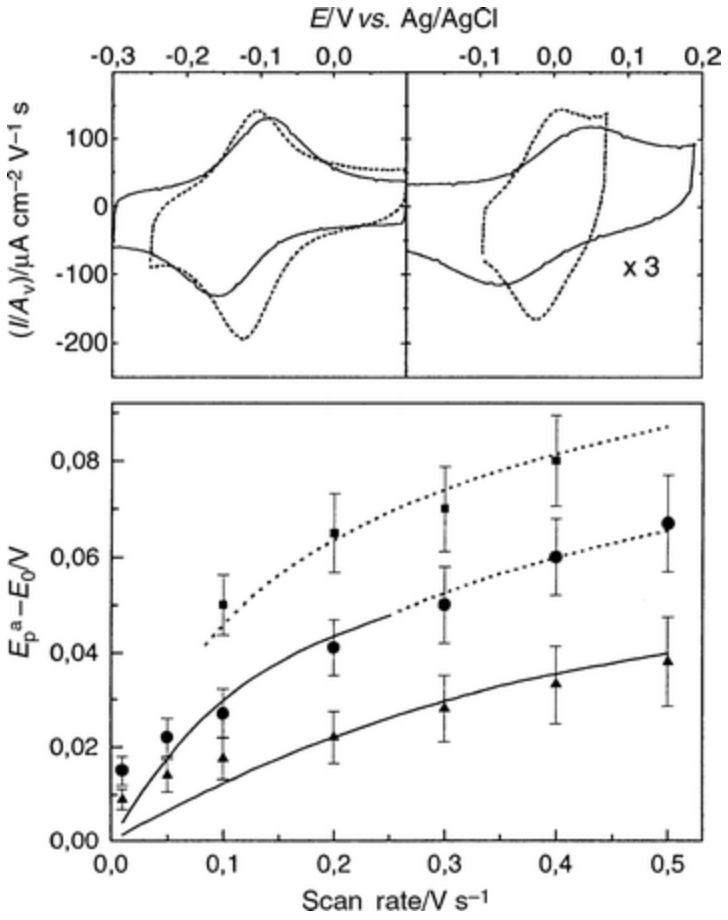


Figure 1-2. Electro-oxidation/reduction of PQQ/PQQ^{2-} functions bound to the termini of monolayers on gold. *Top*: Cyclic voltammograms of Au–S–$(CH_2)_2$–NH–PQQ (left) and Au–S–$(CH_2)_2$-*ds-oligo*-NH–PQQ/Au–S–CH_2–CH_2–OH (right), normalized for scan rate: 10 mV s^{-1} (····) and 500 mV s^{-1} (—). *Bottom*: Dependence of peak separation E_p^a–E_0 on scan rate of Au–S–$(CH_2)_2$–NH–PQQ (▲), 12 base pair duplex Au–S–$(CH_2)_2$-*ds-oligo*-NH–PQQ (●), and 12 base pair duplex with two mismatches (■). Fit to the two domains of the theoretical model of Laviron (—····) (from ref. (12)).

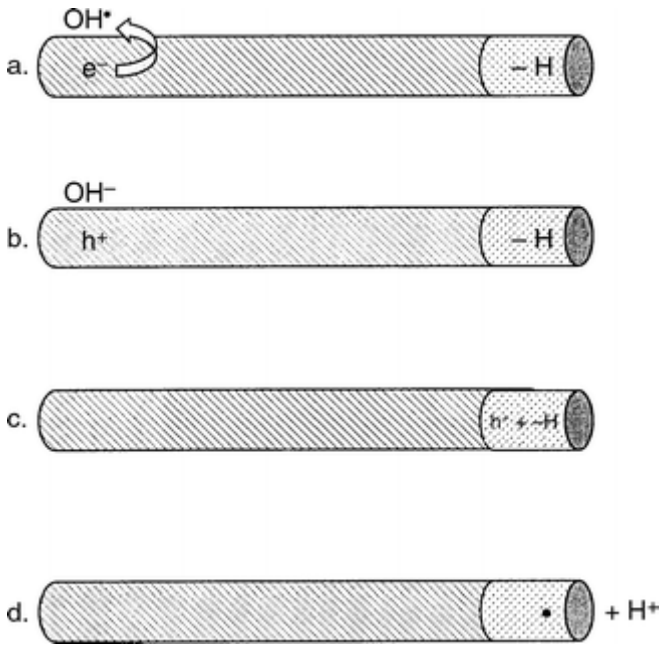


Figure 1-3. When DNA conducts, a gene could be cathodically protected by sacrificial oxidation of a neighboring G-rich domain. The gene (darker shaded) is attacked by an $^{\bullet}OH$ radical that captures an electron and is reduced to an OH^- anion (a). The capture of the electron leaves a mobile electron vacancy (hole, h^+) in the gene (b). The hole diffuses to and is captured in the more reducing G-rich domain (lighter shaded) (c), where it reacts by releasing a proton and forming a radical (d).

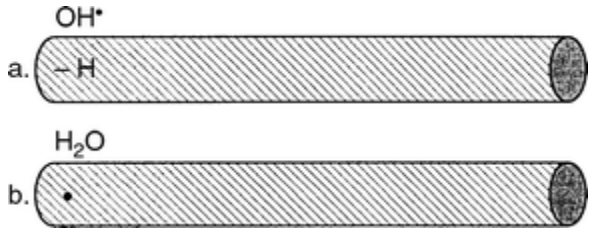


Figure 1-4. When DNA does not conduct electrons or holes, the oxidation site is the site attacked by the oxidizing agent. An •OH radical abstracts a hydrogen atom (a), producing a reactive radical, locally (b).

Chapter 2: The Non-Uniform Distribution of Guanine in Introns of Human Genes: Possible Protection of Exons against Oxidation by Proximal Intron Poly-G Sequences

ABSTRACT

Earlier studies of oligonucleotides have shown that the rate of oxidation of GGG sequences is faster than that of other nucleotide sequences. Recent studies have shown that non-dissolved, double-stranded DNA is a one-dimensional conductor of holes or electrons. GGG and longer poly-G sequences could, therefore, act as sacrificially oxidizable sinks for holes injected remotely into the DNA strand by oxidizing agents. This could cathodically protect the most essential parts of genes, their protein-coding exons. The protection of exons would be optimal if GGG sequences were concentrated near the termini of introns, flanking exons. We find, indeed, that GGG sequences are non-uniformly distributed in introns, that they are much more frequent near 5' intron termini, which flank the 3' ends of exons. We conclude that introns contain sacrificially oxidizable GGG sequences that are optimally positioned both to absorb holes injected directly into exons, and to intercept holes that could diffuse to exons from introns, which are much larger targets for oxidizing agents.

INTRODUCTION

Essential parts of metallic structures commonly are protected against corrosion. An expendable, sacrificially oxidized conductor protects an essential, preserved conductor when the two are electrically connected and residing in the same pool of electrolyte. For example, zinc coating (galvanizing) protects fuel storage tanks, steel roofs, and even nails (Figure 2-1). This long-range protection, termed "cathodic protection" by electrochemists and materials scientists, was introduced in 1824, when Sir Humphrey Dave attached zinc plates to the steel hulls of British warships to prevent their corrosion in seawater (1, 2). In the zinc/aerated seawater/steel electrochemical cell, the zinc plate is the

anode and the steel hull is the (inert) cathode. When oxygen in the seawater captures electrons from the steel cathode, electron vacancies ("holes") in the steel drift to the zinc anode. There the holes oxidize the zinc metal to Zn^{2+}, while the steel remains intact.

There are three requirements for cathodic protection. First, the anode and the cathode must be electrically connected, so that electrons or holes freely diffuse between them. (Insertion of an insulator between the anode and the cathode precludes cathodic protection.) Second, the anode and the cathode must be in electrolytic contact through an ion-transporting electrolyte. Third, the rate of oxidation of the sacrificial electrode (anode) by holes must exceed that of the protected electrode (cathode). The electrode with the fastest oxidation rate usually is the one with the lowest (most reducing) half-cell potential, because the (thermodynamic) redox potential and the (kinetic) corrosion rate usually are related through the Tafel and Butler-Volmer equations (3). According to these equations, the rate of a corrosion reaction increases exponentially with its overpotential, which is the excess potential driving it. When the anode is more reducing, the corrosion reaction is driven by a larger potential.

Earlier studies have shown that DNA domains vary in their oxidation resistance, and that non-dissolved double-stranded DNA is an electronic conductor, conducting holes and/or electrons (see Discussion). It is, therefore, plausible that some genome domains could be protected microcathodes, while others could be protective microanodes. If the protective microanodes are in introns, while the protected microcathodes are exons (Figures 2-1), DNA that does not code for proteins could be sacrificed to protect DNA that does (Figure 2-2). Because introns comprise 95% of human genes (4), it is likely that they, not exons, comprise most of the sites at which oxidizing agents inject holes. If chromosomal DNA conducts holes over tens of base pairs, then holes diffusing or drifting from introns to exons could oxidize exons. This indirect attack could be mitigated if there are sacrificially oxidizable domains at the termini of introns, proximal to exons. Then, many of the holes injected into introns could be intercepted and trapped by the pro-

tective domains (which would be oxidized) before they could reach exons (Figure 2-3). If some of the holes injected into exons could reach and oxidize the proximal protective domains, the damage to exons from direct attack by oxidizing agents could be reduced. The damage to the sacrificial domains eventually would be recognized and repaired by well-known systems (5, 6).

Heller has hypothesized that genes could be protected against oxidation-caused mutation by sacrificially oxidizable, neighboring G-rich domains (7). In this hypothesis, the resting genome is envisioned as a two-component (adenine-thymine (AT) and guanine-cytosine (GC)), electronically conductive filamentary alloy. If genes conduct holes, and if holes injected into their exons (when oxidants capture electrons) drift to and are trapped by oxidizing neighboring G-rich domains, then the likelihood of mutation could be reduced, according to this hypothesis.

In this study we show that triplet guanine sequences (-GGG-), which are reducing with respect to all other nucleotide triplets, doublets and singlets, are non-uniformly distributed in the introns of human genes. They are concentrated near the termini of introns, flanking exons. This is especially true in shorter introns. Thus, they are optimally situated to prevent many of the holes injected into introns from reaching exons and to act as effective sinks of holes injected directly into exons by oxidizing agents.

METHODS

The August 2001 release of the human genome (8) was analyzed. For each exon (coding domain sequences or CDS), the locations of its first and last bases within its contig (contiguous region), its strand (the listed strand or its complement), and its position (first, middle or last) within its gene were extracted from the GenBank files. Within each contig, these coordinates were sorted in ascending order by location, and overlapping or duplicated exons were merged where possible. (Less than 1% of overlapping exons were on different strands or in different positions and could not be counted.)

Using these coordinates, each nucleotide of the genome was classified as gene or non-gene, and if gene, as exon or intron. Each nucleotide was identified as adenine (A), thymine (T), guanine (G), cytosine (C) or other/unknown (N). The guanine nucleotides were further classified as singlet, doublet or triplet. Singlet G were G not neighboring another G; doublet G were G bounded by only one other G, as in TGGA; triplet G were central G bounded on each side by G, as in TGGGA. The sum of singlet, doublet and triplet G was recorded as total G. The exon and intron totals of total, doublet and triplet Gs were recorded for averages (weighted by nucleotides, not segment). The distances to the nearest 5' and 3' exon boundaries were calculated for each nucleotide. For intron nucleotides within 250 bp of their 5' or 3' boundaries (the 3' and 5' boundaries of their adjacent exons) and for exon-nucleotides within 250 bp of their boundaries, the base, the minimum distance and the closest boundary were recorded. When an intron or exon nucleotide was within 100 bp of both of its boundaries, it was considered proximal to both boundaries. (This occurred in (rare) introns and (common) exons that were less than 200 bp long.) Such nuclides were recorded in a special category: short. The results were summed over the genome to derive the average mol fractions of total, doublet and triplet G, at each recorded position. The average mol fraction of, for example, doublet G at 50 bp from the 5' end of introns was calculated as the number of G in doublet G sequences at this position divided by the number of all nucleotides at the same position. The overall mol fraction of, for example, doublet G in introns was calculated as the number of doublet G in all intron positions divided by the number of all intron-nucleotides. While the average mol fraction was a function of position, the overall mol fraction was not. Results for the coding DNA strand (the strand that matches the pre-mRNA produced) were used to infer the template DNA strand (the strand that templates pre-mRNA production). One implied the other, because G on one strand pairs with C on the other strand, and the 5' end of one strand pairs with the 3' end of the other strand.

RESULTS AND DISCUSSION

Distribution of Guanine in Introns. The average mol fractions of total G, doublet G and triplet G are elevated, relative to their overall intron values, within 100 bp of the 5' terminus of introns on the coding (+) DNA strand and near both the 5' and the 3' termini of introns on the template (–) DNA strand (Table 2-1 and Figures 2-4 to 2-11). They are depressed near the 3' terminus of introns on the coding strand. (This intron terminus is rich in pyrimidine nucleotides (T and C) (5, 6).) The relative elevations increase in the order G<GG<GGG (Figures 2-4 to 2-11). The average mol fraction of doublet G is above and that of triplet G is substantially above what would be calculated from probability and the average mol fraction of total G. The numbers of total G and doublet G in 100 bp of exons overall exceed those of the 5' end of introns, but the number of triplet G in 100 bp of the 5' end of introns exceed those of exons overall (Table 2-1). The number triplet G in the ends of short introns substantially exceed those of exons overall. (For the definitions of average and overall mol fractions, see the methodology section.)

Table 2-1 lists the average numbers of total G, doublet G and triplet G in 100 bp of intron ends, introns overall, exon ends and exons overall. The length, 100 bp, is roughly half of the average length of exons on chromosome 1 (247 bp). The average number of triplet G in 100 bp at the 5' end of introns, for example, is the sum of the average mol fractions of triplet G from position 1 through 100. The average number of triplet G in 100 bp of introns overall, for example, is overall mol fraction of triplet G multiplied by 100. Figures 2-4 to 2-11 show the average mol fractions of total G, singlet G, doublet G and triplet G in introns as functions of the distance, expressed in base pairs, from the intron/exon boundary.

On both DNA strands, the average mol fractions of total G, doublet G and triplet G within 100 bp (near) the 5' terminus of introns are elevated relative to the overall mol fractions of introns. On the coding DNA strand, the average mol fraction of total G is a broad plateau over at least 200 bp that slowly decays toward the overall intron value

(Figure 2-4). The first six nucleotides at the 5' end scatter from this trend, because they are part of the consensus sequence for pre-mRNA splicing (5, 6). On the template DNA strand, total G peaks at about 7 bp from the boundary, slopes down to the overall intron value at about 60 bp and then is flat (Figure 2-5). This purine-rich region coincides with the pyrimidine-rich region on the coding strand. On both strands, the average mol fractions of total G and doublet G near the 3' end of introns do not differ significantly from the overall intron values; the average mol fraction of triplet G is elevated (Table 2-1). On the coding strand, total G increases from a low near the boundary to a broad peak at about 60 bp that tails toward the overall intron value (Figure 2-4). The low corresponds to the pyrimidine-rich (A- and G-poor) region. The first three nucleotides at the 3' end scatter from this trend, because they are part of the pre-mRNA splicing sequence. On the template strand, total G rises from a low near the boundary to a small peak at about 12 bp that tails toward the overall intron value (Figure 2-5). On both strands, the average mol fractions of total G, doublet G and triplet G near the ends of short introns are elevated relative to the overall mol fractions of introns (Table 2-1). This elevation is substantially greater than that at the 5' end. On the coding strand, total G is a broad plateau over at least 100 bp that slowly decays toward the overall intron value (Figure 2-4). On the template strand, total G has a peak at about 8 bp from the boundary, slopes down to the overall intron value at about 60 bp and then is flat (Figure 2-5).

The average mol fractions of doublet G and triplet G generally vary with distance like the corresponding average mol fractions of total G, e.g., Figures 2-4, 2-8 and 2-10. However, the relative elevation of doublet G is greater than that of total G, and the relative elevation of triplet G is greater than that of doublet G. Specifically, the average mol fractions of total G, doublet G and triplet G at 30 bp from the 5' terminus of introns are 17%, 28% and 130% greater than the overall intron values on the coding strand. The ends of short introns show even greater elevation. On all intron ends, the average mol fraction of doublet G (X_{GG}) is above what would be calculated from probability (eq. 2-1)

and the total G mol fraction ($X_{\Sigma G}$). This difference is small on the coding strand (Figure 2-8) and appreciable on the template strand (Figure 2-9). On all intron ends, the triplet G mol fraction (X_{GGG}) exceeds what would be calculated (eq. 2-2). This difference is large on both strands (Figures 2-10 and 2-11). Doublet G and triplet G are above what would be calculated on introns overall, also.

$$X_{GG} = 2(1 - X_{\Sigma G})(X_{\Sigma G})^2 \qquad (2\text{-}1)$$

$$X_{GGG} = (X_{\Sigma G})^3 \qquad (2\text{-}2)$$

Table 2-1 lists the average numbers of total G, doublet G and triplet G in 100 bp of intron ends, introns overall, exon ends and exons overall. The numbers of total G and doublet G of exons overall exceed those of the 5' end of introns by 4% and 6%, respectively, on the coding strand and by 15% and 24% on the template strand. However, the number of triplet G of the 5' end of introns exceed those of exons overall by 56% and 7% on the coding and template strands, respectively. The numbers of total G, doublet G and triplet G of the ends of short introns exceed those of exons overall by 9%, 21% and 154%, respectively, on the coding strand and by 0%, 4% and 67% on the template strand. The numbers of total G, doublet G and triplet G of exons overall exceed those of the introns overall and the 3' end of introns by 25%, 39% and 5%, respectively, averaged over the two strands.

Guanine Oxidation and Electronic Conduction in DNA. Evidence for preferential oxidation of guanine nucleotides comes from studies of the electro-oxidation potentials and kinetics of nucleotides and polynucleotides, the ionization potentials of these molecules, the chemical and photochemical oxidation of 30 to 100 base oligonucleotides, the oxidation of long DNA strands in vitro, and the oxidation of genes in living cells. These studies show that G is the most readily oxidized mononucleotide, but with standard electrode potential of 1.04 V, only strong oxidizers oxidize G. GG is more readily oxi-

dized than G, and GGG is the most readily oxidized nucleotide triplet. G is reducing by approximately 250 mV relative to T, 280 mV relative to A and 400 mV relative to C (9). Theory suggests that in the G:C base pair G is even more reducing (10). Correspondingly, G is electro-oxidized at a more reducing potential and more rapidly than other nucleotides (11, 12). The rate of redox-couple mediated electro-oxidation of poly-G sequences exceeds that of isolated G (13). This is also true in long DNA strands (14). In vitro, DNA is cleaved preferentially at G by the Fenton reagent (H_2O_2–Fe^{2+}), which produces hydroxyl radicals (15). In living fibroblasts, G of G:C base pairs are preferentially oxidized by peroxyl radicals (16). O'Neill et al. state that "damage (by 193 nm light and photooxidants) at -GG- sites is significantly greater than at single guanine sites, presumably reflecting the lower ionization potentials of -GG- sites" (17). In a study of the relative reactivates of 5'-TXGYT-3' sequences in 30-mer B-DNA to photo-induced one-electron oxidation, Saito et al. state that the sequence TGGGT is 1.4 to 3.9 times more reactive then the various TGGYT sequences, which are 1.8 to 20 times more reactive than the various TXGYT sequences (18). Yoshioka et al. state that "5'-GGG-3' triplets act as a more effective trap in hole migration than 5'-GG-3'doublets" (19).

Holes injected at sites of oxidative attack react with G of G:C base pairs of dissolved DNA approximately 30 bp remote from the attack site (20-26). This suggests that double-stranded DNA is an electronic conductor, albeit not necessarily a good one, even when dissolved in water. Fink and Schönenberger (27), Kasumov et al. (28) and Okahata et al (29) state that solid (non-dissolved) DNA is a one-dimensional conductor of holes and/or electrons. The electronic conductivity of non-dissolved DNA is not without controversy (30, 31), possibly because the measurement is very difficult (see (32)).

Oxidative Attack on Genes. Powerful oxidizers, like nitric oxide, hydrogen peroxide and peroxonitrite ($ONOO^-$) abound in cells (33). They are well distributed, because their diffusion distance of $>10^{-2}$ cm (for $D \sim 10^{-5} cm^2 s^{-1}$ and $t \sim 20s$) is longer than the diameter of many cells. Amatore et al. state that the transient concentration of nitric

oxide reaches 0.4 mM in fibroblasts, and this high concentration is sustained for >20 s (33). May et al. state that nitric oxide permeates rapidly through the lipid bilayer membrane of red blood cells (34), making it unlikely that the nuclear lipid bilayer membrane can shield the genome from this oxidant. Measurements of the reactions of nucleotides with strong oxidants produced by flash photolysis indicate bimolecular rate constants of 10^7 to 10^9 M^{-1} s^{-1} (35).

The impact of G oxidation depends on the nature of the damage. If, as occurs more frequently (36), G is oxidized to 8-oxo-7,8-dihydroguanine (8-oxoG), the lesion blocks mRNA transcription in mammalian cells and signals for rapid transcription-coupled repair (37). 8-oxoG is a mutagenic lesion, because it does not block DNA polymerases in prokaryotes and eukaryotes, and it pairs with A as well as C(38). If the G is oxidized not to 8-oxoG, but to 2,6-diamino-4-hydroxy-5-formamidopyrimidine (Fapy-G), the lesion may block DNA replication. (Methylated Fapy-G, a related nucleotide, is a potent inhibitor of DNA polymerases in E. Coli and phage T4 (38).) Oxidation of G to Fapy-G usually is a lethal mutation, whether in exons or introns.

Boiteux et al. state that the lesions resulting from attack by reactive oxygen species under physiological conditions threaten genome integrity (38). According to Sekiguchi et al. "oxygen radicals are produced through normal cellular metabolism and formation of such radicals is further enhanced by ionizing radiation and by various chemicals. The oxygen radicals attack nucleic acids and generate various modified nucleotides in DNA. Among them, 8-oxo-7,8-dihydroguanine (8-oxoG) is the most abundant, and appears to play critical roles in carcinogenesis and in aging. 8-OxoG related mutagenesis may account for a considerable number of spontaneous mutations in mammalian cells" (36). Site-specific DNA damage at the GGG sequence accelerates telomere shortening, associated with the aging of cells (39). Oxidative stress of cells correlates with their aging (40).

CONCLUSIONS

Unintended oxidation, hydrolysis and methylation produce 10^3 to 10^4 DNA lesions per day in human cells (5). If 10^3 of these are oxidative, then in the human genome, where genes comprising 28% of nucleotides (4), ~300 oxidative lesions occur in genes each day. Because the number of human genes is ~30000, the average period between oxidative lesions in a particular gene is ~100 days. With ~10 introns per gene, the average period between lesions in a particular intron is ~1000 days. (Assuming 10^4 oxidative lesions per day, the average period between lesions in a particular intron is ~100 days.) Thus, one triplet G, serving as a sacrificial anode, could protect an exon for 100 days or longer by trapping one hole.

The average mol percentage of triplet G, the most reducing of the mono-, di- and tri-nucleotides, is elevated near the 5' termini of introns (proximal to the 3' termini of exons) on both strands. Within 100 bp of the 5' termini, it averages 2.7% and 2.2% on the coding and template strands, respectively. On short introns (<100 bp), it is even higher: 4.4% and 3.5%. It exceeds the overall mol percentages of triplet G of introns (1.4%) and exons (1.7% and 2.1%). The termini of introns are the optimal locations for triplet G acting as sacrificial anodes protecting protein-coding exons from oxidative damage. In these locations, triplet G could intercept holes injected into introns from reaching exons and act as effective sinks of holes injected directly into exons by oxidizing agents. Though there are only about two triplet G within 100 bp of the 5' terminus of an average intron, intron oxidations likely are separated by hundreds of days. Thus, a small number of triplet G could provide long-term protection. The damaged triplet G would be recognized and repaired by well-known systems. Shifting oxidative damage to triplet G near intron termini could adversely affect pre-mRNA transcription and splicing and DNA replication and repair.

ACKNOWLEDGMENTS

Ting Chen of the University of Texas at Austin did a preliminary survey of the human genome for poly-G sequences. Kristine McAndrews of the University of Texas at Austin helped with data analysis. Professors George Georgiou, Brent L. Iverson, G. Barrie Kitto and Edward M. Marcotte of the University of Texas at Austin, and the anonymous reviewers made very helpful comments. The National Science Foundation provided financial support for A.H. The National Science Foundation and the Richard J. Lee Endowed Graduate Fellowship in Engineering provided financial support for K.A.F. The National Institutes of Health Biotechnology Training Grant provided the computer used for data analysis.

REFERENCES

1. Shackelford, J. F. (1996) *Introduction to Materials Science for Engineers* (Prentice-Hall, Upper Saddle River, NJ).

2. Stansbury, E. E. & Buchanan, R. A. (2000) *Fundamentals of electrochemical corrosion* (ASM International, Materials Park, OH).

3. Bard, A. J. & Faulkner, L. R. (2001) *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York).

4. Baltimore, D. (2001) *Nature* **409,** 814-816.

5. Alberts, B., Bray, D., Lewis, J., et al. (1994) *Molecular Biology of the Cell* (Garland, New York).

6. Lodish, H., Berk, A., Zipursky, S. L., et al. (1999) *Molecular Cell Biology* (W.H. Freeman & Co., New York).

7. Heller, A. (2000) *Faraday Discuss.* **116,** 1-13.

8. Collaborators, I. H. G. (2001) (National Center for Biotechnology Information, Washington, DC).

9. Faraggi, M., Broitman, F., Trent, J. B., et al. (1996) *J. Phys. Chem.* **100,** 14751-14761.

10. Hutter, M. & Clark, T. (1996) *J. Am. Chem. Soc.* **118,** 7574-7577.

11. Brett, C. M. A., Brett, A. M. O. & Serrano, S. H. P. (1994) *J. Electroanal. Chem.* **366,** 225-31.

12. Tomschik, M., Jelen, F., Havran, L., et al. (1999) *J. Electroanal. Chem.* **476,** 71-80.

13. Sistare, M. F., Codden, S. J., Heimlich, G., et al. (2000) *J. Am. Chem. Soc.* **122,** 4742-4749.

14. Szalai, V. A. & Thorp, H. H. (2000) *J. Phys. Chem. B* **104,** 6851-6859.

15. Henle, E. S., Han, Z., Tang, N., et al. (1999) *J. Biol. Chem.* **274,** 962-971.

16. Rodriguez, H., Valentine, M. R., Holmquist, G. P., et al. (1999) *Biochemistry* **38,** 16578-16588.

17. O'Neill, P., Parker, A. W., Plumb, M. A., et al. (2001) *J. Phys. Chem. B* **105,** 5283-5290.

18. Saito, I., Nakamura, T., Nakatani, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 12686-12687.

19. Yoshioka, Y., Kitagawa, Y., Takano, Y., et al. (1999) *J. Am. Chem. Soc.* **121,** 8712-8719.

20. Bixon, M., Giese, B., Wessely, S., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11713-11716.

21. Lewis, F. D., Liu, X., Liu, J., et al. (2000) *J. Am. Chem. Soc.* **122,** 12037-12038.

22. Meggers, E., Kusch, D., Spichty, M., et al. (1998) *Angew. Chem., Int. Ed.* **37,** 460-462.

23. Nunez, M. E., Noyes, K. T., Gianolio, D. A., et al. (2000) *Biochemistry* **39,** 6190-6199.

24. Nunez, M. E., Rajski, S. R. & Barton, J. K. (2000) *Methods Enzymol.* **319,** 165-188.

25. Sanii, L. & Schuster, G. B. (2000) *J. Am. Chem. Soc.* **122,** 11545-11546.

26. Schuster, G. B. (2000) *Acc. Chem. Res.* **33,** 253-260.

27. Fink, H.-W. & Schonenberger, C. (1999) *Nature* **398,** 407-410.

28. Kasumov, A. Y., Kociak, M., Gueron, S., et al. (2001) *Science* **291,** 280-282.

29. Okahata, Y., Kobayashi, T., Tanaka, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 6165-6166.

30. Braun, E., Eichen, Y., Sivan, U., et al. (1998) *Nature* **391,** 775-778.

31. Porath, D., Bezryadin, A., De Vries, S., et al. (2000) *Nature* **403,** 635-638.

32. Friedman, K. A. & Heller, A. (2001) *J. Phys. Chem. B* **105,** 11859-11865.

33. Amatore, C., Arbault, S., Bruce, D., et al. (2000) *Faraday Discuss.* **116,** 319-333.

34. May, J. M., Qu, Z.-C., Xia, L., et al. (2000) *Am. J. Physiol.* **279,** C1946-C1954.

35. Rogers, J. E. & Kelly, L. A. (1999) *J. Am. Chem. Soc.* **121,** 3854-3861.

36. Sekiguchi, M. & Hayakawa, H. (1998) *Contemp. Cancer Res.* **2,** 85-93.

37. Le Page, F., Kwoh, E. E., Avrutskaya, A., et al. (2000) *Cell* **101,** 159-171.

38. Boiteux, S. & Laval, J. (1997) *Base Excision Repair DNA Damage***,** 31-44.

39. Oikawa, S., Tada-Oikawa, S. & Kawanishi, S. (2001) *Biochemistry* **40,** 4763-4768.

40. Finkel, T. & Holbrook, N. J. (2000) *Nature* **408,** 239-247.

Table 2-1. Mean mol percentages of total G, G, GG, and GGG in introns and exons overall, and in ends of long and short introns and exons.

		Overall	5' End (Long)	3' End (Long)	5' or 3' End (Short)
Total G	+	21.	25.	21.	29.
Intron	–	20.	22.	21.	26.
Total G	+	26.	27.	26.	26.
Exon	–	26.	27.	26.	25.
G	+	12.	12.	11.	11.
Intron	–	11.	11.	10.	11.
G	+	13.	14.	13.	13.
Exon	–	13.	13.	13.	12.
GG	+	8.3	11.	7.8	14.
Intron	–	8.0	9.1	8.5	12.
GG	+	11.	12.	11.	11.
Exon	–	11.	12.	12.	11.
GGG	+	1.4	2.7	1.7	4.4
Intron	–	1.4	2.2	2.0	3.5
GGG	+	1.7	1.8	1.7	1.6
Exon	–	2.1	2.2	2.2	1.8

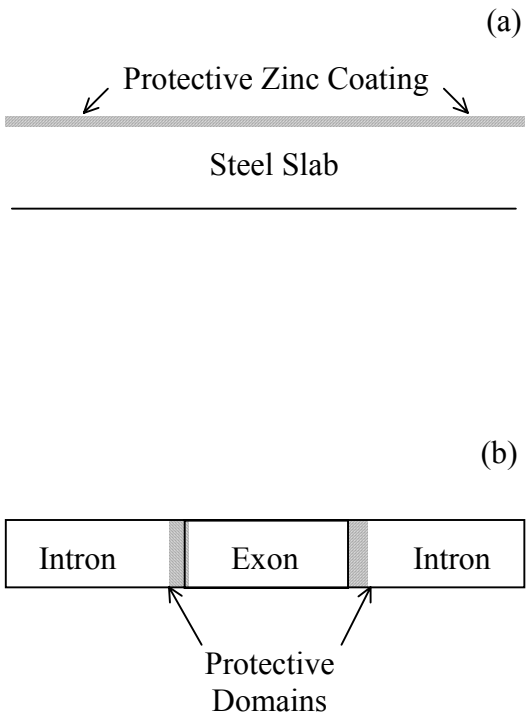


Figure 2-1. Cathodic protection of (a) steel by zinc coating (shaded) and (b) exon by flanking intron domains (shaded).

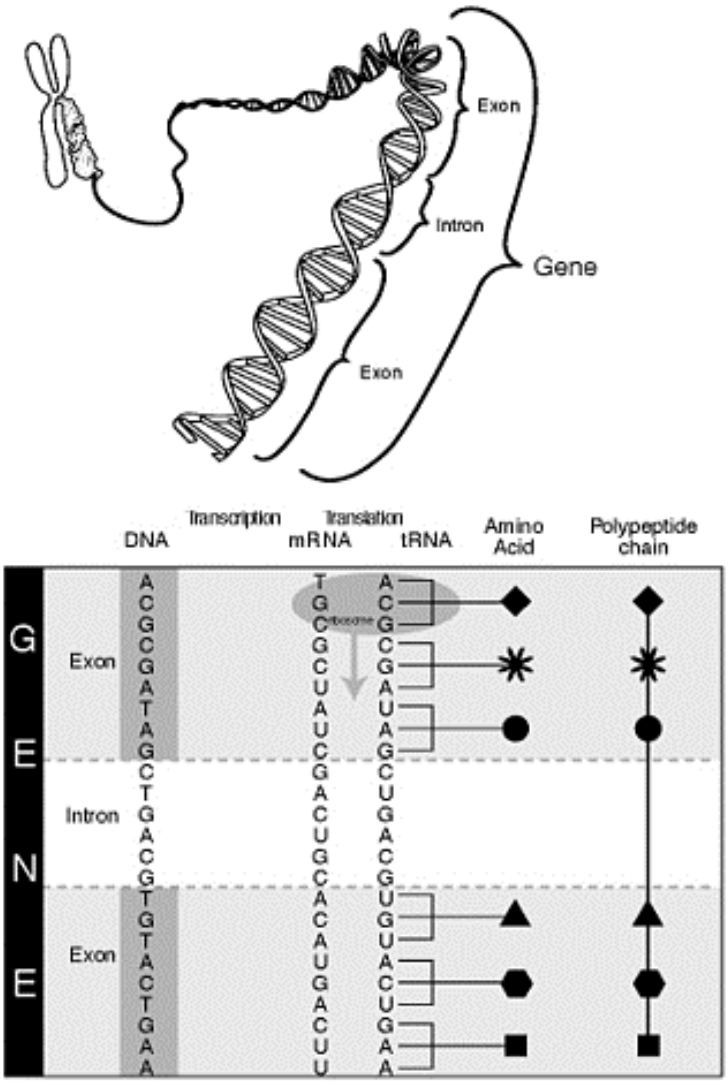


Figure 2-2. A strand of DNA, schematically magnified to show the relationships of chromosome to gene to exons and introns to DNA bases that code for protein (courtesy NIH).

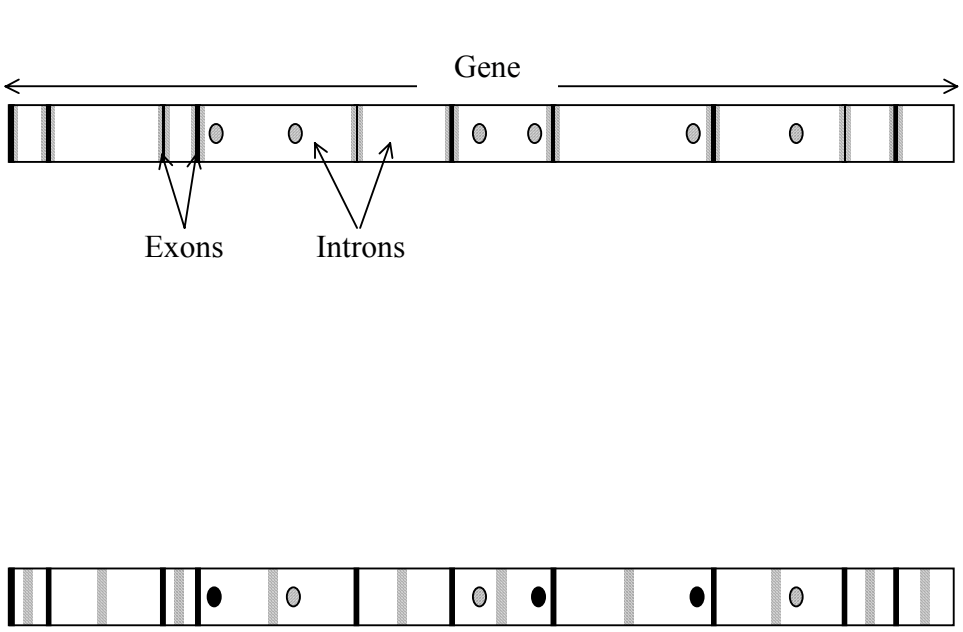


Figure 2-3. Proposed protection of exons from holes injected into a gene attacked by oxidizing agents when protective domains are (a) proximal and (b) remote from exons. Notes: Non-shaded areas represent introns; shaded areas are protective microanodes; solid black areas represent exons. Solid black circles represent holes reaching exons; shaded circles are holes intercepted by protective domains.

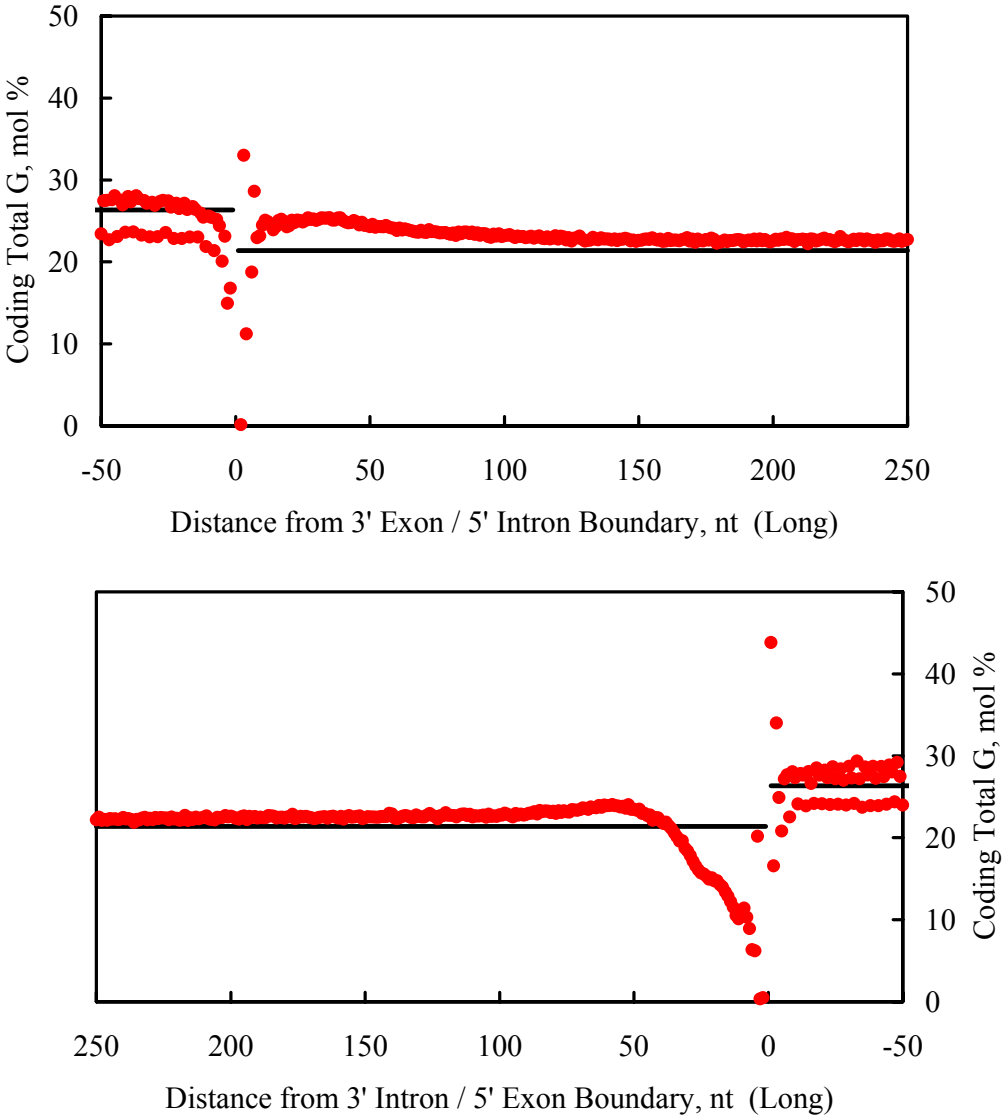


Figure 2-4. Local average (red circles) and overall average (black lines) mol percentages of total G on the coding strand vs. distance from exon/intron boundaries for long (≥100 nt) and short introns and exons.

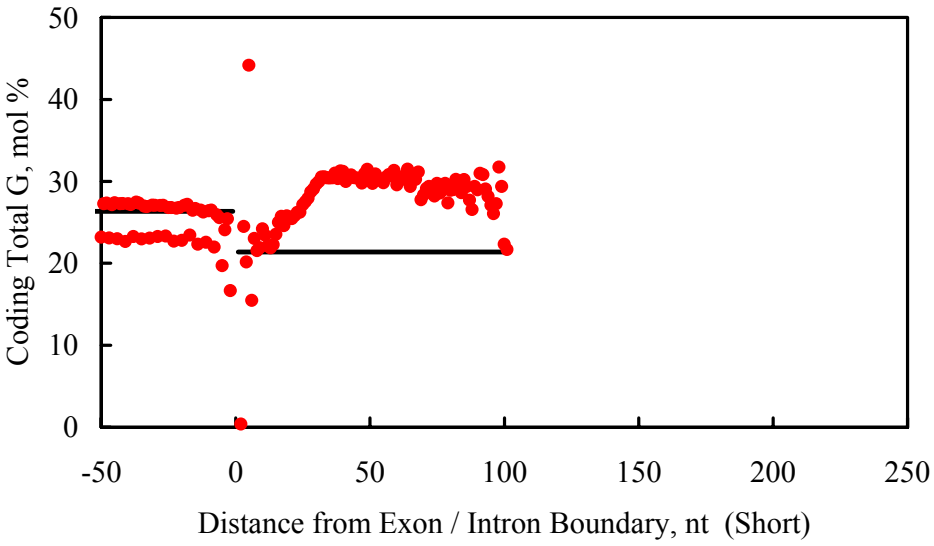


Figure 2-4. Continued.

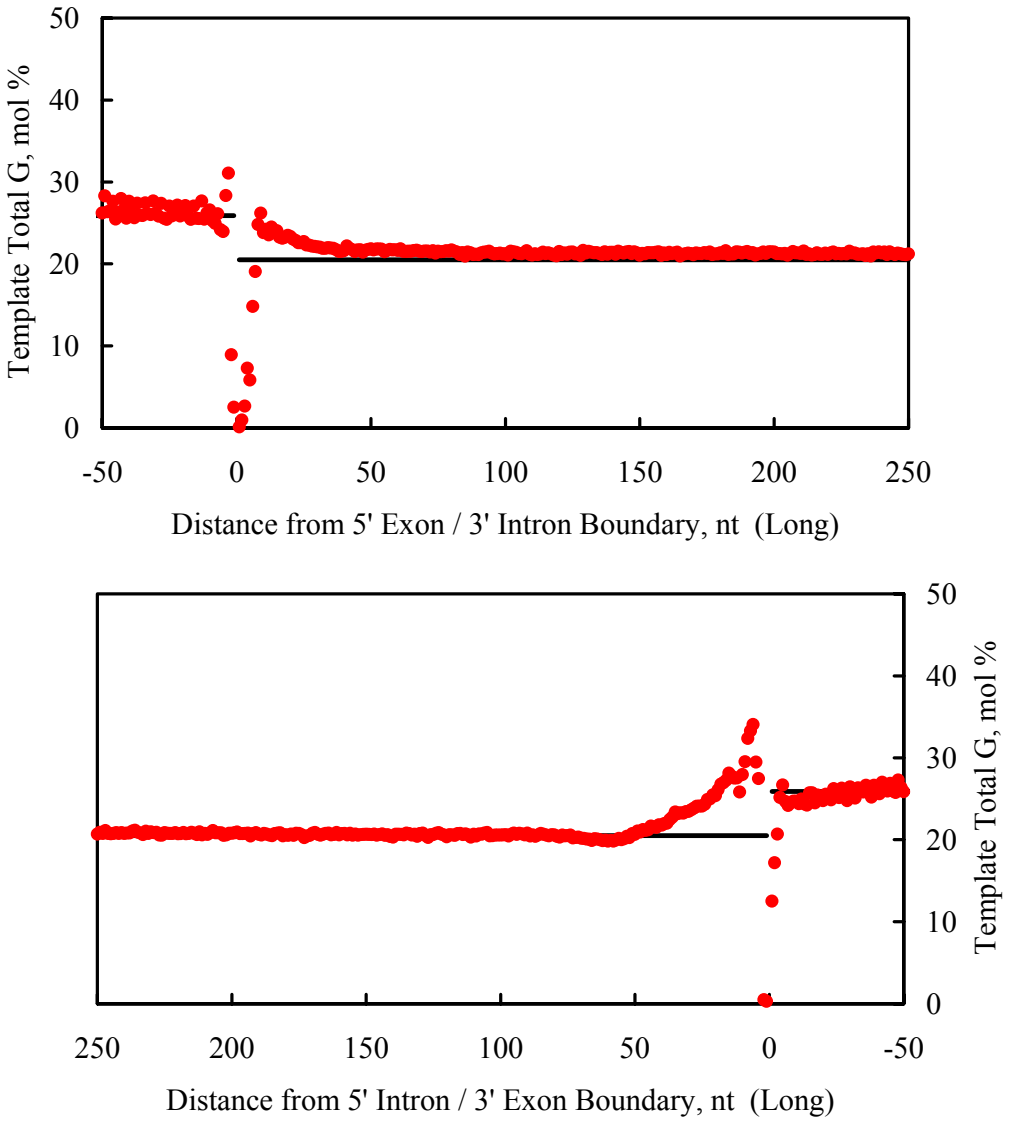


Figure 2-5. Local average and overall average mol percentages of total G on the template strand vs. distance from exon/intron boundaries for long and short introns and exons.

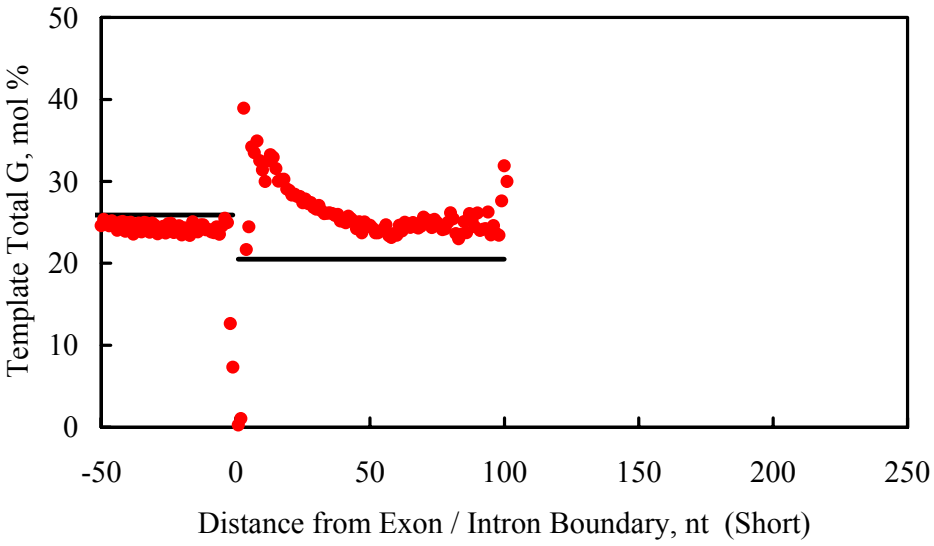


Figure 2-5. Continued.

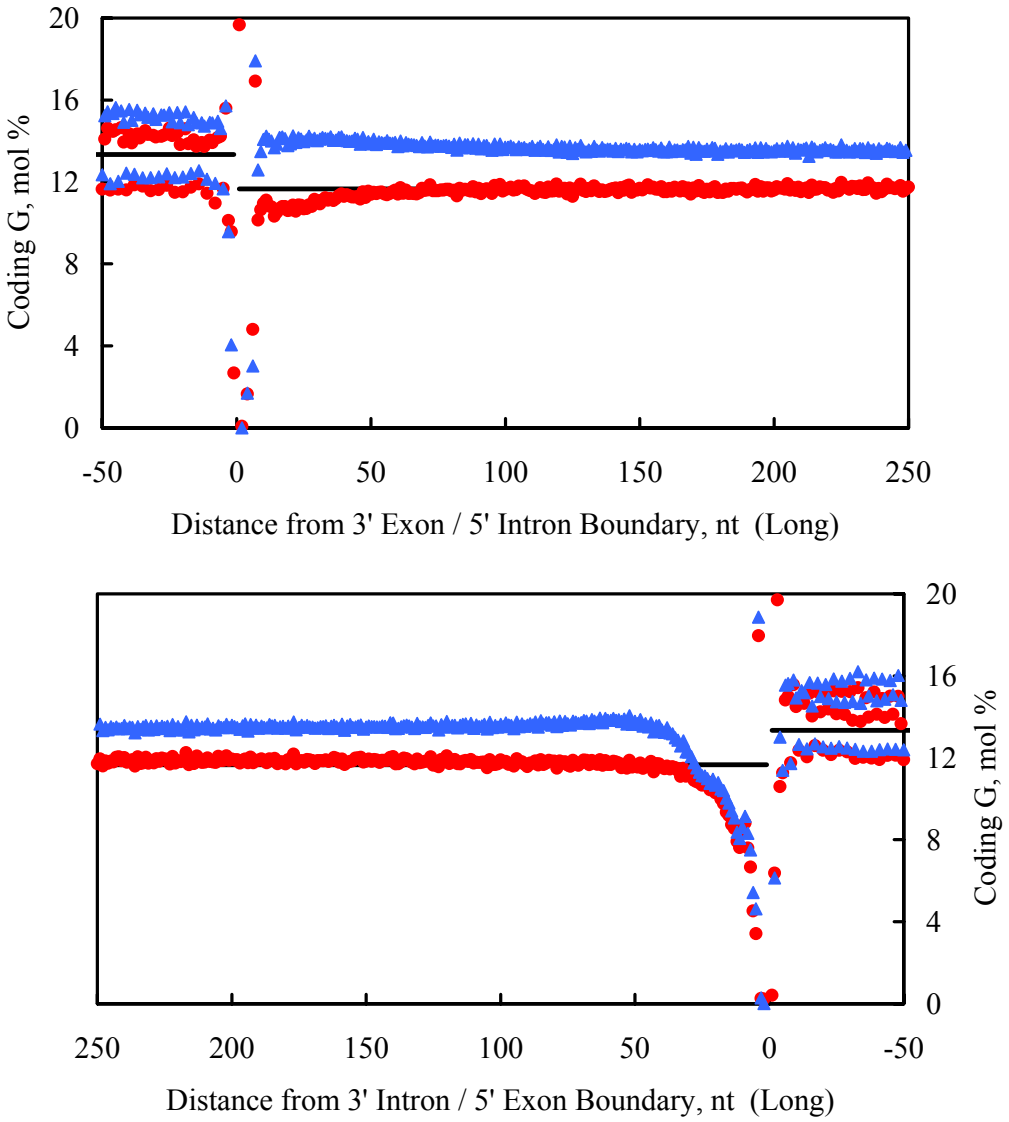


Figure 2-6. Observed (red circles) and probability-predicted (blue triangles) local average and observed overall average (black lines) mol percentages of G on the coding strand vs. distance from exon/intron boundaries for long (≥100 nt) and short introns and exons.

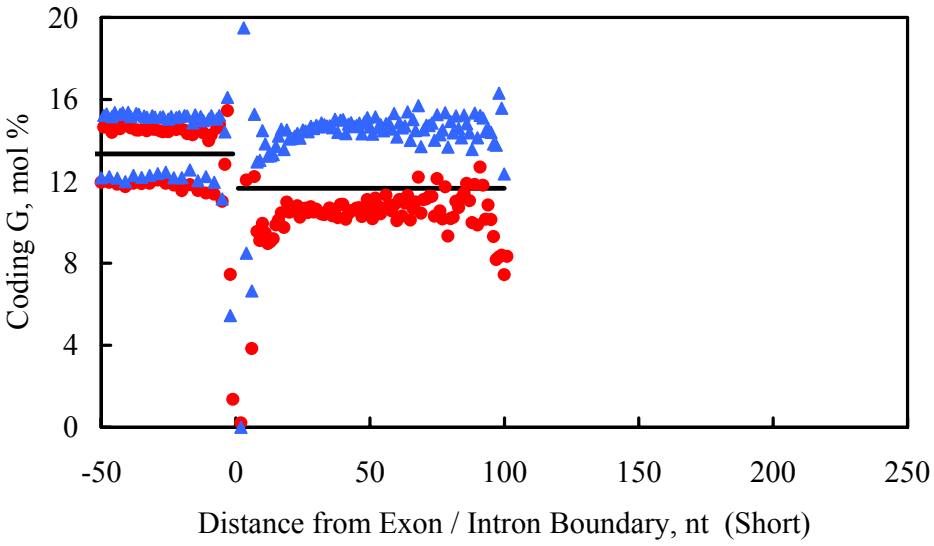


Figure 2-6. Continued.

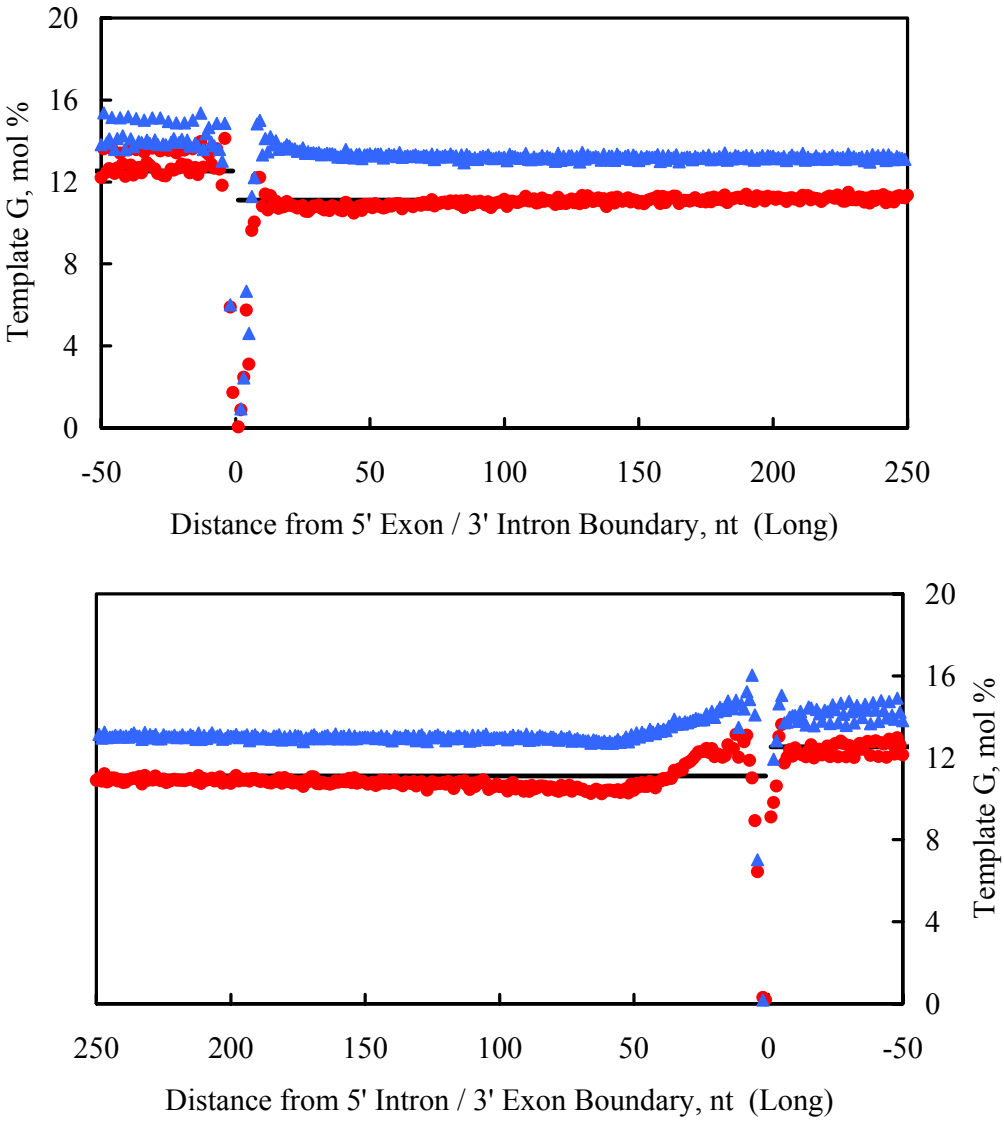


Figure 2-7. Observed and probability-predicted local average and observed overall average mol percentages of G on the template strand vs. distance from exon/intron boundaries for long and short introns and exons.

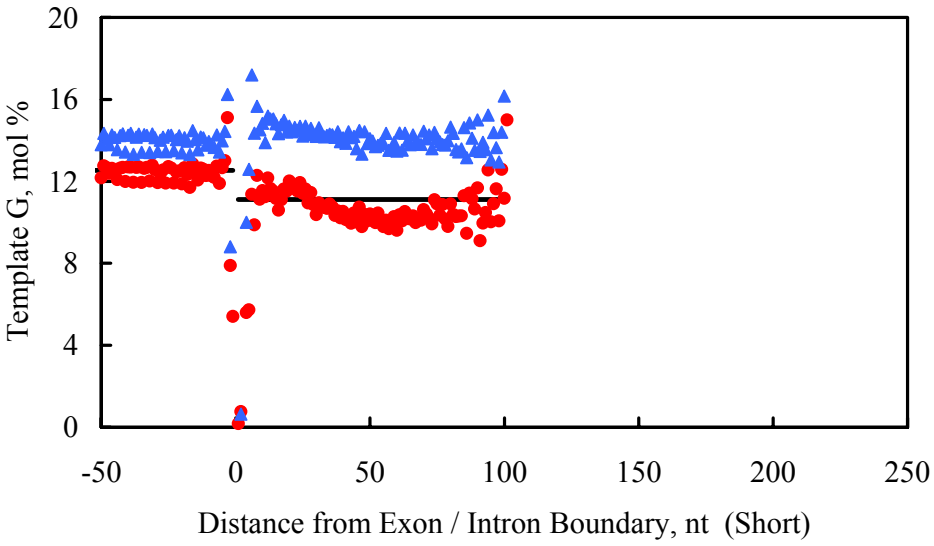


Figure 2-7. Continued.

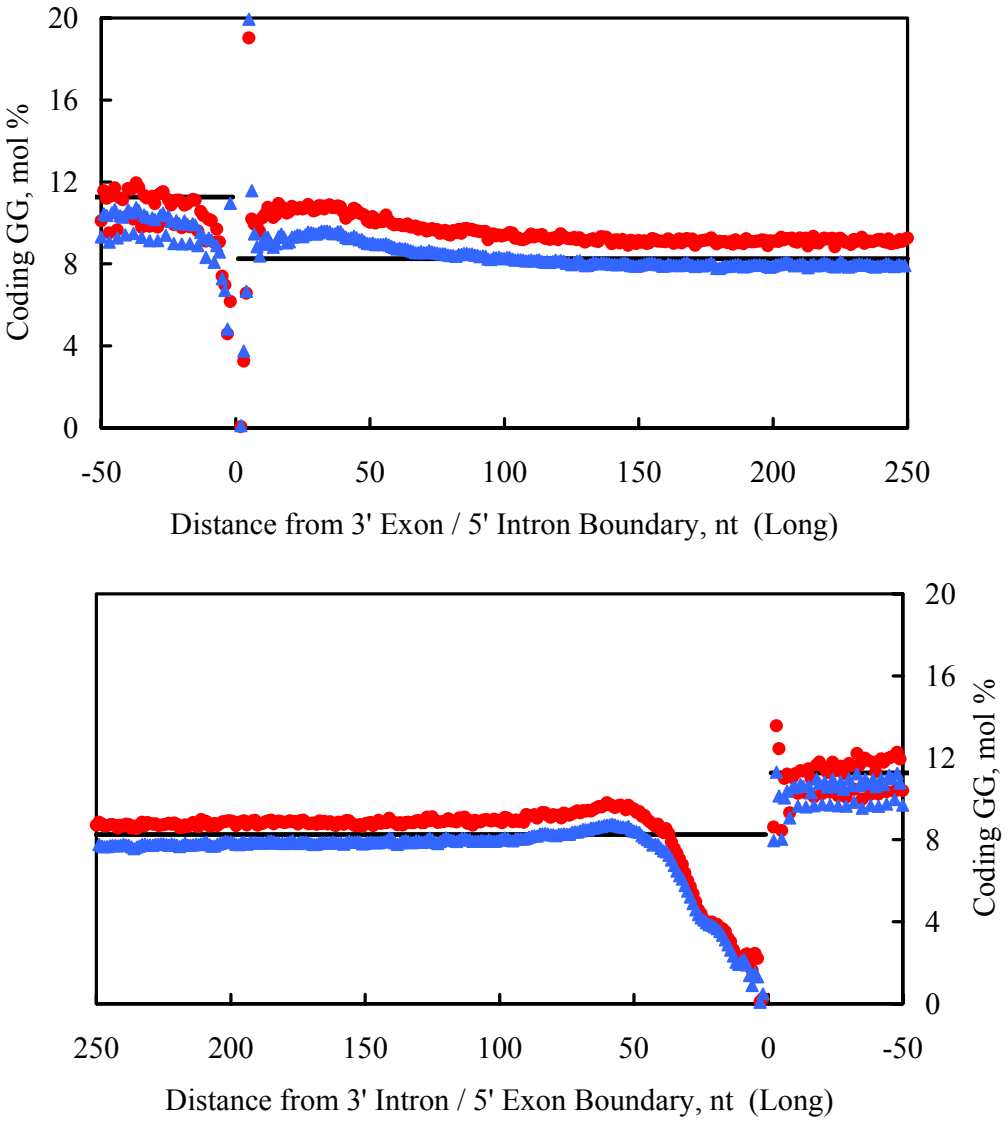


Figure 2-8. Observed and probability-predicted local average and observed overall average mol percentages of GG on the coding strand vs. distance from exon/intron boundaries for long and short introns and exons.

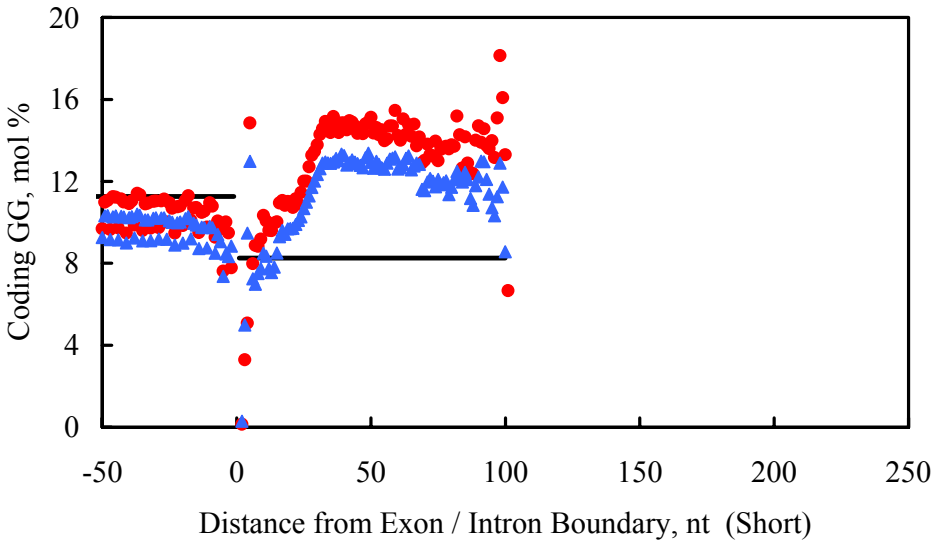


Figure 2-8. Continued.

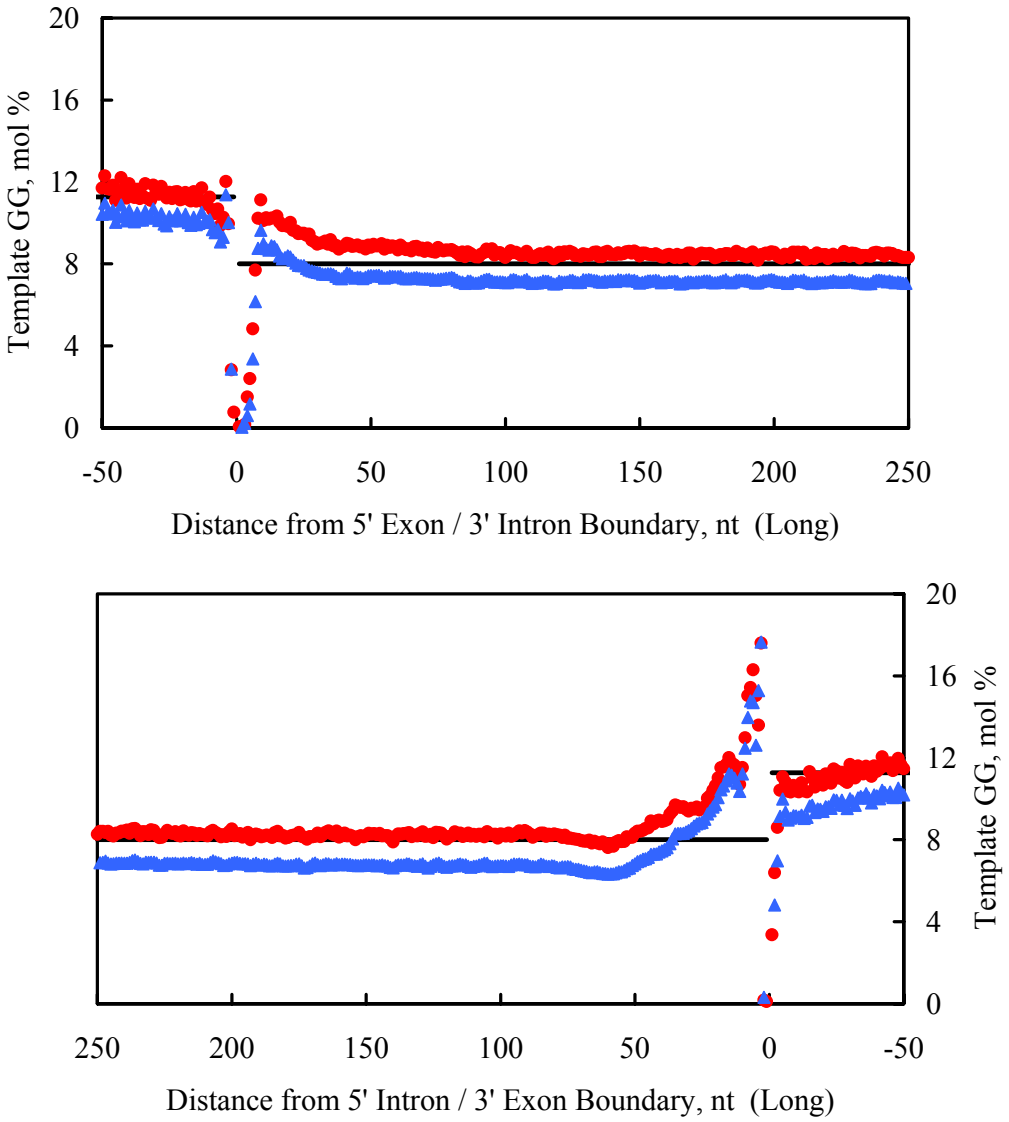


Figure 2-9. Observed and probability-predicted local average and observed overall average mol percentages of GG on the template strand vs. distance from exon/intron boundaries for long and short introns and exons.

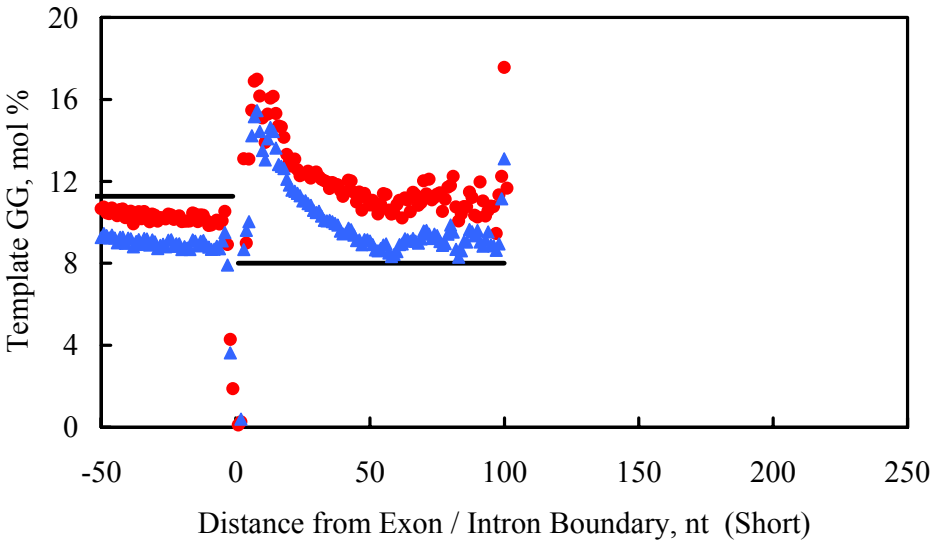


Figure 2-9. Continued.

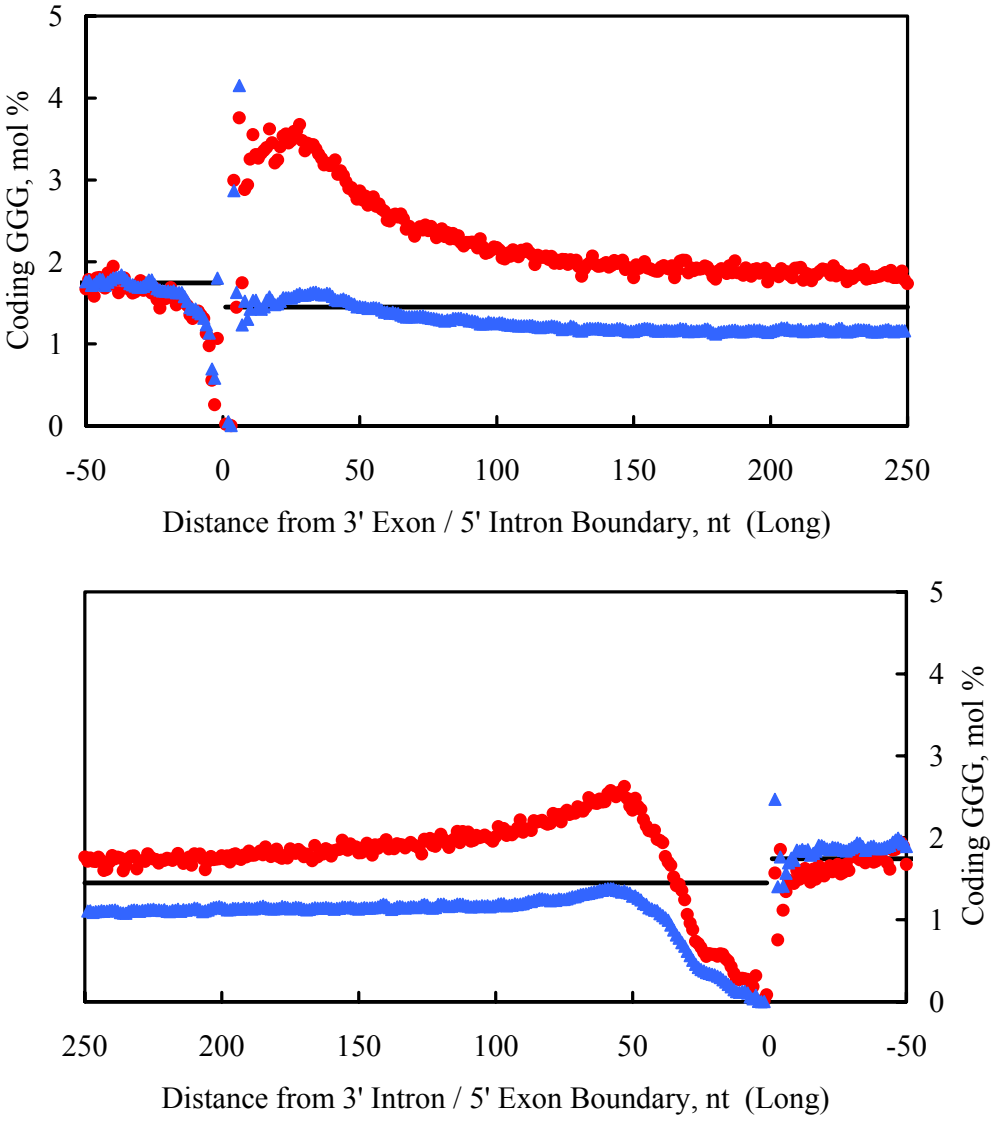


Figure 2-10. Observed and probability-predicted local average and observed overall average mol percentages of GGG on the coding strand vs. distance from exon/intron boundaries for long and short introns and exons.

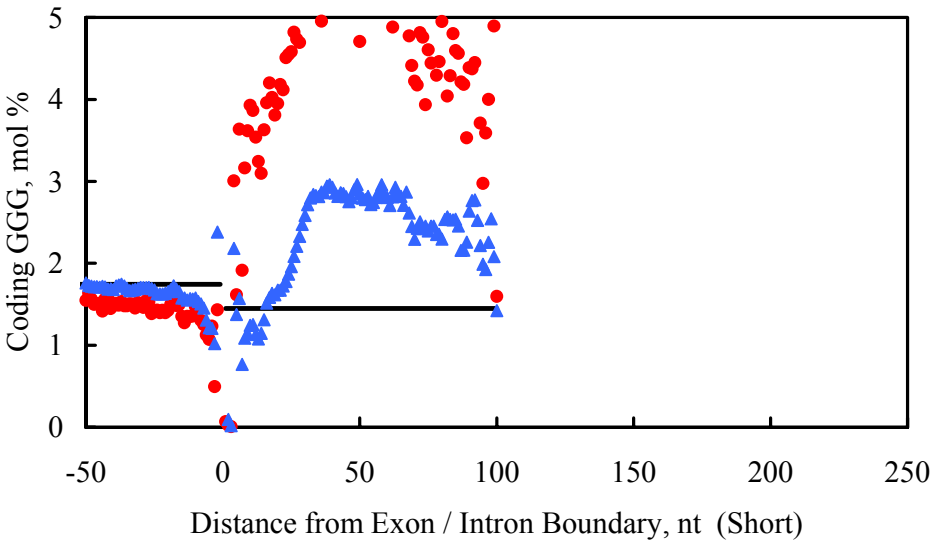


Figure 2-10. Continued.

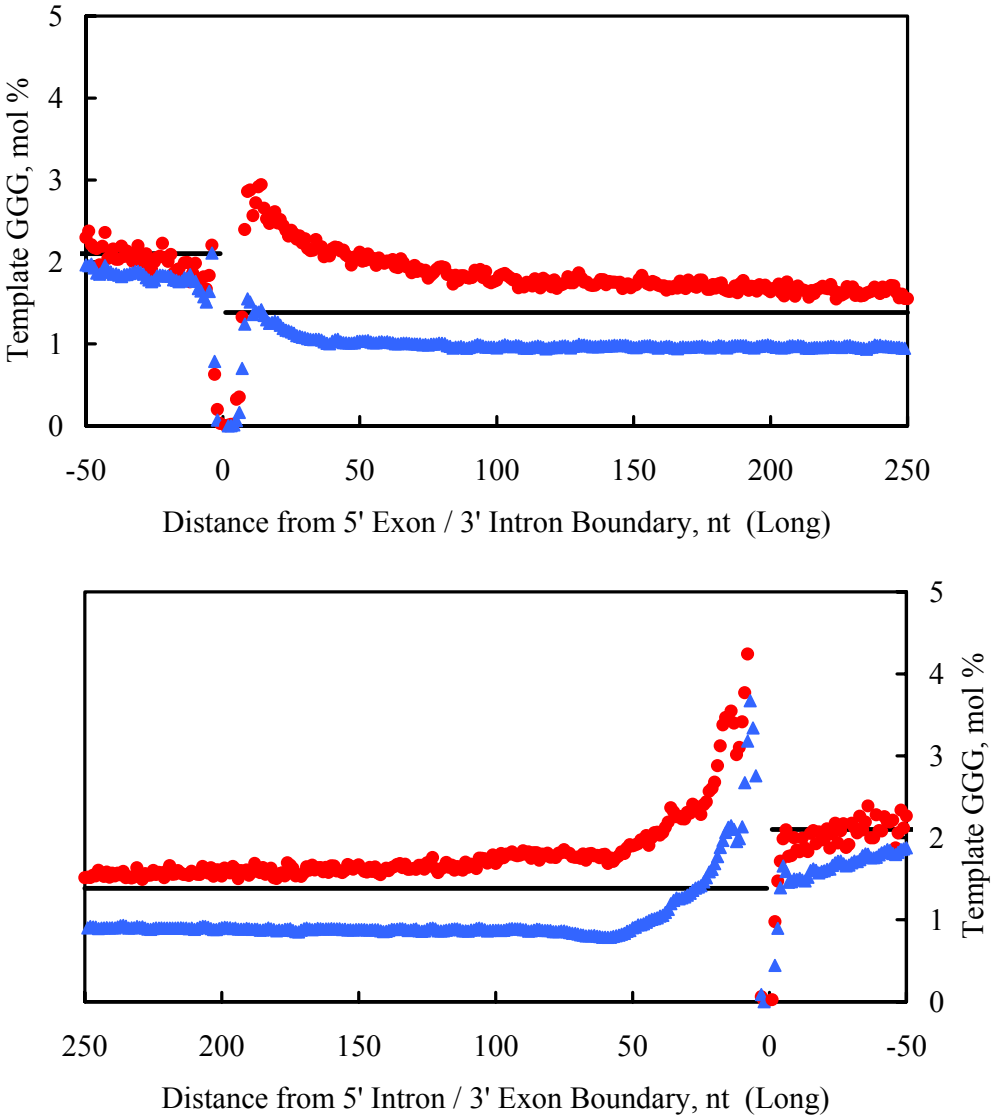


Figure 2-11. Observed and probability-predicted local average and observed overall average mol percentages of GGG on the template strand vs. distance from exon/intron boundaries for long and short introns and exons.

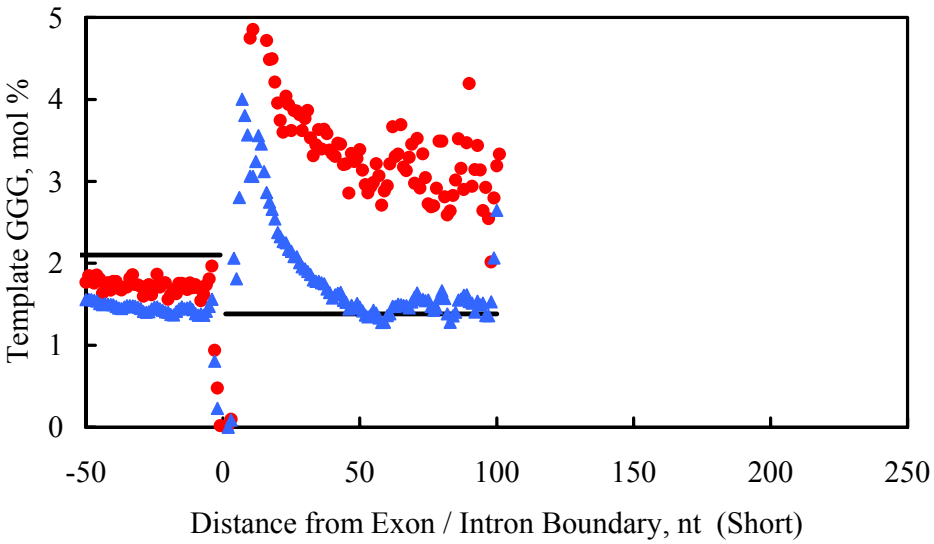


Figure 2-11. Continued.

Chapter 3: The Impact of Selective Oxidation on GGG and GG Levels and Oxidation Resistance in Eight Model Genomes

ABSTRACT

The genome's environment contains strong oxidizers, some of which selectively attack guanine, the most readily oxidized nucleotide. The ranking of guanine oxidation rates is central G in GGG (G**G**G) ≥ 5' G in GG (**G**G) > isolated or 3' G. Vulnerability to selective oxidants puts mutation pressure on guanine. This is apparent in the differences between observed levels of G**G**G and levels predicted by probability from total G. G**G**G is statistically under-represented in *H. sapiens* exons, but over-represented in *H. sapiens* introns and inter-gene domains. G**G**G is below probability predictions in *D. melanogaster, C. elegans, A. thaliana, S. cerevisiae, S. pombe* and. It is not under-represented in *E. cuniculi*. It is over-represented *P. falciparum* chromosomes 2 and 3, but this organism's total G levels are extremely low. **G**G generally is not under-represented in these genomes. The exceptions and other factors underlying guanine distributions are discussed. Guanine minimization, by whatever means, makes genomes more noble, plausibly reducing their vulnerability to selective oxidants.

INTRODUCTION

Oxidation of Guanine in DNA. The DNA in each human cell undergoes ~10^4 oxidative attacks each day and requires continual repair (1, 2). At steady state, each human cell has ~10^4 to ~10^5 oxidative lesions, including ~10^4 guanine nucleotides (nt) oxidized to 8-oxo-2-deoxyguanosine (oxo8dG) (3). Oxo8dG, a mutagenic lesion that is the principal product of guanine oxidation (4), is rapidly removed by base-excision repair, transcription-coupled repair and other systems (3, 5). Oxidative damage of DNA has been implicated as a cause of mutation, cancer (2, 6, 7) and aging (8-10).

The genome's environment contains strong but selective oxidizers, and precursors of extremely strong and non-selective oxidizers: hydrogen peroxide, nitric oxide, superoxide radicals, molecular oxygen, etc. (11-16). Extremely strong oxidizers react with any base they attack, while strong oxidizers are selective toward more readily oxidized guanine bases (17, 18).

The ranking of nucleotide oxidation potentials is G < A ≈ T < C, and the ranking of guanine oxidation rates is G**G**G ≥ **G**G > **G**, where G**G**G is the central G in GGG, **G**G is the left or 5' G in HGG, **G** is the isolated G in HGH and H is any base but G. (A nucleotide triplet with a bold G denote one specific nucleotide, e.g., G**G**G refers to the central G in the sequence GGG, whereas a triplet without a bold G refers to the whole sequence, e.g., GGG refers to the whole sequence GGG.) The one-electron oxidation potentials at pH 7 are 1.04 V for G < 1.32 V for A ≈ 1.29 V for T < 1.44 V vs. NHE for C (19). These potentials are comparable to those of noble metals: 0.92 V for Pd $\rightarrow$ Pd^{2+} + e^{-}, 1.19 V for Pt $\rightarrow$ Pt^{2+} + e^{-} and 1.36 V for Au^{+} $\rightarrow$ Au^{3+} + $2e^{-}$ (20). The relative rates of photoinduced oxidation are G**G**G (2.7) > **G**G (0.7 to 2.0) > **G** (0.1 to 0.4), where the rates vary with the neighboring nucleotides (21). Comparison of these rates with rates averaged over all guanines in GGG and GG sequences show that G**G**G and **G**G are the hot spots (22). While the chemical oxidation rate of **G**G is greater than that of **G** (23, 24), the reactivities of G**G**G and **G**G with peroxyl radicals are sufficiently similar for G**G**G to appear more reactive in one study (25), but less reactive in another (26). In HGGG, the central G (a G**G**G) and the 5' G (a **G**G) are similarly reactive (27).

Selectivity for G**G**G and **G**G is based on their lower oxidation potentials and facilitated by positive charge (hole) transfer in DNA that channels holes to them (22). Hole transport in DNA has been demonstrated experimentally and explained theoretically (23, 28-31). Injected holes by oxidants selectively react with remote G**G**G and **G**G, both *in vitro* and *in vivo* (21, 32-37).

These are imperfect generalizations about guanine oxidation. Preferential oxidation at GGG sequences also reflects sensitizer binding to DNA (38, 39). Oxidation of G in CG and GC by peroxyl radicals appears higher that of **G**G in one study (26), but not in another (25). (CG and GC denotes the sequences, not the paired bases.) CG, but not GC, is highly suppressed in vertebrates, for reasons unrelated to oxidation (40). Oxo8dG has been called "the ultimate sink of oxidizing equivalents in DNA", because the reduction potential of oxo8dG is 0.55 V lower than that of G (41, 42). The clinically observed frequencies of oxidative mutations (G→T or C→A transversions) are comparable at G**G**G, **G**G and **G** in a large human database, but clinical observations may not reflect underlying oxidation rates (43, 44).

Selective guanine oxidation is crudely modeled by eggs (oxidizing agents) dropped on the ground (DNA). Extremely strong oxidants are like eggs thrown hard at the ground: they crack (abstract an electron) and splatter (damage) wherever they hit. Strong selective oxidants are like eggs dropped softly. When they hit grass (A, T or C), they usually do not crack. When they hit dirt (**G**), they often crack, but they may roll along the ground (charge transfer) until they hit rock (G**G**G or **G**G). If they hit rock they usually crack and splatter.

Impact of Selective Guanine Oxidation. Vulnerability to selective oxidants is one of the factors driving guanine concentrations and distributions in genomes. Oxidation eliminates guanine: the principal product of guanine oxidation is oxo8dG which, unrepaired, leads to G→T transversion, because oxo8dG pairs with A (45, 46). This should occur more frequently in more oxidizing environments, which could accompany oxidant generation and/or anti-oxidant poverty. Under these conditions, reducing vulnerability to selective oxidants should confer selective advantage. Decreasing guanine levels could make DNA less prone to damage by selective oxidants, just as decreasing copper levels makes gold-copper alloys more corrosion resistant (47). Thus, the contributions of muta-

tional bias for guanine elimination and selection for oxidant resistance are complementary and difficult to distinguish.

If selective oxidation drives guanine levels, they should be systematically related to oxidation vulnerability or resistance, rather than random. Total guanine (ΣG) levels should be lower in organisms that evolved in more oxidizing environments. This comparison obviously is confounded if different organisms are compared. It could be clearer if the same organism, e.g., a rapidly mutating strain of E. coli (48), is experimentally evolved in more and less oxidizing environments. Levels of more readily oxidized guanine nucleotides, G**G**G and **G**G, should be lower than probability predicts from ΣG levels. These effects should be greater in domains with more mutations. ΣG, G**G**G and **G**G levels should be lower in genome domains with fewer functional constraints, because they accumulate mutations more rapidly. Introns and pseudogenes in inter-gene domains (IGD) mutate more rapidly than exons, and the third codon position mutates more rapidly than the first two positions (49). Guanine levels should be lower in genome domains where oxidation rates are higher and/or repair rates are lower. The repair rates of coding strands could be lower than those of template strands, because they lack transcription-coupled repair (5, 50-52).

While these propositions can be compared with actual guanine levels, this comparison is indicative, but not a definitive test of the hypothesis that selective oxidation drives guanine levels. Some of these propositions could be violated, but this could indicate that factors unrelated to oxidation dominate guanine levels in these cases. Some of these propositions could be validated, but results consistent with the hypothesis do not indicate that selective oxidation is the only or the dominant factor. However, their widespread violation or validity would be suggestive. These propositions do not distinguish the contributions of mutational bias for guanine elimination and selection for oxidant resistance. This could be addressed by propositions such as: guanine levels should be

lower where selection for oxidation resistance should be higher, such as genes whose products fight oxidation (anti-oxidants and oxo8dG repair enzymes).

Total Guanine Levels. Total guanine plus cytosine levels reflect many factors (see reviews (53-55)). ΣG mol percentages are less than 25% in seven of the eight model organisms (Table 3-S1). While this suggests that mechanisms eliminate more ΣG and/or ΣC than ΣA and/or ΣT, it does not imply that the principal mechanism is guanine oxidation followed by G→T transversion. ΣG mol percentages are lower in introns and IGD than in exons in the eight model genomes (Table 3-S1). Oxidative elimination accelerated by reduced constraints is only one of the factors that could generate different guanine levels in exons and introns and IGD. For example, the relative A and T richness of plant and invertebrate introns facilitates intron recognition (56, 57).

The association of survival in highly oxidizing environments and total guanine plus total cytosine (ΣG+ΣC) poverty is suggestive, but confounded by differences between organisms and counter examples. (The mol percentage of ΣG+ΣC is twice the mol percentage of ΣG on the summed coding and template strands.) *Plasmodium falciparum* is extremely poor (<20%) in ΣG+ΣC (58), and it metabolizes hemoglobin from red blood cells and produces reactive oxygen species (ROS) (59). *Entamoeba hystolytica*, an anaerobic parasite of that causes dysentery in aerobic hosts, is very poor (25%) in ΣG+ΣC (60, 61). However, it is highly vulnerable to exogenous H_2O_2, and it may use the bacteria and red blood cells it engulfs to detoxify ROS (62, 63). *Dictyostelium discoideum*, a cellular slime mold that is highly resistant to DNA damage by radiation and H_2O_2, is very poor (22%) in ΣG+ΣC (64, 65). However, *Deinococcus radiodurans*, a bacterium that survives massive doses of ionizing and short UV radiation, is ΣG+ΣC rich (~65%), emphasizing efficient repair and detoxification (66-69), not invulnerability.

Among the eight model genomes, only *Encephalitozoon cuniculi* has approximately 25% ΣG (genome references and Table 3-S1 in supplemental material). Minimizing ΣG probably is evolutionarily penalizing in exons, because deviation from a 1:1

AT:GC ratio sharply reduces the possible number of different arrangements of nucleotides (Figure 3-S5). At 13% ΣG, as in *P. falciparum* exons, the number of combinations possible in a hypothetical 100 bp genome is only 0.001% of the number at 25% ΣG. (Note: A hypothetical 100 bp genome grossly under-states the loss of combinations in larger genomes.) This cost could be too high for *E. cuniculi*, an organism that rigorously minimizes its genome size (70). Glutathione, thioredoxin and superoxide dismutase antioxidants might account for the high oxygen tolerance suggested by the development of microsporidia in various aerobic host cells (71).

GGG and GG Functions. GGG and GG have crucial biological functions in certain locations. The signal sequences for splicing almost all introns (GT-AG or GC-AG motifs) contain GG ((72-74) and from the Intron Sequence Information System (ISIS) (75) and the SpliceDB (76, 77)). The poly-pyrimidine track flanking the 3' splice site is cytosine-rich in *Homo sapiens* and somewhat cytosine-rich in Arthropoda (phylum of *Drosophila melanogaster*) ((72, 73, 78) and from ISIS)), so the 5' flank of the template strand is guanine-rich in *Homo sapiens*. Sequences containing GGG in the 5' flank of vertebrate introns enhance splicing, particularly of small exons and introns (78-83). Telomers, DNA and protein structures that protect the ends of eukaryotic chromosomes, are rich in G clusters, containing the TTAGGG repeat in vertebrates (84).

Examples of GGG and GG in crucial roles emphasize that vulnerability to selective oxidants is only one of the factors driving guanine distribution, and not always a dominant one. CG methylation for gene silencing and genome imprinting in vertebrates is another example of a mutation target with a critical biological function (85). The problem with GGG and GG is the risk that their signaling functions will be compromised by oxidation. This risk is <1 in 10000 at each site in human cells, because they have ~10^4 oxo8dG at steady-state and 3.6×10^8 G**G**G plus **G**G (from Table 3-S1). The risk is higher in organisms with higher oxo8dG levels and lower **G**G plus G**G**G levels (the model unicellular organisms have <10^6 nt G**G**G plus **G**G).

To assess the relationship between selective oxidation and guanine levels, we compared these propositions with published guanine distributions in eukaryote genomes and determined the distributions of **G**G and G**G**G in eight model genomes: *Homo sapiens* (*Hsa*), *Drosophila melanogaster* (*Dme*, fruit fly), *Caenorhabditis elegans* (*Cel*, nematode worm), *Arabidopsis thaliana* (*Ath*, flowering plant), *Saccharomyces cerevisiae* (*Sce*, budding yeast), *Schizosaccharomyces pombe* (*Spo*, fission yeast), *Encephalitozoon cuniculi* (*Ecu*, intracellular parasite) and *Plasmodium falciparum* (*Pfa*, malaria parasite).

METHODS

Determination of Nucleotide Distributions from Genome Data. We analyzed genome sequences from GenBank (86) for *Hsa* (Feb. 2002 release) (87, 88), *Dme* (Oct. 2000) (89), *Cel* (Dec. 2001) (90), *Ath* (Jan. 2002) (91), *Sce* (Mar. 2002) (92), *Spo* (Mar. 2002) (93), *Ecu* (Mar. 2002) (71), and *Pfa* chromosomes 2 (Nov. 1998) (58) and 3 (Apr. 1999) (94). Their GenBank (GBK) file names are listed at the end of supplemental data. Chromosomes 2 and 3 constitute only 7% of the *Pfa* genome, but their total G plus total C mol percentage (20%) is typical of the whole genome (18%) (58, 94). Preliminary analysis of *Hsa* (Dec. 2001), *Cel* (May 1999), *Ath* (Aug. 2001) and *Sce* (Nov. 2001) genome sequences yielded similar results.

For each annotated exon (coding sequence or CDS), we extracted the following: its position within its gene (first, intermediate or last), the coordinates of its first and last bases within its contiguous sequence (contig), and its strand (the given strand or its complement). Within each contig, these coordinates were sorted and overlapping or duplicated exons were merged. Sequences between exons within genes were identified as introns, and sequences next to the 5' end of first exons and/or the 3' end of last exons were identified as IGD. Using these coordinates, each nucleotide was identified as part of an exon, intron or IGD.

Each guanine nucleotide on the given strand in the GBK file was denoted **G** if it had unlike neighbors, "paired" if it was in HGGH and G**G**G if it was central in a triplet.

The paired count was divided by two for the **G**G count. Nucleotides on the complementary strand were inferred from base pairing. For example, -AGCGGCCGGGCCCA- had one **G**, two **G**G and one G**G**G on the coding strand and the same on the template strand.

For each exon, intron and IGD, we calculated the mol percentages of ΣG, **G**, **G**G and G**G**G in the segment overall. We weighted each segment equally, regardless of length, to calculate mean mol percentages. These segment-weighted averages described average exons, introns and IGD; nucleotide-weighted averages would have described average nucleotides within exons, introns and IGD. The Codon Database (95), ISIS (75) and Karlin et al. (40) use nucleotide-weighted averages, so segment-weighted averages offered a somewhat different perspective.

Probability-Prediction of G**G**G**, G**G **and G Mol Percentages.** Equations 3-1 to 3-3 give the mean mol percentages of G**G**G**, G**G and **G** predicted by probability from ΣG in a large sample of DNA single strands. The **G**G prediction is half of the prediction for paired GG (both G in HGGH). The equation for paired GG differs from that used by Karlin et al. (96) and in ISIS for dinucleotides to distinguishing guanines with one like neighbor (H**G**GH) from those with two (G**G**G). The probability predictions for double strands are the sums, not the averages, of predictions for single strands.

$$\text{Prob. G}\mathbf{G}\text{G \%} = (\Sigma\text{G \%}/100)^3 \qquad \text{(Eq. 3-1)}$$

$$\text{Prob. }\mathbf{G}\text{G \%} = (1 - (\Sigma\text{G \%}/100))\,(\Sigma\text{G \%}/100)^2 \qquad \text{(Eq. 3-2)}$$

$$\text{Prob. }\mathbf{G}\text{ \%} = (1 - (\Sigma\text{G \%}/100))^2\,(\Sigma\text{G \%}/100) \qquad \text{(Eq. 3-3)}$$

The equations for the means of large samples provided no information about specific distributions. For example, they did not predict the percentages of segments with no G**G**G. To obtain this information, probability predictions were made by Monte Carlo simulations. For each actual segment, a probability-predicted counterpart, of equal length and equal total A, T, G, C and N mol percentages, was predicted by assigning a random

number (0 to 1) to each nucleotide, and then assigning a type (G**G**G, **G**G, **G**, etc.) to each nucleotide by dividing the (0 to 1) interval among types according to probabilities, such as those of equations 3-1 to 3-3 for G**G**G, **G**G and **G**. Averages and standard deviations of mol percentages were calculated for the predicted segments, and, because there was some variation between successive simulations of small genomes, the means of ten rounds of simulations were used. Table 3-S2 lists percentage standard deviations, expressing the uncertainty of the predicted means, counterparts of the actual means in Tables 3-1 and 3-S3. Because the numbers of samples (exons, introns or IGD) were very large, even small differences between actual and probability-predicted mol percentages were statistically significant.

RESULTS

Tables and Figures in this section highlight the minimization of G**G**G in whole exons, introns and IGD. **G**G results are in the supplemental materials in Appendix 1 (Table numbers with "S"), because G**G**G levels deviate much more from probability predictions. In these Tables, single DNA strands are denoted "+" for coding or "–" for template, and hybridized double strands are denoted "&".

Tables 3-1 and 3-S3 list the mean mol percentages of G**G**G and **G**G in whole exons, introns and IGD. Mol percentages in single strands are mean values of guanine nucleotides per total nucleotides, but mol percentages in double strands are mean values of guanine nucleotides per total base pairs. Single strand values are mol percentages of guanine nucleotides, while double strand values are mol percentages of base pairs containing guanine. The percentage standard deviations (100%(std dev/mean)) were ~30% to ~300% (Table 3-S4 and Figures 3-1 and 3-S1 to 3-S4). They reflected true excursions of mol percentages within populations, not statistical uncertainties because of small sample sizes.

The mean mol percentages of G**G**G were <0.5% in single strands of several organisms, notably, in introns of *Cel*, *Ath*, *Spo* and *Pfa*. Multiplying these values by me-

dian intron lengths (Table 3-S1) showed that most of these introns had less than one G**G**G in two strands. In *Hsa*, *Dme* and *Ecu*, the mean mol percentages of **G**G averaged 3.5 times those of G**G**G; in other genomes, the **G**G values averaged 5.5 time those of G**G**G. In all genomes except *Hsa* and *Ecu*, the mean mol percentages of both G**G**G and **G**G in exons exceeded those in introns or in IGD. In *Hsa*, the mean mol percentage of **G**G in exons exceeded somewhat that in introns, but the reverse was decisively true for G**G**G.

Tables 3-2 and 3-S5 list the percentages of the exon, intron and IGD populations with no G**G**G or no **G**G. The percentages of single strands with no G**G**G averaged 42% in *Cel*, *Ath* and *Pfa* exons, 25% in *Hsa*, *Dme* and *Spo* exons, 71% in non-*Hsa* introns, and 21% in *Sce*, *Spo*, *Ecu* and *Pfa* IGD. These percentages were small in *Sce* and *Ecu* exons, *Hsa* introns, and *Hsa*, *Dme*, *Cel* and *Ath* IGD. G**G**G exclusion was less probable in these segments, because their median lengths were much greater than those of corresponding segments in other organisms (Table 3-S1). The percentages of double strands with no G**G**G were about half the averages of their single strands. The percentages of strands with no **G**G were about an order of magnitude less than the percentages of strands with no G**G**G. **G**G exclusions were improbable relative to G**G**G exclusions, because **G**G mol percentages were roughly an order of magnitude larger than those of G**G**G. The percentages of exon and intron single strands with no **G**G were generally a factor of two greater on coding than on template strands.

The percentage differences (100%(observed − predicted)/predicted) between observed and probability-predicted percentages of single strands with no G**G**G averaged 68% in *Dme* and *Cel* exons, 46% in *Ath* and *Ecu* exons, and 24% in *Hsa*, *Sce* and *Spo* exons (Table 3-S6). Except in *Ath*, these percentage differences seldom were ≥33% in introns and IGD. The percentage differences of single strands with no **G**G were ≥33% in more than half the species and segments, especially in exon and intron template strands,

but many are not underlined in Tables 3-2, 3-S5 and 3-S6, because many of the underlying percent exclusions were <5%.

Tables 3-3 and 3-S7 lists the percentage differences between observed and probability-predicted mean mol percentages of G**G**G and **G**G in exons, introns and IGD with G**G**G and **G**G, respectively. (Calculations for G**G**G only included segments that contained G**G**G. Calculations for **G**G only included segments that contained **G**G, regardless of their G**G**G content.) Table 3-S8 lists the absolute differences (observed − predicted) between observed and probability-predicted mean numbers of G**G**G and **G**G in exons, introns and IGD with G**G**G and **G**G, respectively. (Mean numbers were calculated by multiplying mean lengths by mean mol fractions.)

G**G**G percentage differences of single strands averaged 36% in *Hsa* introns, 41% in *Hsa* IGD, 19% in *Cel* IGD, and 17%, 27% and 130% in *Pfa* exons, introns and IGD, respectively. *Hsa* introns and IGD were longer and G**G**G-richer than those of other organisms, so these percentage differences translated to substantial absolute differences. The absolute differences averaged 3 G**G**G on single strands of *Dme*, *Cel* and *Pfa* IGD. Except in *Hsa* and *Pfa*, G**G**G percentage differences of exon single strands were substantially negative, averaging −30% in *Dme*, *Cel* and *Ath*, and −16% in *Sce*, *Spo* and *Ecu*. These percentage differences translated to absolute differences smaller than −3 G**G**G in *Dme*, *Cel*, *Ath*, *Sce*, *Spo* and *Ecu*.

GG percentage differences, unlike those of G**G**G, were almost uniformly positive for exons, predominately positive for introns, and almost uniformly positive for IGD. The magnitudes of **G**G percentage differences usually were less than half the magnitudes of G**G**G percentage differences. The absolute differences on single strands were ≤3 **G**G, except on *Sce*, *Ecu* and *Pfa* (long) exons, *Hsa* introns and IGD, and *Dme*, *Ecu* and *Pfa* IGD.

Tables 3-4 and 3-S9 list the percentages of segments with sub-mean mol percentages of G**G**G and **G**G in the exon, intron and IGD populations with G**G**G and **G**G, re-

spectively. Virtually all of the distributions were positively skewed: more than 50% of the segments had G**G**G and **G**G mol percentages below mean values and tailed toward higher values (Figures 3-1 and 3-S1 to 3-S4). The sub-mean percentages for G**G**G generally were 60% to 70%, while those for **G**G generally were 50% to 60%. While the percentages of (long) *Hsa* introns and IGD with no G**G**G were negligible, their percentages with sub-mean G**G**G averaged 67%. The percentages of single strands with sub-mean G**G**G were substantial in other cases, averaging 67% in *Cel* introns and 74% in *Pfa* exons, a substantial percentage of which also had no GG. All of the percentage differences between the observed and the probability-predicted values were between ≤33% and ≥−33%.

DISCUSSION

Single strand G**G**G mol percentages are significantly less than 1.6% (equal allotment from 25% ΣG) except in *Hsa* and *Ecu*, and **G**G mol percentages are significantly less than 4.7% except in *Hsa*, *Ecu* and exons of *Dme* (Tables 3-1 and 3-S3). These G**G**G and **G**G levels are the probable results of ΣG < 0.25 (effective minimization), and, for G**G**G in seven genomes, the results of specific minimization beyond what probability predicts from ΣG. *Hsa* exons, *Dme*, *Cel Ath*, *Sce* and *Spo* specifically exclude and suppress G**G**G, but *Hsa* introns and IGD, *Ecu* and *Pfa* do not. Specific skew of G**G**G is small.

Intron and IGD G**G**G **and G**G **Levels.** Among genomes that specifically minimize G**G**G, specific exclusion is ~10% in introns and IGD, and specific suppression is ≥−10% in introns and negligible in IGD (Table 3-5). Exclusion in *Ath* introns and IGD is exceptionally strong.

Only one of the more readily oxidized guanine nucleotides, G**G**G but not **G**G, is lower than probability predicts from ΣG levels. **G**G exclusion generally exceeds what probability predicts, except in *Ecu* and *Pfa*, but most segments contain **G**G. In segments containing **G**G, **G**G mol percentages generally exceed what probability predicts (Table 3-

5). This contradicts one of the predictions of the hypothesis that selective oxidation drives guanine levels, but it may be rationalized by differences between G**G**G and **G**G elimination. GGH offers one hot G as a target for elimination via oxidation and transversion, whereas GGG offers two. Oxidative elimination of the 5' G of GGG, a **G**G, actually eliminates a G**G**G and produces TGG, with a **G**G. If TGG is over-represented, it may account for the lack of **G**G under-representation. (With oxidative elimination, over-representation of TGG and GTG should accompany under-representation of G**G**G, and over-representation of TG should accompany under-representation of **G**G.) In exons, GGG minimization involves two single codons (GGG and CCC), but GGH minimization involves 12 single codons, including all codons for glycine and proline.

G**G**G and **G**G are not specifically minimized in *Hsa* introns and IGD, *Ecu* and *Pfa*. The exception genome compaction of *Ecu* that could contraindicate ΣG minimization, and the exceptionally low ΣG mol percentage of *Pfa* that could obviate specific minimization, cannot be invoked to explain the absence of specific minimization in *Hsa*. This suggests that factors opposing oxidative elimination dominate this guanine distribution. We will argue that G**G**G and **G**G are deployed as scavengers in *Hsa* introns and IGD, to protect essential domains. This argument follows the suggestions of Barton et al. (38, 97), Giese et al. (29, 98), Kawanishi et al. (99), Schuster et al. (100), Thorp et al. (101) and Heller et al. (47, 102) that genomes exploit hole conduction in DNA for oxidation protection, among other functions (103).

Guanine and cytosine dinucleotides (NGGN and NCCN) are over-represented (observed/expected ≥ 1.2) in *Hsa* and Aveolata (phylum of *Pfa*) introns ((40) and from ISIS). (N represents any base, so NGGN includes both GG and GGG.) The over-representation of GG and CC could be related to the very high suppression of CG in *Hsa* (96). While the G of CG is an oxidation hot-spot (26), CG suppression in nuclear genomes is usually ascribed to a methylation-deamination-mutation mechanism causing C→T transition and other factors unrelated to oxidation (96).

Exon GG**G and G**G **Levels.** In exons of *Hsa*, *Dme*, *Cel Ath*, *Sce* and *Spo*, specific exclusion of G**G**G is ~20 to ~70% and specific suppression of G**G**G is ~–10 to ~–30% (Table 3-5). Specific minimization of G**G**G in exons parallels codon usage that is biased against GGG and GG in some eukaryotes. The GGG codon for glycine and the CCC codon for proline each would be 25% of codons used for these amino acids without bias, but GGG usage is 33% in *Ecu*, 25% in *Hsa*, 16% in *Ath*, and ≤12% in *Dme*, *Cel*, *Sce*, *Spo* and *Pfa*, and CCC usage is 33% in *Hsa* and *Dme*, 23% in *Ecu*, ~16% in *Sce* and *Spo*, and <12% in *Cel*, *Ath* and *Pfa* (from the Codon Usage Database (95)). (CC and CCC on the coding strand imply GG and GGG on the template strand.) GGG is underrepresented (observed/expected ≤ 0.8) in Arthropoda (*Dme*), Nematoda (*Cel*), Magnoliopsida (*Ath*), Ascomycota (*Sce* and *Spo*) and Aveolata (*Pfa*) exons, and CCC is underrepresented in Nematoda (*Cel*) and Magnoliopsida (*Ath*) exons (from ISIS). Biases against codons containing GG and CC are similarly ordered, but weaker.

Synonymous codon usage is biased by considerations of gene expression, transcription and translation efficiency, mutational bias, local and global ΣG+ΣC level, DNA structure, and other factors (104-106). Probability prediction approximated exons as collections of nucleotides; they are better modeling them as codon sequences, as in the synonymous-sites approach (107, 108). Because of these confounding factors and simplifications, the lower levels of specific minimization in introns may be better estimate of the effects of selective oxidation.

Strand Asymmetry of Guanine. Table 3-6 lists the percent asymmetries (100%(coding − template)/(coding + template)) of ΣG, G**G**G and **G**G in exons and introns. These values are the local asymmetries of exons and introns accumulated over the genomes. They are not large-scale strand asymmetries which eukaryotic chromosomes lack (96, 109). Codon biases also indicate ΣG asymmetries in exon strands (110). The asymmetries generally are positive, indicating more guanine on the coding strand, except for that of intron G**G**G which is negative (zero in *Hsa*). While G**G**G in introns presents

the cleanest conditions for observing selective oxidation, this result is ambiguous because many factors contribute to strand asymmetry (111).

Defenses against Genome Oxidation. Organisms generally employ multiple defensives against oxidation. Enzymes like catalases, superoxide dismutases, glutathione transferases and heme oxygenase, and protective molecules like bilirubin, melatonin, carotene and glutathione detoxify strong oxidizing agents and are the first line of defense against them (69, 112, 113). Next, DNA protects itself: the bases' oxidation potentials are comparable to those of noble metals, the most readily oxidized nucleotide, guanine, is minimized and the DNA is wrapped in and around proteins (chromatin (114)) that substantially but imperfectly shield it (35, 115, 116). The last lines of defense are repair of oxidative lesions (52, 85) and apoptosis (117).

Guanine minimization can be seen as an element of genome defense. Because G**G**G ≥ **G**G > **G** are the most rapidly oxidized nucleotides, a simple way to increase genomic oxidation resistance is to minimize ΣG mol percentage (47). This minimizes G**G**G > **G**G > **G**: probability predicts (equations 3-1 and 3-2) that **G**G and **G**GG mol percentages in the coding strands of *Pfa* exons (0.22% and 2.9%) that are seven and three times lower than those in *Hsa* exons (1.6% and 9.4%), because the exon ΣG mol percentage in *Pfa* (13%) is a factor of two lower than that in *Hsa* (25%). Seven of the eight model genomes augment ΣG minimization by specifically minimizing G**G**G, but not **G**G, below probability-predicted levels in exons. The absence of specific **G**G minimization undercuts the contribution of specific G**G**G minimization to overall genome resistance to selective oxidants, because **G**G and G**G**G oxidation rates are similar, **G**G is 3.5 to 5.5 times more prevalent and the oxidation frequency is small relative to the numbers **G**G and G**G**G.

Alternately, guanine minimization can be seen as a measure of other genome defenses. Selective oxidation eliminates guanine until its levels that can be maintained by genome defenses and mutations producing guanine. *D. radiodurans* is ΣG+ΣC rich

(~65%) and an expert at repairs (69). Both views could be instructive, because, one way and/or the other, selective oxidation is a factor in the complex function that establishes guanine levels.

ACKNOWLEDGMENTS

Professors Richard Hallick of the University of Arizona, Kensal van Holde of Oregon State University, George Georgiou, Brent Iverson and Edward Marcotte of the University of Texas, Dr. Jonathan Heller of Optiscan Corp. and anonymous reviewers made very helpful comments. The National Science Foundation, the Robert A. Welch Foundation, National Institutes of Health Biotechnology Training Grant and the Richard J. Lee Endowed Graduate Fellowship in Engineering provided financial support for A.H. and K.A.F.

REFERENCES

1. Helbock, H. J., Beckman, K. B., Shigenaga, M. K., et al. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 288-293.
2. Setlow, R. B. (2001) *Mutat. Res.* **477,** 1-6.
3. Beckman, K. B. & Ames, B. N. (1997) *J. Biol. Chem.* **272,** 19633-19636.
4. Sekiguchi, M. & Hayakawa, H. (1998) *Contemp. Cancer Res.* **2,** 85-93.
5. Hanawalt, P. C. (2001) *Mutat. Res.* **485,** 3-13.
6. Ambrosone, C. B. (2000) *Antioxidants & Redox Signaling* **2,** 903-917.
7. Jackson, A. L. & Loeb, L. A. (2001) *Mutat. Res.* **477,** 7-21.
8. Beckman, K. B. & Ames, B. N. (1998) *Physiological Rev.* **78,** 547-581.
9. Bohr, V. A. & Anson, R. M. (1995) *Mutat. Res.* **338,** 25-34.
10. Finkel, T. & Holbrook, N. J. (2000) *Nature* **408,** 239-247.
11. Amatore, C., Arbault, S., Bruce, D., et al. (2000) *Faraday Discuss.* **116,** 319-333.
12. Kawanishi, S., Oikawa, S. & Hiraku, Y. (2000) *Free Radicals in Chemistry, Biology and Medicine***,** 85-91.
13. May, J. M., Qu, Z.-C., Xia, L., et al. (2000) *Am. J. Physiol.* **279,** C1946-C1954.

14. Newcomb, T. G. & Loeb, L. A. (1998) *DNA Damage and Repair* **1,** 65-84.

15. Poulsen, H. E., Jensen, B. R., Weimann, A., et al. (2000) *Free Radical Research* **33,** S33-S39.

16. Wei, Y.-H., Pang, C.-Y., Lee, H.-C., et al. (1998) *Current Science* **74,** 887-893.

17. Kawanishi, S., Hiraku, Y., Murata, M., et al. (2002) *Free Radical Biology & Medicine* **32,** 822-832.

18. Kawanishi, S., Hiraku, Y. & Oikawa, S. (2001) *Mutat. Res.* **488,** 65-76.

19. Faraggi, M., Broitman, F., Trent, J. B., et al. (1996) *J. Phys. Chem.* **100,** 14751-14761.

20. Bard, A. J. & Faulkner, L. R. (2001) *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York).

21. Saito, I., Nakamura, T., Nakatani, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 12686-12687.

22. Lewis, F. D., Liu, X., Liu, J., et al. (2000) *J. Am. Chem. Soc.* **122,** 12037-12038.

23. Sistare, M. F., Codden, S. J., Heimlich, G., et al. (2000) *J. Am. Chem. Soc.* **122,** 4742-4749.

24. Sugiyama, H. & Saito, I. (1996) *J. Am. Chem. Soc.* **118,** 7063-7068.

25. Kawanishi, S., Oikawa, S., Murata, M., et al. (1999) *Biochemistry* **38,** 16733-9.

26. Rodriguez, H., Valentine, M. R., Holmquist, G. P., et al. (1999) *Biochemistry* **38,** 16578-16588.

27. Yoshioka, Y., Kitagawa, Y., Takano, Y., et al. (1999) *J. Am. Chem. Soc.* **121,** 8712-8719.

28. Bixon, M., Giese, B., Wessely, S., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11713-11716.

29. Giese, B. (2002) *Annu. Rev. Biochemistry* **71,** 51-70.

30. Nunez, M. E., Hall, D. B. & Barton, J. K. (1999) *Chem. Biol.* **6,** 85-97.

31. Treadway, C. R., Hill, M. G. & Barton, J. K. (2002) *Chemical Physics* **281,** 409-428.

32. Boone, E. & Schuster, G. B. (2002) *Nucleic Acids Res.* **30,** 830-837.

33. Meggers, E., Michel-Beyerle, M. E. & Giese, B. (1998) *J. Am. Chem. Soc.* **120,** 12950-12955.

34. Nunez, M. E., Holmquist, G. P. & Barton, J. K. (2001) *Biochemistry* **40,** 12465-12471.

35. Nunez, M. E., Noyes, K. T. & Barton, J. K. (2002) *Chemistry & Biology* **9,** 403-415.

36. O'Neill, P., Parker, A. W., Plumb, M. A., et al. (2001) *J. Phys. Chem. B* **105,** 5283-5290.

37. Sanii, L. & Schuster, G. B. (2000) *J. Am. Chem. Soc.* **122,** 11545-11546.

38. Hall, D. B., Holmlin, R. E. & Barton, J. K. (1996) *Nature* **382,** 731-735.

39. Henle, E. S. & Linn, S. (1997) *J. Biol. Chem.* **272,** 19095-19098.

40. Karlin, S. & Marazek, J. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94,** 10227-10232.

41. Hickerson, R. P., Prat, F., Muller, J. G., et al. (1999) *J. Am. Chem. Soc.* **121,** 9423-9428.

42. Steenken, S., Jovanovic, S. V., Bietti, M., et al. (2000) *J. Am. Chem. Soc.* **122,** 2373-2374.

43. Krawczak, M., Ball, E. V. & Cooper, D. N. (1998) *Am. J. Hum. Genet.* **63,** 474-488.

44. Krawczak, M. & Cooper, D. N. (1997) *Trends Genetics* **13,** 121-122.

45. Shibutani, S., Takeshita, M. & Grollman, A. P. (1991) *Nature* **349,** 431-4.

46. Wood, M. L., Dizdaroglu, M., Gajewski, E., et al. (1990) *Biochemistry* **29,** 7024-32.

47. Heller, A. (2000) *Faraday Discuss.* **116,** 1-13.

48. Sueoka, N. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 2653-7.

49. Cooper, D. N. (2000) *Human Gene Evolution* (BIOS Scientific, Oxford).

50. Francino, M. P., Chao, L., Riley, M. A., et al. (1996) *Science* **272,** 107-9.

51. Francino, M. P. & Ochman, H. (1997) *Trends Genetics* **13,** 240-245.

52. Nouspikel, T. & Hanawalt, P. C. (2002) *DNA Repair* **1,** 59-75.

53. Bernardi, G. (2000) *Gene* **259,** 31-43.

54. Eyre-Walker, A. & Hurst, L. D. (2001) *Nature Rev. Genetics* **2,** 549-555.

55. Sueoka, N. (1992) *J. Mol. Evol.* **34,** 95-114.

56. Brendel, V., Kleffe, J., Carle-Urioste, J. C., et al. (1998) *J. Mol. Biol.* **276,** 85-104.

57. McCullough, A. J. & Schuler, M. A. (1997) *Nucleic Acids Res.* **25,** 1071-1077.

58. Gardner, M. J., Tettelin, H., Carucci, D. J., et al. (1998) *Science* **282,** 1126-1132.

59. Francis, S. E., Sullivan, D. J., Jr. & Goldberg, D. E. (1997) *Annu. Rev. Microbiology* **51,** 97-123.

60. Romero, H., Zavala, A. & Musto, H. (2000) *Gene* **242,** 307-311.

61. Serrano-Luna, D. J., Negrete, E., Reyes, M., et al. (1998) *Experimental Parasitology* **89,** 71-77.

62. Bracha, R. & Mirelman, D. (1984) *J. Exp. Med.* **160,** 353-68.

63. Tekwani, B. L. & Mehlotra, R. K. (1999) *Microbes and Infection* **1,** 385-394.

64. Garcia, M. X. U. (2000) *Regulation and role of catalases during development and oxidative stress in Dictyostelium discoideum (PhD Thesis)* (Univ. of Missouri, Columbia), pp. 226 pp.

65. Sharp, P. M. & Devine, K. M. (1989) *Nucleic Acids Res.* **17,** 5029-39.

66. Karlin, S. & Mrazek, J. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 5240-5245.

67. Makarova, K. S., Aravind, L., Wolf, Y. I., et al. (2001) *Microbiol. Mol. Biol. Rev.* **65,** 44-79.

68. Narumi, K., Kikuchi, M., Funayama, T., et al. (1999) *Hoshasen Seibutsu Kenkyu* **34,** 401-418.

69. White, O., Eisen, J. A., Heidelberg, J. F., et al. (1999) *Science* **286,** 1571-1577.

70. Vivares, C. P. & Metenier, G. (2000) *Curr. Opin. Microbiology* **3,** 463-467.

71. Katinka, M. D., Duprat, S., Cornillot, E., et al. (2001) *Nature* **414,** 450-453.

72. Lim, L. P. & Burge, C. B. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11193-11198.

73. Reddy, A. S. N. (2001) *Crit. Rev. Plant Sciences* **20,** 523-571.

74. Rogozin, I. B. & Milanesi, L. (1997) *J. Mol. Evol.* **45,** 50-59.

75. Croft, L., Schandorff, S., Clark, F., et al. (2000) *Nature Genetics* **24,** 340-341.

76. Burset, M., Seledtsov, I. A. & Solovyev, V. V. (2000) *Nucleic Acids Res.* **28,** 4364-4375.

77. Burset, M., Seledtsov, I. A. & Solovyev, V. V. (2001) *Nucleic Acids Res.* **29,** 255-259.

78. Zhang, M. Q. (1998) *Human Molecular Genetics* **7,** 919-932.

79. Carlo, T., Sierra, R. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 3988-3995.

80. Carlo, T., Sterner, D. A. & Berget, S. M. (1996) *RNA* **2,** 342-353.

81. McCullough, A. J. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 9225-9235.

82. McCullough, A. J. & Berget, S. M. (1997) *Mol. Cell. Biol.* **17,** 4562-4571.

83. Sirand-Pugnet, P., Durosay, P., Brody, E., et al. (1995) *Nucleic Acids Res.* **23,** 3501-7.

84. McEachern, M. J., Krauskopf, A. & Blackburn, E. H. (2000) *Annu. Rev. Genetics* **34,** 331-358.

85. Alberts, B., Johnson, A., Lewis, J., et al. (2002) *Molecular Biology of the Cell* (Garland Science, New York).

86. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., et al. (2002) *Nucleic Acids Res.* **30,** 17-20.

87. International Human Genome Sequencing Consortium (2001) *Nature* **409,** 860-921.

88. Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) *Science* **291,** 1304-1351.

89. Adams, M. D., Celniker, S. E., Holt, R. A., et al. (2000) *Science* **287,** 2185-2195.

90. Caenorhabditis elegans Sequencing Consortium (1998) *Science* **282,** 2012-2018.

91. Arabidopsis Genome Initiative (2000) *Nature* **408,** 796-815.

92. Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996) *Science* **274,** 546, 563-567.

93. Wood, V., Gwilliam, R., Rajandream, M. A., et al. (2002) *Nature* **415,** 871-880.

94. Bowman, S., Lawson, D., Basham, D., et al. (1999) *Nature* **400,** 532-538.

95. Nakamura, Y., Gojobori, T. & Ikemura, T. (2000) *Nucleic Acids Res.* **28,** 292.

96. Karlin, S., Campbell, A. M. & Mrazek, J. (1998) *Annu. Rev. Genetics* **32,** 185-225.

97. Boon, E. M. & Barton, J. K. (2002) *Curr. Opin. Structural Biology* **12,** 320-329.

98. Giese, B. (2000) *Chemistry in Britain* **36,** 44-46.

99. Oikawa, S., Tada-Oikawa, S. & Kawanishi, S. (2001) *Biochemistry* **40,** 4763-4768.

100. Kanvah, S. & Schuster, G. B. (2002) *J. Am. Chem. Soc.* **124,** 11286-11287.

101. Szalai, V. A., Singer, M. J. & Thorp, H. H. (2002) *J. Am. Chem. Soc.* **124,** 1625-1631.

102. Friedman, K. A. & Heller, A. (2001) *J. Phys. Chem. B* **105,** 11859-11865.

103. Rajski, S. R., Jackson, B. A. & Barton, J. K. (2000) *Mutat. Res.* **447,** 49-72.

104. Karlin, S. & Mrazek, J. (1996) *J. Mol. Biol.* **262,** 459-472.

105. Musto, H., Romero, H., Zavala, A., et al. (1999) in *J. Mol. Evol.*, Vol. 49, pp. 27-35.

106. Smith, N. G. C. & Eyre-Walker, A. (2001) *Mol. Biol. Evol.* **18,** 982-986.

107. Jermiin, L. S., Foster, P. G., Graur, D., et al. (1996) *J. Mol. Evol.* **42,** 476-480.

108. Jermiin, L. S., Graur, D., Lowe, R. M., et al. (1994) *J. Mol. Evol.* **39,** 160-73.

109. Francino, M. P. & Ochman, H. (2000) *Mol. Biol. Evol.* **17,** 416-422.

110. Sueoka, N. & Kawanishi, Y. (2000) *Gene* **261,** 53-62.

111. Frank, A. C. & Lobry, J. R. (1999) *Gene* **238,** 65-77.

112. Reiter, R. J., Acuna-Castroviejo, D., Tan, D.-X., et al. (2001) *Ann. NY Acad. Sci.* **939,** 200-215.

113. Talalay, P. (2000) *BioFactors* **12,** 5-11.

114. Zlatanova, J., Leuba, S. H. & Van Holde, K. (1998) *Biophys. J.* **74,** 2554-2566.

115. Lodish, H., Berk, A., Zipursky, S. L., et al. (1999) *Molecular Cell Biology* (W.H. Freeman & Co., New York).

116. Smerdon, M. J. & Thoma, F. (1998) *Contemp. Cancer Res.* **2,** 199-222.

117. Higami, Y. & Shimokawa, I. (2000) *Cell & Tissue Res.* **301,** 125-132.

Table 3-1. Mean mol percentages of GGG in exons, introns and IGD.

GGG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	1.8	1.1	0.60	0.98	0.78	0.68	1.8	**0.41**
Exon	–	1.9	1.2	0.58	0.55	0.70	0.64	0.83	**0.28**
	&	3.7	2.3	1.2	1.5	1.5	1.3	2.6	0.69
	+	2.0*	**0.50**	**0.31**	**0.30**	**0.47**	**0.19**		**0.07**
Intron	–	2.0*	0.95	**0.37**	**0.34**	0.53	**0.20**		**0.13**
	&	4.0*	1.5#	0.68#	0.64#	1.0	**0.39**		**0.20**
	+	1.7	0.77	0.51	**0.46**	0.55	**0.35**	1.5	**0.17**
IGD	–	1.7	0.83	0.67*	**0.47**	0.57	**0.39**	1.7*	**0.17**
	&	3.4	1.6	1.2	0.94#	1.1	0.73#	3.2*	**0.34**

Notes: Mol percentages are highlighted when ≤0.5% (bold red). They are underlined when their percentage differences (100%(actual – prob.)/prob.) from probability-predicted values are ≥33% (solid blue) or ≤–33% (dotted red), except when actual mol percentages are ≤0.5%. Intron or IGD mol percentages are stared when their percentage differences (100%(intron – exon)/exon) from corresponding exon values are ≤–33% (# red) or ≥0% (* blue), except when intron or IGD mol percentages are ≤0.5%.

Table 3-2. Percentages of exon, intron and IGD populations with no GGG.

GGG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	26.	20.	**45.**	32.	5.8	28.	2.6	**36.**
Exon	–	28.	19.	**48.**	**49.**	6.1	31.	3.7	**40.**
	&	12.	8.5	25.	20.	2.9	17.	0.61	24.
	+	4.0	**60.**	**67.**	**70.**	**46.**	**86.**		**88.**
Intron	–	3.5	**46.**	**65.**	**69.**	**43.**	**87.**		**83.**
	&	1.7	**35.**	**55.**	**52.**	27.	**77.**		**76.**
	+	0.47	5.5	13.	6.6	22.	17.	32.	16.
IGD	–	0.43	4.8	10.	5.9	23.	16.	24.	17.
	&	0.26	3.4	6.6	3.6	14.	10.	16.	8.0

Notes: Population percentages are highlighted when ≥33% (bold blue). They are underlined when their percentage differences from probability-predicted values are ≥33% (solid blue) or ≤–33% (dotted red), except when actual population percentages are ≤5%

Table 3-3. Percentage differences between observed and probability-predicted mean mol percentages of G**G**G in exons, introns and IGD with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	–8.8	–40.	–31.	–25.	–20.	–25.	–20.	–0.29
Exon	–	1.0	–33.	–26.	–25.	–9.3	–10.	–9.1	34.
	&	–4.7	–40.	–36.	–30.	–16.	–20.	–17.	13.
	+	33.	–7.7	–14.	–17.	–2.9	–20.		5.8
Intron	–	38.	9.9	7.2	–5.8	5.6	–2.9		48.
	&	36.	2.2	–1.1	–15.	1.6	–13.		45.
	+	40.	5.6	16.	2.8	6.6	–6.0	–6.4	140.
IGD	–	42.	8.4	23.	3.1	2.5	–6.1	4.0	120.
	&	41.	8.0	21.	3.4	8.5	–4.9	7.3	190.

Notes: Percentage differences are underlined when ≥33% (solid blue) or ≤–33% (dotted red), except when actual mol percentages are ≤0.5%.

Table 3-4. Positive skew of distributions: percentages of segments with sub-mean mol percentages of G**G**G in exon, intron and IGD populations with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	62	63	64	64	61	**70**	56	**76**
Exon	–	62	61	65	63	60	64	62	**72**
	&	60	58	62	62	58	64	55	**72**
	+	**69**	58	**67**	58	65	58		61
Intron	–	**69**	**70**	**67**	60	**68**	60		**66**
	&	**67**	65	60	**68**	65	58		61
	+	**66**	53	61	60	61	58	62	65
IGD	–	**67**	53	63	60	61	60	63	63
	&	**66**	52	60	58	58	57	60	60

Notes: Population percentages are highlighted when ≥66% (bold blue).

Table 3-5. Percentage differences between observed and probability-predicted values of G**G**G and **G**G exclusion (Table 3-2), suppression (Table 3-3) and skew (Table 3-4) on single strands (avg. coding and template).

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	2	**18.**	**75.**	**61.**	**48.**	**30.**	**24.**	44.	–5.6
Exon	3	–3.9	**–36.**	**–28.**	**–25.**	**–14.**	**–18.**	**–14.**	**17.**
	4	–1.2	3.7	1.5	0.32	–3.1	–1.6	–3.5	–0.19
	2	–10.	8.4	8.2	**26.**	**13.**	**14.**		–7.5
Intron	3	**36.**	1.1	–3.3	**–11.**	1.4	**–11.**		**27.**
	4	–2.3	5.1	2.3	–8.1	**10.**	8.8		5.0
	2	–2.6	9.9	1.2	**46.**	3.9	**21.**	**–19.**	**–69.**
IGD	3	**41.**	7.0	**19.**	2.9	4.6	–6.1	–1.2	130.
	4	–1.5	–4.6	1.7	–1.6	–2.1	–3.8	–7.6	–4.0

Notes: Entries are color-coded purple ≥50%; blue ≥33%; green ≥10%; black <10% and >–10%; orange ≤–10%; red ≤–33%; pink ≤–50%, except when they would not be underlined in Table 3-2 or highlighted in Table 3-3.

Table 3-5. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	2	14.	41.	110.	66.	9.2	37.	–34.	4.8
Exon	3	**11.**	5.8	1.2	4.4	9.1	7.1	5.1	**26.**
	4	–0.67	1.8	2.6	1.9	–1.6	2.0	1.1	–0.74
	2	32.	**59.**	42.	**110.**	**180.**	**69.**		–20.
Intron	3	**14.**	6.6	**11.**	–1.4	–2.9	3.5		**51.**
	4	–2.2	3.8	5.2	0.10	4.8	3.5		–0.70
	2	46.	48.	80.	120.	110.	32.	16.	–27.
IGD	3	**15.**	1.3	2.4	2.3	1.0	–0.59	**15.**	**60.**
	4	–2.7	2.1	0.68	–0.70	–3.1	2.8	–0.67	–4.0

Table 3-6. Percentage asymmetry of ΣG, G**G**G and **G**G in exons and introns.

		Hsa	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Exon	ΣG	1.1	–0.28	1.9	8.0	3.2	1.5	17.	16.
	ΣG	2.8	0.91	2.0	9.3	4.3	5.1	16.	15.
Exon	G**G**G	–3.1	–5.2	1.2	28.	5.2	2.8	36.	20.
	GG	4.4	4.2	8.0	20.	7.3	11.	25.	18.
	ΣG	1.3	–3.4	5.2	3.8	0.25	8.1		8.0
Intron	G**G**G	0.32	–31.	–8.0	–6.0	–6.2	–3.8		–29.
	GG	0.77	–12.	6.1	12.	–2.8	17.		10.

Notes: First exon ΣG asymmetry calculated from the Codon Usage Database.

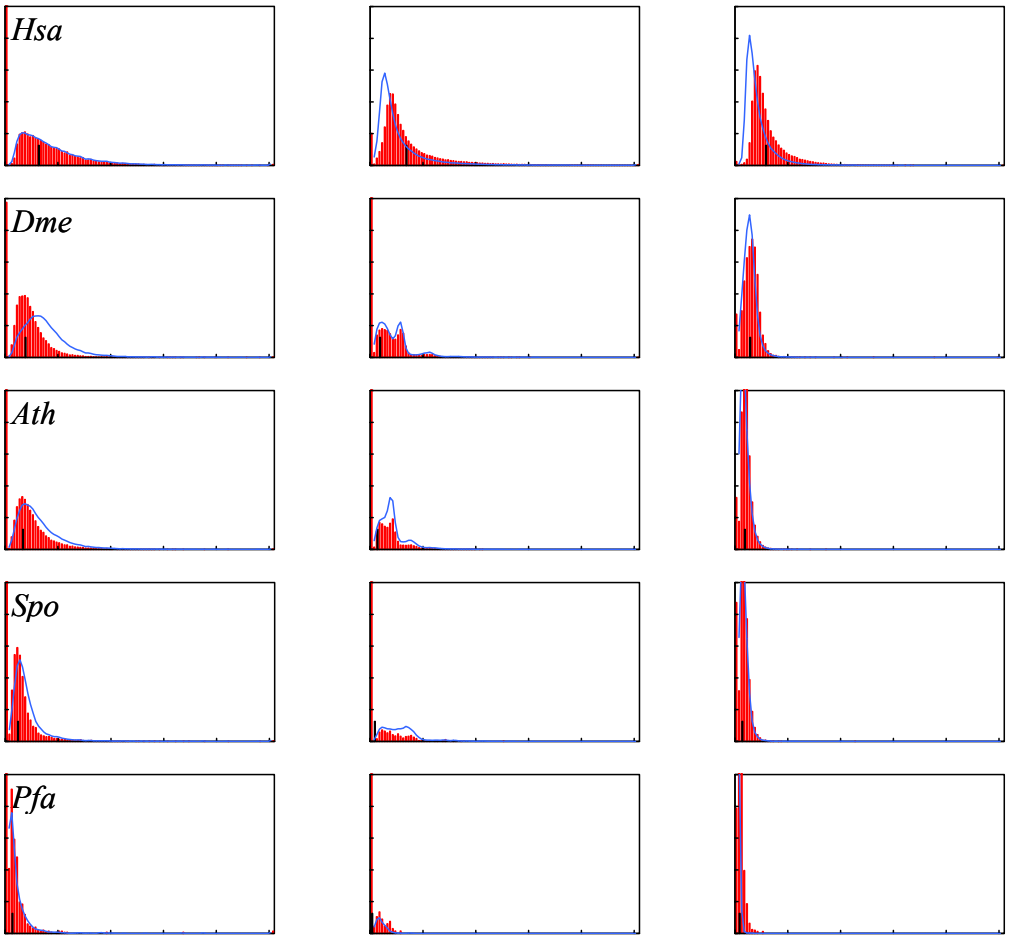


Figure 3-1. Observed (red bars) and probability-predicted (blue lines) frequencies, and observed means (black bars) of G**G**G mol percentages on coding strands. Notes: See Table 3-2 for frequencies at 0 %, which often are off scale. Charts in each row show exons, introns and IGD, from left to right. Vertical full scales are 20% of population. Horizontal full scales are 15% and intervals are 0.15%.

Chapter 4: Evaluation of Genome Designs for Oxidation Resistance: Guanine Minimization and Scavenger Guanine

ABSTRACT

The genome's environment contains strong oxidizers, some of which selectively attack guanine, the most readily oxidized nucleotide. The ranking of guanine oxidation rates is central G in GGG (G**G**G) ≥ 5' G in GG (**G**G) > isolated or 3' G. Beyond enzymatic elimination of the oxidizing agents and their precursors, and excision and repair of oxidative lesions, we propose that genomes are built to mitigate damage to essential domains.

Resistance to oxidation could be enhanced by making genomes more "noble" by reducing the fractions of total G, **G**G and particularly G**G**G. Alternately, if the duplex conducts electron vacancies (holes) over ~100 bp, oxidation could be shifted from essential domains to sacrificially oxidizable G**G**G and **G**G in nonessential domains. The distribution of G**G**G and **G**G in exons, introns and intergenic domains of eight model genomes suggests ennoblement in six, protection by sacrificial anodes in one, and no guanine-based protection in one (*E. cuniculi*). G**G**G triads are excluded or are statistically underrepresented in exons and short splicing-controlling introns of *D. melanogaster, C. elegans, A. thaliana, S. cerevisiae, S. pombe* and *P. falciparum* chromosomes 2 and 3. The introns of *H. sapiens*, which are about twenty times longer than those of the other organisms, are rich in sacrificially oxidizable G**G**G triads that are 50-100 bp from the exons. Their frequency correlates with the presence of protection-requiring G**G**G triads in the exons.

INTRODUCTION

Oxidation and Protection of DNA. The DNA in each human cell undergoes ~10^4 oxidative attacks each day (1, 2) resulting in ~10^4 to ~10^5 oxidative lesions at steady

state (3) and requires many levels of protection and repair. Enzymes and antioxidants detoxify strong oxidizing agents and are the first line of defense against them (4-6). Next, DNA protects itself: the bases' oxidation potentials are comparable to those of noble metals (data in (7, 8)), the DNA is wrapped in and around histone proteins in chromatin (9) which substantially but imperfectly shield it (10-12), and, we propose, the guanine distribution mitigates oxidation. The last line of defense before apoptosis (13) is repair of oxidative lesions (14, 15). Rapid removal by base-excision repair, transcription-coupled repair and other systems limit the number of 8-oxo-2-deoxyguanosine (8oxodG), the initial and principle product of guanine oxidation, to ~10^4 lesions at steady state (3, 16).

The genome's environment contains extremely strong but short-lived oxidizers that attack indiscriminately any base, and strong and more persistent oxidizers that attack selectively the more readily oxidized bases (17). The ranking of nucleotide oxidation potentials is C > A ≈ T > G (8), and the ranking for guanine oxidation rates is **G** ≈ G**G** << **G**G ≤ G**G**G (18-23), where **G** is the isolated G in HGH, G**G** is the right or 3' G in HGG, **G**G is the left or 5' G in GGH, G**G**G is the central G in GGG and H is any base but G. (A nucleotide triplet with a bold G denotes one specific nucleotide, G**G**G refers to the central G in the sequence GGG, whereas a triplet without a bold G refers to the whole sequence, e.g., GGG refers to the triad GGG. The sum of **G**, **G**G, G**G** and G**G**G is ΣG, the total guanine.) Selectivity for **G**G and G**G**G is based on their lower oxidation potentials and is enhanced by positive charge (hole) transport in DNA (21, 24-27) that channels holes to them, both *in vitro* and *in vivo* (11, 28-32).

These are imperfect generalizations about guanine oxidation. The oxidation rate of a particular guanine can reflect the site's affinity for the oxidants and or sensitizers (33, 34) as well as its oxidation potential. Guanine in CG sequences and 8oxodG also are readily oxidized (19, 35). Clinical frequencies of oxidative mutations are comparable at **G**, **G**G and G**G**G in a large human database (clinical observations may not reflect underlying oxidation rates) (36, 37).

Guanine Distribution and Oxidation Resistance. The selectivity of selective oxidizers for guanine, especially **G**G and G**G**G, implies that the guanine distribution affects genomic resistance to selective oxidants. The guanine distribution can be characterized compositionally and positionally by the mol percentages of ΣG, **G**, **G**G and G**G**G in exons, introns and inter-gene domains (IGD) and the mean mol percentages of these nucleotides at locations in these segments. In a hypothetical random distribution, the mean mol percentage of ΣG is 25%, the mean mol percentages of **G**, **G**G and G**G**G equal what probability predicts from ΣG (Equations 4-1 to 4-4), and guanine nucleotides are randomly dispersed over the genome. Averaged over sufficiently large numbers (>1000) of exons, introns and IGD, random dispersion gives mean mol percentages that are approximately uniform, varying little with location. (Note: Randomly dispersed ΣG with a mean of 25% and a standard deviation of 43%, has a standard error of the sampled mean of 1.3% for 1000 segments (see (38)). The standard deviation of ΣG is $(100\%)\times((\Sigma A\ \%/100)(0-\Sigma G)^2 + (\Sigma T\ \%/100)(0-\Sigma G)^2 + (\Sigma G\ \%/100)(100\%-\Sigma G)^2 + (\Sigma C\ \%/100)(0-\Sigma G)^2)^{0.5}$, where ΣA is the mol percentage of adenine.) With hypothetical random distributions considered neutral, a real distribution would be considered favorable for oxidation resistance if it is systematic such that the impact of selective oxidation is mitigated and unfavorable if it is systematic such that the impact is exacerbated. Evaluation of a guanine distribution by inspection is indicative but provisional, because we have sequenced genomes, substantial understanding of DNA oxidation and established engineering practices against corrosion, but we have only evolving understanding of the complex biological gestalten.

$$\text{Prob. } \mathbf{G}\ \% = (100\%) \times (1 - (\Sigma G\ \%/100))^2 \times (\Sigma G\ \%/100) \quad \text{(Eq. 4-1)}$$

$$\text{Prob. } \mathbf{G}G\ \% = (100\%) \times (1 - (\Sigma G\ \%/100)) \times (\Sigma G\ \%/100)^2 \quad \text{(Eq. 4-2)}$$

$$\text{Prob. } G\mathbf{G}\ \% = \text{Prob. } \mathbf{G}G\ \% \quad \text{(Eq. 4-3)}$$

$$\text{Prob. } G\mathbf{G}G\ \% = (100\%) \times (\Sigma G\ \%/100)^3 \quad \text{(Eq. 4-4)}$$

Guanine distributions can be evaluated against engineering principles for corrosion prevention through materials selection, structural design and removal of corrosives (39, 40). Materials with greater oxidation resistance make structures more oxidation resistant. Genomes and domains within genomes that minimize ΣG and minimize **G**G and G**G**G relative to other guanine nucleotides should resist selective oxidants better than those that do not. ΣG, **G**G and G**G**G minimization is constrained in domains where the sequence is the function, e.g., in exons where nucleotide triplets code for amino acids and in intron domains that signal splicing (10, 15). Electrical connection between an essential part that is more readily oxidized and an expendable part that is more oxidation resistant increases corrosion, because it allows the former to be corroded by oxidants attacking the latter (corrosion cell) (41). Conversely, electrical connection between an essential part that is more oxidation resistant and an expendable part that is more readily oxidized reduces corrosion, because it allows the latter to act as a sacrificial anode cathodically protecting the former. Essential genome domains that have lower guanine levels, especially **G**G and G**G**G levels, than neighboring expendable domains, e.g., exons with their splice sites that have lower levels than neighboring introns, should mitigate oxidative damage better than those that have the reverse. Exons that have more **G**G and G**G**G than their neighboring introns, but which are electrically insulated by the intervening DNA, as described by Barton et al. (26, 42-44) (also (28)), should be more oxidation resistant than those that are electrically connected. Removing oxidants from the environment or blocking them with a coating reduces corrosion. Genomes that maximize guanine levels in expendable domains to do oxidant scavenging in parallel with antioxidants and enzymes in the nucleoplasm reduce selective oxidation better than those that do not.

A guanine distribution with all of the characteristics favorable for oxidation resistance goes against some genomic propensities. The guanine minima possible in an expendable domain may be lower than those possible in a neighboring essential domain,

because the essential domain inherently is more functionally constrained. While some of their sequences are functional, introns and IGD do not have to be translated *in toto* into efficient enzymes, exons do. Mutational bias due to the differences in oxidation resistance among guanine nucleotides is a plausible mechanism for minimizing **G**G and G**G**G relative to other guanine nucleotides (Chapter 3), but this mechanism should be more active in non-functional domains of introns and IGD, which mutate faster than exons (45). The ideal distribution minimizes guanine in essential domains to make them more noble and maximizes it in neighboring expendable domains to make them better oxidant scavengers. Two simpler distributions that emphasize guanine minimization or guanine scavengers should resist selective oxidants better than a hypothetical random distribution.

Guanine Minimization. Heller (46) suggested that minimizing guanine levels could make DNA more resistant to oxidation, just as minimizing copper levels makes gold-copper alloys more corrosion resistant, e.g. 14 vs. 18 karat gold. Genomes and domains within genomes with lower ΣG mol percentages should resist selective oxidation better than those with higher ΣG mol percentages, if **G**G and G**G**G mol percentages equal what probability predicts from ΣG (Equations 4-2 and 4-4), because A, T and C are more noble than G. Genomes and domains with lower **G**G and/or G**G**G mol percentages should be more oxidation resistant than those with higher **G**G and/or G**G**G mol percentages, if ΣG levels are the same, because other guanine nucleotides are more noble than **G**G and G**G**G. Genomes and domains that minimize **G**G should resist selective oxidation better than those that minimize G**G**G, even though **G**G and G**G**G oxidation rates are similar. If **G**G and G**G**G are minimized such that they both are lower than probability predicts by the same factor, the **G**G minimization eliminates more nucleotides, because probability predicts that **G**G is 3.0 to 6.7 times more prevalent than G**G**G at 25% to 13% ΣG.

Minimizing ΣG minimizes **G**G and G**G**G more rapidly, because their mol percentages scale with $(\Sigma G)^2$ and $(\Sigma G)^3$, respectively (Equations 4-2 and 4-4). Probability

predicts mean **G**G and **G**GG mol percentages in *Pfa* exons (1.5% and 0.22%) that are three and seven times smaller than those it predicts in *Hsa* exons (4.7% and 1.6%), because the mean ΣG mol percentage in *Pfa* exons (13%) is a factor of two lower than that in *Hsa* exons (25%) (Table 4-S1). This simple strategy probably is evolutionarily penalizing, because deviation from a one-to-one A:T-to-G:C ratio sharply reduces the possible number of different arrangements of nucleotides. At 13% ΣG, the number of possible combinations is $>10^{500}$ times smaller than the number at 25% ΣG in a hypothetical 10^4 bp genome (Equation 4-5 (see (47)) using PAPCW (48)).

Non-specific minimization of **G**G and G**G**G, resulting from ΣG minimization, can be augmented by specific minimization below what probability predicts from ΣG. The combinatorial penalties incurred by specifically minimizing **G**G and G**G**G are far smaller than those incurred by non-specifically minimizing **G**G and G**G**G by minimizing ΣG. At 13% ΣG with 1.5% **G**G and 0.22% G**G**G, the number of possible combinations is $>10^{750}$ times smaller than the number at 25% ΣG with equal **G**G and G**G**G mol percentages in a hypothetical 10^4 bp genome (Equation 4-6). **G**G and G**G**G minimization have different effects. In exons, G**G**G minimization involves two single codons (GGG and CCC), but **G**G minimization involves 12 single codons, including all of the codons for glycine and proline.

$$\text{Combinations} = ((N_{\Sigma A} + N_{\Sigma T} + N_{\Sigma G} + N_{\Sigma C})!) \div (N_{\Sigma A}! \times N_{\Sigma T}! \times N_{\Sigma G}! \times N_{\Sigma C}!) \qquad \text{(Eq. 4-5)}$$

$$\text{Combinations with GG and GGG Specified} = ((N_A + N_T + (N_{\Sigma G} - 2N_{GG} - 3N_{GGG}) + N_{GG} + N_{GGG} + (N_{\Sigma C} - 2N_{CC} - 3N_{CCC}) + N_{CC} + N_{CCC})!) \div (N_A! \times N_T! \times (N_{\Sigma G} - 2N_{GG} - 3N_{GGG})! \times N_{GG}! \times N_{GGG}! \times (N_{\Sigma C} - 2N_{CC} - 3N_{CCC})! \times N_{CC}! \times N_{CCC}!) \qquad \text{(Eq. 4-6)}$$

where, $N_{\Sigma G}$ is the number of guanine nucleotides and

N_{GG} is the number of guanine pairs (GG sequences).

Scavenger Guanine. Genomic and nucleoplasmic guanine could be a significant scavenger of selective oxidants in the nucleus. Guanine is selective for strong oxidants: 1.0 V E°' for guanine (8) vs. –0.24 V E°' for glutathione (49), where E°' is the standard reduction potential at pH 7. Genomic guanine is distributed throughout the nucleus, because chromosomes are tightly condensed only during cell division (10, 15). The guanine concentration in the nucleus is comparable to that of other antioxidants: ~30 mM genomic ΣG in *Hsa*, ~1 mM genomic ΣG in *Ath* and ~0.5 mM free ΣG in *Hsa*, and vs. ~1 to ~10 mM glutathione in mammalian cytoplasm (50) and plant chloroplasts (51). (Note: ΣG concentrations are estimated from 3×10^9 bp and 1×10^8 bp with 20% ΣG in the *Hsa* and *Ath* genomes, and 2×10^7 nt ΣG in the *Hsa* nucleotide pool in a ~5 μm diameter nucleus (52).) However, genomic guanine is less reactive, because it is buried in the double helix and shielded by histone proteins in nucleosomes (10-12). Guanine oxidation is slower in DNA than in free dGTP (53); purine oxidation by singlet oxygen is ~10 times less in nucleosomes than in denatured DNA with (54); and reaction of hydroxyl radicals with backbone sugars occurs primarily at ~3 of every 10 bp when DNA is in nucleosomes whereas it occurs generally when DNA is in solution (55, 56).

The guanine in neighboring introns and/or IGD could be the best scavenger of an oxidizing hole in the DNA duplex or an oxidant molecule near an exon and its splice sites. An oxidant strong enough to oxidize guanine rarely travels far before reacting, e.g., $OH^{\bullet}$ averages 30 Å (~9 bp along the helix) in mammalian cells (57-59), and the hole it injects travels ≤200 bp (see below). Global scavengers can reduce the oxidant concentrations and the global rate of exon oxidation, but only local scavengers can protect individual exons and splice sites. Individual guanine nucleotides, especially **G**G and G**G**G, can scavenge for whole exons, because their oxidation rate (years per exon) is much lower than their repair rate (days per nucleotide), and un-repaired 8oxodG can be oxidized fur-

ther and more readily than guanine (35, 60). (Note: In the human genome, ~10^4 oxidations daily (1, 2) on 3×10^9 bp means ~2000 days between attacks on a average 160 bp exon. ~10^4 8oxodG at steady state (3, 16) means ~1 day to repair, assuming the oxidation rate is independent of and the repair rate is first-order in 8oxodG concentration.)

Local scavenging could be enhanced by deploying guanine strategically rather than uniformly in introns and IGD. Genomes and domains within genomes that over-represent **G**G and G**G**G in introns and IGD should reduce exon and splice site oxidation more than those that have only the **G**G and G**G**G mol percentages that probability predicts from ΣG, because **G**G and G**G**G are the most readily oxidized guanine nucleotides. Genomes that increasing intron and IGD ΣG mol percentages when adjacent exon mol percentages increase should reduce exon and splice site oxidation more than those that do not, because they keep introns and IGD competitive with exons for selective oxidants. Genomes and domains that increase guanine levels in intron and IGD flanks relative to overall levels should reduce exon and splice site oxidation more than those that distribute guanine uniformly, because they deploy scavengers closer to targets. Genomes and domains that use charge transport in DNA to make scavengers function as sacrificial anodes should mitigate oxidative damage better than those that do not.

Sacrificial Anodes in Intron Flanks. Given a conductive path, ΣG, **G**G and G**G**G in intron and IGD flanks could act as sacrificial anodes, like zinc plates protecting steel ships, drawing holes out of exons (46, 61). Barton et al. (31, 33, 62, 63) have investigated this chemistry for at least a decade, and write: "It will also be important to determine whether organisms have evolved to protect their genomes from long-range damage (via charge transport). Perhaps radical damage is funneled to or insulated from specific sites within the genome. One could, however, consider that segments throughout the genome may encode “sinks” for damage, and that other segments could serve as buffers as a result of local sequence-dependent or protein-dependent structural deformations to protect critical regions." Giese (25) writes "Thus, a hole injected by oxidation processes into

a G of an encoding area has a high probability of migrating into the G:C-rich sequence (outside of the encoding area), so that the mutation occurs in the non-encoding part of the DNA." Kawanishi et al. (64) write: "GGG triplets can act as traps in oxidative damage to double-stranded DNA caused by long-range electron transfer." and "It is convenient to imagine that non-coding (GGG-rich) regions such as telomeres and introns may protect chromosomes against oxidative stress-induced toxicity." Thorp et al. (65) write "These results raise the possibility that guanine triplets in telomeric sequences act as sinks for oxidative damage in vivo, although direct evidence -- including information on site-specificity and products -- is not yet available." Kawanishi et al., Thorp et al. and Heller (46) consider both sacrificial anodes and simple scavengers, because CpG islands are up-stream of genes and telomeres are at the ends of chromosomes, out of range of charge transport (31, 63). Schuster et al. (66) have shown that disulfides on molecules intercalated in DNA can serve as sacrificial anodes.

ΣG, **G**G and G**G**G in intron and IGD flanks can be sacrificial anodes for neighboring exons only if the intervening DNA electrically connects them, and its conductivity is limited. Giese et al. (25), Jortner, Bixon et al. (67, 68) and Schuster et al. (69, 70) are developing theoretical models to explain *in vitro* experiments that show hole transport over 40 to 200 Å, 10 to 60 bp in DNA. Barton et al. (11, 31, 63) have shown remote oxidation over 24 bp in nucleosome core particles, and write "Our studies on long-range damage on restriction fragments suggest that the physiological range of charge migration may be on the order of 100 bp, but probably not longer." Guanine sacrificial anodes cannot protect the central nucleotides of exons over ≥120 bp long. Charge transport can be obstructed by the sequence or structure of the intervening DNA (for examples, see (26, 30, 71)). The putative conduction path needs to be evaluated in each case to firmly suggest sacrificial anode functionality.

Intron Flanks and Exon Splicing. GG and GGG sequences in intron domains flanking exons participate in pre-mRNA splicing, an essential function unrelated to and

apparently conflicting with their putative role as sacrificial anodes. The signal sequences for splicing almost all introns (GT-AG or GC-AG motifs) contain GG at most 5' splice sites (3' exon-AG/GYRAGT-intron 5', where R is A or G, and Y is C or T) in *Hsa*, *Dme*, *Cel*, *Ath* and *Sce* and at a plurality of 3' splice sites (3' intron-YAG/N-exon 5') in *Hsa* and *Ath*, but not in *Dme*, *Cel* and *Sce* ((72-74) and from ISIS (75) and SpliceDB (76, 77)). The signal sequences for splicing a tiny fraction of introns (AT-AC) do not have GG at either splice site (15). The poly-pyrimidine track in introns between the branch site and the 3' splice site is cytosine-rich in *Hsa* and somewhat cytosine-rich in Arthropoda (*Dme*), but cytosine-poor (thymine-rich) in Nematoda (*Cel*), Magnoliopsida (*Ath*), and Aveolata (*Pfa*) ((72, 73, 78) and from ISIS)). Ascomycota (*Sce* and *Spo*) lack a strong poly-pyrimidine track. Hence, the 5' flank of the template strand of introns is guanine-rich in *Hsa*. The branch site, the eye of the lariat formed during splicing, is 15 bp to 50 bp from the 3' intron splice site and its consensus sequence lacks GG and GGG (CTRAY in *Hsa*, *Dme*, *Ath* and *Sce*) (72, 73, 78). GGG is over-represented in the 5' flank of the coding strand of human introns (79-81), particularly in short introns (82). This feature facilitates prediction of 5' splice sites (83). GGG sequences in the 5' flank of vertebrate introns enhance splicing, particularly of small exons and introns (78, 82, 84-87).

About 15% of human genetic diseases are due to mutations that generate a new splice site or destroy functional splice sites (88). "Most of these mutations directly affect the canonical consensus sequences that define exon-intron boundaries", but mutations at exonic or intronic splicing enhancers also cause faulty splicing (82, 89).

The involvement of **G**G and G**G**G in pre-mRNA splicing compromises the guanine scavenger emphasis. The **G**G nucleotides at the splice sites are perfectly positioned to act as sacrificial anodes for the adjacent exon, but they are as essential as any **G**G within the exon. Genomes and domains within genomes that deploy guanine scavengers to protect splice sites as well as exons should mitigate oxidative damage better than those that use the **G**G nucleotides at splice sites as sacrificial anodes. The guanine nucleotides

of the poly-pyrimidine track in the 3' flank of *Hsa* and *Dme* introns and the GGG sequences of splicing enhances in the 5' flank of some introns of *Hsa* are well positioned to act as oxidant scavengers for adjacent exons, but they also promote accurate splicing. Genomes and domains using these guanine nucleotides as scavengers shift the oxidative damage from essential exons to important splicing domains, and the benefits of this are ambiguous.

Employing **G**G and G**G**G for biological functions obviously exacerbates the impact of oxidation. Other considerations apparently can trump mutation resistance in these and other cases. For example, the CG sequence is highly mutable via a methylation-deamination mechanism causing C→T transition (90), but "CpG islands" are promoters for some genes and CG methylation is an essential element of the control of gene expression (10, 15).

We analyzed the guanine distributions in eight model genomes to assess them against these criteria. Guanine minimization results were presented earlier, in a discussion of selective oxidation as a source of mutational bias for **G**G and G**G**G elimination (Chapter 3). Results related to guanine scavengers are presented here.

METHODS

As previously described (Chapter 3), we analyzed genome sequences from GenBank (91) for *Hsa* (Feb. 2002 release) (92, 93), *Dme* (Oct. 2000) (94), *Cel* (Dec. 2001) (95), *Ath* (Jan. 2002) (96), *Sce* (Mar. 2002) (97), *Spo* (Mar. 2002) (98), *Ecu* (Mar. 2002) (99), and *Pfa* chromosomes 2 (Nov. 1998) (100) and 3 (Apr. 1999) (101). Each guanine nucleotide on the given strand in the GBK file was identified as isolated with non-G neighbors (**G**), one of a pair with non-G neighbors, or central in a triplet (G**G**G). The count of paired nucleotides was divided by two for the **G**G and G**G** counts. Nucleotides on the complementary strand were inferred from base pairing. For example, -TCTGGAGGGTCCTGT- had one **G**, two **G**G, two G**G** and one G**G**G on the coding strand, and one **G**, one **G**G and one G**G** on the template strand.

Each nucleotide was identified as part of an annotated exon (coding sequence or CDS), intron or IGD, and its distances from the 5' and 3' boundaries of its segment were calculated. If an intron or IGD nucleotide was ≥6 bp and ≤105 bp from its boundaries, it was further specified as part of a flank. Flanks were defined as ≤100 bp long segments of introns or IGD found ≥6 bp and ≤105 nt from their ends. Thus, each intron or IGD had four flanks: (1) one at the 5' end of its coding strand, flanking the 3' end of the adjacent exon, (2) one at the 3' end of its coding strand, flanking the 5' end of the adjacent exon, (3) one at the 3' end of its template strand, complementary to the 5' flank of its coding strand, and (4) one at the 5' end of its template strand, complementary to the 3' flank of its coding strand. When an intron or (rarely) IGD was <112 bp long, its flanks were identified as the domain ≥6 bp from both of its ends. The five end nucleotides were omitted from flanks, but the branch site nucleotides were included.

For each exon, intron and IGD, we calculated the mol percentages of total ΣG, **G**, **G**G and G**G**G in the segment overall and in its flanks. We weighted each segment equally, regardless of length, to calculate mean mol percentages. These segment-weighted averages described average exons, introns and IGD; nucleotide-weighted averages would have described average nucleotides within exons, introns and IGD. We plotted the mean mol percentages of each nucleotide type vs. distance from the 3' or 5' segment boundaries. The number of segments in these averages decreased as length increased. Introns shorter than 100 bp were plotted separately, and exons shorter than 25 bp were not plotted.

RESULTS

The Tables and Figures in this section highlight the distribution of G**G**G, because this distribution is more systematic, less uniform. **G**G results are in the supplemental materials in Appendix 2 (Table and Figure numbers with "S"). In these tables, single DNA strands are denoted (+) for coding or (–) for template, and hybridized double strands are denoted (&). Ends of single strands are denoted (5) for 5’ and (3) for 3’, and ends of

double strands are denoted (53) for hybridized coding 5' end and template 3' end, and (35) for hybridized coding 3' end and template 5' end.

Correlation of Intron or IGD and Exon G**G**G **and** G**G** **Levels.** Tables 4-1 and 4-S2 list η^2 values from analysis of variance (ANOVA) for correlation between intron or IGD and neighboring exon mol percentages of G**G**G and **G**G. ANOVA η^2, like regression r^2, is the fraction of variance in the dependent variable explained, linearly or nonlinearly, by the independent variable. ($\eta^2 = 0.23$ for correlation of G**G**G in *Hsa* intron and exon combined strands means that 23% of the G**G**G variation in a typical *Hsa* intron was related to variation in the mean G**G**G of its two neighboring exons.) The G**G**G η^2 values of combined strands generally were comparable or greater than the values of their component strands. The η^2 values for intron/exon correlation were roughly comparable to those for IGD/exon correlation. Almost uniformly, **G**G η^2 were greater than corresponding G**G**G η^2.

Only *Hsa* η^2 values exceeded 0.2 (20% correlation); others were ≤0.03 for G**G**G and ≤0.11 for **G**G. ANOVA f-test values were >500 for *Hsa* intron/exon correlations and >100 for *Hsa* IGD/exon correlations, indicating statistical significance (Table 4-S3). Figures 4-1, 4-S1 and 4-S2 are intron vs. exon and IGD vs. exon scatter plots of G**G**G and **G**G mol percentages. As these figures show, only small percentages of the intron and IGD populations were neighbors of exons with the highest G**G**G and **G**G. In *Hsa*, average introns neighboring exons with the lowest, the mean and the highest combined strand G**G**G and **G**G mol percentages had 2.2%, 4.0%, 9.2% G**G**G, and 6.5%, 9.4%, 13.% **G**G, respectively. Average *Hsa* IGD neighboring exons with the lowest, the mean and the highest combined strand G**G**G and **G**G had 2.4%, 3.4%, 5.8% G**G**G, and 7.%, 8.8%, 12.% **G**G, respectively.

Distributions of G**G**G **and** G**G** **in Flanks of Introns and IGD.** Tables 4-2 and 4-S4 list and Figures 4-2 and 4-S3 to 4-S6 show the mean mol percentages of G**G**G and **G**G in intron and IGD flanks, ≤100 bp long segments at ≥6 bp from their ends. Mol per-

centages in single strands are guanine nucleotides per hundred total nucleotides, but mol percentages in combined strands are guanine nucleotides per hundred total base pairs. The mean mol percentages in combined strands are the means of the sum of coding and template mol percentages. The percentage standard deviations were ~50% to ~500%, reflecting true excursions, not statistical uncertainties (Table 4-S5). *Ecu* introns were not analyzed, because there were fewer than 100 of them.

The mean mol percentages of G**G**G and **G**G in flanks declined in the order *Hsa* > *Ecu* > *Dme* > *Cel*, *Ath*, *Sce*, *Spo* > *Pfa*. In the latter two groups, the mean mol percentages of G**G**G in single strands of intron flanks were nearly always ≤0.5%, less than one-fourth of those of *Hsa*. Flank G**G**G values in *Hsa* introns and IGD, and *Cel* and *Ecu* IGD template strands equaled or exceeded their exon G**G**G values. At many other positions in Tables 4-2 and 4-S4, the percentage differences between flank and overall G**G**G and **G**G were ≤–33%.

Tables 4-3 and 4-S6 list the percentages of intron and IGD flanks with no G**G**G or no **G**G. The percentages of single strand flanks with no G**G**G were ≥66% in many positions, notably in *Cel*, *Ath*, *Sce*, *Spo* and *Pfa*. An average of 34% of the single strand flanks of *Hsa* introns and IGD had no G**G**G. The percentages of double strand flanks with no G**G**G were about half the averages of their single strand values in *Hsa* and *Ecu*, but about three-quarters of the averages of their single strand values in *Cel*, *Ath*, *Sce*, *Spo* and *Pfa*. The percentages of single strand flanks with no **G**G averaged a factor of 9 less than the percentages with no G**G**G in *Hsa*, a factor of 4 less in *Cel*, *Ath*, *Sce*, *Spo* and *Ecu*, and a factor of 2 less in *Pfa*.

Tables 4-4 and 4-S7 list the percentage differences between flank and overall mean mol percentages of G**G**G and **G**G in introns and IGD with G**G**G and **G**G, respectively. Tables 4-5 and 4-S8 list the absolute differences between the flank and the overall mean numbers of G**G**G and **G**G in 100 nt (nominal flank length) of introns and IGD with G**G**G and **G**G, respectively. In *Hsa*, G**G**G and **G**G were particularly elevated in the 5'-

flanks of introns and in all IGD flanks. The 5' flank of intron template strands contains the poly-purine track, complementary to the poly-pyrimidine track. These elevations typically added ~1 G**G**G and ~1 **G**G for every two intron 5'-flanks, and ~1 G**G**G and ~2 **G**G for every two IGD flanks. Comparable elevations were not found in other organisms. G**G**G and **G**G were elevated in the 5'-flanks of IGD template strands in *Dme*, *Ath* and *Spo*, the increment being on average ~0.4 G**G**G and ~2 **G**G for every two elevated flanks.

Table 4-S9 lists the mean mol percentages of G**G**G and **G**G in intron flanks and whole introns with G**G**G and **G**G, as a function of their length. IGD were omitted, because very few were shorter than 200 bp. *Sce* and *Pfa* introns 100 bp or shorter were omitted, because there were fewer than 100 of them. In all eight organisms, introns shorter than 200 bp (two flank lengths) were substantially richer in G**G**G than the average of all introns. G**G**G enrichment averaged 2-fold in *Hsa*, *Cel* and *Sce*, and 1.3-fold in *Dme*, *Ath*, *Spo* and *Pfa*. Introns shorter than 200 bp were richer than average in **G**G only in *Hsa* (1.3-fold).

DISCUSSION

Guanine Distributions Unrelated to Oxidation. The distribution of guanine in a genome is only one of many factors that affect its resistance to selective oxidants, and oxidation is only one of many factors that affect the guanine distribution within and between genomes (see reviews regarding ΣG (102-104)). For example, the association of survival in highly oxidizing environments and total guanine poverty is suggestive, but confounded by differences between organisms and counter examples. *Plasmodium falciparum* metabolizes hemoglobin releasing reactive oxygen species (105), and it has <10% ΣG (100). However, *Deinococcus radiodurans* survives massive doses of ionizing radiation by emphasizing repair and detoxification, and it has ~32% ΣG (6, 106-108).

ΣG mol percentages vary widely, not only between, but also within some eukaryote genomes. (Note: Originally stated in ΣG+ΣC, this information is re-stated in ΣG,

because in eukaryotes ΣG = ΣC in long DNA strands (no asymmetry) (90, 109), though not in human exons at the third codon position (110).) Vertebrate genomes are mosaics of long (>300 kb), compositionally homogeneous (averaged over 3 kb) segments (isochores) whose ΣG mol percentages fit into a small number of families (111). Light isochores (ΣG < 22%) comprise 63% of *Hsa*, ~80% of *Dme* and ~100% of *Cel*, *Ath* and *Sce* (112, 113). Compositional homogeneity within isochores implies positive correlation between exons and their neighboring introns and/or IGD in ΣG levels.

ΣG+ΣC-rich exons generally are surrounded by ΣG+ΣC-rich introns and IGD, though the mean ΣG+ΣC mol percentage of exons is higher than that of introns or IGD (Table 4-S1). The coefficient (r) of linear correlation between *Hsa* exon and neighboring intron ΣG+ΣC levels is 0.78 (111). The correlation between the GC_3 level (ΣG+ΣC level at the third codon nucleotide) in exons and the ΣG+ΣC levels in their flanking domains is 0.56 to 0.65 for *Hsa* and 0.38 to 0.55 for *Dme*, varying with flank length (1 to 20 kb) (114). The correlation between the ΣG+ΣC level in exons and the level in the 5' and 3' flanks (50 bp, excluding splice site and poly-pyrimidine tract) of their neighboring introns is ~0.63 in *Hsa*, 0.24 (5') and 0.17 (3') in *Dme*, 0.14 (5') and 0.06 (3') in *Cel*, and <0.10 in *Ath*, *Sce*, *Spo* and *Pfa* (115). The correlation between the ΣG+ΣC level in exons and that in 50 bp segments in the middle of their neighboring introns was much smaller. "Thus, in genomes with a high global heterogeneity there seems to be a selective force for compliance of intron base composition with the adjacent exons. This force is stronger in those parts of the intron that are closer to exons" (115).

Guanine Minimization. ΣG, **G**G and G**G**G are minimized in seven of the eight model genomes. ΣG averages ≤22% in the exons of *Caenorhabditis elegans* (*Cel*, nematode worm), *Arabidopsis thaliana* (*Ath*, flowering plant), *Saccharomyces cerevisiae* (*Sce*, budding yeast), *Schizosaccharomyces pombe* (*Spo*, fission yeast) and *Plasmodium falciparum* (*Pfa*, malaria parasite) chromosomes 2 and 3 (Table 4-S1). ΣG averages 24% to 26% in the exons of *Homo sapiens* (*Hsa*), *Drosophila melanogaster* (*Dme*, fruit fly) and

Encephalitozoon cuniculi (*Ecu*, intracellular parasite). ΣG averages ≤20% in the introns and IGD of all model genomes except *Hsa*. Minimizing ΣG minimizes **G**G and G**G**G more rapidly, because their mol percentages scale with $(\Sigma G\ \%)^2$ and $(\Sigma G\ \%)^3$, respectively (Equations 4-2 and 4-4). Probability predicts mean **G**G and **G**GG mol percentages in *Pfa* exons (2.9% and 0.22%) that are three and seven times lower than those in *Hsa* exons (9.4% and 1.6%), because the mean ΣG mol percentage in *Pfa* exons (13%) is a factor of two lower than that in *Hsa* exons (25%) (Table 4-S1). This simple strategy probably is evolutionarily penalizing, because deviation from a one-to-one A:T-to-G:C ratio sharply reduces the possible number of different arrangements of nucleotides. At 13% ΣG, the number of possible combinations is only 0.001% of the number at 25% ΣG. (The number of combinations is $((N_A + N_T + N_G + N_C)!) \div (N_A! \times N_T! \times N_G! \times N_C!)$, where N_A is the number of adenine nucleotides.)

GG and G**G**G minimization can be non-specific, resulting from ΣG minimization, or specific, below what probability predicts from ΣG. **G**G and G**G**G are non-specifically minimized and G**G**G is specifically minimized in exons, introns and IGD of *Dme*, *Cel*, *Ath*, *Sce*, *Spo* and *Pfa* (Table 3-5 in Chapter 3). G**G**G is statistically under-represented in *Hsa* exons, but over-represented in *Hsa* introns and IGD. It is not minimized in *Ecu* exons, introns and IGD. **G**G generally is not under-represented in *Hsa* and *Ecu*. **G**G and G**G**G minimization conflicts with the genome compaction by constraining codon and hence amino acid selection. This conflict could explain its absence in *Ecu* which rigorously minimizes its genome size (116) (2.5×10^6 bp, 86% exons (Table 4-S1)), but not in *Hsa* ((2.8×10^9 bp, 1.5% exons).

The over-representation of **G**G and G**G**G in *Hsa* introns and IGD suggests the third recommendation for mitigating oxidative damage: deploy ΣG, **G**G and G**G**G as oxidant scavengers and sacrificial anodes.

Scavenger Guanine. For successful global scavenging **G**G and G**G**G in introns and IGD must greatly outnumber **G**G and G**G**G in exons, and they must vastly outnum-

ber oxidants that evade other genome defenses. The intron plus IGD to exon ratios of **G**G plus G**G**G numbers are 56 in *Hsa*, 2.8 in *Dme*, 1.9 in *Cel*, 1.4 in *Ath*, 0.26 in *Sce*, 0.48 in *Spo*, 0.17 in *Ecu* and 0.45 in *Pfa* (from Table 4-S1). In a human cell, DNA oxidations are minuscule (~10^4 per day (1, 2)) relative to the numbers of **G**G and G**G**G in introns and IGD ($3.6x10^8$ nt (from Table 4-S1)).

Introns and IGD, and exons have similar G**G**G levels, and putatively similar vulnerability to selective oxidants, only in the *Hsa* and *Ecu* genomes. Alternately, in *Hsa* and *Ecu*, introns and IGD are composed so that they can compete equally with exons for selective oxidizing agents. IGD are only 14% of the *Ecu* genome and some of their domains are involved in gene regulation, so they cannot quantitatively compete with exons for holes. At 98.5% of the *Hsa* genome, introns and IGD eclipse exons as oxidation targets. The benefits of this are obvious: oxidative damage and mutation are much better tolerated in nonessential ("junk") DNA than in protein-coding exons. Employing introns and IGD as oxidant sinks does not require decreasing ΣG levels, and thus does not reduce the options for protein design.

In *Hsa*, the mean mol percentage of G**G**G in introns exceed that in exons, and the G**G**G level in IGD nearly equal that exons (Table 4-S1). This occurs despite the mean mol percentages of ΣG in introns and IGD being below that in exons (Table 4-S1), for three reasons. Introns and IGD with no G**G**G are rare in *Hsa* (Table 3-2 in Chapter 3). The differences between actual and probability-predicted G**G**G levels in introns and IGD are substantially positive in *Hsa* (Table 3-5 in Chapter 3). In *Hsa*, G**G**G levels in introns and IGD increase as G**G**G levels in their neighboring exons increase (Table 4-1). In *Ecu*, mean mol percentages of G**G**G in IGD exceed those in exons, but G**G**G exclusions are not rare, differences between observed and probable are positive but not substantial, and the G**G**G in IGD is not correlated with the G**G**G in exons.

In *Hsa*, the mean mol percentages of G**G**G in intron 5' flanks and IGD flanks are elevated above those in introns and IGD overall and exceed those in exons (Tables 4-2, 4-

4 and 4-5). Thus, intron 5' flanks and IGD flanks are better competitors for oxidizing agents and holes than introns and IGD overall, which are better competitors than exons. However, a substantial percentage of *Hsa* flanks have no G**G**G (Table 4-3), and cannot compete with introns and IGD overall which rarely have no G**G**G (Table 3-2 in Chapter 3). Similar elevation is absent from introns of other organisms, perhaps because most of their median-length introns have one or fewer G**G**G in a single strand. Similar elevation is seen in *Dme*, *Ath* and *Spo* on the 5' flank of the IGD template strands.

In the human genome, G**G**G is minimized in exons, but not in introns or IGD. We propose that G**G**G in nonessential intron and IGD domains, especially in their flanks, protected essential exons by trapping holes. For a hole in an exon to diffuse to and be captured by an intron G**G**G, the intervening DNA must conduct holes efficiently. Hole conduction, by the proposed mechanism of hopping from G to G (24, 70, 117, 118), requires a sufficiently high ΣG mol percentage. Hole trapping also requires exons of modest length and introns and IGD with non-essential DNA. Only in the human genome are these three factors common.

ACKNOWLEDGMENTS

Professors George Georgiou, Brent Iverson and Edward Marcotte of the University of Texas, Dr. Jonathan Heller of Optiscan Corp. and anonymous reviewers made very helpful comments. The National Science Foundation, the Robert A. Welch Foundation, National Institutes of Health Biotechnology Training Grant and the Richard J. Lee Endowed Graduate Fellowship in Engineering provided financial support for A.H. and K.A.F.

REFERENCES

1. Helbock, H. J., Beckman, K. B., Shigenaga, M. K., et al. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 288-293.

2. Setlow, R. B. (2001) *Mutat. Res.* **477,** 1-6.

3. Beckman, K. B. & Ames, B. N. (1997) *J. Biol. Chem.* **272,** 19633-19636.

4. Reiter, R. J., Acuna-Castroviejo, D., Tan, D.-X., et al. (2001) *Ann. NY Acad. Sci.* **939,** 200-215.

5. Talalay, P. (2000) *BioFactors* **12,** 5-11.

6. White, O., Eisen, J. A., Heidelberg, J. F., et al. (1999) *Science* **286,** 1571-1577.

7. Bard, A. J. & Faulkner, L. R. (2001) *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York).

8. Faraggi, M., Broitman, F., Trent, J. B., et al. (1996) *J. Phys. Chem.* **100,** 14751-14761.

9. Zlatanova, J., Leuba, S. H. & Van Holde, K. (1998) *Biophys. J.* **74,** 2554-2566.

10. Lodish, H., Berk, A., Zipursky, S. L., et al. (1999) *Molecular Cell Biology* (W.H. Freeman & Co., New York).

11. Nunez, M. E., Noyes, K. T. & Barton, J. K. (2002) *Chemistry & Biology* **9,** 403-415.

12. Smerdon, M. J. & Thoma, F. (1998) *Contemp. Cancer Res.* **2,** 199-222.

13. Higami, Y. & Shimokawa, I. (2000) *Cell & Tissue Res.* **301,** 125-132.

14. Nouspikel, T. & Hanawalt, P. C. (2002) *DNA Repair* **1,** 59-75.

15. Alberts, B., Johnson, A., Lewis, J., et al. (2002) *Molecular Biology of the Cell* (Garland Science, New York).

16. Hanawalt, P. C. (2001) *Mutat. Res.* **485,** 3-13.

17. Kawanishi, S., Hiraku, Y. & Oikawa, S. (2001) *Mutat. Res.* **488,** 65-76.

18. Kawanishi, S., Oikawa, S., Murata, M., et al. (1999) *Biochemistry* **38,** 16733-9.

19. Rodriguez, H., Valentine, M. R., Holmquist, G. P., et al. (1999) *Biochemistry* **38,** 16578-16588.

20. Saito, I., Nakamura, T., Nakatani, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 12686-12687.

21. Sistare, M. F., Codden, S. J., Heimlich, G., et al. (2000) *J. Am. Chem. Soc.* **122,** 4742-4749.

22. Sugiyama, H. & Saito, I. (1996) *J. Am. Chem. Soc.* **118,** 7063-7068.

23. Yoshioka, Y., Kitagawa, Y., Takano, Y., et al. (1999) *J. Am. Chem. Soc.* **121,** 8712-8719.

24. Bixon, M., Giese, B., Wessely, S., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11713-11716.

25. Giese, B. (2002) *Annu. Rev. Biochemistry* **71,** 51-70.

26. Nunez, M. E., Hall, D. B. & Barton, J. K. (1999) *Chem. Biol.* **6,** 85-97.

27. Treadway, C. R., Hill, M. G. & Barton, J. K. (2002) *Chemical Physics* **281,** 409-428.

28. Boone, E. & Schuster, G. B. (2002) *Nucleic Acids Res.* **30,** 830-837.

29. Lewis, F. D., Liu, X., Liu, J., et al. (2000) *J. Am. Chem. Soc.* **122,** 12037-12038.

30. Meggers, E., Michel-Beyerle, M. E. & Giese, B. (1998) *J. Am. Chem. Soc.* **120,** 12950-12955.

31. Nunez, M. E., Holmquist, G. P. & Barton, J. K. (2001) *Biochemistry* **40,** 12465-12471.

32. Sanii, L. & Schuster, G. B. (2000) *J. Am. Chem. Soc.* **122,** 11545-11546.

33. Hall, D. B., Holmlin, R. E. & Barton, J. K. (1996) *Nature* **382,** 731-735.

34. Henle, E. S., Han, Z., Tang, N., et al. (1999) *J. Biol. Chem.* **274,** 962-971.

35. Steenken, S., Jovanovic, S. V., Bietti, M., et al. (2000) *J. Am. Chem. Soc.* **122,** 2373-2374.

36. Krawczak, M., Ball, E. V. & Cooper, D. N. (1998) *Am. J. Hum. Genet.* **63,** 474-488.

37. Krawczak, M. & Cooper, D. N. (1997) *Trends Genetics* **13,** 121-122.

38. Witte, R. S. (1985) *Statistics* (Holt, Reinhard & Winston, New York).

39. Askeland, D. R. & Editor (1996) *The Science and Engineering of Materials*.

40. Shackelford, J. F. (2000) *Introduction to Materials Science for Engineers* (Prentice-Hall, Upper Saddle River, NJ).

41. Stansbury, E. E. & Buchanan, R. A. (2000) *Fundamentals of electrochemical corrosion* (ASM International, Materials Park, OH).

42. Hall, D. B., Kelley, S. O. & Barton, J. K. (1998) *Biochemistry* **37,** 15933-15940.

43. Holmlin, R. E., Dandliker, P. J. & Barton, J. K. (1998) *Angew. Chem., Int. Ed.* **36,** 2715-2730.

44. Rajski, S. R. & Barton, J. K. (2000) *Proc. Conversation in Biomolecular Stereodynamics* **11,** 285-291.

45. Cooper, D. N. (2000) *Human Gene Evolution* (BIOS Scientific, Oxford).

46. Heller, A. (2000) *Faraday Discuss.* **116,** 1-13.

47. Davidson, N. R. (1962) *Statistical Mechanics* (McGraw-Hill, New York).

48. Olasky, S. J. (1995) *Programmable arbitrary precision calculator for windows* (http://www.alberts.com/authorpages/00013288/prod_132.htm, 2003).

49. Schafer, F. Q. & Buettner, G. R. (2001) *Free Radical Biology & Medicine* **30,** 1191-1212.

50. Dringen, R. (2000) *Progress in Neurobiology* **62,** 649-671.

51. Noctor, G., Arisi, A.-C. M., Jouanin, L., et al. (1998) *J. Exp. Botany* **49,** 623-647.

52. Cooper, G. M. (1997) *The Cell: A Molecular Approach*.

53. Maki, H. & Sekiguchi, M. (1992) *Nature* **355,** 273-5.

54. Hogan, M. E., Rooney, T. F. & Austin, R. H. (1987) *Nature* **328,** 554-7.

55. Hayes, J. J., Tullius, T. D. & Wolffe, A. P. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87,** 7405-9.

56. Tullius, T. D., Dombroski, B. A., Churchill, M. E. A., et al. (1987) *Methods Enzymol.* **155,** 537-58.

57. Chatterjee, A. & Holley, W. R. (1993) *Adv Radiat Biol* **17,** 181-226.

58. Luo, Y., Han, Z., Chin, S. M., et al. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 12438-42.

59. Woodmansee, A. N. & Imlay, J. A. (2002) *J. Biol. Chem.* **277,** 34055-34066.

60. Hickerson, R. P., Prat, F., Muller, J. G., et al. (1999) *J. Am. Chem. Soc.* **121,** 9423-9428.

61. Friedman, K. A. & Heller, A. (2001) *J. Phys. Chem. B* **105,** 11859-11865.

62. Murphy, C. J., Arkin, M. R., Jenkins, Y., et al. (1993) *Science* **262,** 1025-9.

63. Boon, E. M. & Barton, J. K. (2002) *Curr. Opin. Structural Biology* **12,** 320-329.

64. Oikawa, S., Tada-Oikawa, S. & Kawanishi, S. (2001) *Biochemistry* **40,** 4763-4768.

65. Szalai, V. A., Singer, M. J. & Thorp, H. H. (2002) *J. Am. Chem. Soc.* **124,** 1625-1631.

66. Kanvah, S. & Schuster, G. B. (2002) *J. Am. Chem. Soc.* **124,** 11286-11287.

67. Bixon, M. & Jortner, J. (2001) *J. Am. Chem. Soc.* **123,** 12556-12567.

68. Jortner, J., Bixon, M., Voityuk, A. A., et al. (2002) *J. Phys. Chem. A* **106,** 7599-7606.

69. Henderson, P. T., Jones, D., Hampikian, G., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 8353-8358.

70. Schuster, G. B. (2000) *Acc. Chem. Res.* **33,** 253-260.

71. Hall, D. B. & Barton, J. K. (1997) *J. Am. Chem. Soc.* **119,** 5045-5046.

72. Lim, L. P. & Burge, C. B. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11193-11198.

73. Reddy, A. S. N. (2001) *Crit. Rev. Plant Sciences* **20,** 523-571.

74. Rogozin, I. B. & Milanesi, L. (1997) *J. Mol. Evol.* **45,** 50-59.

75. Croft, L., Schandorff, S., Clark, F., et al. (2000) *Nature Genetics* **24,** 340-341.

76. Burset, M., Seledtsov, I. A. & Solovyev, V. V. (2000) *Nucleic Acids Res.* **28,** 4364-4375.

77. Burset, M., Seledtsov, I. A. & Solovyev, V. V. (2001) *Nucleic Acids Res.* **29,** 255-259.

78. Zhang, M. Q. (1998) *Human Molecular Genetics* **7,** 919-932.

79. Brudno, M., Gelfand, M. S., Spengler, S., et al. (2001) *Nucleic Acids Res.* **29,** 2338-2348.

80. Engelbrecht, J., Knudsen, S. & Brunak, S. (1992) *J. Mol. Biol.* **227,** 108-13.

81. Nussinov, R. (1989) *J. Biomol. Struct. Dyn.* **6,** 985-1000.

82. McCullough, A. J. & Berget, S. M. (1997) *Mol. Cell. Biol.* **17,** 4562-4571.

83. Solovyev, V. V., Salamov, A. A. & Lawrence, C. B. (1994) *Nucleic Acids Res.* **22,** 5156-63.

84. Carlo, T., Sierra, R. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 3988-3995.

85. Carlo, T., Sterner, D. A. & Berget, S. M. (1996) *RNA* **2,** 342-353.

86. McCullough, A. J. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 9225-9235.

87. Sirand-Pugnet, P., Durosay, P., Brody, E., et al. (1995) *Nucleic Acids Res.* **23,** 3501-7.

88. Cooper, T. A. & Mattox, W. (1997) *Am. J. Hum. Genet.* **61,** 259-266.

89. Cartegni, L., Chew, S. L. & Krainer, A. R. (2002) *Nature Rev. Genetics* **3,** 285-298.

90. Karlin, S., Campbell, A. M. & Mrazek, J. (1998) *Annu. Rev. Genetics* **32,** 185-225.

91. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., et al. (2002) *Nucleic Acids Res.* **30,** 17-20.

92. International Human Genome Sequencing Consortium (2001) *Nature* **409,** 860-921.

93. Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) *Science* **291,** 1304-1351.

94. Adams, M. D., Celniker, S. E., Holt, R. A., et al. (2000) *Science* **287,** 2185-2195.

95. C. elegans Sequencing Consortium (1998) *Science* **282,** 2012-2018.

96. Arabidopsis Genome Initiative (2000) *Nature* **408,** 796-815.

97. Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996) *Science* **274,** 546, 563-567.

98. Wood, V., Gwilliam, R., Rajandream, M. A., et al. (2002) *Nature* **415,** 871-880.

99. Katinka, M. D., Duprat, S., Cornillot, E., et al. (2001) *Nature* **414,** 450-453.

100. Gardner, M. J., Tettelin, H., Carucci, D. J., et al. (1998) *Science* **282,** 1126-1132.

101. Bowman, S., Lawson, D., Basham, D., et al. (1999) *Nature* **400,** 532-538.

102. Bernardi, G. (2000) *Gene* **259,** 31-43.

103. Eyre-Walker, A. & Hurst, L. D. (2001) *Nature Rev. Genetics* **2,** 549-555.

104. Sueoka, N. (1992) *J. Mol. Evol.* **34,** 95-114.

105. Francis, S. E., Sullivan, D. J., Jr. & Goldberg, D. E. (1997) *Annu. Rev. Microbiology* **51,** 97-123.

106. Karlin, S. & Mrazek, J. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 5240-5245.

107. Makarova, K. S., Aravind, L., Wolf, Y. I., et al. (2001) *Microbiol. Mol. Biol. Rev.* **65,** 44-79.

108. Narumi, K., Kikuchi, M., Funayama, T., et al. (1999) *Hoshasen Seibutsu Kenkyu* **34,** 401-418.

109. Francino, M. P. & Ochman, H. (2000) *Mol. Biol. Evol.* **17,** 416-422.

110. Sueoka, N. & Kawanishi, Y. (2000) *Gene* **261,** 53-62.

111. Bernardi, G. (2000) *Gene* **241,** 3-17.

112. Nekrutenko, A. & Li, W.-H. (2000) *Genome Res.* **10,** 1986-1995.

113. Oliver, J. L., Bernaola-Galvan, P., Carpena, P., et al. (2001) *Gene* **276,** 47-56.

114. Jabbari, K. & Bernardi, G. (2000) *Gene* **247,** 287-292.

115. Vinogradov, A. E. (2001) *Gene* **276,** 143-151.

116. Vivares, C. P. & Metenier, G. (2000) *Curr. Opin. Microbiology* **3,** 463-467.

117. Giese, B., Wessely, S., Spormann, M., et al. (1999) *Angew. Chem., Int. Ed.* **38,** 996-998.

118. Meggers, E. & Giese, B. (1999) *Nucleosides Nucleotides* **18,** 1317-1318.

Table 4-1. ANOVA η^2 for correlation between G**G**G mol percentages in introns or IGD and in neighboring exons.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	0.087	0.003	0.003	0.002	0.008	0.001		0.017
Intron	–	0.13	0.005	0.005	0.003	0.021	0.002		0.004
	&	**0.23**	0.010	0.009	0.002	0.024	0.001		0.001
	+	0.093	0.002	0.002	0.006	0.008	0.001	0.029	0.002
IGD	–	0.13	0.004	0.007	0.004	0.019	0.007	0.011	0.001
	&	**0.21**	0.007	0.008	0.008	0.017	0.003	0.031	0.019

Notes: η^2 are highlighted when ≥0.20 (bold blue).

Table 4-2. Mean mol percentages of G**G**G in intron and IGD flanks.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	2.6*	0.52#	**0.33**	**0.29**	0.54	**0.19**		**0.069**
Flank	– 3	2.2*	0.91	**0.33**	**0.35**	0.56	**0.19**		**0.080**
	& 53	4.7*	1.4#	0.67#	0.64#	1.1	**0.38**		**0.15**
Intron	+ 3	1.9*	**0.35**	**0.24**	**0.29**	**0.38**	**0.16**		**0.039**
Flank	– 5	2.4*	0.96	**0.42**	**0.32**	**0.50**	**0.20**		**0.17**
	& 35	4.3*	1.3#	0.66#	0.61#	0.88#	**0.36**		**0.21**
IGD	+ 5	1.9*	0.69#	**0.35**	**0.48**	**0.38**	**0.35**	1.0#	**0.24**
Flank	– 3	2.1*	0.92	0.79*	**0.37**	0.52	**0.38**	1.0*	**0.29**
	& 53	4.0*	1.6	1.1	0.85#	0.90#	0.73#	2.1	0.53
IGD	+ 3	2.1*	0.67#	**0.37**	**0.48**	0.52#	**0.35**	1.2	**0.13**
Flank	– 5	2.3*	1.1	0.76*	0.68*	**0.47**	0.55	1.4*	**0.28**
	& 35	4.4*	1.7	1.1	1.2	0.99#	0.89	2.7*	**0.41**

Notes: Mol percentages are highlighted when ≤0.5% (bold red). Flank mol percentages are underlined when their percentage differences (100%(flank–all)/all) from overall values are ≥20% (solid blue) or ≤–20% (dotted red), except when flank mol percentages are ≤0.5%. Intron or IGD mol percentages are stared when their percentage differences (100%(intron – exon)/exon) from corresponding exon values are ≤–33% (# red) or ≥0% (* blue), except when intron or IGD mol percentages are ≤0.5%.

Table 4-3. Percentages of flanks with no G**G**G in intron and IGD populations.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	32	**75**	**82**	**80**	**66**	**89**		**95**
Flank	– 3	36	61	**83**	**77**	**67**	**90**		**94**
	& 53	16	47	**69**	62	44	**80**		**89**
Intron	+ 3	40	**81**	**87**	**80**	**76**	**90**		**96**
Flank	– 5	33	59	**79**	**78**	**68**	**89**		**88**
	& 35	18	49	**70**	63	49	**81**		**85**
IGD	+ 5	36	63	**76**	**69**	**74**	**76**	58	**85**
Flank	– 3	34	55	64	**75**	**69**	**75**	56	**89**
	& 53	16	35	48	52	51	56	34	**76**
IGD	+ 3	32	62	**77**	**70**	**67**	**75**	50	**91**
Flank	– 5	31	49	60	60	**69**	65	42	**86**
	& 35	13	30	46	41	45	48	22	**77**

Notes: Population percentages are highlighted when ≥+66% (bold blue)

Table 4-4. Percentage differences between flank and overall mean mol percentages of G**G**G in introns and IGD with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	28.	2.6	6.0	–4.0	16.	4.8		–2.2
Flank	– 3	8.0	–4.4	–10.	3.3	5.8	–5.5		–37.
	& 53	18.	–2.0	–2.7	–0.14	10.	–0.55		–25.
Intron	+ 3	–6.5	–30.	–25.	–3.4	–19.	–13.		–45.
Flank	– 5	21.	0.94	14.	–4.7	–4.9	–1.3		32.
	& 35	7.2	–9.8	–4.0	–4.1	–11.	–6.9		5.2
IGD	+ 5	11.	–9.6	–30.	5.7	–25.	5.7	–21.	43.
Flank	– 3	23.	12.	21.	–21.	–3.0	1.8	–34.	75.
	& 53	17.	1.7	–0.77	–8.0	–14.	3.6	–29.	59.
IGD	+ 3	26.	–16.	–29.	–0.36	–15.	–7.0	–26.	–28.
Flank	– 5	36.	24.	11.	38.	–25.	33.	–25.	52.
	& 35	31.	5.2	–5.8	20.	–18.	16.	–24.	14.

Notes: Percentage differences are underlined when ≥+20% (solid blue) or ≤–20% (dotted red), except when flank mol percentages are ≤0.5%.

Table 4-5. Differences between flank and overall mean numbers of G**G**G in 100 nt of introns and IGD with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	0.60	0.032	0.06	–0.041	0.14	0.064		–0.012
Flank	– 3	0.16	–0.08	–0.11	0.035	0.054	–0.083		–0.28
	& 53	0.75	–0.044	–0.041	–0.002	0.14	–0.009		–0.20
Intron	+ 3	–0.14	–0.38	–0.24	–0.035	–0.16	–0.17		–0.26
Flank	– 5	0.43	0.017	0.15	–0.050	–0.045	–0.020		0.24
	& 35	0.29	–0.22	–0.060	–0.054	–0.16	–0.11		0.04
IGD	+ 5	0.19	–0.079	–0.17	0.028	–0.18	0.024	–0.46	0.087
Flank	– 3	0.40	0.10	0.16	–0.11	–0.02	0.008	–0.78	0.16
	& 53	0.59	0.028	–0.010	–0.078	–0.19	0.030	–1.1	0.22
IGD	+ 3	0.44	–0.13	–0.17	–0.002	–0.11	–0.03	–0.55	–0.058
Flank	– 5	0.61	0.21	0.081	0.19	–0.19	0.15	–0.57	0.11
	& 35	1.1	0.087	–0.073	0.19	–0.24	0.13	–0.89	0.052

Notes: Differences are underlined when corresponding values are in Table 4-4.

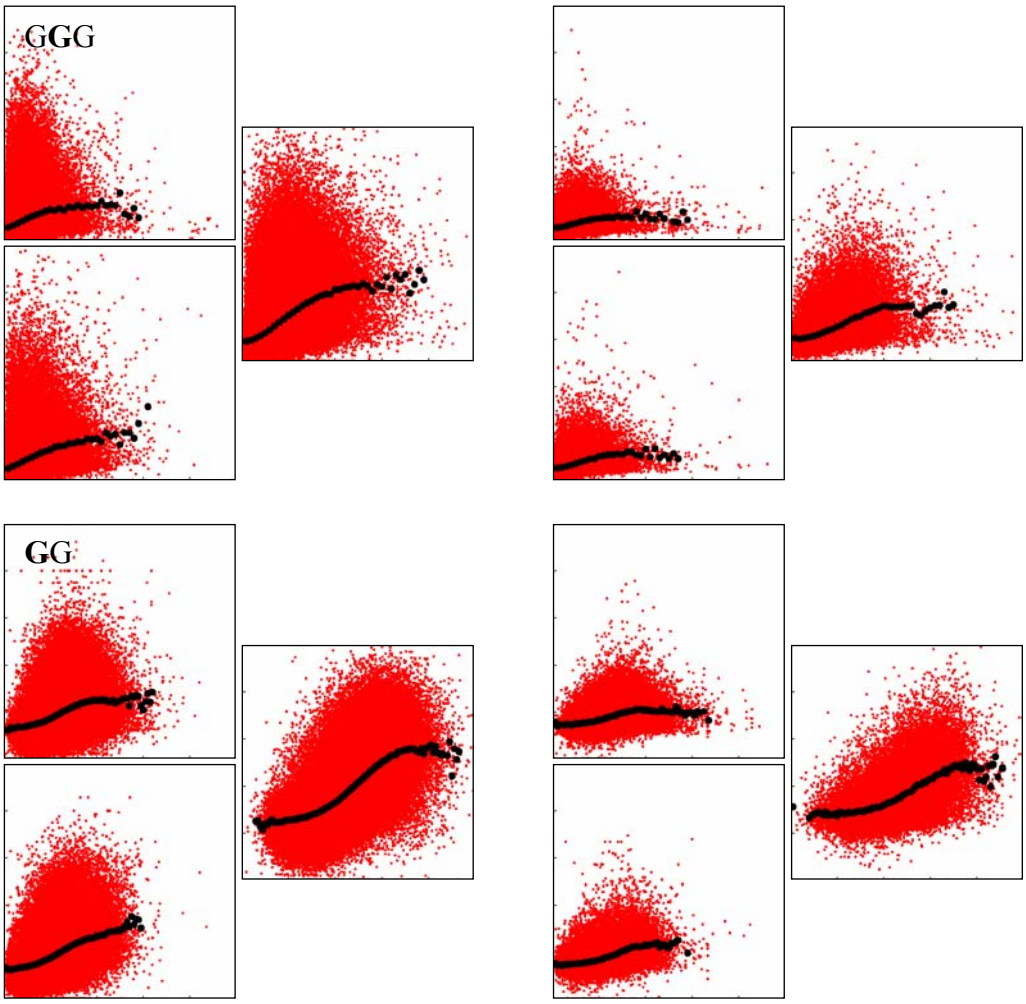


Figure 4-1. Scatter plots of G**G**G and **G**G percentages of individual (red dots) and mean (black circles) introns (left) or IGD (right) vs. adjacent exons for *Hsa*. Notes: Charts in each triplet show coding (top), template (bottom) and combined (middle) strands. Vertical (intron or IGD) and horizontal (exon) full scales are 25% for G**G**G and **G**G.

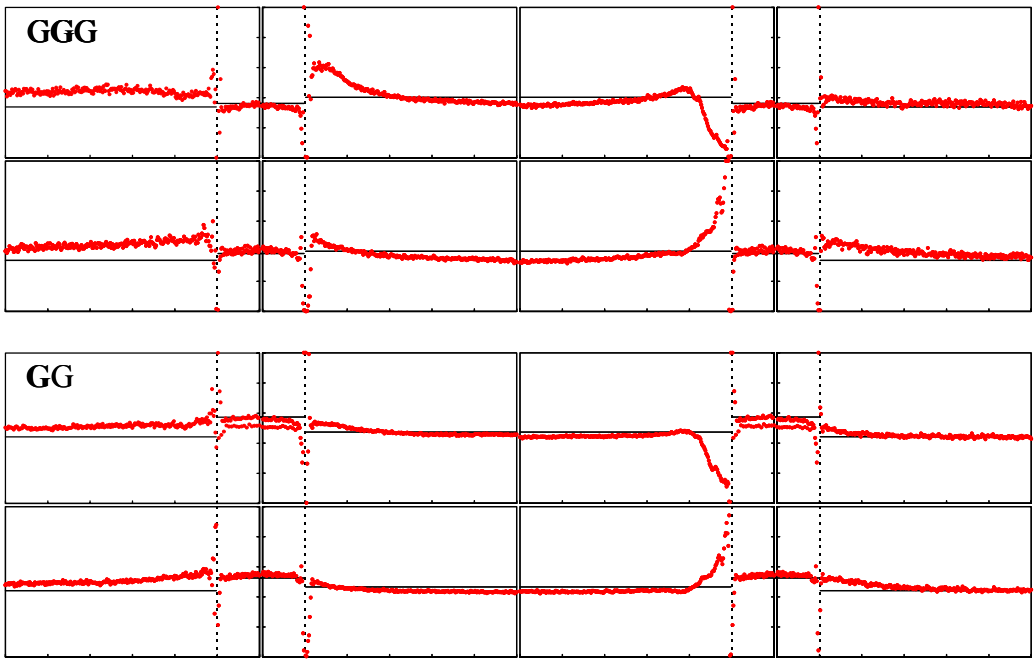


Figure 4-2. Mean percentages of G**G**G and **G**G vs. distance from exon, intron or IGD boundaries (red dots) and overall mean percentages of G**G**G and **G**G (black lines) in *Hsa*. Introns and IGD longer than 100 bp are shown. Notes: The four charts in the top row of each pair of rows show, from left to right, the ends of the 3' IGD and 5' exon, the 3' exon and 5' intron, the 3' intron and 5' exon, and the 3' exon and 5' IGD on the coding strand. Dashed vertical lines separate introns and IGD from exons. Each pair of rows shows the coding (top) and template strands (bottom). Vertical full scale is 5% for G**G**G and 10% for **G**G. Horizontal full scale is 300 bp on all charts.

Chapter 5: Summary and Recommendations

INTRODUCTION

From the perspective of the materials scientist, a genome is a double-stranded, linear copolymer composed of four nucleotide bases, adenine, thymine, guanine and cytosine (Figure 5-1), arranged in the familiar double helix. Only one of the genome's two strands, the coding strand, comprises nucleotide triplets or codons that specify the amino acids. Codons are organized into exons and then into genes which code for proteins (Figure 5-2). The information on the coding strand is backed-up by the genome's second strand, the template strand, for repair and replication. The backup and repair systems are necessary, because the genome operates in an oxidizing environment that contains nitric oxide, hydrogen peroxide, singlet oxygen, hydroxyl radicals, etc. Oxidants inject holes which hop between the genome's guanines. The double-stranded DNA is a one-dimensional electronic conductor, albeit a highly resistive one, with unevenly distributed conducting and resistive domains. Hence, the materials scientist sees a corrosion problem: a conductive composite in an electrolyte containing oxidizers. The corrosion science of the past 150 years can illuminate aspects of DNA oxidation and, thereby, shed light on a particular aspect of mutation.

DISCUSSION

Guanine Oxidation in DNA. The genome in each human cell undergoes ~10^4 oxidative attacks each day (5, 6) resulting in ~10^4 to ~10^5 oxidized bases at steady state (7). They include ~10^4 guanine bases oxidized to 8-oxo-2-deoxyguanosine (8oxoG) (7), a mutagenic lesion (8). Oxidative damage of DNA contributes to mutation, cancer (6, 9, 10) and aging (11-13).

The genome's environment contains strong oxidizers that are selective for guanine bases and extremely strong oxidizers that react with any base they attack (14, 15). Com-

mon cellular oxidizers include hydrogen peroxide, nitric oxide, superoxide radicals, singlet oxygen and molecular oxygen (16-21). Some are precursors of strong and extremely strong oxidizers: hydroxyl radicals (OH•) are formed by hydrogen peroxide and reduced transition metal cations, and peroxynitrite anions ($OONO^-$) are formed by nitric oxide and superoxide radicals.

The ranking of nucleotide oxidation potentials is G < A ≤ T < C, and the ranking of guanine oxidation rates is G**G**G ≥ **G**G > **G**. (G**G**G denotes the one central G in the GGG sequence. **G**G denotes the left or 5' G in HGG, where H is a base other than G. **G** denotes the isolated G in HGH. H is any base but G.) At pH 7, the one-electron oxidation potentials of G, A, T and C, respectively, are 1.04 V, 1.32 V, 1.29 V and 1.44 V vs. NHE (22). These potentials are comparable to those of noble metals, such as palladium, platinum and gold (23). The relative rates of oxidation are reported as GGG (2.7) > GG (0.7 to 2.0) > G (0.1 to 0.4) (24) and GGG (5.3) > GG (3.7) > G (1.0) (25) in two independent studies. The **G**G and **G** rates are ranges, because they vary with the neighboring non-guanine nucleotides. While the chemical oxidation rates of **G**G and G**G**G are greater than that of **G** (26-28), the reactivities of **G**G and G**G**G are similar (29-31).

Selectivity for **G**G and G**G**G is based on their lower oxidation potentials, that is enhanced by hole transport or conduction in DNA (26). Holes injected by oxidants react with remote **G**G and G**G**G, in oligonucleotides (4, 24, 32-35) as well as in nucleosome core particles, which are the DNA-protein complexes found in chromosomes (36, 37)

These are imperfect generalizations about guanine oxidation. Preferential oxidation at GGG sequences also reflects sensitizer binding to DNA (38, 39). G in CG and GC sequences (30) and 8oxoG (25, 40, 41) also are oxidation targets. The clinically observed frequencies of oxidative mutations are comparable at **G,** **G**G and G**G**G in a large human database, but clinical observations may not reflect underlying oxidation rates (42, 43).

Guanine Distributions and Genomic Oxidation Resistance. The selectivity of strong oxidizers for guanine, especially **G**G and G**G**G, means that the guanine distribu-

tion affects genomic resistance to selective oxidants. Guanine distributions can be evaluated against engineering principles for corrosion prevention through materials selection, structural design and removal of corrosives (44, 45).

- Elements with greater oxidation resistance generally make alloys more oxidation resistant. Genomes and domains within genomes that minimize total G (ΣG) and minimize **G**G and G**G**G relative to other guanine nucleotides should resist selective oxidants better than those that do not.
- Electrical (electronic, not ionic) contact between an essential part that is more oxidation resistant and an expendable part that is more readily oxidized reduces corrosion of the essential part, because it allows the expendable part to act as a sacrificial anode, cathodically protecting the essential part (46). Essential genome domains that have lower guanine levels, especially lower **G**G and G**G**G levels, than neighboring expendable domains, e.g., exons that have lower levels than neighboring introns, should suffer less oxidation than those that have the reverse.
- Conversely, electrical contact between an essential part that is more readily oxidized and an expendable part that is more oxidation resistant increases the corrosion of the essential part, because it allows the former to be corroded by oxidants attacking the latter (corrosion cell), the oxidant capturing electrons and thereby injecting mobile holes (47). Exons having more **G**G and G**G**G than their neighboring introns, but which are electrically insulated by the intervening DNA, as described by Barton et al. (48-51) (also (32)), should be more oxidation resistant than those that are electrically connected.
- Removing oxidants from the environment or blocking their access by coating reduces corrosion. Genomes that maximize guanine levels in expendable domains to scavenge oxidants (in parallel with antioxidants and enzymes in the nucleoplasm) should reduce selective oxidation better than those that do not.

Obviously, a genome is not designed primarily to minimize its own oxidation; its primary function, direction of protein synthesis, constrains its guanine distribution. In exons the sequence is the function, because their nucleotide triplets code for the amino acids of proteins (Figure 5-2). In the nucleus, the multiple exons that comprise a gene are spliced together to make a functional messenger RNA. The **G**G at most exon/intron boundaries are part of the sequence that signals exon splicing (52-54). While they are perfectly positioned to act as sacrificial anodes, they are as essential as any **G**G within an exon. About 15% of human genetic diseases are associated with mutations that generate new splice sites or destroy functional splice sites (55).

The ideal distribution minimizes guanine in essential domains to make them more noble and maximizes guanine in neighboring expendable domains to make them better oxidant scavengers. Two simpler distributions that emphasize guanine minimization or guanine scavengers should resist selective oxidants better than a hypothetical random distribution. Our genome evaluations to date, described in this dissertation, have considered these simpler distributions.

Guanine Distributions in Model Genomes. We examined guanine minimization in Chapter 3. ΣG, **G**G and G**G**G levels are lower than statistically expected in seven of eight model genomes. Exon ΣG averages ≤22% in five genomes and 24% to 26% the other three genomes, including the human genome. ΣG averages ≤20% in the introns and inter-gene domains (IGD) of all of the model genomes except human. Minimizing ΣG minimizes **G**G and G**G**G more rapidly, because their mol percentages scale with $(\Sigma G)^2$ and $(\Sigma G)^3$, respectively. This simple strategy probably is evolutionarily penalizing, because deviation from a one-to-one A:T-to-G:C ratio sharply reduces the possible number of different arrangements of nucleotides. At 13% ΣG, the level in the malaria parasite, the number of possible combinations is $>10^{500}$ times smaller than the number at 25% ΣG in a hypothetical 10^4 bp genome (see Note 1).

GG and G**G**G minimization can be non-specific, resulting from ΣG minimization, or specific, below what probability predicts from ΣG. **G**G and G**G**G are non-specifically minimized and G**G**G is specifically minimized in exons, introns and IGD of six of the eight genomes. For example, G**G**G is extensively excluded from exons and introns (Table 5-1). G**G**G is statistically under-represented in human exons, but it is over-represented in human introns and IGD. **G**G generally is not under-represented in the human genome.

The over-representation of **G**G and G**G**G in human introns and IGD suggests the second strategy for mitigating oxidative damage: deployment of guanine nucleotides as oxidant scavengers and sacrificial anodes. We examined this in Chapters 2 and 4. **G**G and G**G**G are not uniformly distributed in human introns; they are over-represented near their ends, exceeding the levels that probability predicts from ΣG (Figure 5-3). In the 100 bp of intron and IDG DNA adjacent to exons, these elevations above overall levels typically added ~1 G**G**G and ~0 **G**G to introns, and ~1.5 G**G**G and ~2 **G**G to IGD.

GG and G**G**G in the flanks of introns and IGD are optimally positioned both to absorb holes that oxidizing agents inject directly into exons, and to intercept holes that could diffuse to exons from introns. Individual guanine nucleotides can scavenge for whole exons, because the oxidation rate constant (~1/years) in human cells is much lower than the repair rate constant (~1/days) (see Note 2), and un-repaired 8oxodG can be oxidized further and more readily than guanine (25, 41). Further, the ΣG mol percentages of human introns and IGD increase when adjacent exon mol percentages increase, keeping introns and IGD competitive with exons for selective oxidants (Figure 5-4).

We consider our finding that **G**G and G**G**G levels are elevated in the flanks of human introns to be only the first essential step in showing that these nucleotides function as sacrificial anodes. GGG sequences in the 5' flank of introns enhance exon splicing in vertebrates (56-61). Therefore, some or all of them could serve this function independently of providing cathodic protection. Furthermore, **G**G and G**G**G in intron and IGD

flanks may or may not be electrically connected to their neighboring exon. If they were electrically isolated, they could not cathodically protect the exons.

Sacrificial Anodes and Conduction in DNA. Given a conductive path, ΣG, **G**G and G**G**G in intron and IGD flanks could act as sacrificial anodes, like zinc plates protecting steel ships, drawing holes out of neighboring exons (62, 63). Barton et al. (36, 38, 64, 65) have investigated this chemistry for at least a decade, and write: "It will also be important to determine whether organisms have evolved to protect their genomes from long-range damage (via charge transport). Perhaps radical damage is funneled to or insulated from specific sites within the genome. One could, however, consider that segments throughout the genome may encode "sinks" for damage, and that other segments could serve as buffers as a result of local sequence-dependent or protein-dependent structural deformations to protect critical regions." Giese (4), Kawanishi et al. (66), Thorp et al. (67), and Heller (63) have considered both sacrificial anodes and guanine scavengers. Schuster et al. (68) have extended this concept, showing that disulfides on molecules intercalated in DNA can serve as sacrificial anodes.

Holes can travel limited distances in DNA, but they can be impeded by some DNA sequences. *In vitro* experiments show hole transport over 40 to 200 Å or 10 to 60 bp in DNA (4, 69, 70). Barton et al. (36, 37, 65) have shown remote oxidation over 24 bp in nucleosome core particles, and write "Our studies on long-range damage on restriction fragments suggest that the physiological range of charge migration may be on the order of 100 bp, but probably not longer." Charge transport can be obstructed by DNA sequences and structures such as bulges, mismatches and multiple TA steps (49, 71, 72). Thus, the putative conduction path needs to be evaluated in each case, because guanine nucleotides in intron and IGD flanks can be sacrificial anodes for neighboring exons only if the intervening DNA electrically connects them.

Based on these experiments, Bixon and Jortner et al. (3, 69, 73-80), Berlin and Ratner et al. (81-86), Giese et al. (4, 87, 88) and Schuster et al. (40, 70, 89, 90) are developing models to predict hole movement by hopping from G to G and reaction on guanine.

- Holes "rest" on guanines as guanine radical cations, and **G**G and G**G**G act as shallow traps for holes (26, 76).
- A guanine radical cation can undergo several reactions (91). Absolute rates for these reactions are rarely available (87), though the rate of hydration has been estimated (88). The periods for hole movement (tens of nanoseconds (92, 93)) and the lifetimes of guanine radical cations (microseconds (94, 95)) have been measured, and the maximum distances of hole travel before reaction have been bounded (references above).
- If the destination guanine is the next nucleotide in the sequence, hole re-distribution is extremely rapid and reflects thermal equilibrium. The probability of a hole resting on a particular guanine can be estimated from the Boltzmann distribution based on energetic data (69, 96) and from measured equilibrium constants (26).
- When moving to a group of remote guanines, a hole moves to the proximal guanine, before re-distributing among the adjoining nucleotides (97).
- If the source and destination are separated by 3 to 4 bp or less, hole movement is predominantly by superexchange (tunneling). Rate constants can be calculated from the intervening DNA sequence (3, 69, 78, 80).
- If the separation is greater than 3 to 4 bp, hole movement is thermally-induced hopping via the intervening adenine nucleotides. Rate constants can be extracted from experimental data (e.g., Figure 5-5), and do not depend strongly on the intervening sequence (3). Current models do not explain all hopping experiments (3, 98), and they are only beginning to include the thermal fluctuation of the double helix, which may strongly influence charge movement (70, 86).

Recommendations. Having found elevated **G**G and G**G**G in the flanks of human introns, the next task is to test whether they are sacrificial anodes by probing whether or not there are hole conduction paths between them and their putatively protected counterparts in adjacent exons.

Beyond fundamental understanding of nature, such studies will contribute to understanding why some genes mutate more rapidly than others, and why multiple mutations are frequent in some genes, like the breast cancer linked BRCA, but not in others. Identification of oxidation hot-spots could pinpoint targets of chemotherapeutic drugs that are oxidation catalysts, such as cisplatin or doxorubicin. DNA has been proposed as a building block of future nanostructured devices. If such devices are to operate reliably in environments lacking the anti-oxidants and the DNA repair systems of living cells, tools to design their DNA to resist oxidation by holes will be required.

NOTES

1. The number of combinations is $((N_A + N_T + N_G + N_C)!) \div (N_A! \times N_T! \times N_G! \times N_C!)$, where N_A is the number of adenine nucleotides (from combinatorial formulas in (99)). PAPCW (100) is used to calculate these very large numbers.
2. In the human genome, $\sim 10^4$ oxidations daily (5, 6) on 3×10^9 bp means ~2000 days between attacks on a average 160 bp exon. $\sim 10^4$ 8oxodG at steady state (7, 101) means ~1 day to repair, assuming the oxidation rate is independent and the repair rate is first-order in 8oxodG concentration.

REFERENCES

1. Alberts, B., Johnson, A., Lewis, J., et al. (2002) *Molecular Biology of the Cell* (Garland Science, New York).
2. Witherly, J. (2003) *Talking Glossary of Genetics, Intron* (http://www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/intron.shtml, National Human Genome Research Institute)
3. Bixon, M. & Jortner, J. (2002) *Chemical Physics* **281,** 393-408.
4. Giese, B. (2002) *Annu. Rev. Biochemistry* **71,** 51-70.

5. Helbock, H. J., Beckman, K. B., Shigenaga, M. K., et al. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 288-293.

6. Setlow, R. B. (2001) *Mutat. Res.* **477,** 1-6.

7. Beckman, K. B. & Ames, B. N. (1997) *J. Biol. Chem.* **272,** 19633-19636.

8. Sekiguchi, M. & Hayakawa, H. (1998) *Contemp. Cancer Res.* **2,** 85-93.

9. Ambrosone, C. B. (2000) *Antioxidants & Redox Signaling* **2,** 903-917.

10. Jackson, A. L. & Loeb, L. A. (2001) *Mutat. Res.* **477,** 7-21.

11. Beckman, K. B. & Ames, B. N. (1998) *Physiological Rev.* **78,** 547-581.

12. Bohr, V. A. & Anson, R. M. (1995) *Mutat. Res.* **338,** 25-34.

13. Finkel, T. & Holbrook, N. J. (2000) *Nature* **408,** 239-247.

14. Kawanishi, S., Hiraku, Y., Murata, M., et al. (2002) *Free Radical Biology & Medicine* **32,** 822-832.

15. Kawanishi, S., Hiraku, Y. & Oikawa, S. (2001) *Mutat. Res.* **488,** 65-76.

16. Amatore, C., Arbault, S., Bruce, D., et al. (2000) *Faraday Discuss.* **116,** 319-333.

17. Kawanishi, S., Oikawa, S. & Hiraku, Y. (2000) *Free Radicals in Chemistry, Biology and Medicine***,** 85-91.

18. May, J. M., Qu, Z.-C., Xia, L., et al. (2000) *Am. J. Physiol.* **279,** C1946-C1954.

19. Newcomb, T. G. & Loeb, L. A. (1998) *DNA Damage and Repair* **1,** 65-84.

20. Poulsen, H. E., Jensen, B. R., Weimann, A., et al. (2000) *Free Radical Research* **33,** S33-S39.

21. Wei, Y.-H., Pang, C.-Y., Lee, H.-C., et al. (1998) *Curr. Science* **74,** 887-893.

22. Faraggi, M., Broitman, F., Trent, J. B., et al. (1996) *J. Phys. Chem.* **100,** 14751-14761.

23. Bard, A. J. & Faulkner, L. R. (2001) *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York).

24. Saito, I., Nakamura, T., Nakatani, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 12686-12687.

25. Hickerson, R. P., Prat, F., Muller, J. G., et al. (1999) *J. Am. Chem. Soc.* **121,** 9423-9428.

26. Lewis, F. D., Liu, X., Liu, J., et al. (2000) *J. Am. Chem. Soc.* **122,** 12037-12038.

27. Sistare, M. F., Codden, S. J., Heimlich, G., et al. (2000) *J. Am. Chem. Soc.* **122,** 4742-4749.

28. Sugiyama, H. & Saito, I. (1996) *J. Am. Chem. Soc.* **118,** 7063-7068.

29. Kawanishi, S., Oikawa, S., Murata, M., et al. (1999) *Biochemistry* **38,** 16733-9.

30. Rodriguez, H., Valentine, M. R., Holmquist, G. P., et al. (1999) *Biochemistry* **38,** 16578-16588.

31. Yoshioka, Y., Kitagawa, Y., Takano, Y., et al. (1999) *J. Am. Chem. Soc.* **121,** 8712-8719.

32. Boone, E. & Schuster, G. B. (2002) *Nucleic Acids Res.* **30,** 830-837.

33. Meggers, E., Kusch, D., Spichty, M., et al. (1998) *Angew. Chem., Int. Ed.* **37,** 460-462.

34. O'Neill, P., Parker, A. W., Plumb, M. A., et al. (2001) *J. Phys. Chem. B* **105,** 5283-5290.

35. Sanii, L. & Schuster, G. B. (2000) *J. Am. Chem. Soc.* **122,** 11545-11546.

36. Nunez, M. E., Holmquist, G. P. & Barton, J. K. (2001) *Biochemistry* **40,** 12465-12471.

37. Nunez, M. E., Noyes, K. T. & Barton, J. K. (2002) *Chemistry & Biology* **9,** 403-415.

38. Hall, D. B., Holmlin, R. E. & Barton, J. K. (1996) *Nature* **382,** 731-735.

39. Henle, E. S. & Linn, S. (1997) *J. Biol. Chem.* **272,** 19095-19098.

40. Ly, D., Sanii, L. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 9400-9410.

41. Steenken, S., Jovanovic, S. V., Bietti, M., et al. (2000) *J. Am. Chem. Soc.* **122,** 2373-2374.

42. Krawczak, M., Ball, E. V. & Cooper, D. N. (1998) *Am. J. Hum. Genet.* **63,** 474-488.

43. Krawczak, M. & Cooper, D. N. (1997) *Trends Genetics* **13,** 121-122.

44. Askeland, D. R. (1996) *The Science and Engineering of Materials*.

45. Shackelford, J. F. (2000) *Introduction to Materials Science for Engineers* (Prentice-Hall, Upper Saddle River, NJ).

46. Morgan, J. H. (1987) *Cathodic Protection* (NACE, Houston, TX).

47. Stansbury, E. E. & Buchanan, R. A. (2000) *Fundamentals of electrochemical corrosion* (ASM International, Materials Park, OH).

48. Hall, D. B., Kelley, S. O. & Barton, J. K. (1998) *Biochemistry* **37,** 15933-15940.

49. Nunez, M. E., Hall, D. B. & Barton, J. K. (1999) *Chem. Biol.* **6,** 85-97.

50. Holmlin, R. E., Dandliker, P. J. & Barton, J. K. (1998) *Angew. Chem., Int. Ed.* **36,** 2715-2730.

51. Rajski, S. R. & Barton, J. K. (2000) *Proc. Conversation in Biomolecular Stereodynamics* **11,** 285-291.

52. Lim, L. P. & Burge, C. B. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11193-11198.

53. Reddy, A. S. N. (2001) *Crit. Rev. Plant Sciences* **20,** 523-571.

54. Rogozin, I. B. & Milanesi, L. (1997) *J. Mol. Evol.* **45,** 50-59.

55. Cooper, T. A. & Mattox, W. (1997) *Am. J. Hum. Genet.* **61,** 259-266.

56. Carlo, T., Sierra, R. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 3988-3995.

57. Carlo, T., Sterner, D. A. & Berget, S. M. (1996) *RNA* **2,** 342-353.

58. McCullough, A. J. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 9225-9235.

59. McCullough, A. J. & Berget, S. M. (1997) *Mol. Cell. Biol.* **17,** 4562-4571.

60. Sirand-Pugnet, P., Durosay, P., Brody, E., et al. (1995) *Nucleic Acids Res.* **23,** 3501-7.

61. Zhang, M. Q. (1998) *Human Molecular Genetics* **7,** 919-932.

62. Friedman, K. A. & Heller, A. (2001) *J. Phys. Chem. B* **105,** 11859-11865.

63. Heller, A. (2000) *Faraday Discuss.* **116,** 1-13.

64. Murphy, C. J., Arkin, M. R., Jenkins, Y., et al. (1993) *Science* **262,** 1025-9.

65. Boon, E. M. & Barton, J. K. (2002) *Curr. Opin. Struct. Biol.* **12,** 320-329.

66. Oikawa, S., Tada-Oikawa, S. & Kawanishi, S. (2001) *Biochemistry* **40,** 4763-4768.

67. Szalai, V. A., Singer, M. J. & Thorp, H. H. (2002) *J. Am. Chem. Soc.* **124,** 1625-1631.

68. Kanvah, S. & Schuster, G. B. (2002) *J. Am. Chem. Soc.* **124,** 11286-11287.

69. Jortner, J., Bixon, M., Voityuk, A. A., et al. (2002) *J. Phys. Chem. A* **106,** 7599-7606.

70. Schuster, G. B. (2000) *Acc. Chem. Res.* **33,** 253-260.

71. Hall, D. B. & Barton, J. K. (1997) *J. Am. Chem. Soc.* **119,** 5045-5046.

72. Treadway, C. R., Hill, M. G. & Barton, J. K. (2002) *Chemical Physics* **281,** 409-428.

73. Bixon, M., Giese, B., Wessely, S., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11713-11716.

74. Bixon, M. & Jortner, J. (2000) *J. Phys. Chem. B* **104,** 3906-3913.

75. Bixon, M. & Jortner, J. (2001) *J. Am. Chem. Soc.* **123,** 12556-12567.

76. Bixon, M. & Jortner, J. (2001) *J. Phys. Chem. A* **105,** 10322-10328.

77. Jortner, J., Bixon, M., Langenbacher, T., et al. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 12759-12765.

78. Voityuk, A. A., Roesch, N., Bixon, M., et al. (2000) *J. Phys. Chem. B* **104,** 9740-9745.

79. Voityuk, A. A., Jortner, J., Bixon, M., et al. (2000) *Chemical Physics Letters* **324,** 430-434.

80. Voityuk, A. A., Jortner, J., Bixon, M., et al. (2001) *J. Chem. Phys.* **114,** 5614-5620.

81. Berlin, Y. A., Burin, A. L. & Ratner, M. A. (2000) *Journal of Physical Chemistry A* **104,** 443-445.

82. Berlin, Y. A., Burin, A. L. & Ratner, M. A. (2000) *Superlattices and Microstructures* **28,** 241-252.

83. Berlin, Y. A., Burin, A. L. & Ratner, M. A. (2001) *Journal of the American Chemical Society* **123,** 260-268.

84. Berlin, Y. A., Burin, A. L., Siebbeles, L. D. A., et al. (2001) *Journal of Physical Chemistry A* **105,** 5666-5678.

85. Berlin, Y. A., Burin, A. L. & Ratner, M. A. (2002) *Chemical Physics* **275,** 61-74.

86. Grozema, F. C., Siebbeles, L. D. A., Berlin, Y. A., et al. (2002) *ChemPhysChem* **3,** 536-539.

87. Meggers, E., Michel-Beyerle, M. E. & Giese, B. (1998) *J. Am. Chem. Soc.* **120,** 12950-12955.

88. Giese, B. & Spichty, M. (2000) *ChemPhysChem* **1,** 195-198.

89. Barnett, R. N., Cleveland, C. L., Joy, A., et al. (2001) *Science* **294,** 567-571.

90. Henderson, P. T., Jones, D., Hampikian, G., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 8353-8358.

91. Burrows, C. J. & Muller, J. G. (1998) *Chemical Reviews (Washington, D. C.)* **98,** 1109-1151.

92. Giese, B., Amaudrut, J., Kohler, A. K., et al. (2001) *Nature* **412,** 318-320.

93. Lewis, F. D., Liu, X., Liu, J., et al. (2000) *Nature* **406,** 51-53.

94. Armitage, B. (1998) *Chemical Reviews (Washington, D. C.)* **98,** 1171-1200.

95. Armitage, B., Yu, C., Devadoss, C., et al. (1994) *Journal of the American Chemical Society* **116,** 9847-59.

96. Moore, W. J. (1972) *Physical Chemistry* (Prentice-Hall, Englewood Cliffs, NJ).

97. Davis, W. B., Naydenova, I., Haselsberger, R., et al. (2000) *Angewandte Chemie, International Edition* **39,** 3649-3652.

98. Pascaly, M., Yoo, J. & Barton, J. K. (2002) *Journal of the American Chemical Society* **124,** 9083-9092.

99. Davidson, N. R. (1962) *Statistical Mechanics* (McGraw-Hill, New York).

100. Olasky, S. J. (1995) *Programmable arbitrary precision calculator for windows* (http://www.alberts.com/authorpages/00013288/prod_132.htm,

101. Hanawalt, P. C. (2001) *Mutat. Res.* **485,** 3-13.

Table 5-1. The percentages of exons, introns and IGD with no GGG.

GGG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	26.	20.	**45.**	32.	5.8	28.	2.6	**36.**
Exon	–	28.	19.	**48.**	**49.**	6.1	31.	3.7	**40.**
	&	12.	8.5	25.	20.	2.9	17.	0.61	24.
	+	4.0	**60.**	**67.**	**70.**	**46.**	**86.**		**88.**
Intron	–	3.5	**46.**	**65.**	**69.**	**43.**	**87.**		**83.**
	&	1.7	**35.**	**55.**	**52.**	27.	**77.**		**76.**
	+	0.47	5.5	13.	6.6	22.	17.	32.	16.
IGD	–	0.43	4.8	10.	5.9	23.	16.	24.	17.
	&	0.26	3.4	6.6	3.6	14.	10.	16.	8.0

Notes: Population percentages are highlighted when ≥33% (bold blue). They are underlined when their percentage differences from probability-predicted values are ≥33% (solid blue) or ≤–33% (dotted red), except when actual population percentages are ≤5%. Genomes: *Homo sapiens* (*Hsa*), *Drosophila melanogaster* (*Dme*, fruit fly), *Caenorhabditis elegans* (*Cel*, nematode worm), *Arabidopsis thaliana* (*Ath*, flowering plant), *Saccharomyces cerevisiae* (*Sce*, budding yeast), *Schizosaccharomyces pombe* (*Spo*, fission yeast), *Encephalitozoon cuniculi* (*Ecu*, intracellular parasite) and *Plasmodium falciparum* (*Pfa*, malaria parasite).

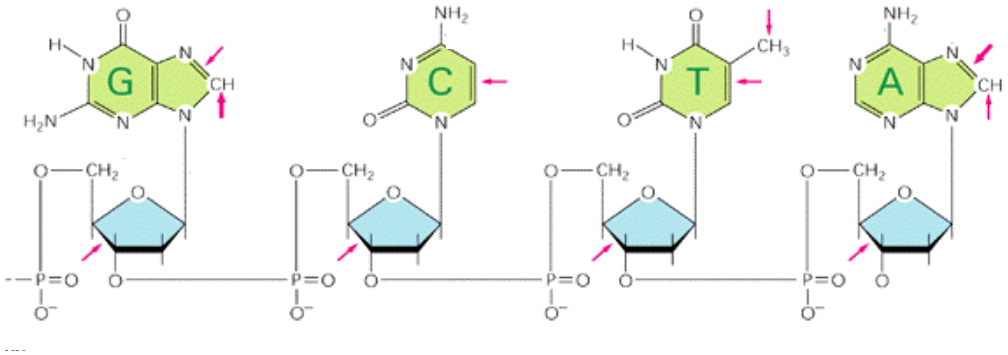


Figure 5-1. The four bases of DNA. The red arrows indicate targets for oxidative damage, with the size of each arrow indicating the relative frequency of each attack. Other types of damage omitted. (from (1))

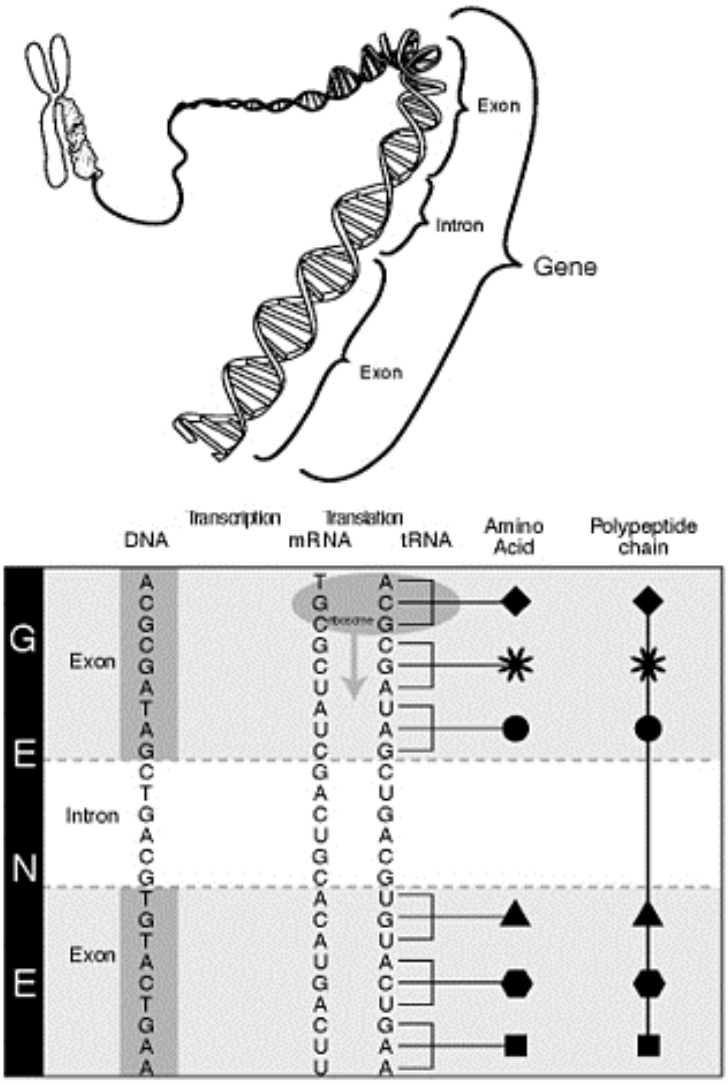


Figure 5-2. A strand of DNA, schematically magnified to show the relationships of chromosome to gene to exons and introns to DNA bases that code for protein (from (2)).

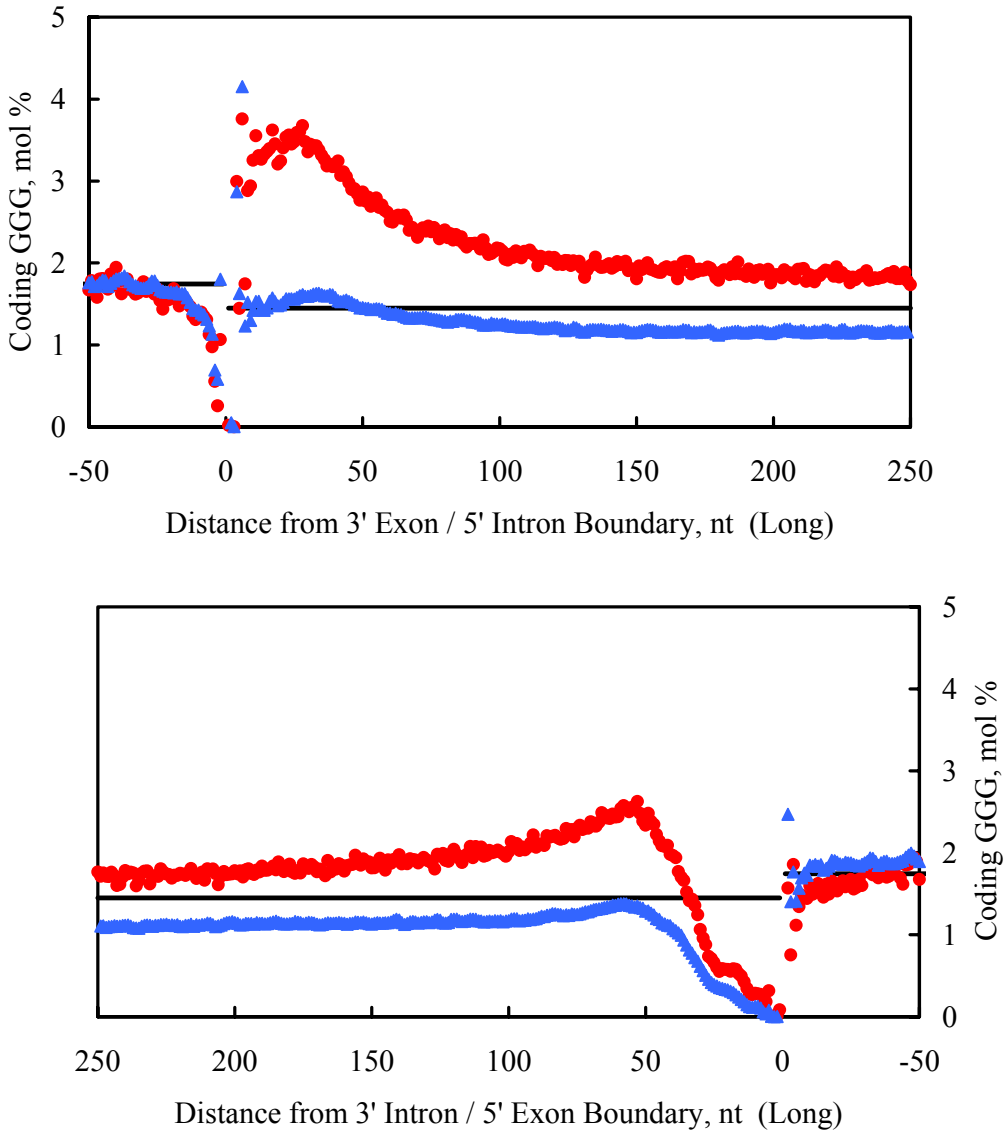


Figure 5-3. The average mol percentage of G**G**G in human exons and introns vs. distance from the exon/intron boundary. Data include the observed local average (red circles), the probability-predicted local average (blue triangles) and the observed overall average (black lines).

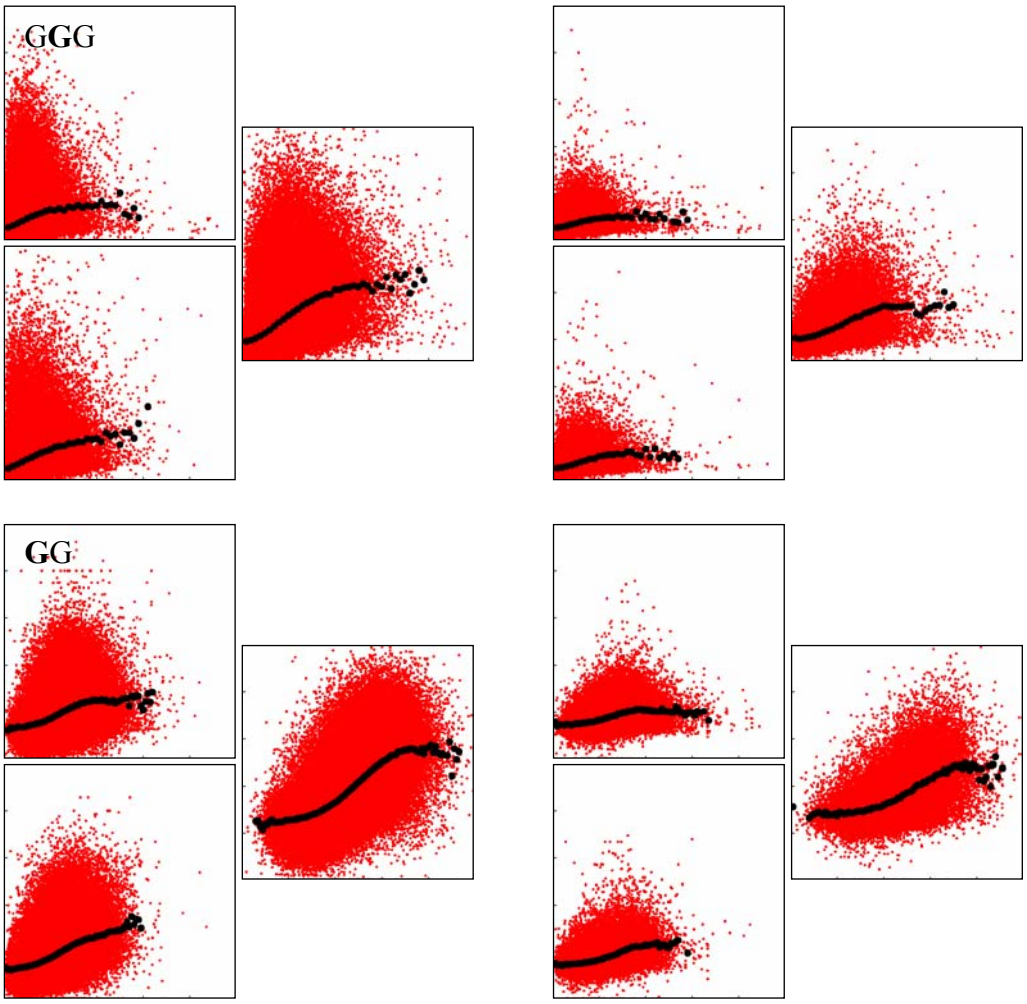
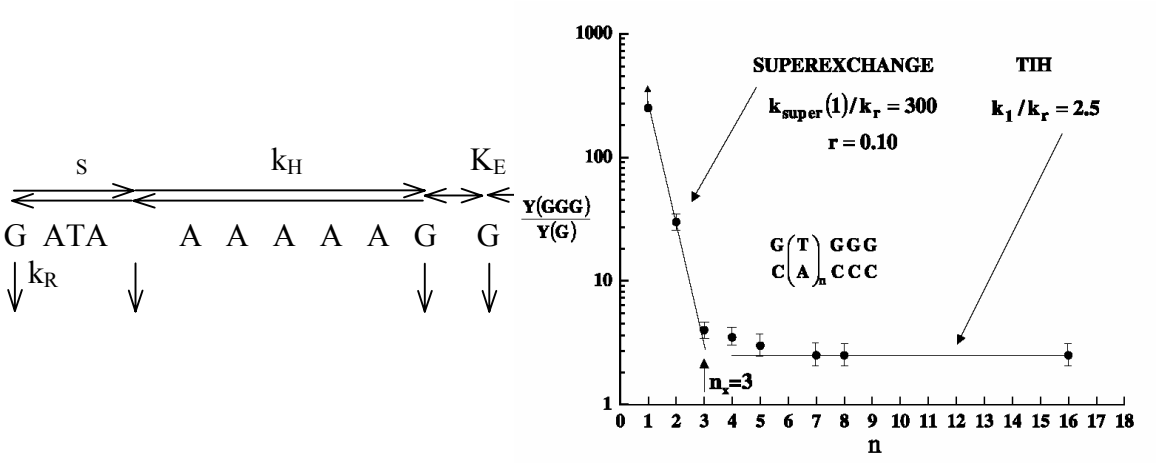


Figure 5-4. Scatter plots of G**G**G and **G**G percentages of individual (red dots) and mean (black circles) introns (left) or IGD (right) vs. adjacent exons for the human genome. Notes: Charts in each triplet show coding (top), template (bottom) and combined (middle) strands. Vertical (intron or IGD) and horizontal (exon) full scales are 25% for G**G**G and **G**G.



Rate and Equilibrium Constants

k_S: Superexchange to near ($n \leq 3$ bp) G.

k_H: Thermally-induced hopping to far ($n \geq 4$) G via A.

k_R: Competing Reactions.

K_E: Thermal equilibrium between adjacent G.

rate = k[hole]

(All rate laws are first-order in hole conc'n.)

Probabilities of Hole Moves and Reaction

Left = $(k_S/k_R) / ((k_S/k_R) + (k_H/k_R) + 1)$

Right = $(k_H/k_R) / ((k_S/k_R) + (k_H/k_R) + 1)$

React = $1 / ((k_S/k_R) + (k_H/k_R) + 1)$

Figure 5-5. A simple kinetic scheme for hole movement and reaction of a hole ($\mathbf{G}^{\bullet+}$). Graph of relative yield (Y_{GGG}/Y_G) vs. separation (n) from (3, 4).

Appendix 1: Supplemental Material for Chapter 3

Table 3-S1. Compositions and structures of model genomes.

Exon	*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Genome, bp	2.8×10^9	1.2×10^8	1.0×10^8	1.2×10^8	1.2×10^7	1.2×10^7	2.5×10^6	2.0×10^6
% of Genome	1.5	17.	25.	29.	73.	55.	86.	49.
Number	2.6×10^5	5.0×10^4	1.2×10^5	1.4×10^5	6.1×10^3	9.6×10^3	1.9×10^3	7.6×10^2
ΣG, %	0.26	0.26	0.21	0.22	0.20	0.20	0.24	0.13
Mean Length, bp	170.	400.	210.	250.	1400.	710.	1100.	1300.
Median. Len., bp	130.	230.	150.	140.	1200.	340.	890.	590.
Len. <200 bp, %	81.	45.	66.	66.	4.7	39.	0.91	33.

Notes: %ΔEx is (100%)(Intron – Exon)/Exon or (100%)(IGD – Exon)/Exon.

Table 3-S1. (continued)

Intron	*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
% of Genome	47.	15.	27.	15.	0.48	3.1	0.02	3.5
Number	2.1×10^5	3.6×10^4	1.0×10^5	1.1×10^5	2.3×10^2	4.7×10^3	1.5×10^1	3.4×10^2
Number / Gene	4.0	2.6	4.2	5.0	0.039	0.94	0.008	0.79
ΣG, %	0.23	0.18	0.15	0.16	0.17	0.14	0.20	0.065
ΣG, %ΔEx	−10.	−31.	−31.	−25.	−16.	−27.	−15.	−49.
Mean Length, bp	6300.	490.	270.	160.	250.	82.	36.	210.
Median. Len., bp	1900.	71.	65.	99.	170.	56.	33.	160.
Len. <200 bp, %	12.	72.	68.	80.	51.	94.	100.	66.

Table 3-S1. (continued)

IGD	*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
% of Genome	51.	68.	48.	56.	27.	42.	14.	48.
Number	5.3×10^4	1.4×10^4	2.0×10^4	2.6×10^4	5.8×10^3	5.0×10^3	1.9×10^3	4.3×10^2
ΣG, %	0.22	0.19	0.16	0.16	0.17	0.15	0.22	0.071
ΣG, %ΔEx	–13.	–28.	–22.	–26.	–17.	–22.	–7.6	–44.
Mean Length, bp	27000.	5800.	2400.	2500.	560.	1100.	190.	2300.
Median. Len., bp	13000.	1600.	1100.	1300.	370.	730.	92.	2100.
Len. <200 bp, %	1.4	5.0	7.7	1.9	18.	6.6	76.	0.71

Table 3-S2. Percentage standard deviations in ten simulations of probability-predicted mean mol percentages of G**G**G and **G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	0.088	0.37	0.26	0.28	1.7	1.7	1.2	3.2
Exon	–	0.082	0.23	0.19	0.25	0.75	1.3	0.52	7.3
	&	0.062	0.21	0.12	0.22	0.87	1.2	0.80	3.3
	+	0.13	0.56	0.61	0.45	6.5	3.1		27.
Intron	–	0.12	0.49	0.74	0.48	10.	3.3		37.
	&	0.060	0.35	0.50	0.33	6.3	2.6		28.
	+	0.10	0.28	0.39	0.24	1.2	1.2	3.1	7.8
IGD	–	0.14	0.33	0.42	0.40	1.2	0.98	3.0	3.9
	&	0.082	0.23	0.25	0.19	0.99	0.99	2.1	5.1

Table 3-S2. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	0.080	0.10	0.12	0.088	0.22	0.41	0.34	1.7
Exon	–	0.050	0.067	0.051	0.093	0.13	0.27	0.31	1.5
	&	0.050	0.065	0.068	0.057	0.14	0.23	0.27	1.3
	+	0.014	0.25	0.13	0.091	2.2	0.93		3.7
Intron	–	0.031	0.23	0.12	0.13	3.0	0.64		3.6
	&	0.017	0.19	0.11	0.081	1.1	0.45		2.5
	+	0.026	0.089	0.066	0.077	0.28	0.34	0.99	1.2
IGD	–	0.032	0.098	0.11	0.085	0.24	0.28	0.60	0.81
	&	0.022	0.059	0.058	0.063	0.23	0.16	0.69	0.74

Table 3-S3. Mean mol percentages of G**G**G and **G**G in exons, introns and IGD.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	1.8	1.1	0.60	0.98	0.78	0.68	1.8	**0.41**
Exon	–	1.9	1.2	0.58	0.55	0.70	0.64	0.83	**0.28**
	&	3.7	2.3	1.2	1.5	1.5	1.3	2.6	0.69
	+	2.0*	**0.50**	**0.31**	**0.30**	**0.47**	**0.19**		**0.07**
Intron	–	2.0*	0.95	**0.37**	**0.34**	0.53	**0.20**		**0.13**
	&	4.0*	1.5#	0.68#	0.64#	1.0	**0.39**		**0.20**
	+	1.7	0.77	0.51	**0.46**	0.55	**0.35**	1.5	**0.17**
IGD	–	1.7	0.83	0.67*	**0.47**	0.57	**0.39**	1.7*	**0.17**
	&	3.4	1.6	1.2	0.94#	1.1	0.73#	3.2*	**0.34**

Notes: Mol percentages are highlighted when ≤0.5% (bold red). They are underlined when their percentage differences (100%(actual – prob.)/prob.) from probability-predicted values are ≥33% (solid blue) or ≤–33% (dotted red), except when actual mol percentages are ≤0.5%. Intron or IGD mol percentages are stared when their percentage differences (100%(intron – exon)/exon) from corresponding exon values are ≤–33% (# red) or ≥0% (* blue), except when intron or IGD mol percentages are ≤0.5%.

Table 3-S3. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	5.7	5.7	3.9	4.8	3.8	3.7	5.8	2.2
Exon	–	5.3	5.2	3.3	3.2	3.3	3.0	3.4	1.5
	&	11.	11.	7.2	7.9	7.1	6.7	9.2	3.8
	+	4.7	2.6#	2.1#	2.4#	2.2#	1.9#		0.79#
Intron	–	4.7	3.3#	1.8#	1.9#	2.3	1.4#		0.64#
	&	9.4	5.8#	3.9#	4.4#	4.6#	3.3#		1.4#
	+	4.4	3.0#	2.2#	2.3#	2.4#	1.9#	4.6	0.80#
IGD	–	4.4	3.1#	2.5	2.3	2.4	2.1	4.4	0.78#
	&	8.8	6.0#	4.7#	4.6#	4.8	4.0#	8.9	1.6#

Table 3-S4. Percentage standard deviations of G**G**G and **G**G mol percentages in exons, introns and IGD.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	110	100	150	130	120	170	60	320
Exon	–	110	90	150	150	76	120	81	200
	&	84	67	110	95	71	100	43	210
	+	94	180	220	210	130	310		320
Intron	–	97	150	230	200	150	330		280
	&	79	110	170	140	97	230		220
	+	62	61	85	65	95	79	110	99
IGD	–	66	62	84	62	97	80	110	93
	&	60	53	70	54	78	67	84	82

Table 3-S4. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	46	37	50	42	36	53	29	73
Exon	–	53	39	59	57	33	54	38	68
	&	35	26	36	31	23	35	16	52
	+	43	65	74	58	51	77		92
Intron	–	45	65	91	73	50	110		110
	&	37	48	60	44	34	60		69
	+	30	31	38	29	41	37	53	45
IGD	–	31	30	34	28	42	36	53	38
	&	29	27	31	24	34	33	40	37

Table 3-S5. Percentages of exon, intron and IGD populations with no G**G**G or no **G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	26.	20.	**45.**	32.	5.8	28.	2.6	**36.**
Exon	–	28.	19.	**48.**	**49.**	6.1	31.	3.7	**40.**
	&	12.	8.5	25.	20.	2.9	17.	0.61	24.
	+	4.0	**60.**	**67.**	**70.**	**46.**	**86.**		**88.**
Intron	–	3.5	**46.**	**65.**	**69.**	**43.**	**87.**		**83.**
	&	1.7	**35.**	**55.**	**52.**	27.	**77.**		**76.**
	+	0.47	5.5	13.	6.6	22.	17.	32.	16.
IGD	–	0.43	4.8	10.	5.9	23.	16.	24.	17.
	&	0.26	3.4	6.6	3.6	14.	10.	16.	8.0

Notes: Population percentages are highlighted when ≥33% (bold blue). They are underlined when their percentage differences from probability-predicted values are ≥33% (solid blue) or ≤–33% (dotted red), except when actual population percentages are ≤5%.

Table 3-S5. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	0.89	1.2	1.8	0.48	0.78	2.5	0.11	6.2
Exon	–	3.9	2.9	4.7	5.8	1.7	7.3	0.70	7.2
	&	0.35	0.58	0.41	0.15	0.43	1.2	0.11	2.1
	+	0.20	6.5	12.	2.3	5.7	16.		15.
Intron	–	0.53	9.4	29.	15.	4.4	**40.**		39.
	&	0.063	1.0	5.8	0.45	0.44	7.2		6.0
	+	0.10	0.30	1.0	0.11	1.5	0.71	5.2	0.24
IGD	–	0.12	0.26	0.81	0.16	1.6	0.56	3.9	0.47
	&	0.051	0.17	0.54	0.069	1.0	0.50	2.5	0.24

Table 3-S6. Percentage differences between observed and probability-predicted percentages of exon, intron and IGD populations with no G**G**G or no **G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	29.	89.	65.	56.	43.	34.	79.	8.9
Exon	–	6.6	61.	57.	40.	17.	14.	9.1	–20.
	&	48.	100.	130.	110.	31.	40.	17.	–2.2
	+	1.5	17.	13.	34.	17.	19.		–4.1
Intron	–	–22.	–0.54	3.3	18.	8.8	9.0		–11.
	&	18.	15.	17.	60.	35.	29.		–12.
	+	6.5	12.	1.6	48.	0.45	19.	–3.4	–70.
IGD	–	–12.	7.3	0.75	44.	7.3	23.	–35.	–68.
	&	35.	66.	27.	220.	44.	77.	–9.5	–76.

Notes: Percentage differences are underlined when ≥33% (solid blue) or ≤–33% (dotted red), except when actual population percentages are ≤5%.

Table 3-S6. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	–23.	19.	38.	–33.	–4.3	9.4	–69.	13.
Exon	–	52.	64.	180.	160.	23.	65.	0.86	–2.9
	&	–15.	20.	66.	–28.	–18.	14.	–69.	26.
	+	5.5	39.	18.	–11.	92.	36.		–44.
Intron	–	58.	79.	65.	230.	260.	100.		3.7
	&	50.	89.	95.	300.	800.	150.		–49.
	+	48.	37.	90.	100.	99.	44.	35.	–61.
IGD	–	44.	60.	70.	130.	130.	20.	–2.6	5.9
	&	71.	150.	200.	160.	370.	87.	110.	800.

Table 3-S7. Percentage differences between observed and probability-predicted mean mol percentages of G**G**G in exons, introns and IGD with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	–8.8	–40.	–31.	–25.	–20.	–25.	–20.	–0.29
Exon	–	1.0	–33.	–26.	–25.	–9.3	–10.	–9.1	34.
	&	–4.7	–40.	–36.	–30.	–16.	–20.	–17.	13.
	+	33.	–7.7	–14.	–17.	–2.9	–20.		5.8
Intron	–	38.	9.9	7.2	–5.8	5.6	–2.9		48.
	&	36.	2.2	–1.1	–15.	1.6	–13.		45.
	+	40.	5.6	16.	2.8	6.6	–6.0	–6.4	140.
IGD	–	42.	8.4	23.	3.1	2.5	–6.1	4.0	120.
	&	41.	8.0	21.	3.4	8.5	–4.9	7.3	190.

Notes: Percentage differences are underlined when ≥33% (solid blue) or ≤–33% (dotted red), except when actual mol percentages are ≤0.5%.

Table 3-S7. (continued) Percentage differences between observed and probability-predicted mean mol percentages of **G**G in exons, introns and IGD with **G**G.

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	11.	8.4	4.8	7.6	9.1	9.0	4.9	15.
Exon	–	12.	3.2	–2.5	1.2	9.0	5.1	5.2	37.
	&	11.	5.2	–0.29	3.4	8.8	5.8	4.9	23.
	+	13.	–0.53	1.7	–1.5	–5.8	–5.6		36.
Intron	–	16.	14.	20.	–1.2	0.088	13.		66.
	&	14.	4.1	4.3	–6.0	–5.4	–8.4		50.
	+	15.	0.56	2.0	2.3	2.3	–1.0	15.	60.
IGD	–	16.	2.1	2.8	2.3	–0.23	–0.18	15.	59.
	&	15.	1.4	2.4	2.3	1.0	–0.49	16.	60.

Table 3-S8. Differences between actual and probability-predicted mean numbers of GGG in exons, introns and IGD with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	–0.39	–3.7	–1.0	–1.2	–2.9	–2.3	–5.1	–0.024
Exon	–	0.045	–3.0	–0.82	–0.88	–1.1	–0.74	–1.0	1.5
	&	–0.35	–6.8	–1.9	–2.0	–4.1	–2.8	–6.2	1.3
	+	33.	–0.50	–0.41	–0.34	–0.065	–0.28		0.066
Intron	–	36.	0.77	0.19	–0.11	0.12	–0.036		0.50
	&	69.	0.24	–0.045	–0.39	0.05	–0.21		0.52
	+	130.	2.6	1.9	0.34	0.25	–0.29	–1.6	2.7
IGD	–	140.	4.0	3.3	0.38	0.10	–0.32	0.92	2.5
	&	270.	7.2	5.3	0.82	0.57	–0.45	2.7	5.5

Notes: Differences are underlined when corresponding values are in Table 3-S8.

Table 3-S8. (continued) Differences between actual and probability-predicted mean numbers of **G**G in exons, introns and IGD with **G**G.

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	0.96	1.8	0.38	0.83	4.7	2.2	3.1	3.9
Exon	–	0.96	0.67	–0.18	0.10	4.0	1.1	2.0	5.7
	&	1.8	2.2	–0.043	0.64	8.4	2.6	5.0	9.2
	+	34.	–0.071	0.11	–0.063	–0.36	–0.11	0.17	0.50
Intron	–	40.	2.1	1.1	–0.046	0.005	0.21	0.55	0.86
	&	73.	1.1	0.45	–0.45	–0.66	–0.27	–0.48	1.0
	+	106.	1.0	1.0	1.3	0.30	–0.21	6.6	6.8
IGD	–	160.	3.8	1.7	1.3	–0.032	–0.041	6.4	6.6
	&	320.	4.8	2.6	2.6	0.27	–0.21	13.	13.

Table 3-S9. Positive skew of distributions: percentages of segments with sub-mean mol percentages of G**G**G in exon, intron and IGD populations with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	62	63	64	64	61	**70**	56	**76**
Exon	–	62	61	65	63	60	64	62	**72**
	&	60	58	62	62	58	64	55	**72**
	+	**69**	58	**67**	58	65	58		61
Intron	–	**69**	**70**	**67**	60	**68**	60		**66**
	&	**67**	65	60	**68**	65	58		61
	+	**66**	53	61	60	61	58	62	65
IGD	–	**67**	53	63	60	61	60	63	63
	&	**66**	52	60	58	58	57	60	60

Notes: Population percentages are highlighted when ≥66% (bold blue).

Table 3-S9. (continued) Positive skew of distributions: percentages of segments with sub-mean mol percentages of **G**G in exon, intron and IGD populations with **G**G.

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	55	53	55	55	56	60	54	60
Exon	–	54	52	57	57	56	57	58	63
	&	52	49	54	53	55	55	53	59
	+	61	57	60	55	57	58		63
Intron	–	61	60	65	57	58	62		61
	&	59	54	55	53	51	54		58
	+	59	49	52	55	51	53	55	56
IGD	–	60	50	54	55	51	55	56	57
	&	59	49	52	55	50	53	53	55

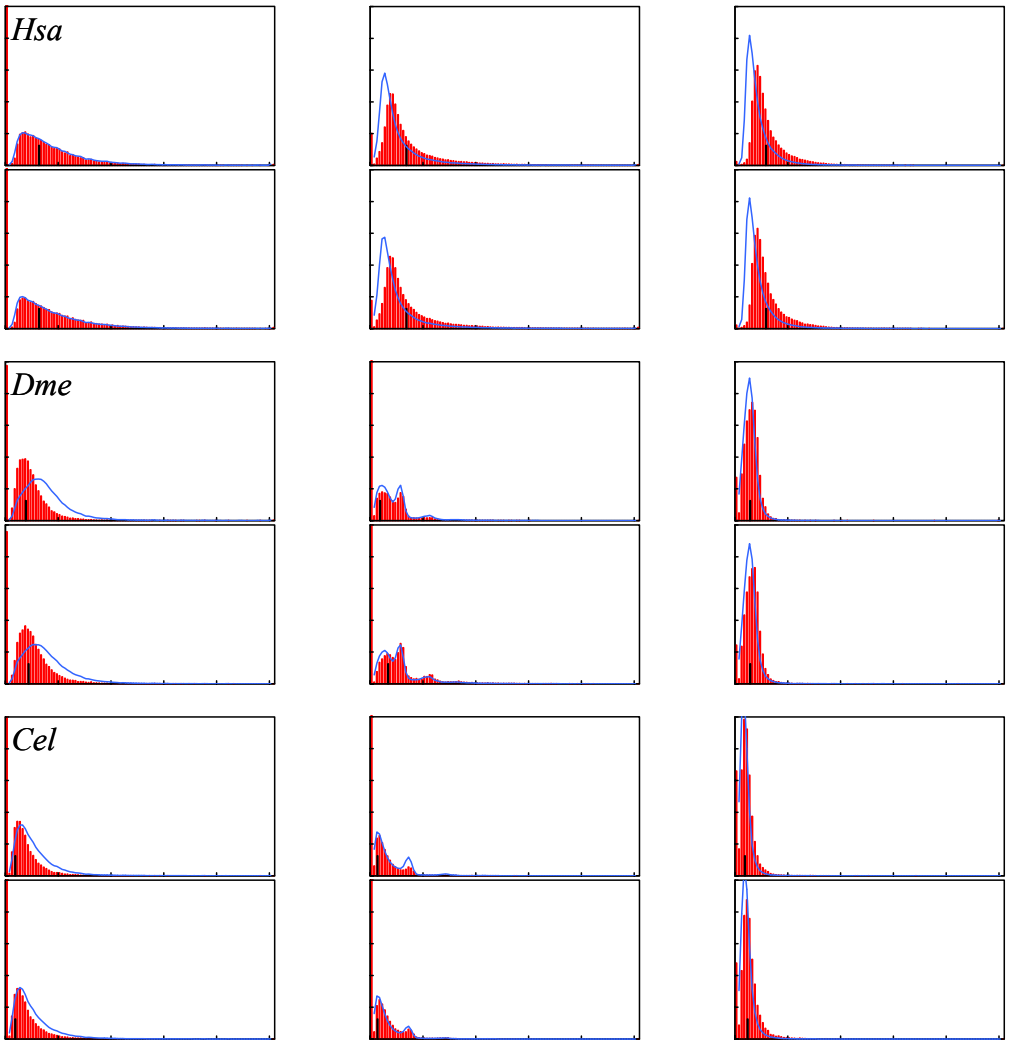


Figure 3-S1. Observed (red bars) and probability-predicted (blue lines) frequencies, and observed means (black bars) of G**G**G mol percentages. Notes: Charts in each pair of rows show, from left to right, exons, introns and IGD on the coding (top) and template (bottom) strands. Vertical full scales are 20% of population. Horizontal full scales are 15% and intervals are 0.15%.

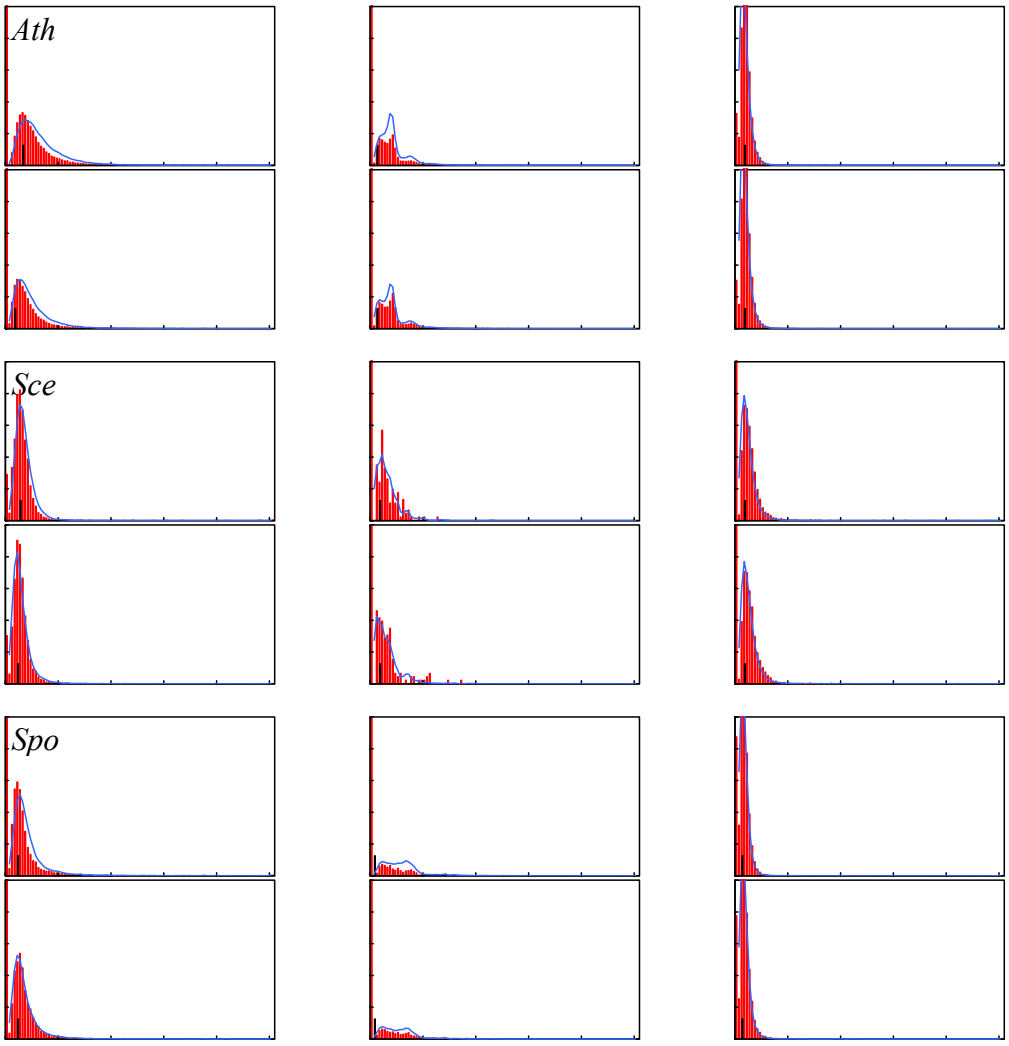


Figure 3-S1. (continued)

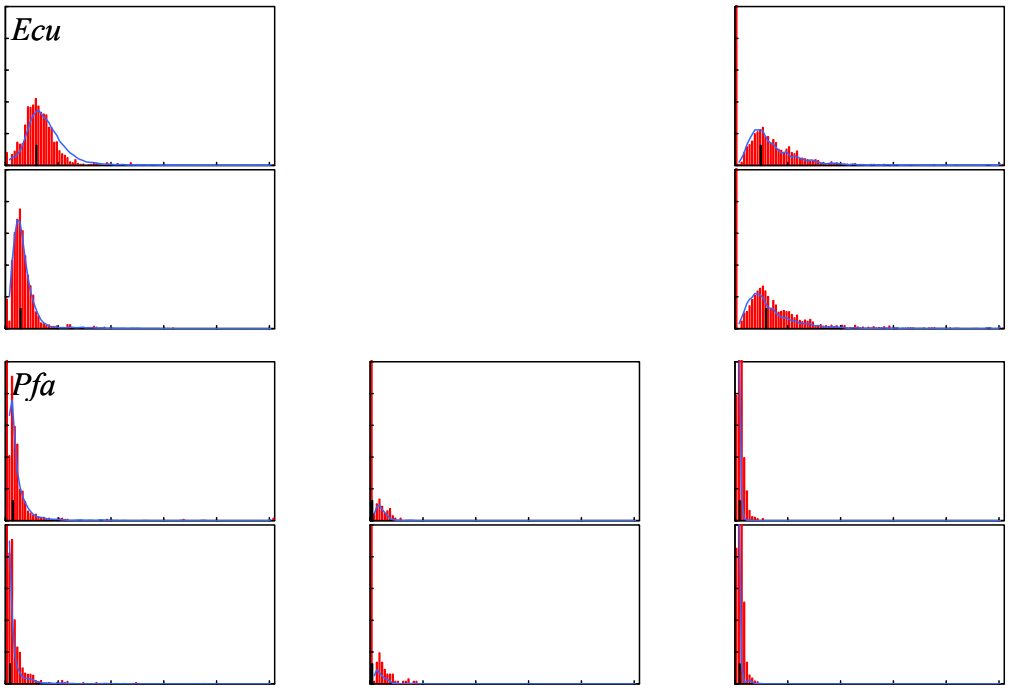


Figure 3-S1. (continued)

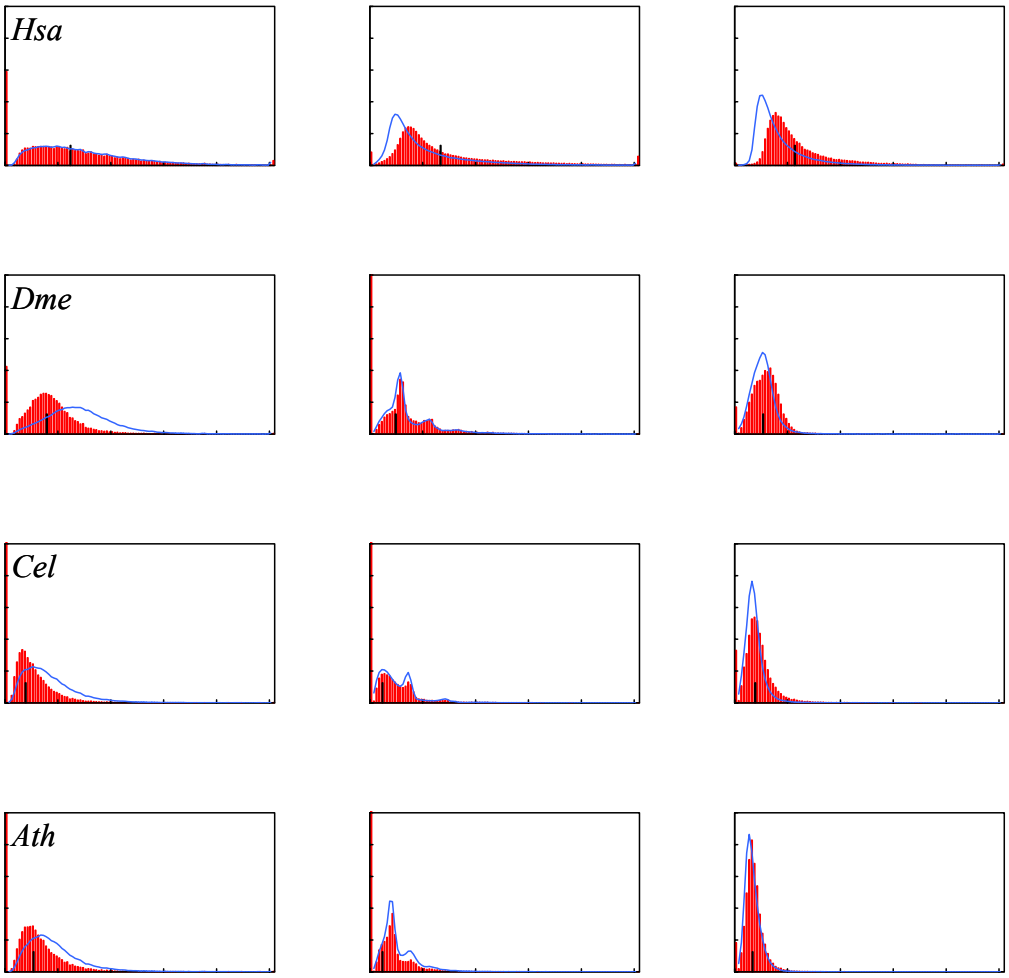


Figure 3-S2. Observed (red bars) and probability-predicted (blue lines) frequencies, and observed means (black bars) of G**G**G mol percentages on combined coding plus template strands. Notes: Charts in each row show exons, introns and IGD, from left to right. Charts are scaled as in Figure 3-S1.

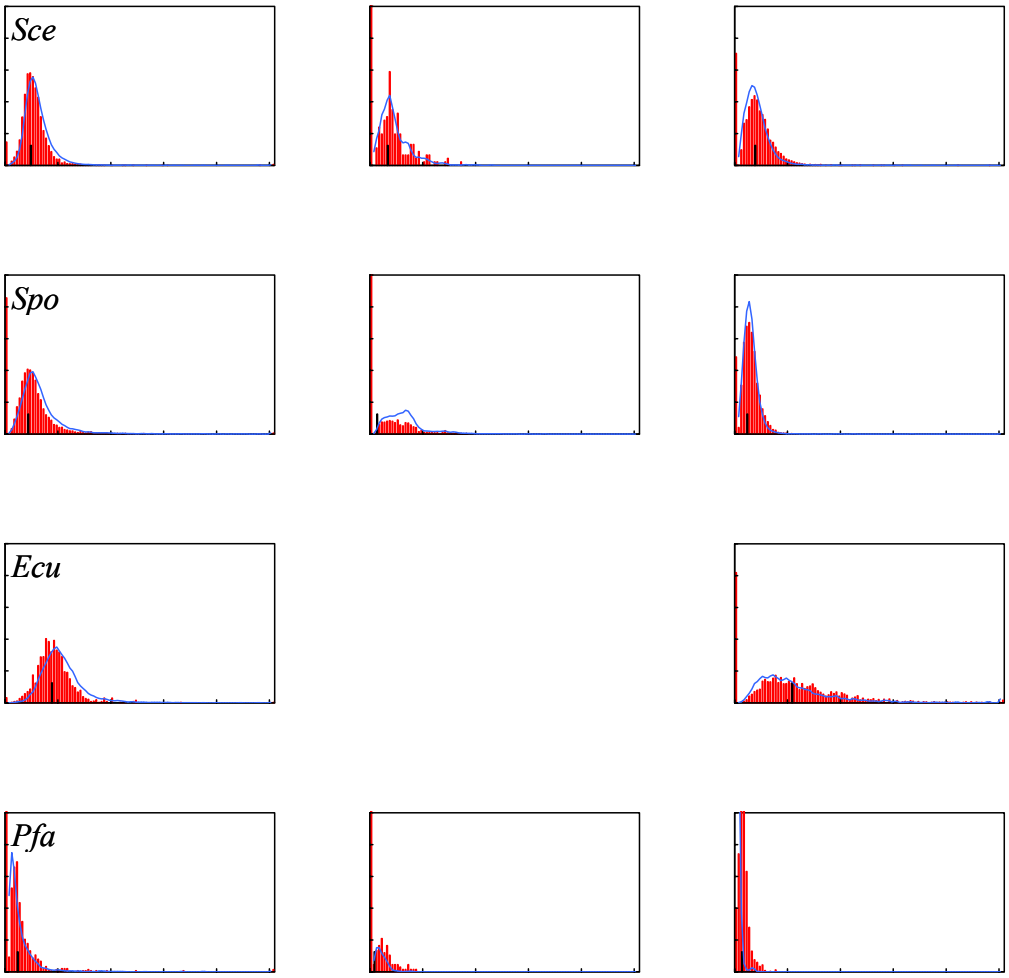


Figure 3-S2. (continued)

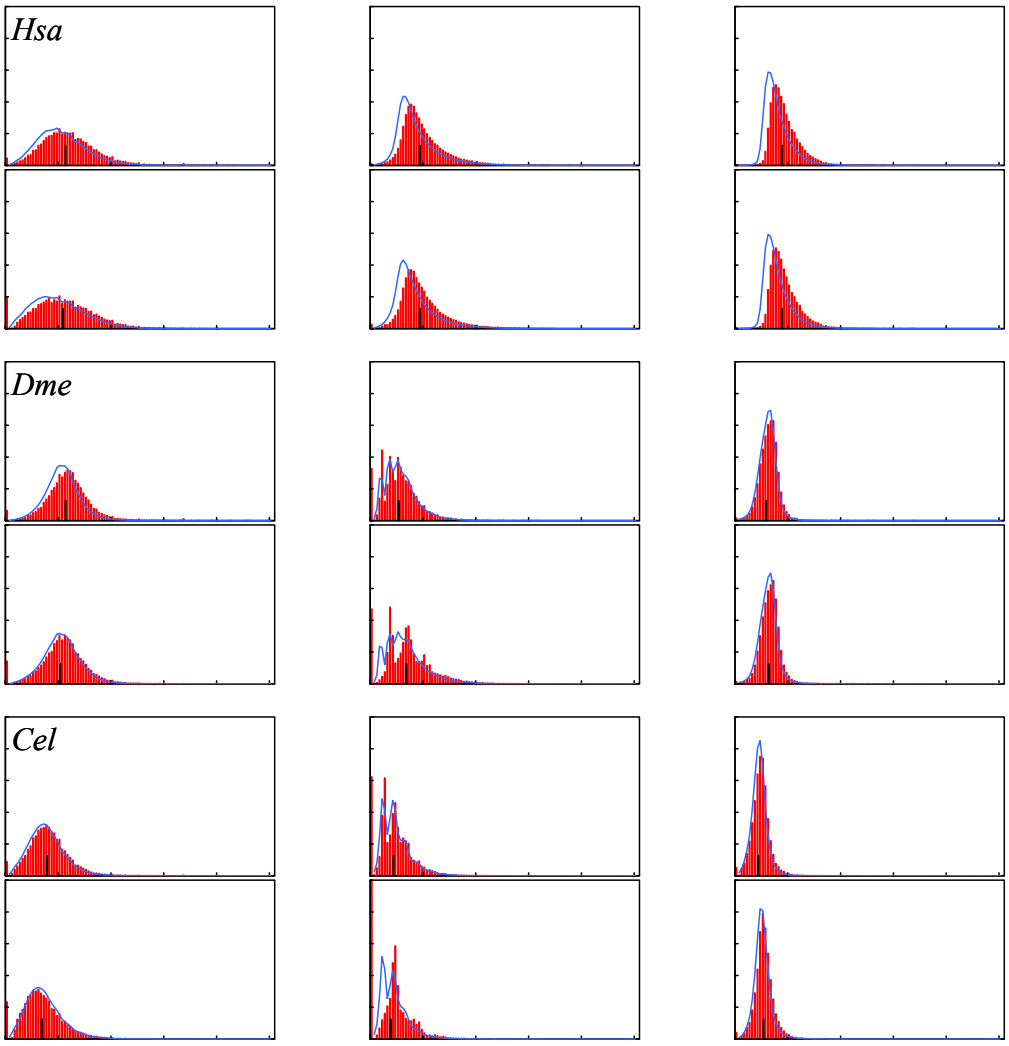


Figure 3-S3. Observed and probability-predicted frequencies, and observed means of **G**G mol percentages. Notes: Charts are organized and color-coded as in Figure 3-S1. Vertical full scales are 20% of population. Horizontal full scales are 25% and intervals are 0.25%.

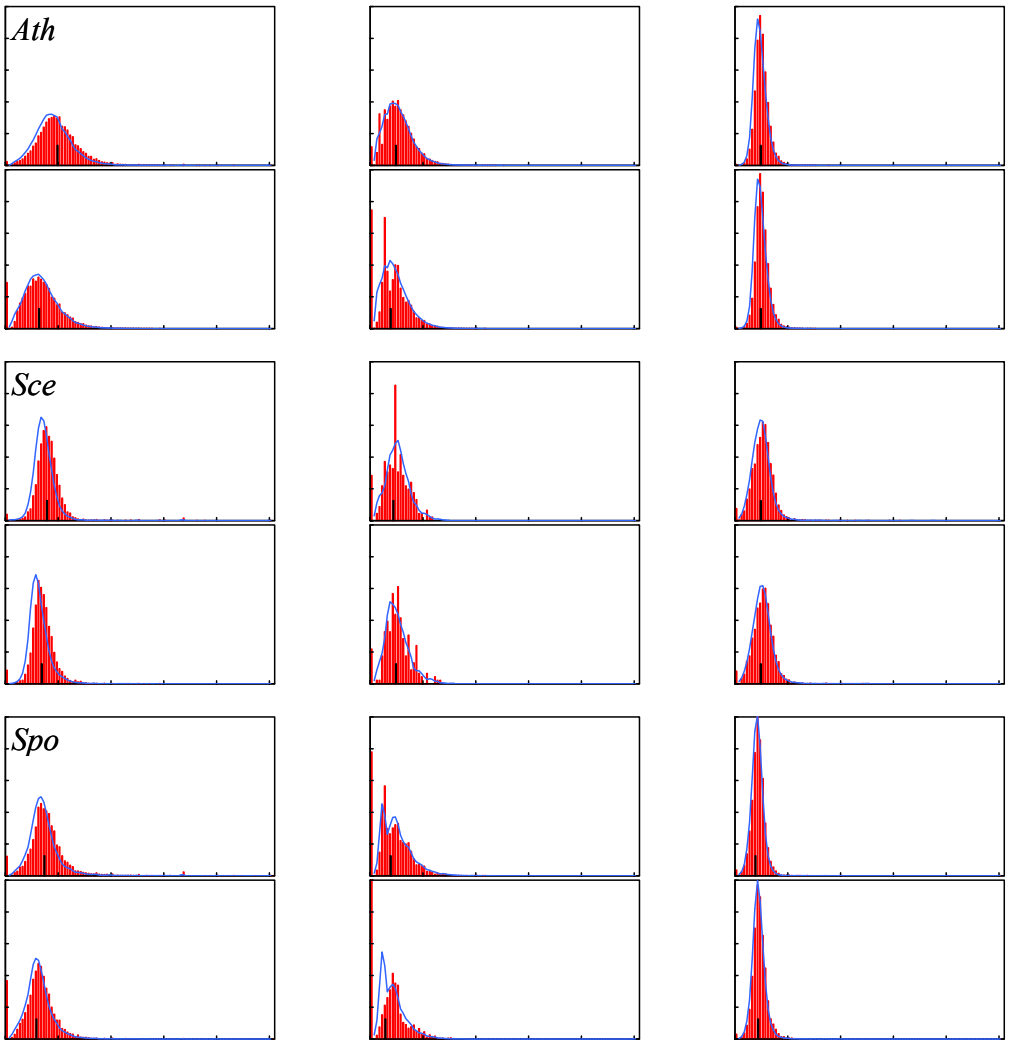


Figure 3-S3. (continued)

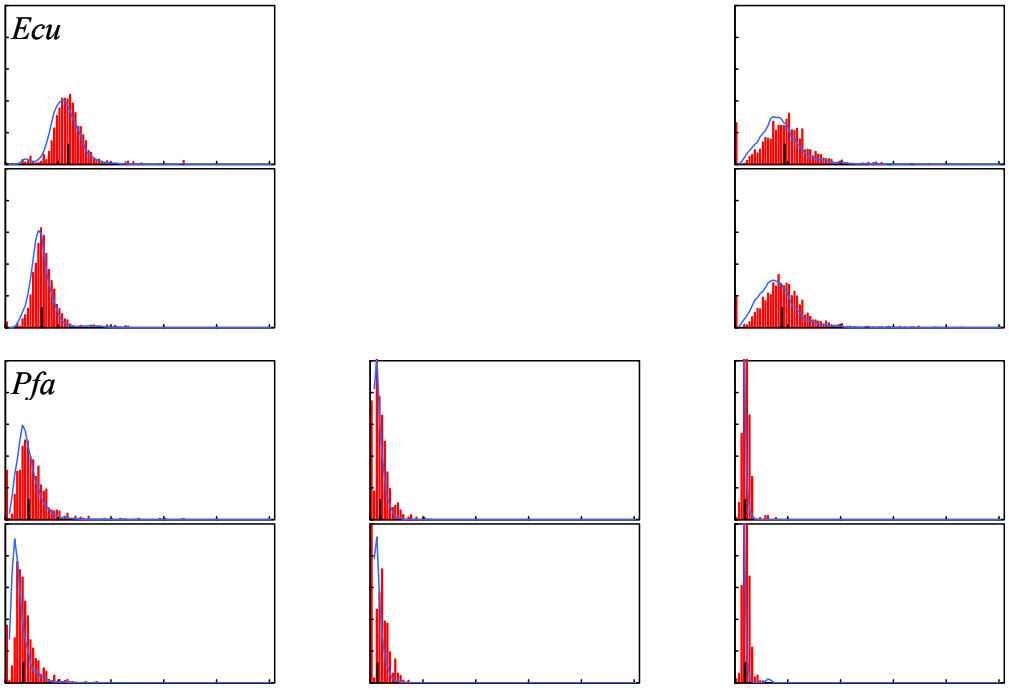


Figure 3-S3. (continued)

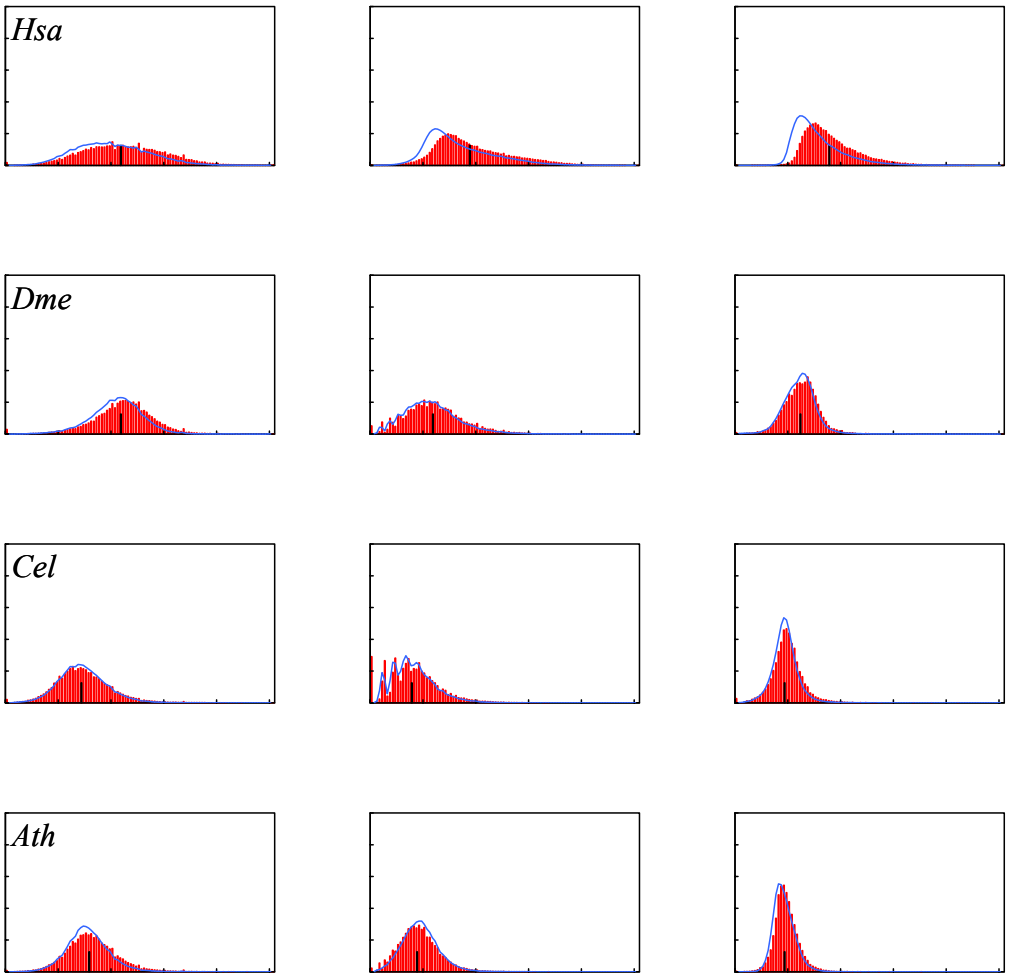


Figure 3-S4. Observed and probability-predicted frequencies, and observed means of **G**G mol percentages on combined coding plus template strands. Notes: Charts are organized and color-coded as in Figure 3-S2. Charts are scaled as in Figure 3-S3.

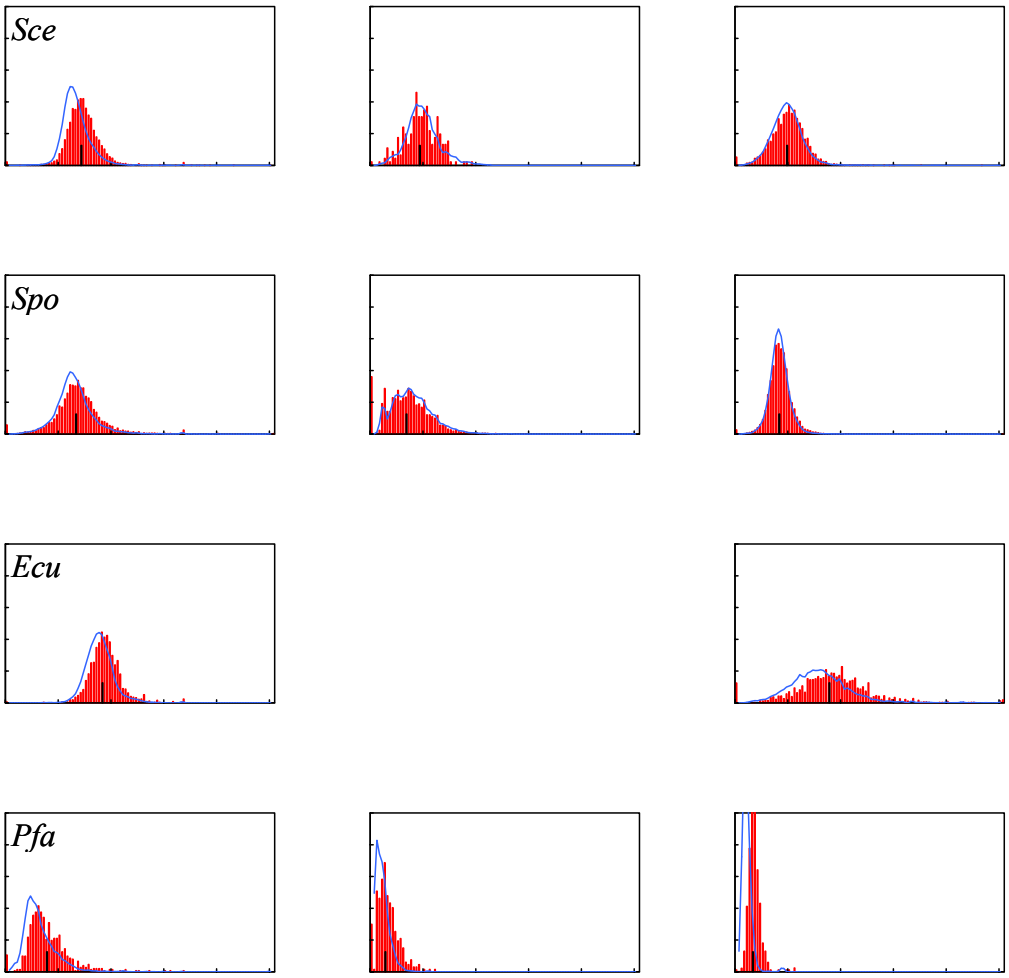


Figure 3-S4. (continued)

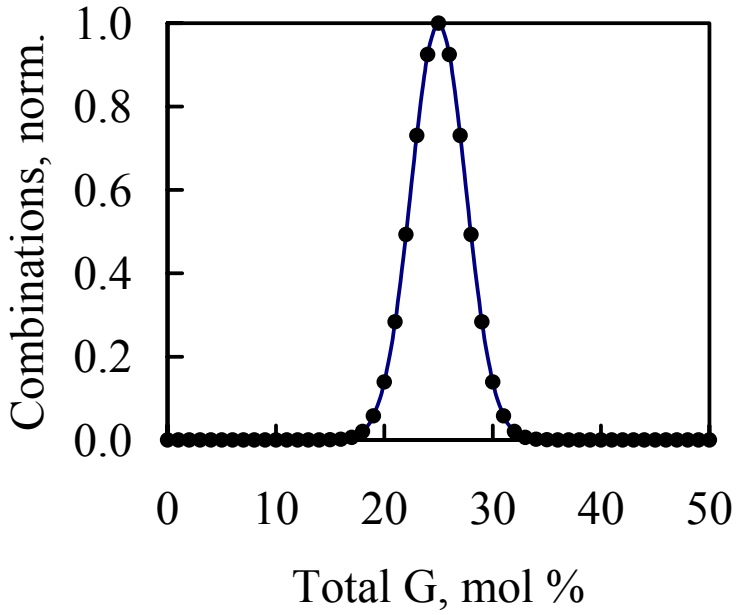


Figure 3-S5. Normalized combinations of 100 nucleotides vs. G mol percentage. Notes: Non-normalized combinations are ~10^{57} at 25% total G and ~10^{29} at 0% total G.

GenBank File Names (.gbk)

Homo sapiens (Feb. 2002 release)
hs_chr1, hs_chr2, hs_chr3, hs_chr4, hs_chr5, hs_chr6, hs_chr7, hs_chr8, hs_chr9, hs_chr10, hs_chr11, hs_chr12, hs_chr13, hs_chr14, hs_chr15, hs_chr16, hs_chr17, hs_chr18, hs_chr19, hs_chr20, hs_chr21, hs_chr22, hs_chrX, hs_chrY.

Drosophila melanogaster (Oct. 2000)
AE002566, AE002575, AE002584, AE002593, AE002602, AE002620, AE002629, AE002638, AE002647, AE002681, AE002690, AE002699, AE002708, AE002725, AE002769, AE002778, AE002787, AE002796, AE002804, small.

Caenorhabditis elegans (Dec. 2001)
NC_003279, NC_003280, NC_003281, NC_003282, NC_003283, NC_003284.

Arabidopsis thaliana (Jan. 2002)
NC_003070, NC_003071, NC_003074, NC_003075, NC_003076.

Saccharomyces cerevisiae (Mar. 2002)
NC_001133, NC_001134, NC_001135, NC_001136, NC_001137, NC_001138, NC_001139, NC_001140, NC_001141, NC_001142, NC_001143, NC_001144, NC_001145, NC_001146, NC_001147, NC_001148.

Schizosaccharomyces pombe (Mar. 2002)
NC003421, NC003423, NC003424.

Encephalitozoon cuniculi (Mar. 2002)
NC003229, NC003230, NC003231, NC003232, NC003233, NC003234, NC003235, NC003236, NC003237, NC003238, NC003242.

Plasmodium falciparum chromosomes 2 (Nov. 1998) and 3 (Apr. 1999)
AE001362, MAL3.

Appendix 2: Supplemental Material for Chapter 4

Table 4-S1. Compositions and structures of model genomes.

Exon	*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Genome, bp	2.8×10^9	1.2×10^8	1.0×10^8	1.2×10^8	1.2×10^7	1.2×10^7	2.5×10^6	2.0×10^6
% of Genome	1.5	17.	25.	29.	73.	55.	86.	49.
Number	2.6×10^5	5.0×10^4	1.2×10^5	1.4×10^5	6.1×10^3	9.6×10^3	1.9×10^3	7.6×10^2
ΣG, %	0.26	0.26	0.21	0.22	0.20	0.20	0.24	0.13
Mean Length, bp	170.	400.	210.	250.	1400.	710.	1100.	1300.
Median. Len., bp	130.	230.	150.	140.	1200.	340.	890.	590.
Len. <200 bp, %	81.	45.	66.	66.	4.7	39.	0.91	33.

Notes: %ΔEx is (100%)(Intron – Exon)/Exon or (100%)(IGD – Exon)/Exon.

Table 4-S1. (continued)

Intron	*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
% of Genome	47.	15.	27.	15.	0.48	3.1	0.02	3.5
Number	2.1×10^5	3.6×10^4	1.0×10^5	1.1×10^5	2.3×10^2	4.7×10^3	1.5×10^1	3.4×10^2
Number / Gene	4.0	2.6	4.2	5.0	0.039	0.94	0.008	0.79
ΣG, %	0.23	0.18	0.15	0.16	0.17	0.14	0.20	0.065
ΣG, %ΔEx	−10.	−31.	−31.	−25.	−16.	−27.	−15.	−49.
Mean Length, bp	6300.	490.	270.	160.	250.	82.	36.	210.
Median. Len., bp	1900.	71.	65.	99.	170.	56.	33.	160.
Len. <200 bp, %	12.	72.	68.	80.	51.	94.	100.	66.

Table 4-S1. (continued)

IGD	*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
% of Genome	51.	68.	48.	56.	27.	42.	14.	48.
Number	5.3×10^4	1.4×10^4	2.0×10^4	2.6×10^4	5.8×10^3	5.0×10^3	1.9×10^3	4.3×10^2
ΣG, %	0.22	0.19	0.16	0.16	0.17	0.15	0.22	0.071
ΣG, %ΔEx	–13.	–28.	–22.	–26.	–17.	–22.	–7.6	–44.
Mean Length, bp	27000.	5800.	2400.	2500.	560.	1100.	190.	2300.
Median. Len., bp	13000.	1600.	1100.	1300.	370.	730.	92.	2100.
Len. <200 bp, %	1.4	5.0	7.7	1.9	18.	6.6	76.	0.71

Table 4-S2. ANOVA η^2 for correlation between G**G**G and **G**G mol percentages in introns or IGD and in neighboring exons.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	0.087	0.003	0.003	0.002	0.008	0.001		0.017
Intron	–	0.13	0.005	0.005	0.003	0.021	0.002		0.004
	&	**0.23**	0.010	0.009	0.002	0.024	0.001		0.001
	+	0.093	0.002	0.002	0.006	0.008	0.001	0.029	0.002
IGD	–	0.13	0.004	0.007	0.004	0.019	0.007	0.011	0.001
	&	**0.21**	0.007	0.008	0.008	0.017	0.003	0.031	0.019

Notes: η^2 are highlighted when ≥0.20 (bold blue).

Table 4-S2. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	0.15	0.013	0.016	0.011	0.048	0.019		0.082
Intron	–	**0.23**	0.021	0.016	0.006	0.061	0.008		0.018
	&	**0.40**	0.039	0.029	0.005	0.11	0.008		0.071
	+	0.12	0.011	0.007	0.005	0.010	0.006	0.074	0.024
IGD	–	0.19	0.018	0.016	0.005	0.056	0.010	0.039	0.018
	&	**0.30**	0.025	0.024	0.007	0.036	0.009	0.041	0.015

Table 4-S3. ANOVA f-test for correlation between G**G**G and **G**G mol percentages in introns or IGD and in neighboring exons.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	**720.**	9.4	27.	14.	0.76	0.43		2.8
Intron	–	**1100.**	15.	41.	25.	2.3	1.0		1.1
	&	**1700.**	22.	59.	14.	1.0	0.49		0.09
	+	**200.**	2.1	4.6	13.	11.	1.0	7.8	0.77
IGD	–	**320.**	6.0	18.	16.	28.	7.4	3.9	0.44
	&	**410.**	6.5	13.	14.	17.	1.8	6.5	2.7

Notes: f-test values are highlighted when ≥100 (bold blue).

Table 4-S3. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
	+	**610.**	10.	33.	24.	0.66	3.0		2.2
Intron	–	**1100.**	18.	39.	17.	1.1	1.5		0.49
	&	**1600.**	25.	45.	10.	1.8	0.94		1.4
	+	**110.**	3.5	3.3	2.7	3.1	1.1	7.4	0.93
IGD	–	**230.**	6.1	8.5	3.6	19.	2.3	4.5	0.78
	&	**280.**	6.3	9.0	3.4	9.3	1.2	3.6	0.52

Table 4-S4. Mean mol percentages of G**G**G and **G**G in intron and IGD flanks.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	2.6*	0.52#	**0.33**	**0.29**	0.54	**0.19**		**0.069**
Flank	– 3	2.2*	0.91	**0.33**	**0.35**	0.56	**0.19**		**0.080**
	& 53	4.7*	1.4#	0.67#	0.64#	1.1	**0.38**		**0.15**
Intron	+ 3	1.9*	**0.35**	**0.24**	**0.29**	**0.38**	**0.16**		**0.039**
Flank	– 5	2.4*	0.96	**0.42**	**0.32**	**0.50**	**0.20**		**0.17**
	& 35	4.3*	1.3#	0.66#	0.61#	0.88#	**0.36**		**0.21**
IGD	+ 5	1.9*	0.69#	**0.35**	**0.48**	**0.38**	**0.35**	1.0#	**0.24**
Flank	– 3	2.1*	0.92	0.79*	**0.37**	0.52	**0.38**	1.0*	**0.29**
	& 53	4.0*	1.6	1.1	0.85#	0.90#	0.73#	2.1	0.53
IGD	+ 3	2.1*	0.67#	**0.37**	**0.48**	0.52#	**0.35**	1.2	**0.13**
Flank	– 5	2.3*	1.1	0.76*	0.68*	**0.47**	0.55	1.4*	**0.28**
	& 35	4.4*	1.7	1.1	1.2	0.99#	0.89	2.7*	**0.41**

Notes: Mol percentages are highlighted when ≤0.5% (bold red). Flank mol percentages are underlined when their percentage differences (100%(flank–all)/all) from overall values are ≥20% (solid blue) or ≤–20% (dotted red), except when flank mol percentages are ≤0.5%. Intron or IGD mol percentages are stared when their percentage differences (100%(intron – exon)/exon) from corresponding exon values are ≤–33% (# red) or ≥0% (* blue), except when intron or IGD mol percentages are ≤0.5%.

Table 4-S4. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	5.0	2.0#	1.5#	1.8#	2.3#	1.2#		0.60#
Flank	– 3	4.7	3.1#	1.6#	1.9#	2.3	1.3#		0.53#
	& 53	9.7	5.1#	3.1#	3.7#	4.6#	2.5#		1.1#
Intron	+ 3	4.1	1.7#	1.3#	1.8#	1.6#	1.0#		**0.39**
Flank	– 5	5.0	3.2#	1.8#	1.8#	2.3	1.3#		0.76#
	& 35	9.1	4.9#	3.0#	3.6#	3.9#	2.4#		1.2#
IGD	+ 5	4.6	2.7#	1.6#	2.4#	1.8#	1.9#	3.5#	0.91#
Flank	– 3	4.8	3.3#	2.4	1.9#	2.1#	1.9#	3.3	0.73#
	& 53	9.4	6.0#	3.9#	4.3#	3.9#	3.8#	6.8	1.6#
IGD	+ 3	5.2	3.2#	1.8#	2.2#	2.5#	1.9#	4.0	0.73#
Flank	– 5	5.2	3.8	2.9	2.9	2.3	2.7	3.8*	0.84#
	& 35	10.	7.0#	4.7#	5.0#	4.8#	4.6	7.8	1.6#

Table 4-S5. Percent standard deviations of G**G**G and **G**G mol percentages in intron and IGD flanks.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	120	220	280	240	160	320		460
Flank	– 3	130	170	280	220	180	340		430
	& 53	95	140	200	160	110	230		320
Intron	+ 3	130	250	310	240	200	340		540
Flank	– 5	120	160	260	230	180	340		300
	& 35	98	140	200	160	130	240		260
IGD	+ 5	130	180	230	190	210	210	170	300
Flank	– 3	130	160	200	210	200	210	150	490
	& 53	97	120	150	140	140	140	110	290
IGD	+ 3	120	170	240	190	180	200	140	400
Flank	– 5	120	140	180	160	190	170	140	360
	& 35	91	110	150	120	130	130	93	270

Table 4-S5. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	66	96	120	84	68	120		160
Flank	– 3	67	76	110	79	63	110		150
	& 53	51	62	86	56	45	82		110
Intron	+ 3	69	98	120	83	85	120		170
Flank	– 5	65	74	110	82	63	110		120
	& 35	53	61	85	56	52	81		100
IGD	+ 5	62	82	98	73	84	78	77	130
Flank	– 3	64	72	82	79	80	80	77	220
	& 53	48	57	61	53	58	52	52	120
IGD	+ 3	61	69	91	84	69	76	68	130
Flank	– 5	62	62	69	71	71	65	66	140
	& 35	47	47	55	49	46	48	44	97

Table 4-S6. Percentages of flanks with no G**G**G or no **G**G in intron and IGD populations.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	32	**75**	**82**	**80**	**66**	**89**		**95**
Flank	– 3	36	61	**83**	**77**	**67**	**90**		**94**
	& 53	16	47	**69**	62	44	**80**		**89**
Intron	+ 3	40	**81**	**87**	**80**	**76**	**90**		**96**
Flank	– 5	33	59	**79**	**78**	**68**	**89**		**88**
	& 35	18	49	**70**	63	49	**81**		**85**
IGD	+ 5	36	63	**76**	**69**	**74**	**76**	58	**85**
Flank	– 3	34	55	64	**75**	**69**	**75**	56	**89**
	& 53	16	35	48	52	51	56	34	**76**
IGD	+ 3	32	62	**77**	**70**	**67**	**75**	50	**91**
Flank	– 5	31	49	60	60	**69**	65	42	**86**
	& 35	13	30	46	41	45	48	22	**77**

Notes: Population percentages are highlighted when ≥+66% (bold blue)

Table 4-S6. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	4.2	28.	42.	22.	13.	49.		60.
Flank	– 3	4.7	13.	38.	19.	9.6	44.		61.
	& 53	0.69	5.0	21.	4.0	1.3	24.		37.
Intron	+ 3	6.2	31.	46.	21.	22.	51.		**70.**
Flank	– 5	4.3	13.	36.	21.	8.7	43.		51.
	& 35	0.80	5.1	21.	4.0	2.2	24.		36.
IGD	+ 5	3.2	12.	27.	12.	21.	17.	19.	48.
Flank	– 3	3.6	8.7	16.	17.	16.	18.	18.	57.
	& 53	0.36	1.7	3.9	1.9	4.2	3.1	8.2	32.
IGD	+ 3	2.5	7.0	21.	18.	10.	16.	12.	52.
Flank	– 5	3.0	3.8	8.6	9.9	12.	7.9	9.2	49.
	& 35	0.33	0.43	1.9	0.91	1.7	1.3	3.7	26.

Table 4-S7. Percentage differences between flank and overall mean mol percentages of G**G**G in introns and IGD with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	28.	2.6	6.0	–4.0	16.	4.8		–2.2
Flank	– 3	8.0	–4.4	–10.	3.3	5.8	–5.5		–37.
	& 53	18.	–2.0	–2.7	–0.14	10.	–0.55		–25.
Intron	+ 3	–6.5	–30.	–25.	–3.4	–19.	–13.		–45.
Flank	– 5	21.	0.94	14.	–4.7	–4.9	–1.3		32.
	& 35	7.2	–9.8	–4.0	–4.1	–11.	–6.9		5.2
IGD	+ 5	11.	–9.6	–30.	5.7	–25.	5.7	–21.	43.
Flank	– 3	23.	12.	21.	–21.	–3.0	1.8	–34.	75.
	& 53	17.	1.7	–0.77	–8.0	–14.	3.6	–29.	59.
IGD	+ 3	26.	–16.	–29.	–0.36	–15.	–7.0	–26.	–28.
Flank	– 5	36.	24.	11.	38.	–25.	33.	–25.	52.
	& 35	31.	5.2	–5.8	20.	–18.	16.	–24.	14.

Notes: Percentage differences are underlined when ≥+20% (solid blue) or ≤–20% (dotted red), except when flank mol percentages are ≤0.5%.

Table 4-S7. (continued) Percentage differences between flank and overall mean mol percentages of **G**G in introns and IGD with **G**G.

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	5.9	–21.	–25.	–28.	1.6	–39.		–24.
Flank	– 3	0.41	–5.3	–14.	–0.25	–0.75	–6.3		–18.
	& 53	3.2	–12.	–20.	–16.	0.37	–25.		–21.
Intron	+ 3	–14.	–34.	–39.	–26.	–26.	–46.		–50.
Flank	– 5	7.2	–2.8	–3.8	–4.8	–3.7	–2.9		19.
	& 35	–3.4	–17.	–23.	–17.	–15.	–28.		–19.
IGD	+ 5	3.9	–7.6	–27.	5.1	–24.	–1.9	–22.	15.
Flank	– 3	8.9	6.1	–5.4	–17.	–12.	–8.7	–24.	–6.9
	& 53	6.4	–0.61	–16.	–5.8	–18.	–5.4	–23.	3.9
IGD	+ 3	17.	6.5	–16.	–5.2	3.6	–2.1	–14.	–8.8
Flank	– 5	18.	24.	14.	24.	–6.6	30.	–15.	7.4
	& 35	17.	15.	0.065	9.4	–1.5	15.	–14.	–0.77

Table 4-S8. Differences between flank and overall mean numbers of G**G**G in 100 nt of introns and IGD with G**G**G.

G**G**G		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	0.60	0.032	0.06	–0.041	0.14	0.064		–0.012
Flank	– 3	0.16	–0.08	–0.11	0.035	0.054	–0.083		–0.28
	& 53	0.75	–0.044	–0.041	–0.002	0.14	–0.009		–0.20
Intron	+ 3	–0.14	–0.38	–0.24	–0.035	–0.16	–0.17		–0.26
Flank	– 5	0.43	0.017	0.15	–0.050	–0.045	–0.020		0.24
	& 35	0.29	–0.22	–0.060	–0.054	–0.16	–0.11		0.04
IGD	+ 5	0.19	–0.079	–0.17	0.028	–0.18	0.024	–0.46	0.087
Flank	– 3	0.40	0.10	0.16	–0.11	–0.02	0.008	–0.78	0.16
	& 53	0.59	0.028	–0.010	–0.078	–0.19	0.030	–1.1	0.22
IGD	+ 3	0.44	–0.13	–0.17	–0.002	–0.11	–0.03	–0.55	–0.058
Flank	– 5	0.61	0.21	0.081	0.19	–0.19	0.15	–0.57	0.11
	& 35	1.1	0.087	–0.073	0.19	–0.24	0.13	–0.89	0.052

Notes: Differences are underlined when corresponding values are in Table 4-S7.

Table 4-S8. (continued) Differences between flank and overall mean numbers of **G**G in 100 nt of introns and IGD with **G**G.

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
Intron	+ 5	0.28	–0.53	–0.52	–0.68	0.035	–0.75		–0.19
Flank	– 3	0.019	–0.17	–0.25	–0.005	–0.018	–0.087		–0.11
	& 53	0.30	–0.71	–0.78	–0.69	0.017	–0.84		–0.30
Intron	+ 3	–0.66	–0.87	–0.81	–0.63	–0.58	–0.88		–0.39
Flank	– 5	0.33	–0.091	–0.070	–0.093	–0.088	–0.040		0.12
	& 35	–0.32	–0.97	–0.88	–0.72	–0.67	–0.92		–0.27
IGD	+ 5	0.17	–0.23	–0.60	0.12	–0.58	–0.043	–1.0	0.11
Flank	– 3	0.40	0.18	–0.14	–0.39	–0.30	–0.19	–1.1	–0.056
	& 53	0.57	–0.044	–0.74	–0.27	–0.88	–0.23	–2.2	0.058
IGD	+ 3	0.73	0.20	–0.35	–0.12	0.10	–0.036	–0.59	–0.068
Flank	– 5	0.79	0.74	0.36	0.55	–0.15	0.64	–0.58	0.060
	& 35	1.5	0.93	0.02	0.43	–0.049	0.61	–1.2	–0.009

Table 4-S9. Mean mol percentages of G**G**G and **G**G in intron flanks and whole introns with G**G**G and **G**G, respectively, analyzed by length (L).

G**G**G			*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
L ≤ 100	+	5	4.9	1.5	1.7	1.4		1.8		
Intron	–	3	4.5	2.3	2.4	1.5		2.0		
Flank	&	53	7.6	2.3	2.2	1.6		2.0		
L < 200	+	5	5.0	1.5	1.6	1.3	1.6	1.6		0.83
Intron	–	3	4.1	2.2	2.0	1.4	1.7	1.7		0.73
Flank	&	53	7.8	2.3	2.0	1.5	1.9	1.8		0.81
L ≥ 200	+	5	2.5	1.1	0.83	0.63	0.83	0.96		0.45
Intron	–	3	2.0	1.2	0.59	0.67	0.60	0.36		0.29
Flank	&	53	4.5	2.1	1.2	0.97	1.2	0.99		0.47

Notes: Mol percentages are underlined when percentage differences between limited-length and all-length values are ≥33% (solid blue) or ≤–33% (dashed red), except when limited-length mol percentages are ≤0.5%.

Table 4-S9. (continued)

GGG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
L ≤ 100	+ 0	5.4	2.0	2.3	1.6		2.0		
Intron	– 0	5.2	2.6	2.5	1.6		2.2		
Overall	& 0	8.6	2.8	2.5	1.7		2.2		
L < 200	+ 0	4.7	1.8	1.8	1.3	1.4	1.6		0.81
Intron	– 0	4.6	2.4	2.0	1.4	1.6	1.8		1.2
Overall	& 0	8.0	2.6	2.2	1.5	1.8	1.8		1.1
L ≥ 200	+ 0	1.8	0.84	0.66	0.64	0.70	0.63		0.48
Intron	– 0	1.8	1.1	0.72	0.59	0.55	0.58		0.45
Overall	& 0	3.6	1.7	1.2	0.93	1.1	0.88		0.60

Table 4-S9. (continued)

GGG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
L ≤ 100	+ 3	4.9	1.5	1.7	1.4		1.8		
Intron	– 5	4.5	2.3	2.4	1.5		2.0		
Flank	& 35	7.6	2.3	2.2	1.6		2.0		
L < 200	+ 3	4.4	1.4	1.4	1.3	1.7	1.4		0.50
Intron	– 5	4.3	2.2	2.0	1.3	1.6	1.7		1.6
Flank	& 35	7.5	2.3	2.0	1.5	1.9	1.7		1.3
L ≥ 200	+ 3	1.7	0.51	0.48	0.59	0.41	0.38		0.24
Intron	– 5	2.3	1.3	0.91	0.55	0.46	0.83		0.57
Flank	& 35	4.0	1.7	1.2	0.86	0.75	0.85		0.55

Table 4-S9. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
L ≤ 100	+ 5	4.7	1.7	1.1	1.6		1.1		
Intron	– 3	5.8	3.4	2.3	2.2		2.3		
Flank	& 53	10.	4.5	2.3	3.3		2.3		
L < 200	+ 5	5.7	1.8	1.3	1.7	1.9	1.3		0.57
Intron	– 3	5.9	3.4	2.3	2.3	2.6	2.2		0.94
Flank	& 53	11.	4.7	2.6	3.5	4.1	2.6		1.0
L ≥ 200	+ 5	4.9	3.0	2.6	2.2	2.9	2.8		0.93
Intron	– 3	4.5	3.5	2.1	2.2	2.3	1.4		0.77
Flank	& 53	9.5	6.4	4.6	4.3	5.1	4.1		1.6

Table 4-S9. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
L ≤ 100	+ 0	6.5	2.8	2.4	2.5		2.3		
Intron	– 0	6.8	3.8	2.8	2.4		2.5		
Overall	& 0	13.	5.9	3.7	4.3		3.5		
L < 200	+ 0	6.3	2.8	2.4	2.5	2.1	2.3		1.0
Intron	– 0	6.4	3.8	2.7	2.3	2.6	2.4		1.2
Overall	& 0	12.	5.9	3.8	4.3	4.3	3.5		1.4
L ≥ 200	+ 0	4.5	2.7	2.4	2.5	2.6	2.2		0.86
Intron	– 0	4.5	3.3	2.4	2.0	2.3	1.7		0.93
Overall	& 0	9.0	5.9	4.8	4.5	4.9	3.9		1.7

Table 4-S9. (continued)

GG		*Hsa*	*Dme*	*Cel*	*Ath*	*Sce*	*Spo*	*Ecu*	*Pfa*
L ≤ 100	+ 3	4.7	1.7	1.1	1.6	1.6	1.1		0.64
Intron	– 5	5.8	3.4	2.3	2.2	2.5	2.3		1.5
Flank	& 35	10.	4.5	2.3	3.3	3.6	2.3		1.4
L < 200	+ 3	5.2	1.8	1.2	1.7	1.9	1.2		0.48
Intron	– 5	6.1	3.4	2.4	2.2	2.5	2.2		1.3
Flank	& 35	11.	4.6	2.6	3.5	4.0	2.5		1.1
L ≥ 200	+ 3	3.9	1.9	1.8	2.3	1.6	1.5		0.43
Intron	– 5	4.9	3.7	2.6	1.9	2.2	2.1		1.1
Flank	& 35	8.8	5.6	4.4	4.2	3.8	3.6		1.4

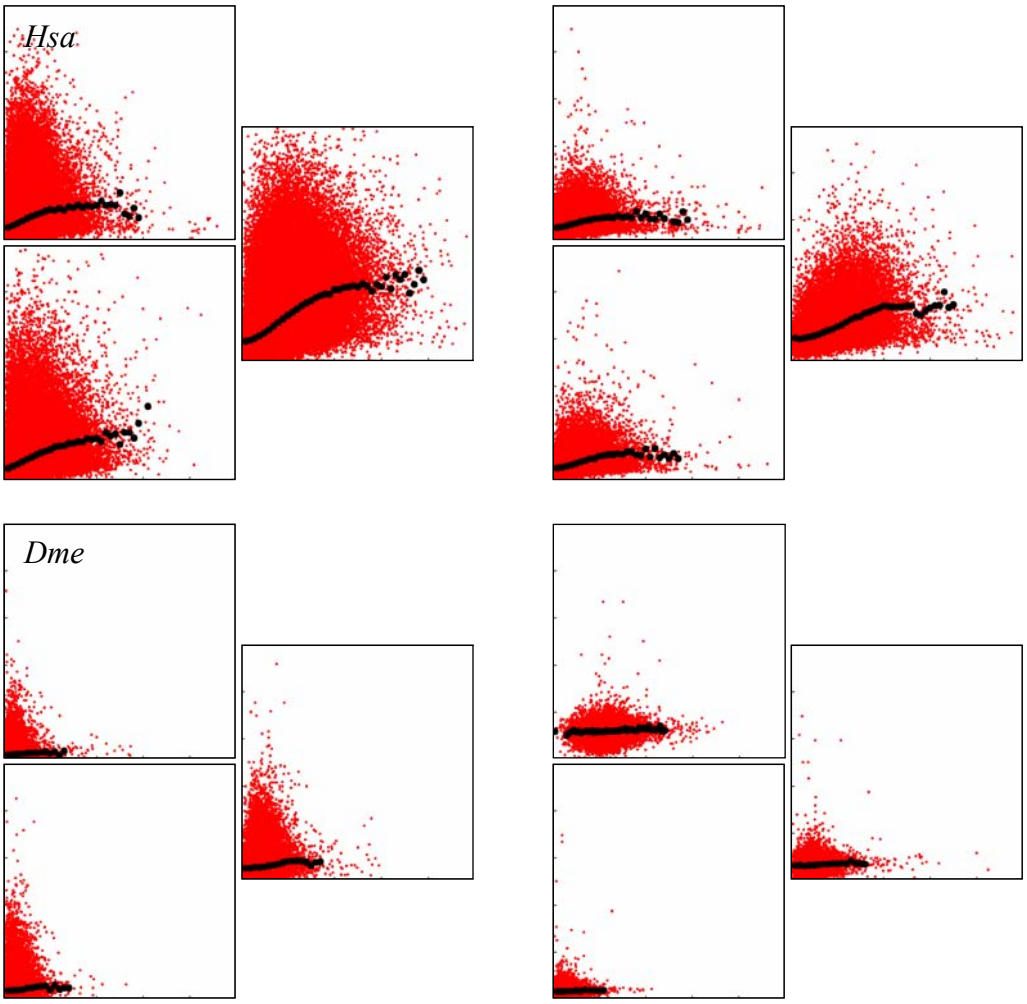


Figure 4-S1. Scatter plots of G**G**G percentages of individual (red dots) and mean (black circles) introns (left) or IGD (right) vs. adjacent exons. Notes: Charts in each triplet show coding (top), template (bottom) and combined (middle) strands. Vertical (intron or IGD) and horizontal (exon) full scales are 25% for G**G**G.

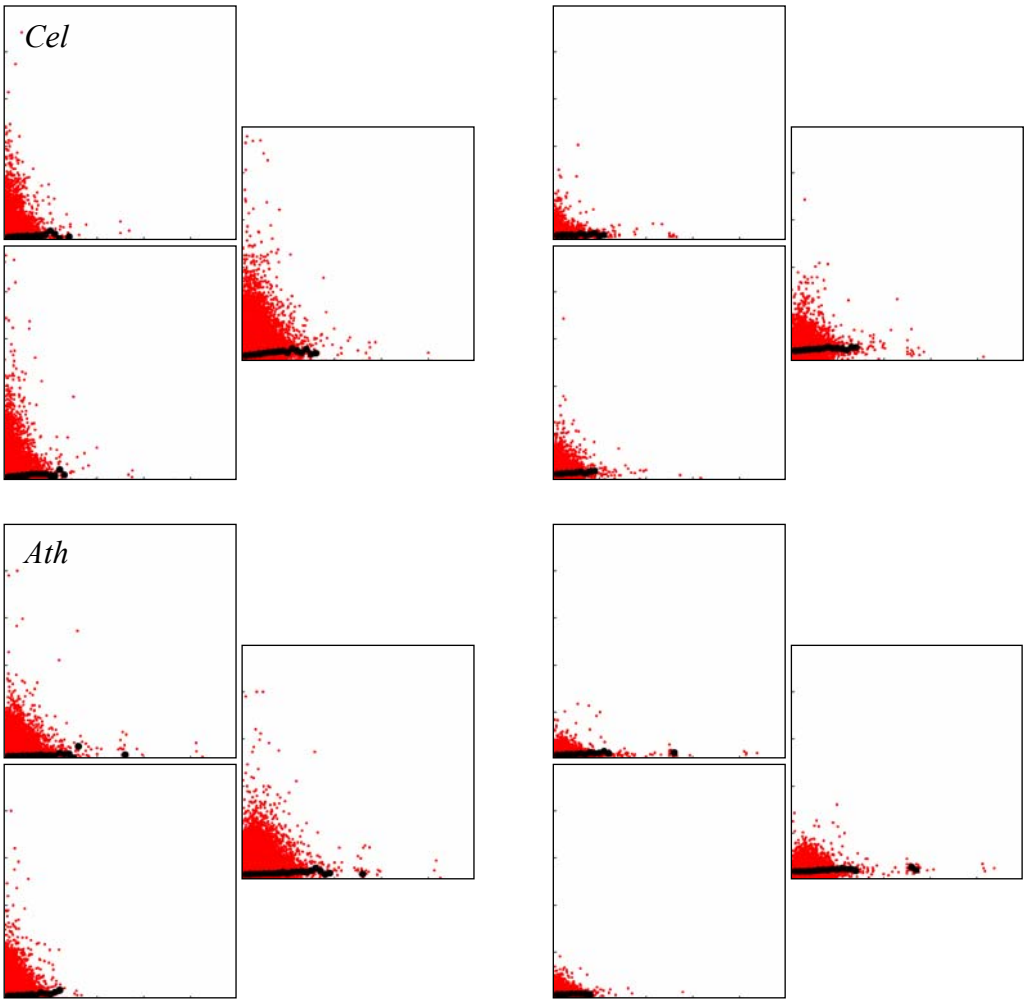


Figure 4-S1. (continued)

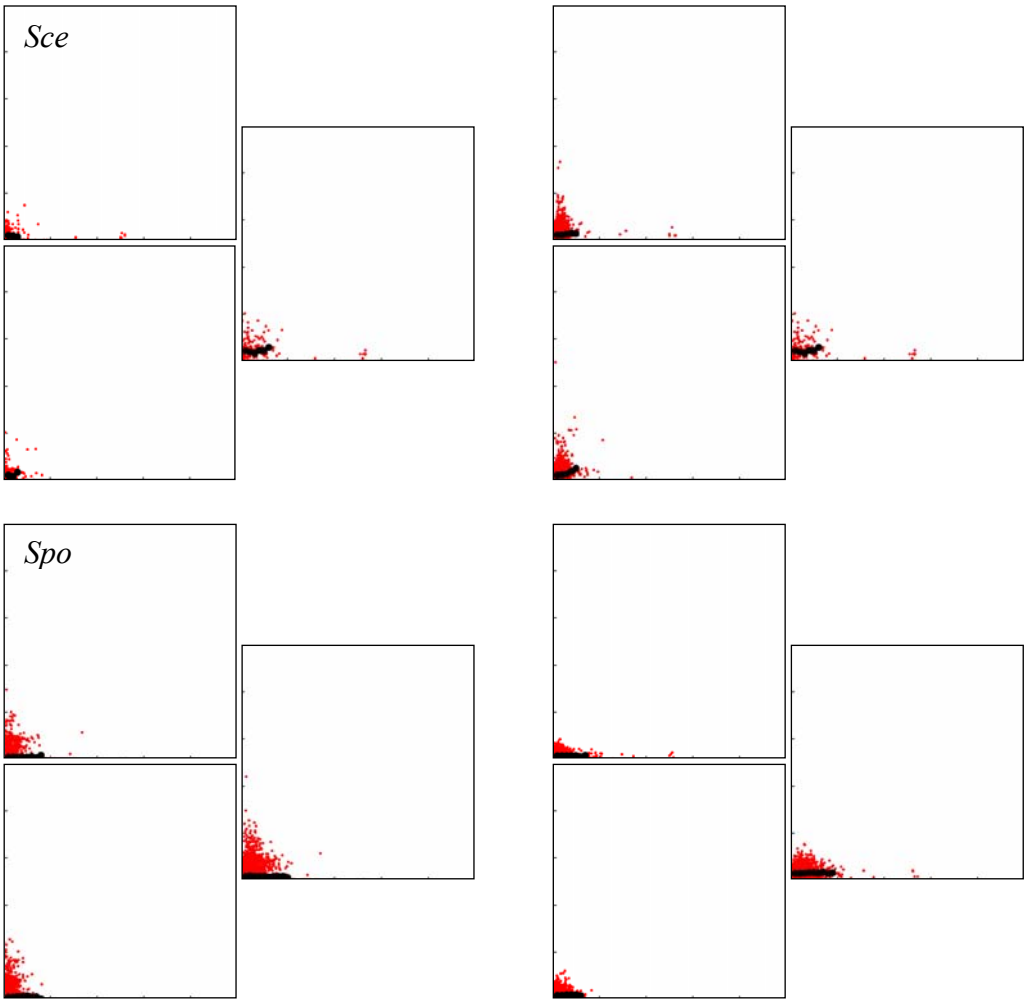


Figure 4-S1. (continued)

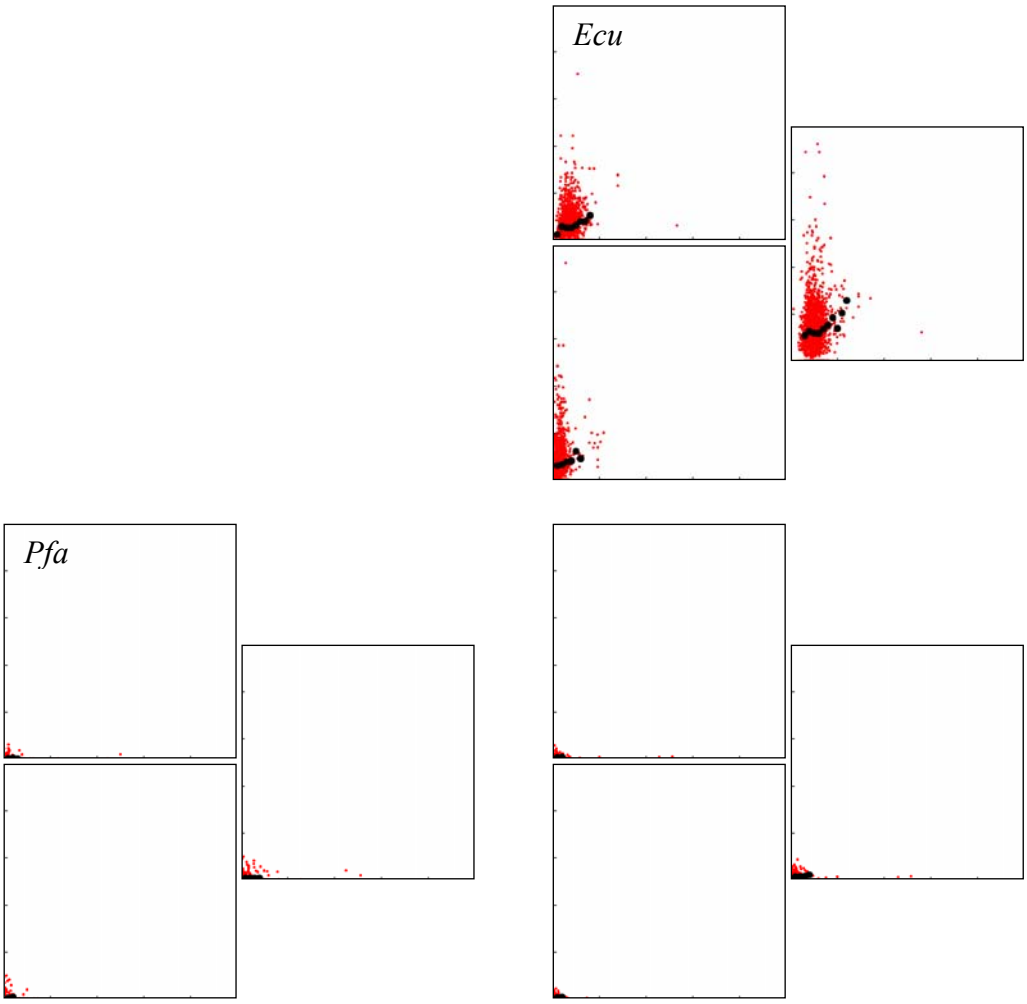


Figure 4-S1. (continued)

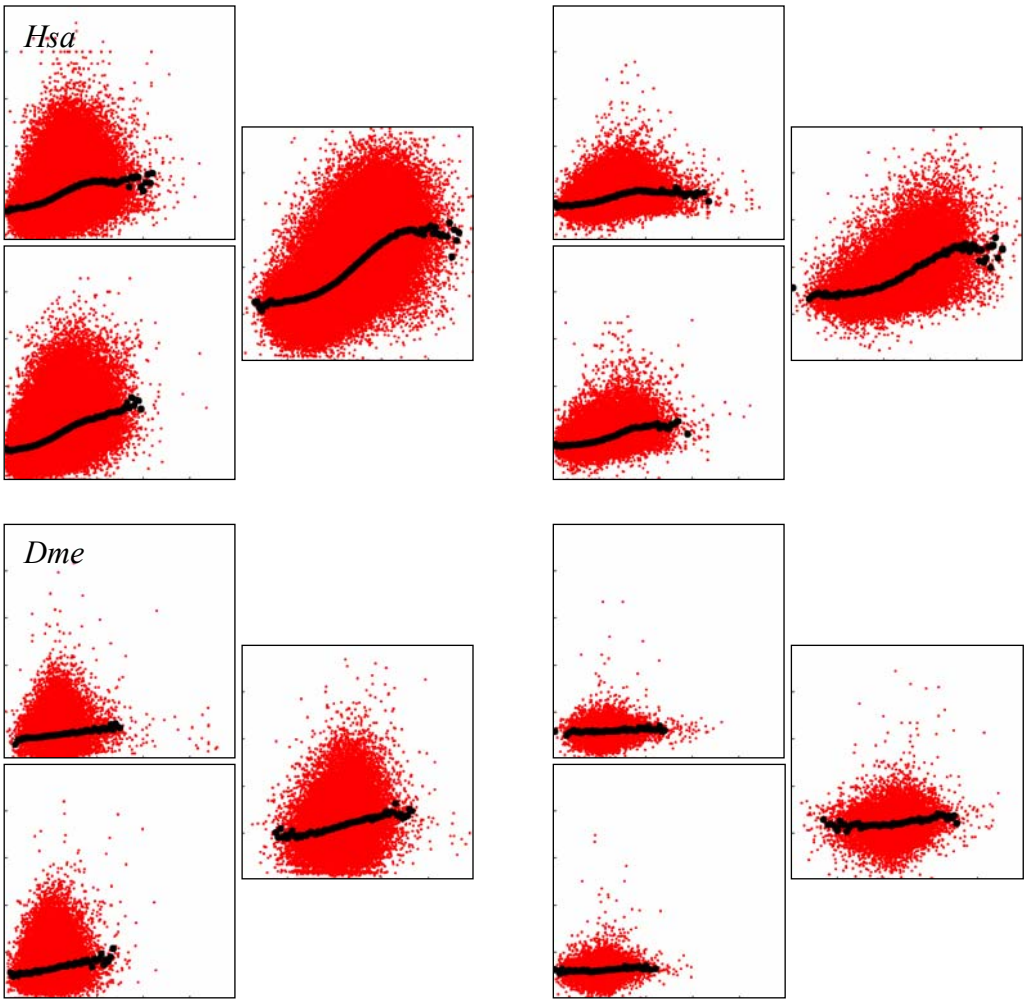


Figure 4-S2. Scatter plots of **G**G percentages of individual and mean introns or IGD vs. adjacent exons. Notes: Charts are organized, color-coded and scaled as in Figure 4-S1.

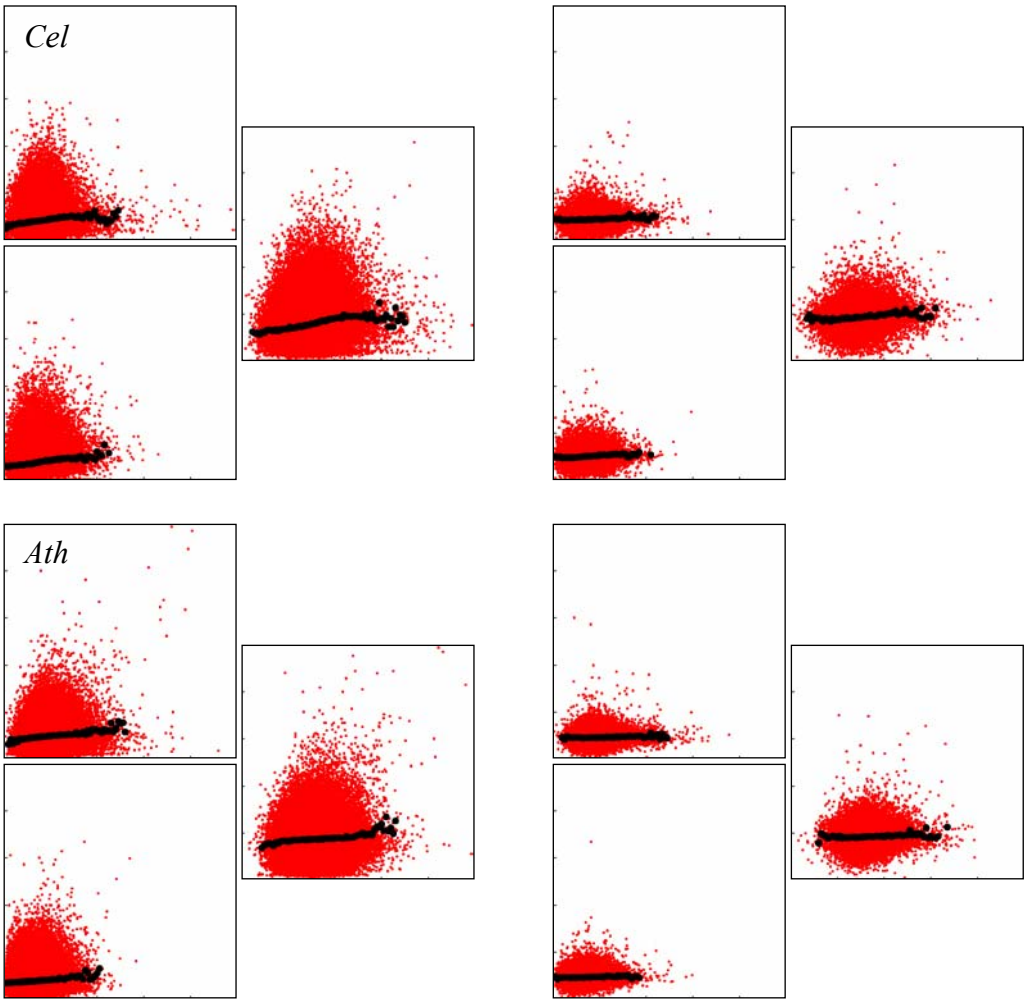


Figure 4-S2. (continued)

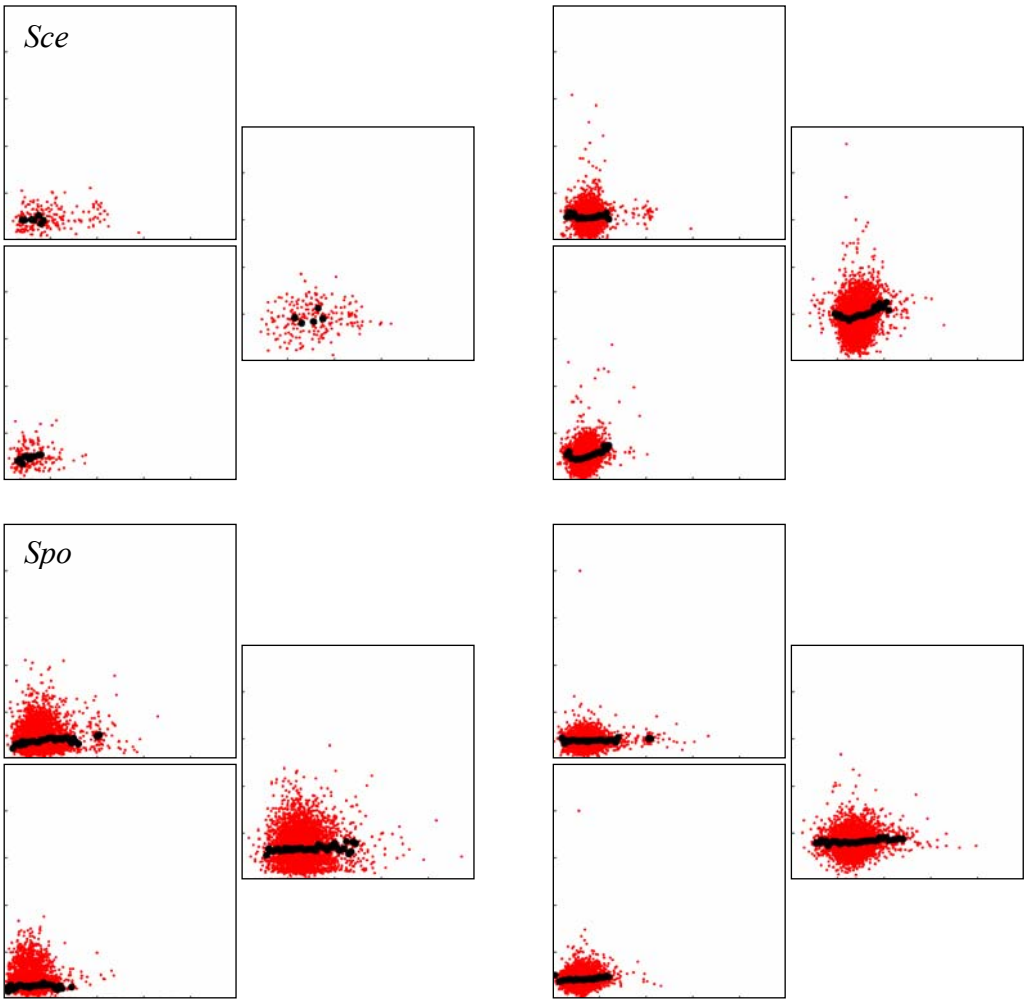


Figure 4-S2. (continued)

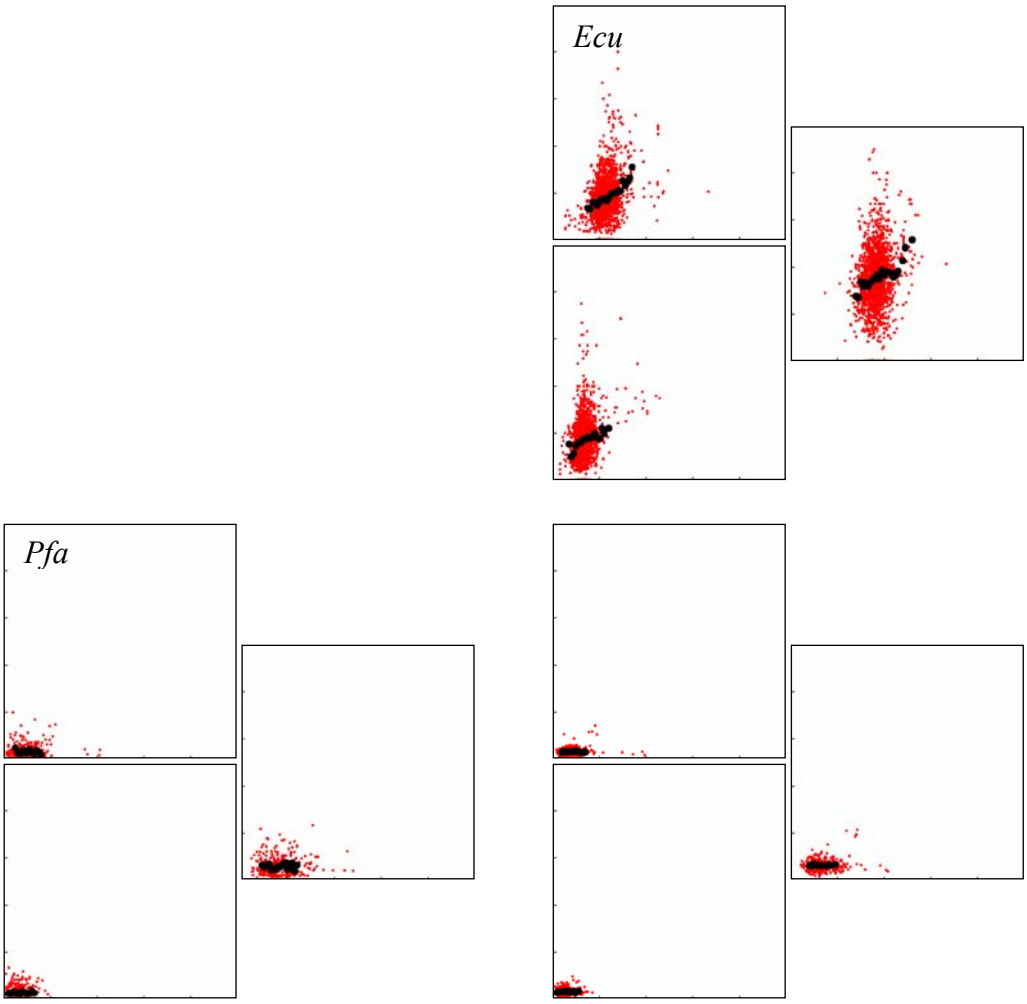


Figure 4-S2. (continued)

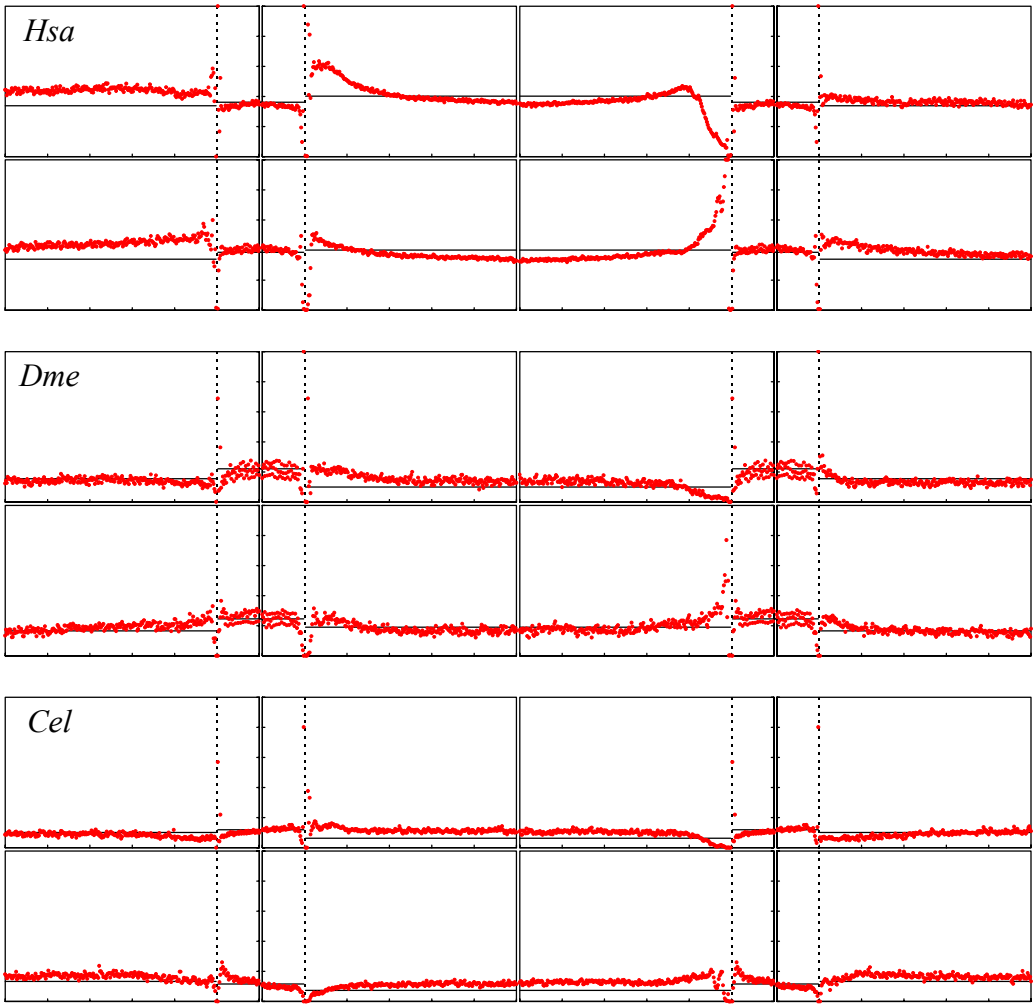


Figure 4-S3. Mean percentages of G**G**G vs. distance from exon, intron or IGD boundaries (red dots) and overall mean percentages of G**G**G (black lines). Introns and IGD longer than 100 bp are shown. Notes: The four charts in the top row of each pair of rows show, from left to right, the ends of the 3' IGD and 5' exon, the 3' exon and 5' intron, the 3' intron and 5' exon, and the 3' exon and 5' IGD on the coding strand.

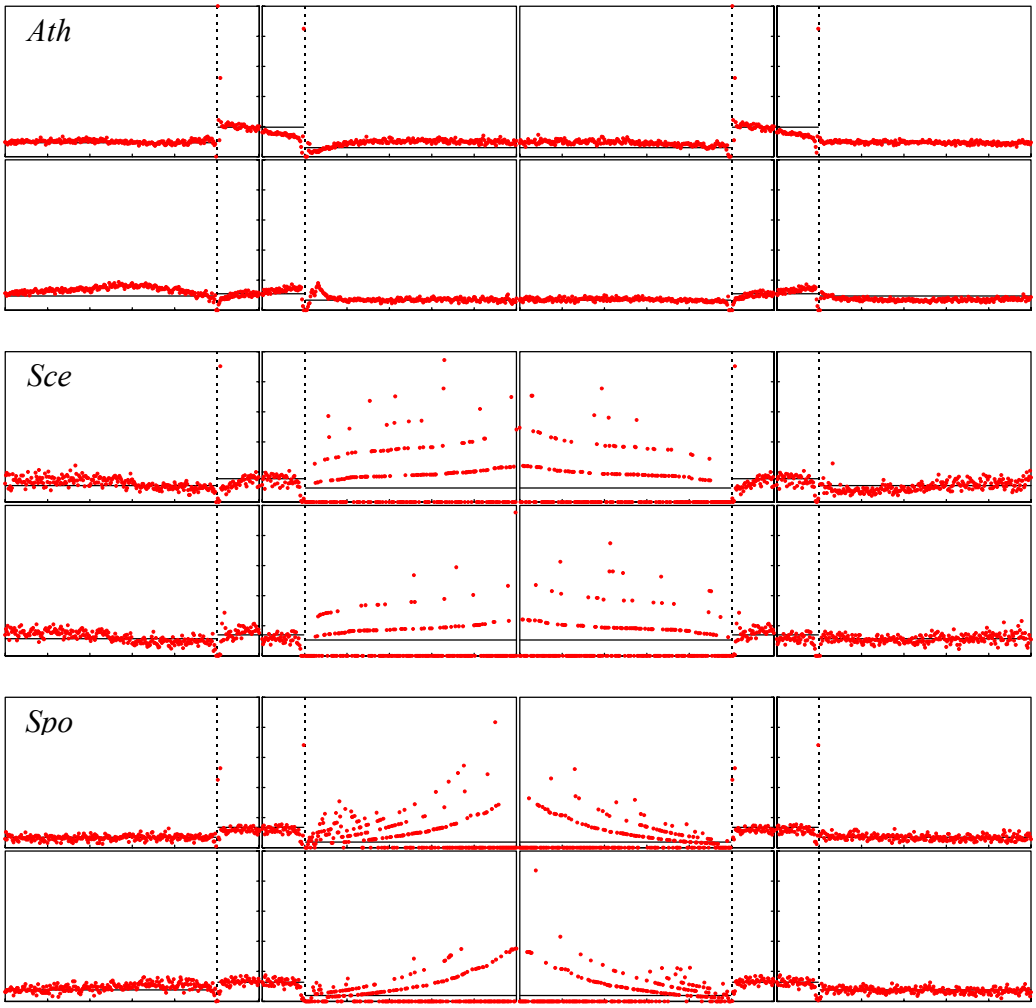


Figure 4-S3. (continued) Dashed vertical lines separate introns and IGD from exons. Each pair of rows shows the coding (top) and template strands (bottom). Vertical full scale is 5% for G**G**G. Horizontal full scale is 300 bp on all charts.

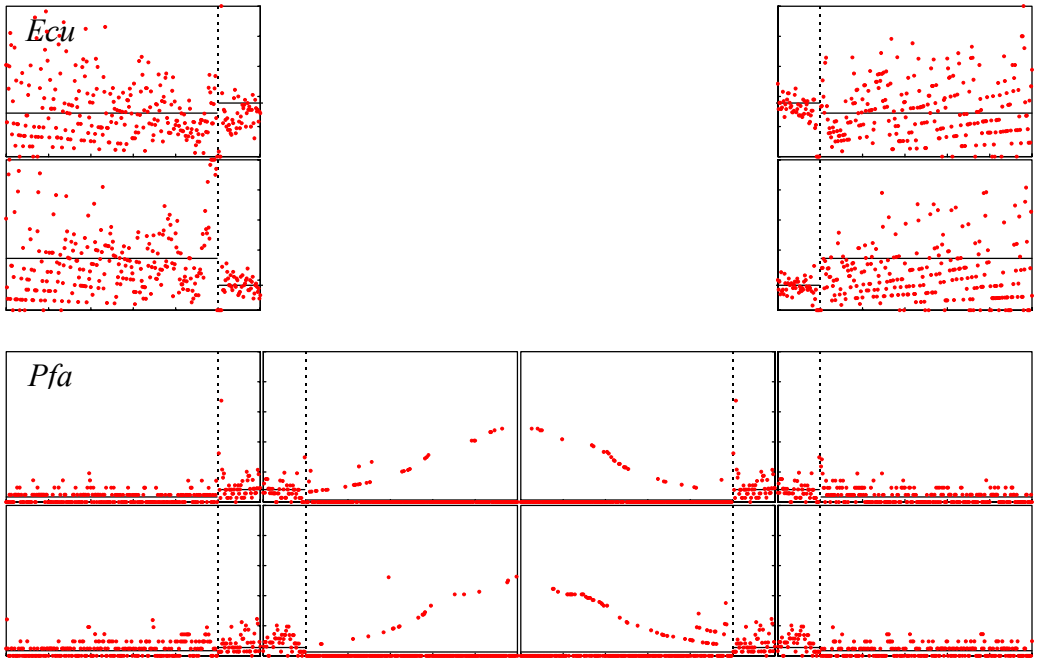


Figure 4-S3. (continued)

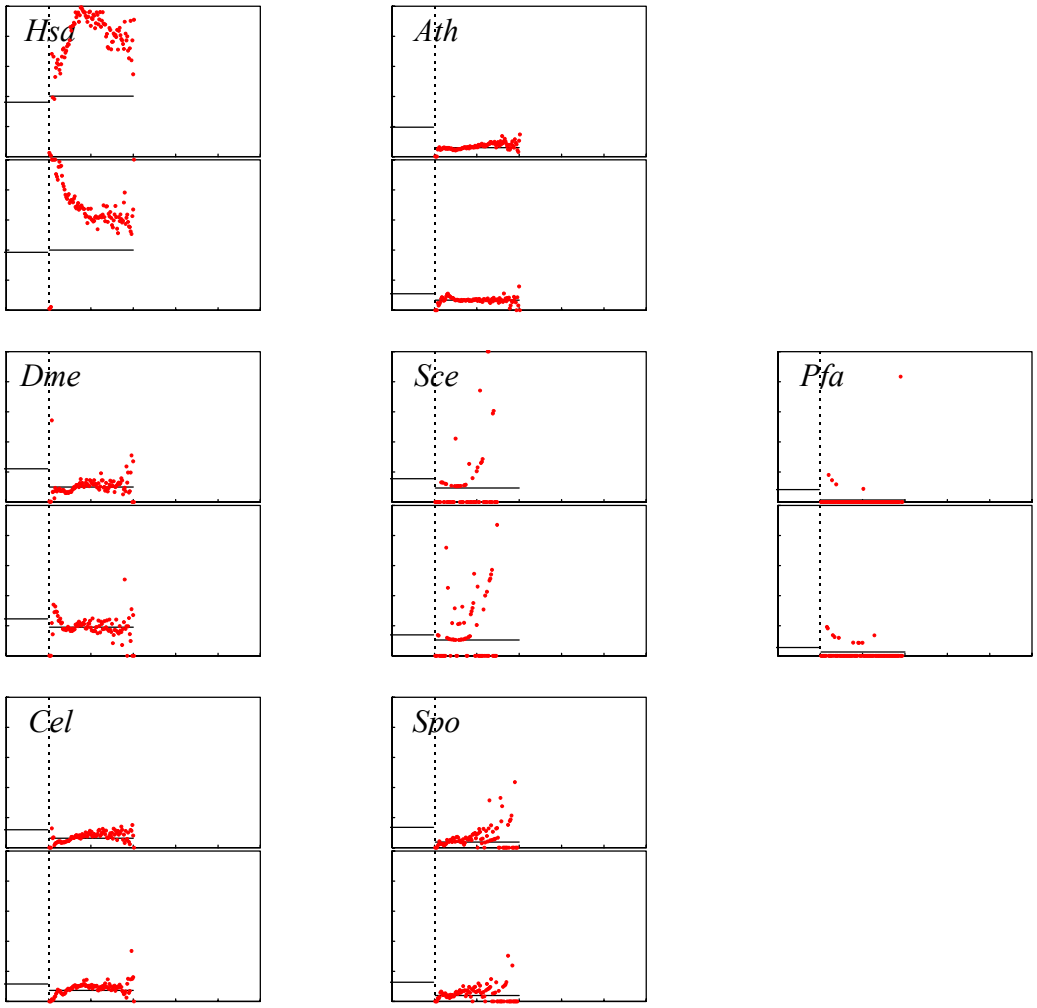


Figure 4-S4. Mean percentages of G**G**G in short (≤100 bp) introns vs. distance from exon/intron boundaries (red dots) and overall mean percentages of G**G**G (black lines). Notes: Each pair of charts shows the coding (top) and template (bottom) strands, without distinguishing the 5' and 3, intron ends. Charts are scaled as in Figure 4-S3.

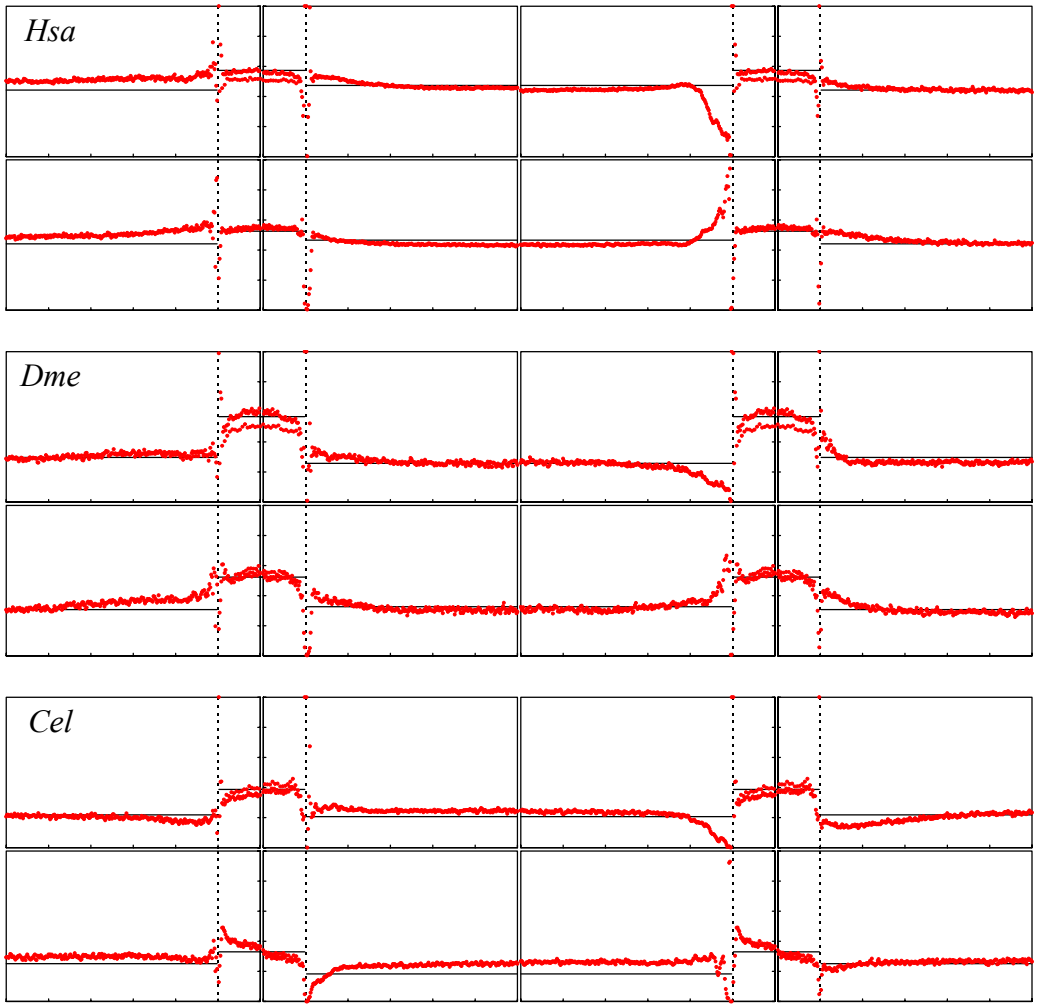


Figure 4-S5. Mean percentages of **G**G vs. distance from exon, intron or IGD boundaries and overall mean percentages of **G**G. Introns and IGD longer than 100 bp are shown. Notes: Charts are organized and color-coded as in Figure 4-S3. Vertical full scale is 10% for **G**G. Horizontal full scale is 300 bp on all charts.

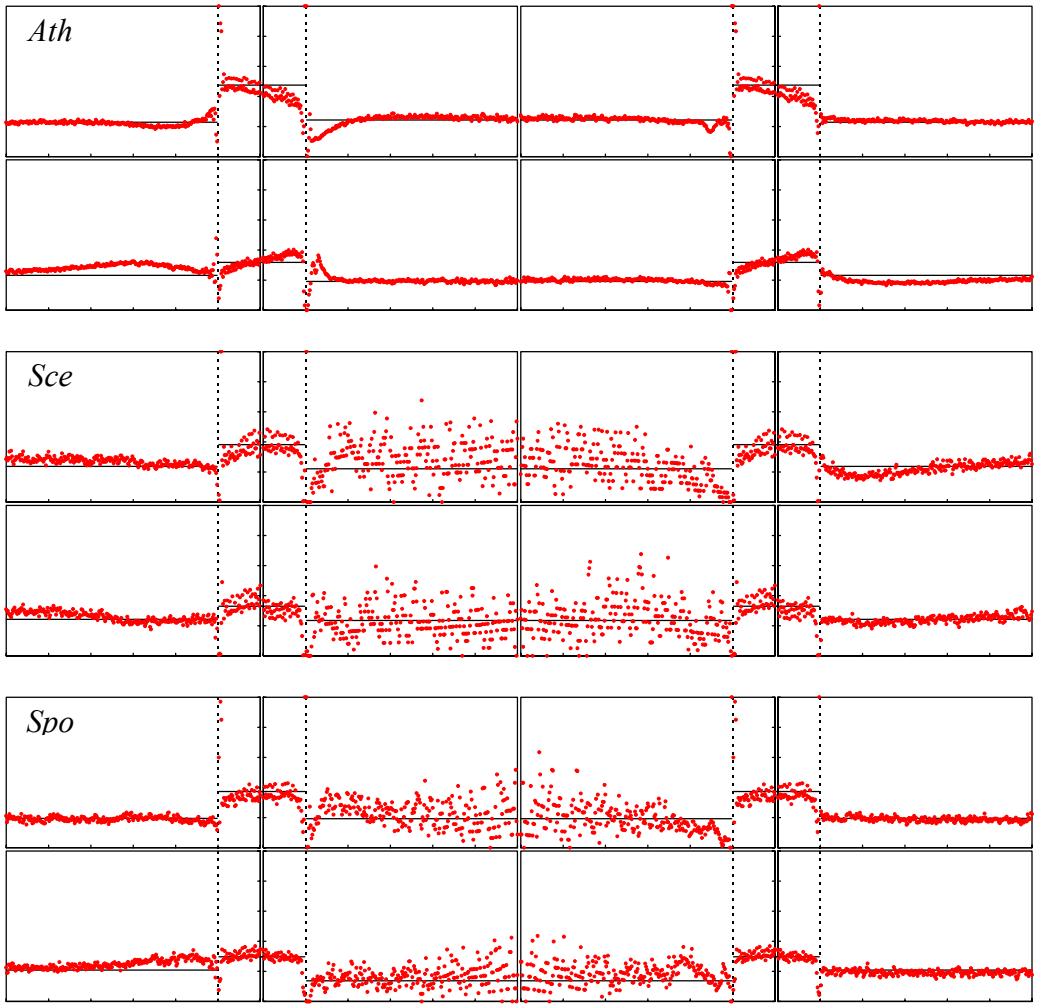


Figure 4-S5. (continued)

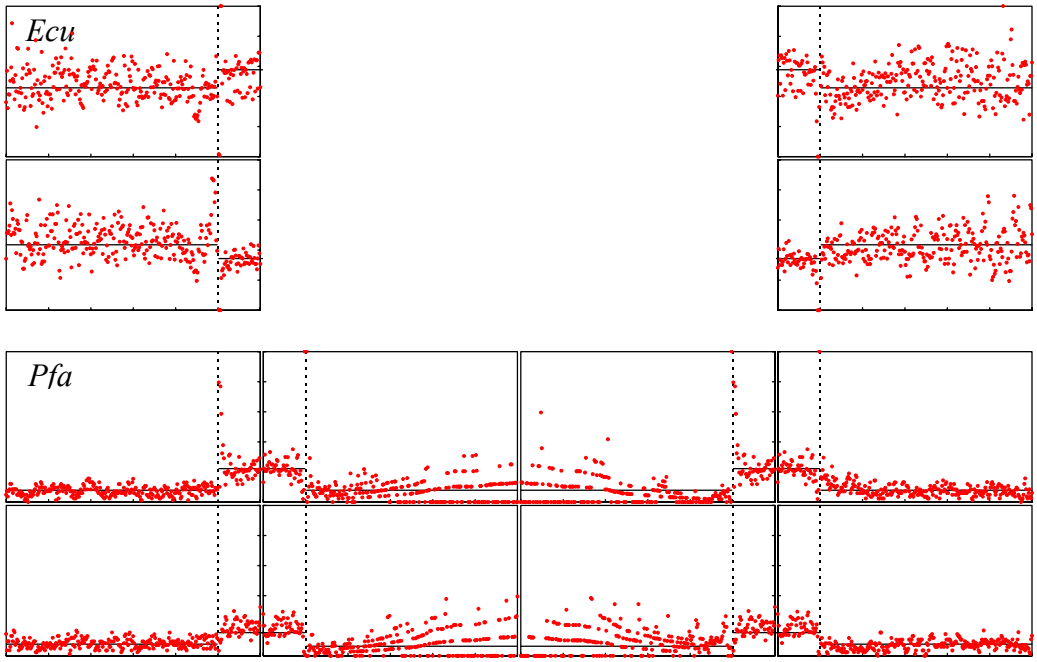


Figure 4-S5. (continued)

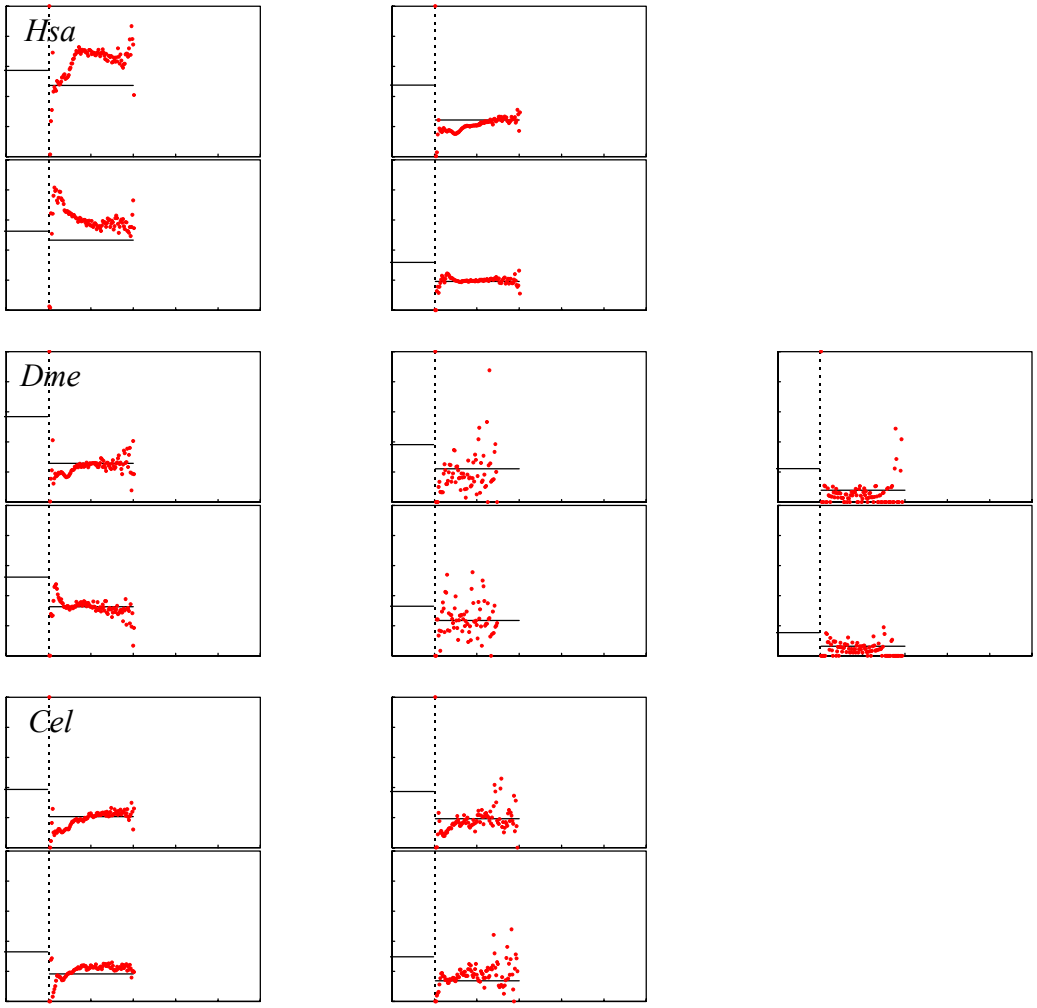


Figure 4-S6. Mean mol percentages of **G**G in short (≤100 nt) introns vs. distance from exon/intron boundaries and overall mean mol percentages of **G**G. Notes: Charts are organized and color-coded as in Figure 4-S4. Charts are scaled as in Figure 4-S5.

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., et al. (2000) *Science* **287,** 2185-2195.

Alberts, B., Bray, D., Lewis, J., et al. (1994) *Molecular Biology of the Cell* (Garland, New York).

Alberts, B., Johnson, A., Lewis, J., et al. (2002) *Molecular Biology of the Cell* (Garland Science, New York).

Amatore, C., Arbault, S., Bruce, D., et al. (2000) *Faraday Discuss.* **116,** 319-333.

Ambrosone, C. B. (2000) *Antioxidants & Redox Signaling* **2,** 903-917.

Arabidopsis Genome Initiative,. (2000) *Nature* **408,** 796-815.

Arkin, M. R., Stemp, E. D. A., Pulver, S. C., et al. (1997) *Chem. Biol.* **4,** 389-400.

Askeland, D. R. & Editor (1996) *The Science and Engineering of Materials*.

Baltimore, D. (2001) *Nature* **409,** 814-816.

Bard, A. J. & Faulkner, L. R. (2001) *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York).

Beckman, K. B. & Ames, B. N. (1997) *J. Biol. Chem.* **272,** 19633-19636.

Beckman, K. B. & Ames, B. N. (1998) *Physiological Rev.* **78,** 547-581.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., et al. (2002) *Nucleic Acids Res.* **30,** 17-20.

Bernardi, G. (2000) *Gene* **259,** 31-43.

Bernardi, G. (2000) *Gene* **241,** 3-17.

Bixon, M., Giese, B., Wessely, S., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11713-11716.

Bixon, M. & Jortner, J. (2001) *J. Am. Chem. Soc.* **123,** 12556-12567.

Bohr, V. A. & Anson, R. M. (1995) *Mutat. Res.* **338,** 25-34.

Boiteux, S. & Laval, J. (1997) *Base Excision Repair DNA Damage***,** 31-44.

Boon, E. M. & Barton, J. K. (2002) *Curr. Opin. Structural Biology* **12,** 320-329.

Boone, E. & Schuster, G. B. (2002) *Nucleic Acids Res.* **30,** 830-837.

Bowman, S., Lawson, D., Basham, D., et al. (1999) *Nature* **400,** 532-538.

Bracha, R. & Mirelman, D. (1984) *J. Exp. Med.* **160,** 353-68.

Braun, E., Eichen, Y., Sivan, U., et al. (1998) *Nature* **391,** 775-778.

Brendel, V., Kleffe, J., Carle-Urioste, J. C., et al. (1998) *J. Mol. Biol.* **276,** 85-104.

Brett, C. M. A., Brett, A. M. O. & Serrano, S. H. P. (1994) *J. Electroanal. Chem.* **366,** 225-31.

Brudno, M., Gelfand, M. S., Spengler, S., et al. (2001) *Nucleic Acids Res.* **29,** 2338-2348.

Burset, M., Seledtsov, I. A. & Solovyev, V. V. (2000) *Nucleic Acids Res.* **28,** 4364-4375.

Burset, M., Seledtsov, I. A. & Solovyev, V. V. (2001) *Nucleic Acids Res.* **29,** 255-259.

Caenorhabditis elegans Sequencing Consortium, (1998) *Science* **282,** 2012-2018.

Carlo, T., Sierra, R. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 3988-3995.

Carlo, T., Sterner, D. A. & Berget, S. M. (1996) *RNA* **2,** 342-353.

Cartegni, L., Chew, S. L. & Krainer, A. R. (2002) *Nature Rev. Genetics* **3,** 285-298.

Chatterjee, A. & Holley, W. R. (1993) *Adv Radiat Biol* **17,** 181-226.

Conklin, K. A. (2000) *Nutrition and Cancer* **37,** 1-18.

Cooper, D. N. (2000) *Human Gene Evolution* (BIOS Scientific, Oxford).

Cooper, G. M. (1997) *The Cell: A Molecular Approach*.

Cooper, T. A. & Mattox, W. (1997) *Am. J. Hum. Genet.* **61,** 259-266.

Croft, L., Schandorff, S., Clark, F., et al. (2000) *Nature Genetics* **24,** 340-341.

Cross, S. H. & Bird, A. P. (1995) *Curr. Opin. Genet. Dev.* **5,** 309-14.

Croteau, D. L. & Bohr, V. A. (1997) *J. Biol. Chem.* **272,** 25409-25412.

Davidson, N. R. (1962) *Statistical Mechanics* (McGraw-Hill, New York).

Dringen, R. (2000) *Progress in Neurobiology* **62,** 649-671.

Engelbrecht, J., Knudsen, S. & Brunak, S. (1992) *J. Mol. Biol.* **227,** 108-13.

Eyre-Walker, A. & Hurst, L. D. (2001) *Nature Rev. Genetics* **2,** 549-555.

Faraggi, M., Broitman, F., Trent, J. B., et al. (1996) *J. Phys. Chem.* **100,** 14751-14761.

Fink, H.-W. & Schonenberger, C. (1999) *Nature* **398,** 407-410.

Finkel, T. & Holbrook, N. J. (2000) *Nature* **408,** 239-247.

Francino, M. P., Chao, L., Riley, M. A., et al. (1996) *Science* **272,** 107-9.

Francino, M. P. & Ochman, H. (1997) *Trends Genetics* **13,** 240-245.

Francino, M. P. & Ochman, H. (2000) *Mol. Biol. Evol.* **17,** 416-422.

Francis, S. E., Sullivan, D. J., Jr. & Goldberg, D. E. (1997) *Annu. Rev. Microbiology* **51,** 97-123.

Frank, A. C. & Lobry, J. R. (1999) *Gene* **238,** 65-77.

Friedman, K. A. & Heller, A. (2001) *J. Phys. Chem. B* **105,** 11859-11865.

Fukui, K. & Tanaka, K. (1998) *Angew. Chem., Int. Ed.* **37,** 158-161.

Garcia, M. X. U. (2000) (Univ. of Missouri, Columbia), pp. 226.

Gardner, M. J., Tettelin, H., Carucci, D. J., et al. (1998) *Science* **282,** 1126-1132.

Gasper, S. M., Armitage, B., Shui, X., et al. (1998) *J. Am. Chem. Soc.* **120,** 12402-12409.

Gasper, S. M. & Schuster, G. B. (1997) *J. Am. Chem. Soc.* **119,** 12762-12771.

Giese, B. (2000) *Chemistry in Britain* **36,** 44-46.

Giese, B. (2002) *Annu. Rev. Biochemistry* **71,** 51-70.

Giese, B., Wessely, S., Spormann, M., et al. (1999) *Angew. Chem., Int. Ed.* **38,** 996-998.

Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996) *Science* **274,** 546, 563-567.

Hall, D. B. & Barton, J. K. (1997) *J. Am. Chem. Soc.* **119,** 5045-5046.

Hall, D. B., Holmlin, R. E. & Barton, J. K. (1996) *Nature* **382,** 731-735.

Hall, D. B., Kelley, S. O. & Barton, J. K. (1998) *Biochemistry* **37,** 15933-15940.

Hanawalt, P. C. (2001) *Mutat. Res.* **485,** 3-13.

Hardin, C. C., Henderson, E., Watson, T., et al. (1991) *Biochemistry* **30,** 4460-72.

Hartwich, G., Caruana, D. J., de Lumley-Woodyear, T., et al. (1999) *J. Am. Chem. Soc.* **121,** 10803-10812.

Hayes, J. J., Tullius, T. D. & Wolffe, A. P. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87,** 7405-9.

Helbock, H. J., Beckman, K. B., Shigenaga, M. K., et al. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 288-293.

Heller, A. (2000) *Faraday Discuss.* **116,** 1-13.

Henderson, P. T., Armitage, B. & Schuster, G. B. (1998) *Biochemistry* **37,** 2991-3000.

Henderson, P. T., Jones, D., Hampikian, G., et al. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 8353-8358.

Henle, E. S., Han, Z., Tang, N., et al. (1999) *J. Biol. Chem.* **274,** 962-971.

Henle, E. S. & Linn, S. (1997) *J. Biol. Chem.* **272,** 19095-19098.

Hickerson, R. P., Prat, F., Muller, J. G., et al. (1999) *J. Am. Chem. Soc.* **121,** 9423-9428.

Higami, Y. & Shimokawa, I. (2000) *Cell & Tissue Res.* **301,** 125-132.

Hogan, M. E., Rooney, T. F. & Austin, R. H. (1987) *Nature* **328,** 554-7.

Holmlin, R. E., Dandliker, P. J. & Barton, J. K. (1998) *Angew. Chem., Int. Ed.* **36,** 2715-2730.

Hutter, M. & Clark, T. (1996) *J. Am. Chem. Soc.* **118,** 7574-7577.

International Human Genome Collaborators, (2001) (National Center for Biotechnology Information, Washington, DC).

International Human Genome Sequencing Consortium, (2001) *Nature* **409,** 860-921.

Jabbari, K. & Bernardi, G. (2000) *Gene* **247,** 287-292.

Jackson, A. L. & Loeb, L. A. (2001) *Mutat. Res.* **477,** 7-21.

Jermiin, L. S., Foster, P. G., Graur, D., et al. (1996) *J. Mol. Evol.* **42,** 476-480.

Jermiin, L. S., Graur, D., Lowe, R. M., et al. (1994) *J. Mol. Evol.* **39,** 160-73.

Jortner, J., Bixon, M., Voityuk, A. A., et al. (2002) *J. Phys. Chem. A* **106,** 7599-7606.

Kan, Y. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 10857-10864.

Kan, Y. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 11607-11614.

Kanvah, S. & Schuster, G. B. (2002) *J. Am. Chem. Soc.* **124,** 11286-11287.

Karlin, S., Campbell, A. M. & Mrazek, J. (1998) *Annu. Rev. Genetics* **32,** 185-225.

Karlin, S. & Marazek, J. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94,** 10227-10232.

Karlin, S. & Mrazek, J. (1996) *J. Mol. Biol.* **262,** 459-472.

Karlin, S. & Mrazek, J. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 5240-5245.

Kasumov, A. Y., Kociak, M., Gueron, S., et al. (2001) *Science* **291,** 280-282.

Katinka, M. D., Duprat, S., Cornillot, E., et al. (2001) *Nature* **414,** 450-453.

Kawanishi, S., Hiraku, Y., Murata, M., et al. (2002) *Free Radical Biology & Medicine* **32,** 822-832.

Kawanishi, S., Hiraku, Y. & Oikawa, S. (2001) *Mutat. Res.* **488,** 65-76.

Kawanishi, S., Oikawa, S. & Hiraku, Y. (2000) *Free Radicals in Chemistry, Biology and Medicine***,** 85-91.

Kawanishi, S., Oikawa, S., Murata, M., et al. (1999) *Biochemistry* **38,** 16733-9.

Krawczak, M., Ball, E. V. & Cooper, D. N. (1998) *Am. J. Hum. Genet.* **63,** 474-488.

Krawczak, M. & Cooper, D. N. (1997) *Trends Genetics* **13,** 121-122.

Laviron, E. (1979) *J. Electroanal. Chem.* **101,** 19-28.

Le Page, F., Kwoh, E. E., Avrutskaya, A., et al. (2000) *Cell* **101,** 159-171.

Leforestier, A. & Livolant, F. (1993) *Biophys. J.* **65,** 56-*72.

Lewis, F. D., Liu, X., Liu, J., et al. (2000) *J. Am. Chem. Soc.* **122,** 12037-12038.

Lim, L. P. & Burge, C. B. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11193-11198.

Livolant, F. (1991) *Physica A* **176,** 117-37.

Lodish, H., Berk, A., Zipursky, S. L., et al. (1999) *Molecular Cell Biology* (W.H. Freeman & Co., New York).

Luo, Y., Han, Z., Chin, S. M., et al. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 12438-42.

Ly, D., Sanii, L. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 9400-9410.

Makarova, K. S., Aravind, L., Wolf, Y. I., et al. (2001) *Microbiol. Mol. Biol. Rev.* **65,** 44-79.

Maki, H. & Sekiguchi, M. (1992) *Nature* **355,** 273-5.

May, J. M., Qu, Z.-C., Xia, L., et al. (2000) *Am. J. Physiol.* **279,** C1946-C1954.

McCullough, A. J. & Berget, S. M. (1997) *Mol. Cell. Biol.* **17,** 4562-4571.

McCullough, A. J. & Berget, S. M. (2000) *Mol. Cell. Biol.* **20,** 9225-9235.

McCullough, A. J. & Schuler, M. A. (1997) *Nucleic Acids Res.* **25,** 1071-1077.

McEachern, M. J., Krauskopf, A. & Blackburn, E. H. (2000) *Annu. Rev. Genetics* **34,** 331-358.

Meggers, E. & Giese, B. (1999) *Nucleosides Nucleotides* **18,** 1317-1318.

Meggers, E., Kusch, D., Spichty, M., et al. (1998) *Angew. Chem., Int. Ed.* **37,** 460-462.

Meggers, E., Michel-Beyerle, M. E. & Giese, B. (1998) *J. Am. Chem. Soc.* **120,** 12950-12955.

Moser, C. C., Keske, J. M., Warncke, K., et al. (1992) *Nature* **355,** 796-802.

Murphy, C. J., Arkin, M. R., Ghatlia, N. D., et al. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 5315-19.

Murphy, C. J., Arkin, M. R., Jenkins, Y., et al. (1993) *Science* **262,** 1025-9.

Musto, H., Romero, H., Zavala, A., et al. (1999) in *J. Mol. Evol.*, Vol. 49, pp. 27-35.

Nakamura, Y., Gojobori, T. & Ikemura, T. (2000) *Nucleic Acids Res.* **28,** 292.

Narumi, K., Kikuchi, M., Funayama, T., et al. (1999) *Hoshasen Seibutsu Kenkyu* **34,** 401-418.

Nekrutenko, A. & Li, W.-H. (2000) *Genome Res.* **10,** 1986-1995.

Newcomb, T. G. & Loeb, L. A. (1998) *DNA Damage and Repair* **1,** 65-84.

Nicholson, R. S. (1965) *Anal. Chem.* **37,** 1351-5.

Noctor, G., Arisi, A.-C. M., Jouanin, L., et al. (1998) *J. Exp. Botany* **49,** 623-647.

Nouspikel, T. & Hanawalt, P. C. (2002) *DNA Repair* **1,** 59-75.

Nunez, M. E., Hall, D. B. & Barton, J. K. (1999) *Chem. Biol.* **6,** 85-97.

Nunez, M. E., Holmquist, G. P. & Barton, J. K. (2001) *Biochemistry* **40,** 12465-12471.

Nunez, M. E., Noyes, K. T. & Barton, J. K. (2002) *Chemistry & Biology* **9,** 403-415.

Nunez, M. E., Noyes, K. T., Gianolio, D. A., et al. (2000) *Biochemistry* **39,** 6190-6199.

Nunez, M. E., Rajski, S. R. & Barton, J. K. (2000) *Methods Enzymol.* **319,** 165-188.

Nussinov, R. (1989) *J. Biomol. Struct. Dyn.* **6,** 985-1000.

Oikawa, S., Tada-Oikawa, S. & Kawanishi, S. (2001) *Biochemistry* **40,** 4763-4768.

Okahata, Y., Kobayashi, T., Tanaka, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 6165-6166.

Olasky, S. J. (1995) (http://www.alberts.com/authorpages/00013288/prod_132.htm), 2003.

Oliveira-Brett, A. M., Vivan, M., Fernandes, I. R., et al. (2002) *Talanta* **56,** 959-970.

Oliver, J. L., Bernaola-Galvan, P., Carpena, P., et al. (2001) *Gene* **276,** 47-56.

O'Neill, P., Parker, A. W., Plumb, M. A., et al. (2001) *J. Phys. Chem. B* **105,** 5283-5290.

Porath, D., Bezryadin, A., De Vries, S., et al. (2000) *Nature* **403,** 635-638.

Pothukuchy, A., Mano, N., Salazar, M., et al. (2002) *submitted*.

Poulsen, H. E., Jensen, B. R., Weimann, A., et al. (2000) *Free Radical Research* **33,** S33-S39.

Rajski, S. R. & Barton, J. K. (2000) *Proc. Conversation in Biomolecular Stereodynamics* **11,** 285-291.

Rajski, S. R., Jackson, B. A. & Barton, J. K. (2000) *Mutat. Res.* **447,** 49-72.

Reddy, A. S. N. (2001) *Crit. Rev. Plant Sciences* **20,** 523-571.

Reiter, R. J., Acuna-Castroviejo, D., Tan, D.-X., et al. (2001) *Ann. NY Acad. Sci.* **939,** 200-215.

Rodriguez, H., Valentine, M. R., Holmquist, G. P., et al. (1999) *Biochemistry* **38,** 16578-16588.

Rogers, J. E. & Kelly, L. A. (1999) *J. Am. Chem. Soc.* **121,** 3854-3861.

Rogozin, I. B. & Milanesi, L. (1997) *J. Mol. Evol.* **45,** 50-59.

Romero, H., Zavala, A. & Musto, H. (2000) *Gene* **242,** 307-311.

Saito, I., Nakamura, T., Nakatani, K., et al. (1998) *J. Am. Chem. Soc.* **120,** 12686-12687.

Saito, I., Takayama, M., Sugiyama, H., et al. (1995) *J. Am. Chem. Soc.* **117,** 6406-7.

Sanii, L. & Schuster, G. B. (2000) *J. Am. Chem. Soc.* **122,** 11545-11546.

Sartor, V., Henderson, P. T. & Schuster, G. B. (1999) *J. Am. Chem. Soc.* **121,** 11027-11033.

Schafer, F. Q. & Buettner, G. R. (2001) *Free Radical Biology & Medicine* **30,** 1191-1212.

Schuster, G. B. (2000) *Acc. Chem. Res.* **33,** 253-260.

Sekiguchi, M. & Hayakawa, H. (1998) *Contemp. Cancer Res.* **2,** 85-93.

Serrano-Luna, D. J., Negrete, E., Reyes, M., et al. (1998) *Experimental Parasitology* **89,** 71-77.

Setlow, R. B. (2001) *Mutat. Res.* **477,** 1-6.

Shackelford, J. F. (1996) *Introduction to Materials Science for Engineers* (Prentice-Hall, Upper Saddle River, NJ).

Shackelford, J. F. (2000) *Introduction to Materials Science for Engineers* (Prentice-Hall, Upper Saddle River, NJ).

Sharp, P. M. & Devine, K. M. (1989) *Nucleic Acids Res.* **17,** 5029-39.

Shibutani, S., Takeshita, M. & Grollman, A. P. (1991) *Nature* **349,** 431-4.

Sirand-Pugnet, P., Durosay, P., Brody, E., et al. (1995) *Nucleic Acids Res.* **23,** 3501-7.

Sistare, M. F., Codden, S. J., Heimlich, G., et al. (2000) *J. Am. Chem. Soc.* **122,** 4742-4749.

Smerdon, M. J. & Thoma, F. (1998) *Contemp. Cancer Res.* **2,** 199-222.

Smith, N. G. C. & Eyre-Walker, A. (2001) *Mol. Biol. Evol.* **18,** 982-986.

Solovyev, V. V., Salamov, A. A. & Lawrence, C. B. (1994) *Nucleic Acids Res.* **22,** 5156-63.

Stansbury, E. E. & Buchanan, R. A. (2000) *Fundamentals of electrochemical corrosion* (ASM International, Materials Park, OH).

Steenken, S., Jovanovic, S. V., Bietti, M., et al. (2000) *J. Am. Chem. Soc.* **122,** 2373-2374.

Strey, H. H., Podgornik, R., Rau, D. C., et al. (1998) *Curr. Opin. Structural Biology* **8,** 309-313.

Sueoka, N. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 2653-7.

Sueoka, N. (1992) *J. Mol. Evol.* **34,** 95-114.

Sueoka, N. & Kawanishi, Y. (2000) *Gene* **261,** 53-62.

Sugiyama, H. & Saito, I. (1996) *J. Am. Chem. Soc.* **118,** 7063-7068.

Sundquist, W. I. & Klug, A. (1989) *Nature* **342,** 825-9.

Szalai, V. A., Singer, M. J. & Thorp, H. H. (2002) *J. Am. Chem. Soc.* **124,** 1625-1631.

Szalai, V. A. & Thorp, H. H. (2000) *J. Phys. Chem. B* **104,** 6851-6859.

Talalay, P. (2000) *BioFactors* **12,** 5-11.

Tekwani, B. L. & Mehlotra, R. K. (1999) *Microbes and Infection* **1,** 385-394.

Tomschik, M., Jelen, F., Havran, L., et al. (1999) *J. Electroanal. Chem.* **476,** 71-80.

Treadway, C. R., Hill, M. G. & Barton, J. K. (2002) *Chemical Physics* **281,** 409-428.

Tullius, T. D., Dombroski, B. A., Churchill, M. E. A., et al. (1987) *Methods Enzymol.* **155,** 537-58.

Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) *Science* **291,** 1304-1351.

Vinogradov, A. E. (2001) *Gene* **276,** 143-151.

Vivares, C. P. & Metenier, G. (2000) *Curr. Opin. Microbiology* **3,** 463-467.

Wei, Y.-H., Pang, C.-Y., Lee, H.-C., et al. (1998) *Curr. Science* **74,** 887-893.

White, O., Eisen, J. A., Heidelberg, J. F., et al. (1999) *Science* **286,** 1571-1577.

Witte, R. S. (1985) *Statistics* (Holt, Reinhard & Winston, New York).

Wolf, S. G., Frenkiel, D., Arad, T., et al. (1999) *Nature* **400,** 83-85.

Wood, M. L., Dizdaroglu, M., Gajewski, E., et al. (1990) *Biochemistry* **29,** 7024-32.

Wood, V., Gwilliam, R., Rajandream, M. A., et al. (2002) *Nature* **415,** 871-880.

Woodmansee, A. N. & Imlay, J. A. (2002) *J. Biol. Chem.* **277,** 34055-34066.

Yokomizo, A., Ono, M., Nanri, H., et al. (1995) *Cancer Research* **55,** 4293-6.

Yoshioka, Y., Kitagawa, Y., Takano, Y., et al. (1999) *J. Am. Chem. Soc.* **121,** 8712-8719.

Zhang, M. Q. (1998) *Human Molecular Genetics* **7,** 919-932.

Zlatanova, J., Leuba, S. H. & Van Holde, K. (1998) *Biophys. J.* **74,** 2554-2566.

Vita

Keith Albert Friedman was born in Pomona, California on June 17, 1957, the son of Robert and Blanche Friedman. He received his BS from the University of California at Berkeley in 1980. He received his MS from San Jose State University with Dr. Michael Jennings in 1996 with a thesis on mathematical modeling of batch distillation. He will receive his PhD from the University of Texas at Austin with Dr. Adam Heller in 2003 with a dissertation on genome design and oxidation. All of these degrees are in chemical engineering.

Friedman worked for NWT Corp., a research and consulting firm involved with water chemistry and corrosion at power plants, and Dionex Corp., a developer and manufacturer of analytical chemistry instruments. He collaborated on large-scale evaluations of water purification systems, modeling of high-temperature water chemistry to establish guidelines for power plants, development of instruments for online ion chromatography and environmental water analysis, and development of the self-regenerating suppressor, a device that made ion chromatography much simpler.

Permanent address: 500 E. Riverside Dr. #208, Austin, TX 98704 USA

This dissertation was typed by the author.