

Clusters of mutations from transient hypermutability

John W. Drake^{*†}, Anna Bebenek^{*‡}, Grace E. Kissling[§], and Shyamal Peddada[§]

^{*}Laboratory of Molecular Genetics and [§]Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709; and [†]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, 02-106 Warsaw, Poland

Edited by Philip C. Hanawalt, Stanford University, Stanford, CA, and approved July 22, 2005 (received for review April 18, 2005)

Collections of mutants usually contain more mutants bearing multiple mutations than expected from the mutant frequency and a random distribution of mutations. This excess is seen in a variety of organisms and also after DNA synthesis *in vitro*. The excess is unlikely to originate in mutator mutants but rather from transient hypermutability resulting from a perturbation of one of the many transactions that maintain genetic fidelity. The multiple mutations are sometimes clustered and sometimes randomly distributed. We model some spectra as populations comprising a majority with a low mutation frequency and a minority with a high mutation frequency. In the case of mutants produced *in vitro* by a bacteriophage RB69 mutator DNA polymerase, mutants with two mutations are in ≈ 10 -fold excess and mutants with three mutations are in even greater excess. However, phenotypically undetectable mutations seen only as hitchhikers with detectable mutations are ≈ 5 -fold more frequent than mutants bearing detectable mutations, indicating that they arose in a subpopulation with a higher mutation frequency. Excess multiple mutations may contribute critically to carcinogenesis and to adaptive mutation, including the adaptations of pathogens as they move from host to host. In the case of the rapidly mutating riboviruses, the viral population appears to be composed of a majority with a mutation frequency substantially lower than the average and a minority with a huge mutational load.

bacteriophage RB69 | multiple mutations | mutation rate | ribovirus

The idea that mutations arise at random is a common motif in descriptions of the mutation process. In contrast, mutations are almost always nonrandomly assorted among the sites at which they can be detected. Here, we describe another kind of nonrandom assortment: often, more mutants contain two or more mutations than expected from a random process.

An excess of mutants with multiple mutations (“multiples,” comprising doubles, triples, and so on) is of interest for at least four reasons. First, it signals an aspect of genetic instability that has received little attention. Second, it provides an insight into the way that genetic diseases that require multiple mutations (such as carcinogenesis) can proceed in the absence of a conventional enhancer of mutagenesis such as a mutator mutation. Third, it similarly provides an insight into the way that pathogens that require multiple mutations to sustain an infection in a new host can accumulate these mutations in bursts rather than serially. Fourth, it provides a mechanism by which adaptive evolution can proceed relatively rapidly when two or more mutations are required, each of which is individually neutral or deleterious.

Widespread Excesses of Mutants Bearing Multiple Mutations

Incidence of Multiples in Mutation Spectra. We examined numerous spectra for their incidence of multiples and calculated the expected frequencies of doubles and higher-order multiples. Many of these spectra were gathered first to characterize the fraction of spontaneous mutations that are base substitutions (1), to which were added other spectra as encountered unsystematically. Three kinds of mutations were excluded from consideration: (i) tandem multiples, because they were deemed likely to arise as single events; (ii) deletions that include most or all of the reporter sequence, because they necessarily preclude

multiple mutations; and (iii) insertions of mobile elements, because they might sometimes arise by mechanisms independent of other kinds of mutagenesis. The experimentally observed variables required to estimate the expected number of multiples are the mutant frequency, F , and the number of sequenced mutants, M_s . The expected average number of mutations per reporter sequence, a , can be estimated from the first term of the Poisson distribution, $P(0) =$ the fraction of the population with no mutations $= e^{-a} = 1 - F$, whence $a = -\ln(1 - F)$. The expected number of doubles among M_s sequenced mutants is $a^2 e^{-a}/2$ times the population size that yielded the sequenced mutants, M_s/F . Thus, the expected number of doubles is $E_2 = M_s a^2 e^{-a}/2F$. The expected number of higher multiples can be calculated sequentially from each previous term; for instance, the number of expected triples among sequenced mutants is $E_3 = E_2 a/3$. The calculation also holds well when the fraction of mutations detected is substantially <1 .

Table 1 lists spectra with multiple mutations. The mutation-reporter sequences represent endogenous genes and transgenes, gene fragments and complete genes, *in vivo* and *in vitro* systems, and phylogenetically diverse organisms. Some spectra could not be considered because a value for F was not provided or it was impossible to discern whether any of the mutations had arisen in multiples. Because multiples usually occur in a small fraction of the mutants, most of the spectra lacking multiples were relatively small. Of the 38 spectra in Table 1 (including two entries that pool spectra), 27 contain multiples in substantial to huge excess over expectations, four have multiples in modest excess, and seven have about the expected numbers of multiples. Excesses of multiples occur in RNA and DNA viruses, cellular microbes, and vertebrate tissues. Multiples are also generated in excess by diverse polymerases *in vitro*. In most of these spectra, the experimental procedures were unlikely to introduce additional mutations during the processing of a mutant isolate (for instance, by errors during amplification). Although we sometimes could not distinguish between mutations that were known to be detectable and those known or suspected to be undetectable in the scoring system, the fraction of undetected base substitutions in most spectra is only a fewfold more than the fraction that does produce a phenotypic change (1), and, thus, the presence of phenotypically silent mutations does not confound our general conclusion: a striking excess of multiple mutations is a widespread phenomenon.

In addition to mutations arising in replicating cells, which dominate the entries in Table 1, mutation also occurs in nondividing stationary-phase cells (28, 29). Some of these mutations also arise as multiples in a transiently hypermutating subpopulation of the mutating cells (30, 31).

Transient or Heritable Hypermutation? An excess of multiples suggests the existence of a hypermutating subfraction of the population. Hypermutation might be driven either by a heritable mutator mutation or by a transient condition, such as the induction of an encoded mutator factor (as in the *Escherichia coli*

This paper was submitted directly (Track II) to the PNAS office.

[†]To whom correspondence should be addressed. E-mail: drake@niehs.nih.gov.

© 2005 by The National Academy of Sciences of the USA

Table 1. Mutants with multiple mutations in mutational spectra

System	Genotype or strain	Reporter gene	F	M_s	E_2	D	T	Notes	Ref.	
Tobacco mosaic virus	WT	<i>MP</i>	4.3×10^{-2}	17	0.36	3	3		2	
HIV-1 RT <i>in vitro</i>	WT	<i>lacZα</i>	6.4×10^{-2}	434	13.9	24		a, b	3	
	WT	<i>lacZα</i>	2.3×10^{-2}	99	1.2	2			4	
			1.3×10^{-1}	97	6.2	19	1		4	
Bacteriophage RB69/T4	Exo ⁻	<i>rl</i>	2.0×10^{-2}	72	0.72	3		c	5	
	Pol ^{A/S/T}	<i>rl</i>	2.7×10^{-2}	147	2.0	3		c	5	
T4 gp43 <i>in vitro</i>	Exo ⁻	<i>lacZα</i>	1.1×10^{-2}	121	0.65	2		b	6	
Herpes simplex virus	WT	<i>supF</i>	4.9×10^{-4}	80	0.020	7	1	d	7	
		<i>tk</i>	6×10^{-5}	66	0.0020	1		d	8	
	PAAr5	<i>supF</i>	1.0×10^{-3}	87	0.045	4	2		7	
	Y7	<i>supF</i>	1.9×10^{-3}	53	0.050	0	1		7	
		<i>tk</i>	4×10^{-2}	66	1.3	6			8	
	Y7 Exo ⁻	<i>supF</i>	4.8×10^{-3}	249	0.60	11			9	
	YD12	<i>supF</i>	1.5×10^{-3}	77	0.059	2			7	
	<i>Escherichia coli</i>	WT	<i>supF</i>	2.1×10^{-7}	38	0.000004	1		e	10
		WT	<i>lacI^d</i>	1.3×10^{-7}	368	0.000024	2			11
<i>mutD5</i>		<i>lacI^d</i>	1.5×10^{-3}	498	0.37	4		e	12	
<i>mutL</i>		<i>lacI^d</i>	3.5×10^{-5}	243	0.0043	2			13	
<i>mutL</i>		<i>lacI^d</i>	1.2×10^{-5}	196	0.0011	1			13	
<i>dnaE911</i>		<i>lacI^d</i>	0.8×10^{-7}	476	0.000019	1			11	
<i>dnaE173</i>		<i>rpsL</i>	9.2×10^{-6}	56	0.00026	1	1		14	
<i>E. coli</i> Pol I(K) <i>in vitro</i>	WT	<i>lacZα</i>	4.7×10^{-3}	118	0.28	3		b, f	15	
	Y766A/S	<i>lacZα</i>	3.8×10^{-2}	224	4.3	5			15	
<i>Saccharomyces cerevisiae</i>	WT	<i>SUP4-o</i>	1.9×10^{-6}	297	0.00028	2			16	
	<i>rad1</i>	<i>SUP4-o</i>	1.3×10^{-5}	242	0.0015	1			16	
Rat cell line	WT	<i>cII</i>	1.3×10^{-4}	99	0.0064	1		g	17	
Mouse cell line	WT	<i>gpt</i>	2×10^{-5}	43	0.00043	0	1 (5)	h	18	
Chinese hamster cell line	WT	<i>gpt</i>	1.2×10^{-4}	18	0.0011	2		i	19	
Human cell line	WT	<i>hprt</i>	9×10^{-6}	200	0.0009	6	1	j	20	
Mouse tissue	WT	<i>cII</i>	9.5×10^{-5}	182	0.0086	1		k	21	
	WT	<i>lacI</i>	4.2×10^{-5}	348	0.0073	2		l	22	
	WT	<i>lacI</i>	2.3×10^{-5}	435	0.0050	7	1 (5)	m	23	
Human tissue	WT	<i>HPRT</i>	1.9×10^{-4}	82	0.0078	5	1 (4)		24	
Rat hepatoma Pol β <i>in vitro</i>	WT	<i>lacZα</i>	10.6×10^{-2}	296	15.7	≤ 16		b, n	25	
Chick embryo Pol β <i>in vitro</i>	WT	<i>lacZα</i>	7.3×10^{-2}	144	5.2	≤ 1		n	25	
Rat Pol β^{*} <i>in vitro</i>	WT	<i>HSV-tk</i>	1.4×10^{-3}	86	0.060	2		o	26	
	T79S	<i>HSV-tk</i>	2.7×10^{-3}	79	0.11	3	9 (≥ 3)	o	27	
	Y265C	<i>HSV-tk</i>	4.4×10^{-2}	79	1.7	31	14 (≥ 3)	o	27	

The reporter gene is natural unless noted to be a transgene. WT, wild type; F , frequency of spontaneous mutants, adjusted where appropriate for the efficiency of detecting mutants and for mutations that do not produce a phenotype; M_s , number of sequenced mutants exclusive of those containing either large deletions or insertions of mobile elements; E_2 , number of doubles expected from a random distribution of mutations; D , observed number of doubles; T , observed number of triples or, for entries of the form $n(m)$, number n of mutants with m mutations.

^aHuman immunodeficiency virus reverse transcriptase.

^bThe *lacZ α* system distinguishes between mutations that are or are not detectable when present as singles. Only detectable mutations are tabulated.

^cBacteriophage T4 43⁻ was replicated by a plasmid-encoded replicase (gp43) from the related bacteriophage RB69. Gp43 has both polymerase (Pol) and exonuclease proofreading (Exo) sites. The Pol^{A/S/T} entry is the sum for three different substitutions at Y567 with very similar mutator properties, F being the value for each mutant weighted by its M_s .

^d*supF* is a tRNA transgene from *E. coli* of a type that may be generally hypermutable, whereas *tk* is an endogenous gene and displays an approximately normal mutation rate.

^eNo other multiples were observed among *lacI* mutants pooled from several wild-type or mutator strains with $F \approx 2 \times 10^{-7}$, $M_s \approx 683$, $E_2 \approx 0.00005$ or with $F \approx 2 \times 10^{-6}$, $M_s \approx 661$, $E_2 \approx 0.0008$.

^fK indicates the Klenow fragment of Pol I. Y766A/S are mutator mutations.

^gEmbryonic fibroblast cell line with a *cII* transgene from phage λ .

^hA9 cell line with a *gpt* transgene from *E. coli*. The five mutations were base substitutions scattered throughout *gpt*.

ⁱOvary cell line with a *gpt* transgene from *E. coli*.

^jTK6 lymphoblastoid cell line with a human *hprt* cDNA transgene at five different sites.

^kMouse liver, lung, and spleen with a *cII* transgene from phage λ .

^lMouse liver with a *lacI* transgene from *E. coli*.

^mNumerous mouse tissues with a *lacI* transgene from *E. coli*.

ⁿBecause doubles were combined with duplications and complex mutations, the D values are " \leq ".

^oRecombinant enzyme made in *E. coli* and bearing an additional Gly-Ser at its 5' end.

SOS system) or an accident in some other DNA transaction; the latter notion was described in ref. 32. Mutator mutations are deleterious in most conditions (33–36) and are found at corre-

spondingly low frequencies in laboratory populations of, for instance, *E. coli* and *Salmonella typhimurium* (37, 38). A general argument suggests that excesses of multiples usually occur

Table 2. Observed and expected distributions of detectable and undetectable mutations among mutants generated by RB69 Exo⁻ Pol^{Y567A} DNA polymerase

Category	No APs		gp32 ± 44/62		gp45 + 44/62		All four APs	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
<i>F_d</i>	0.0153		0.0232		0.0220		0.0220	
Total <i>d</i>	213	213.00	474	474.00	304	304.00	333	333.00
Singles	198	211.36	424	468.46	262	300.63	283	329.31
Doubles	15	1.63	49	5.49	39	3.35	48	3.67
Triples	0	0.01	1	0.04	3	0.02	2	0.03
<i>F_{ud}</i>	0.108		0.177		0.204		0.204	
Δ <i>F_{BPS}</i>	4.00		4.07		5.26		4.96	
Total <i>ud</i>	23	23.00	84	84.00	62	62.00	68	68.00
Singles	22	21.71	76	76.06	54	55.29	61	60.53
Doubles	1	1.24	8	7.42	8	6.30	6	6.91
Triples	0	0.05	0	0.48	0	0.48	1	0.53

See text concerning the four different combinations of polymerase accessory proteins (APs). Obs., observed numbers of mutants; Exp., expected numbers of mutants assuming a Poisson distribution of mutations among mutants; *F*, mutant frequency; *F_d*, frequency of mutants bearing at least one detectable mutation and *F_{ud}*, frequency of mutants bearing at least one undetectable mutation among mutants bearing a detectable mutation (for instance, in column 2, *F_{ud}* = 23/213 = 0.108). Total *d* = total number of mutants with ≥1 detectable mutations. Total *ud* = total number of mutants with ≥1 undetectable mutations among the total *d* mutants. For base pair substitutions (BPS), the target sequence comprises 244 detectable changes and 530 undetectable changes (see text). Δ*F_{BPS}* = factor of increase in BPS mutant frequency in *ud* mutations compared with *d* mutations = (244*F_{ud}**P_{udBPS}*/530*F_d**P_{dBPS}*) where *P_{BPS}* = proportion of BPS among mutations; almost all (0.995) of the *ud* mutations and ≈0.83 of the *d* mutations were base substitutions.

independently of mutator mutations. The observed frequency of doubles equals the mutant frequency times the proportion of doubles among mutants (*F_D* = *FD*/*M_s*). Numerous reports over several decades indicate that a mutator mutation increases the value of *F* averaged across a gene by at most ≈10²-fold, stimulating some sites by much more and many others by much less. At least in *E. coli* and *S. typhimurium*, the frequency of mutator mutants in a laboratory population is ≤10⁻⁵ (36, 37). Therefore, the frequency of doubles generated by mutators will be ≤(10²*F*)²(10⁻⁵) = 10⁻¹*F*², and the fraction of doubles generated by mutators among all doubles will be ≤*FM_s*/10*D*. Inspection of Table 1 reveals that this fraction is <0.1 for all of the entries for which *D* is substantially greater than *E₂*. Thus, mutator mutations probably generate few of the observed multiples in mutational spectra.

Multiples Generated *in Vitro*

The RB69 System. For most spectra, collecting and sequencing mutants requires months and too few multiples accumulate to provide grist for analysis. One useful exception is *lacZα* spectra generated *in vitro* by mutator versions of the bacteriophage RB69 DNA polymerases, with or without polymerase accessory proteins. With an engineered version of the polymerase whose proofreading exonuclease has been inactivated (Exo⁻) and whose insertion accuracy has been compromised by an active-site replacement that stimulates base substitutions (Pol^{Y567A}) (5), *lacZα* mutant frequencies are ≈10⁻² (39, 40). Here, we describe the multiples that arise in these spectra, sometimes in ≈15% of the mutants. The *lacZα* mutants were produced by the mutant polymerase together with one of four combinations of accessory proteins: none, gp32 ± gp44/62, gp45 plus gp44/62, and all four (gp32 being the single-stranded DNA-binding protein or SSB, gp45 the processivity clamp, and gp44/62 the clamp-loading complex.) The *lacZα* system has been used so extensively that experience has revealed which mutations are detectable as singles and which are not (39, 41). Thus, two kinds of mutations can be distinguished in these experiments. Almost all mutants bear a detectable mutation within the target sequence, whereas almost all mutations that are undetectable as singles are seen only in the presence of a detectable mutation and may be considered to be hitchhikers.

The distributions of detectable mutations, and undetectable mutations arising together with one or more detectable mutations, are shown in Table 2. The upper half of the table shows that multiples comprising detectable mutations arose on average in 10.6-fold (range 9.2–12.7) excess of the predictions of a random assortment. Triples, when seen, were in much greater excess. Thus, a DNA polymerase copying a DNA template *in vitro*, with or without accessory proteins, can produce a substantial excess of multiples over the predictions of a random distribution of mutations. The accessory proteins increased total mutant frequencies by ≈1.5-fold and increased the frequency of multiples by up to ≈2-fold (7.0% with no accessory proteins, 10.5% with gp32, 14% with gp45, and 15% with gp32 plus gp45), perhaps reflecting the abilities of the accessory proteins to improve processivity (A.B. and J.W.D., unpublished results).

The bottom half of Table 2 reveals that the distribution changes markedly for undetectable mutations hitchhiking with detectable mutations. The frequencies of mutants with undetectable mutations among mutants with detectable mutations (*F_{ud}*) increased compared with the mutant frequencies of mutants bearing one or more detectable mutations (*F_d*), the average increase being 4.6-fold (range 4.0–5.3). However, the number of pairs of otherwise undetectable mutations was always close to the expectation for a random distribution of mutations, on average only 1.006-fold in excess (range 0.8- to 1.3-fold).

Table 3 describes mutants generated by other RB69 DNA polymerases both *in vitro* and *in vivo* (5, 39). In the *lacZα in vitro* system, the wild-type enzyme is too accurate to produce new mutations. For the Exo⁻ Pol⁺ and Exo⁺ Pol^{Y567A} enzymes, results with and without all four accessory proteins were pooled; detectable doubles were produced in ≈6-fold excess, a little lower than the average 10.6-fold excess produced by the Exo⁻ Pol^{Y567A} polymerase (Table 2). In the *rl in vivo* system, a few doubles were observed with the Exo⁻ Pol⁺ polymerase and with Exo⁺ polymerases bearing various substitutions at the Pol Y567 residue; probably too few mutants generated by the Pol⁺ Exo⁺ enzyme were sequenced to detect multiples. The *D*/*E_D* ratios were ≈5 for the Exo⁻ Pol⁺ enzyme and averaged ≈1.7 for Exo⁺ enzymes with replacements at Y567. Thus, with the RB69 DNA polymerase, more than expected numbers of multiples are seen

Table 3. Multiples in other RB69 DNA polymerase constructs

System	Reporter	Polymerase	F	M_s	E_2	D
<i>In vitro</i>	<i>lacZα</i>	Exo ⁻ Pol ⁺	0.0027	333	0.45	3
		Exo ⁺ Pol ^{Y567A}	0.0066	275	0.91	5
<i>In vivo</i>	<i>rl</i>	Exo ⁺ Pol ⁺	0.000047	79	0.002	0
		Exo ⁻ Pol ⁺	0.027	72	1.0	5
		Exo ⁺ Pol ^{Y567A}	0.036	75	1.4	4
		Exo ⁺ Pol ^{Y567S}	0.039	35	0.7	1
		Exo ⁺ Pol ^{Y567T}	0.051	37	1.0	0

See Table 1 legend for definitions of the parameters F , M_s , E_2 , and D .

with several different versions of the enzyme both *in vitro* and *in vivo*.

Distributions of Multiples in Genetic Space. Different mechanisms can be imagined that would tend to cluster or to scatter the two components of a double within a mutation-reporter sequence. We therefore examined the extent to which the two components of each double were a random sample of the underlying spectrum of detectable mutations. We used the doubles generated *in vitro* by the Exo⁻ Pol^{Y567A} enzyme with or without the four accessory proteins, because that number of doubles was the largest in any set. We explored two measures of genetic distance. In the first, the distance D_s between any pair of mutations was defined as the number of intervening sites at which mutations were detectable by using the observed spectrum rather than all detectable sites because of the strong biases with which mutations are distributed in different spectra. In the second measure, the distance D_m was defined as the number of observed mutations between the two sites in that spectrum, thus reflecting not only the number of mutable sites, but also their summed specific mutabilities. In statistical tests (see Table 4, which is published as supporting information on the PNAS web site), the hypothesis that the two components of each double are distributed uniformly in either kind of genetic space was robust ($P \approx 0.97$).

In two other spectra, the components of multiples were reported to be nonrandomly close to each other. The number of multiples in the *E. coli lacI* gene established as a transgene in the mouse far exceeded the predictions of the mutant frequency (23), and the distances between mutations were distributed exponentially, with half of the mutation-pairs separated by ≤ 120 bases in the $\approx 1,440$ -base mutational target (42). A spectrum of *HPRT* mutations in human kidney epithelial cells was expected to contain < 0.01 doubles but contained six (including one tandem double) plus one quadruple (24). The distributions of numbers of bases between the components of doubles was 0, 1, 6, 13, 5,012, 7,023, and 25,024 (and, for the quadruple, 13, 214 and 4,886), a distribution that is clearly a nonrandom sample from the spectrum ($P = 0.002$). In contrast, among *HPRT* mutations recovered from patients treated for acute lymphocytic leukemia by using agents that included both mutagens and chemicals that select for *HPRT* mutants, $M_s = 182$, $F \approx 0.0003$, $E_2 = 0.03$, and numerous multiples were recorded, but the mutations in a multiple appeared not to be clustered, and other evidence suggested that they arose sequentially rather than in a single cell division (43).

Deconstructing Spectra into Components with Different Mutation Frequencies. An excess of multiples signals the existence of at least two subpopulations with different mutation frequencies. There are several ways to deconstruct a spectrum into hypothetical subpopulations whose parameters comprise the size of each subpopulation and its characteristic mutation frequency.

One of these methods is purely algebraic; for its more detailed derivation, see *Supporting Text*, which is published as supporting

information on the PNAS web site. As before, F = the mutant frequency and $M_s = \sum M_i$ sequenced mutants with i mutations each. Let n = the total population. Let S_1 = a majority subfraction of the population with the lower mutation frequency f_1 and S_2 = a minority subfraction with the higher mutation frequency f_2 . Assuming random distributions of mutations among mutants in each subpopulation, with no significant contribution from the S_1, f_1 class to the multiples, then it follows that $f_2 = 3M_3/M_2$, $S_2 = 2FM_2/M_s f_2^2 e^{-f_2}$, $S_1 = 1 - S_2$, and $f_1 = -\ln[(1 - F - S_2 e^{-f_2})/S_1]$. Similarly, $S_2 = 6FM_3/M_s f_2^3 e^{-f_2}$; if there are only two significantly contributing subfractions in the population, then these two measures of S_2 should agree. The main drawback of this method is the frequent paucity of triples in mutant collections.

The other method is stochastic, a progressive fitting of the parameters S_1, f_1, S_2 , and f_2 (or more if necessary) to the data. Parameters of mixtures of Poisson distributions were estimated by an iterative Expectation-Maximization (E-M) method (44). First, we selected the number of subpopulations, such as two. Based on this number, the likelihood function of the mixture distribution was maximized by using an E-M algorithm (45). The procedure was repeated for all other plausible numbers of populations that could be represented within the data, and the likelihoods were compared. The most parsimonious set of parameter estimates that achieved the maximum among all of the maximized likelihoods was selected as the best representation of the data. The fit of the selected Poisson mixture was then tested with a χ^2 goodness-of-fit test (46). Note that if a fit is disappointing because of a few outlier mutants with a large number of mutations, a third population can be designed with S_3 and f_3 chosen to introduce just those mutants.

Consider the data for tobacco mosaic virus (2), wherein $F = [(43 \text{ mutants})/(1820 \text{ clones})]$ (correction factor of 1.8 for incomplete recovery of mutants) = 0.04253 and $M_s = 17$ comprising $M_1 = 11, M_2 = 3$, and $M_3 = 3$. Application of the algebraic method gives $S_1 = 0.9665, f_1 = 0.01113, S_2 = 0.0335$, and $f_2 = 3$, all biologically reasonable values with no significant contribution (0.001 mutants) from class 1 to the multiples. However, these values predict too few singles (6.3 instead of 11) and too many multiples with four or more mutations (4.7 instead of 0), and the P value for this fit is a modest 0.40. Application of the stochastic method gives $S_1 = 0.9575, f_1 = 0.0131, S_2 = 0.0425$, and $f_2 = 1.2353$, values similar to those provided by the analytical method but predicting 11.0 singles, 3.8 doubles, 1.6 triples, 0.6 multiples with four or more mutations, and an impressive goodness-of-fit $P = 0.98$. Many of the spectra listed in Table 1 can be similarly modeled, but with modest precision when doubles are few and triples are absent.

A different picture emerged from a typical experiment by using the *in vitro lacZ α* system with the Exo⁻ Pol^{Y567A} polymerase and all four accessory proteins (part of the summed results of Table 2), with $F = 0.021, M_s = 149, M_2 = 21$, and $M_3 = 1$ (but all of these values must be corrected for the loss of 40% of the original numbers of mutants in the assay system). The algebraic

model failed because it generated a negative value for f_1 and the two ways of calculating S_2 gave different answers. This failure may reflect a small contribution of multiples from the low-frequency class, the low ratio of triples to doubles, and perhaps, the presence of more than two contributing subpopulations. If this failure was due to a sampling error, and if the observed numbers of triples and doubles were 3 and 19 instead of 1 and 21, then the algebraic method gives $S_1 = 0.936$, $f_1 = 0.0117$, $S_2 = 0.064$, and $f_2 = 0.474$ but with goodness-of-fit $P < 0.0001$. The best stochastic model yielded two subpopulations with $S_1 = 0.864$, $f_1 = 0.000045$, $S_2 = 0.136$, and $f_2 = 0.2984$ with $P = 0.86$. Here, the extremely low value of f_1 is biologically unreasonable, and a higher value that still contributed very few mutations might be more reasonable but would still be unconvincing. However, an important insight into this system was obtained from the distribution of hitchhiker mutations that are undetectable (produce no mutant phenotype) by themselves. These mutants have an average mutant frequency 4.6-fold higher than for mutants bearing detectable mutations and presumably arose preferentially within the subpopulation(s) with one or more higher mutation frequencies. The hitchhiker mutations fit a Poisson distribution well (21.9 doubles expected, 23 observed; 1.54 triples expected, 1 observed), indicating that they arose mainly in a subpopulation with a fairly uniform mutation frequency at least 5-fold higher than that for the average detectable mutation. This result provides some limits to the parameters. Assume only two subpopulations, so that $F = f_1S_1 + f_2S_2$ and $S_1 + S_2 = 1$. Because $F \approx 0.02$ and, from Table 2, $f_2/F \approx 5$, it follows that $f_2 \approx 0.1$ and $f_1 \approx (0.02 - 0.1S_2)/(1 - S_2)$. Nonnegative values of f_1 require that $S_2 < 0.2$, which in turn implies that $f_1 \approx 0.02 - 0.1S_2$. Overall, the difficulty in decomposing the RB69 data is instructive because it signals a poorly understood complexity to at least this example of hypermutability.

The Biology of Multiples

Carcinogenesis. The full course of carcinogenesis requires more mutations than can be attributed to conventional mutation rates, even at the extreme of a Poisson distribution (47). Surveys of tumors often reveal continuing high mutation rates that signal mutator mutations but also usually reveal tumors that do not display high mutation rates. Such tumors might contain heritable mutators not detected by the method of the moment (48, 49), but they may also arise during transient bursts of hypermutability. Clear evidence for transient hypermutability already exists in the large database of *TP53* mutations (50–52). More than 5% of tumors contain two or more base pair substitutions in *TP53* and many of these are silent (synonymous) mutations. Altogether there are $>10^{-4}$ mutations per base pair, a value that would be immediately lethal if sustained throughout the genome. Thus, transient *TP53* hypermutability must also be localized.

Adaptive Evolution. The concept of multiple mutations that are individually deleterious but advantageous when combined seems to have been mentioned first by Haldane (53), who cited what was also perhaps the first experimental example, from an analysis of the genetics of lifespan in *Drosophila* (54). The barrier to moving from a local fitness maximum to a higher adaptive peak (55) could be reduced considerably by mutations that arise in bursts. A similar situation arises in the case of a mutation producing a mixture of beneficial and deleterious effects with a net reduction in fitness that can be ameliorated by modifiers (ref. 56, pp. 463 and 471). In addition, two deleterious alleles that are neutral when combined can become fixed under random drift plus mutation pressure, provided they are tightly linked (57). It was recently estimated that such compensating deleterious mutations are common and tend to reside in the same protein (58),

which would reflect tight linkage and might suggest their origin in a spatially delimited burst of hypermutation.

When the frequency of multiples is substantially higher than the product of the individual frequencies, adaptive leaps requiring multiple mutations can be greatly accelerated. This result is particularly the case when the component mutations are individually deleterious and thus cannot accumulate in the population. Thus, transient phenotypic hypermutability is likely to contribute to adaptive evolution. However, just as mutator mutants are unlikely to become fixed by selection on their adaptive mutations because they are soon separated from their beneficial products and carry a considerable selective disadvantage (34), transient hypermutability is unlikely to be an adaptive trait that has been maintained by selection except in those special cases such as the SOS response where it is specifically encoded. Even in such cases, hypermutability may be an unavoidable inefficiency in a process targeted to survival rather than mutagenesis. On the whole, transient hypermutability may simply represent the numerous ways in which fidelity mechanisms often go awry, the cost of further reducing their incidence exceeding the penalty paid for the resulting deleterious mutations.

Microbial Pathogens. Although the frequency of mutator mutants is very low among laboratory populations of bacteria, numerous examples have accumulated of high frequencies among fresh isolates of human bacterial pathogens (e.g., refs. 59 and 60). This enrichment for mutator mutants has been modeled as the result of selection for very rapid rates of adaptation (e.g., refs. 61 and 62). However, the frequency of such mutator mutants is generally ≤ 0.2 , so that most adaptations to the new host may be achieved without the help of mutator mutations. An alternative hypothesis, reflecting the strong selective disadvantage of mutator mutants in a stable environment, is that many successful nonmutator pathogens recently had a mutator mutation that was expelled by reversion or recombination. However, escape from a mutator mutation within the time frame of most single-host infections seems unlikely. Thus, transient phenotypic hypermutability is likely to contribute to microbial pathogenesis.

RNA Genomes. Based on limited data, riboviruses have genomic mutation rates of 0.1–1.2 with a median of 0.76 per replication or 1.5 per cycle of infection (2, 63). Small increases produce mutational meltdown (64). Tobacco mosaic virus, the only ribovirus for which a probably unbiased mutation spectrum is available, has a genomic mutation rate ≈ 0.1 per replication (2). Our analysis of multiples in tobacco mosaic virus suggests that $\approx 96\%$ of the population has a genomic mutation rate ≈ 0.03 , which means that most of the population may be considerably more genetically stable than previously estimated. Approximately 4% of the tobacco mosaic virus population displays a much higher mutation frequency, one that would correspond to a genomic rate of almost 3, but even a single round of replication at this error frequency would produce mostly dead genomes unless the hypermutability was limited to only a portion of the genome.

Because retroelements (retroviruses and retrotransposons) have mutation rates only a fewfold lower than do riboviruses (65), the above considerations apply to them as well.

Mechanisms. Faulty proteins are inevitable consequences of cellular metabolism. Transient hypermutability might be generated by errors of transcription and/or translation that generate proteins with altered primary sequences, errors of folding or posttranslational modification that generate dysfunctionally active or dominant-inactive proteins, and regulatory errors at any level that overproduce or underproduce proteins, leading to

their incorrect distribution within the cell or between dividing cells. Any protein involved in replication fidelity could be thusly affected, so that hypermutability could be generated by many discrete mechanisms. Because proteins (and particularly faulty proteins) turn over, their mutagenic impact can be brief. A faulty polymerase, for instance, might produce a single, limited tract of error-prone synthesis, thus delimiting the region of hypermutation. Another candidate mechanism would have a segment of DNA synthesis begin in an error-prone manner, or fall into an error-prone state after an initial error, by perpetuating an

anomalous interaction between normal enzyme and normal substrates. The high frequencies of mutants and multiples in the RB69 system may provide the first opportunity to address questions of mechanisms in an efficient manner.

We thank John Cairns, Marilyn Diaz, Pat Foster, and Roel Schaaper for their valuable critiques during the writing of this article. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences.

- Grogan, D. W., Carver, G. T. & Drake, J. W. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 7928–7933.
- Malpica, J. M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W. & Garcia-Arenal, F. (2002) *Genetics* **162**, 1505–1511.
- Bebenek, K., Abbotts, J., Roberts, J. D., Wilson, S. H. & Kunkel, T. A. (1989) *J. Biol. Chem.* **264**, 16948–16956.
- Eckert, K. A. & Kunkel, T. A. (1993) *Nucleic Acids Res.* **21**, 5212–5220.
- Bebenek, A., Dressman, H. K., Carver, G. T., Ng, S., Petrov, V., Yang, G., Konigsberg, W. H., Karam, J. D. & Drake, J. W. (2001) *J. Biol. Chem.* **276**, 10387–10397.
- Kroutil, L. C., Frey, M. W., Kaboord, B. F., Kunkel, T. A. & Benkovic, S. J. (1998) *J. Mol. Biol.* **278**, 135–146.
- Hwang, Y. T., Liu, B. Y., Hong, C. Y., Shillitoe, E. J. & Hwang, C. B. C. (1999) *J. Virol.* **73**, 5326–5332.
- Lu, O., Hwang, Y. T. & Hwang, C. B. C. (2002) *J. Virol.* **76**, 5822–5828.
- Hwang, Y. T. & Hwang, C. B. C. (2003) *J. Virol.* **77**, 2946–2955.
- Akasaka, S., Takimoto, K. & Yamamoto, K. (1992) *Mol. Gen. Genet.* **235**, 173–178.
- Oller, A. R. & Schaaper, R. M. (1994) *Genetics* **138**, 263–270.
- Schaaper, R. M. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 8126–8130.
- Schaaper, R. M. (1993) *Genetics* **134**, 1031–1038.
- Mo, J.-Y., Maki, H. & Sekiguchi, M. (1991) *J. Mol. Biol.* **222**, 925–936.
- Bell, J. B., Eckert, K. A., Joyce, C. M. & Kunkel, T. A. (1997) *J. Biol. Chem.* **272**, 7345–7351.
- Kunz, B. A., Kohalmi, L., Kang, X. L. & Magnusson, K. A. (1990) *J. Bacteriol.* **172**, 3009–3014.
- Watson, D. E., Cunningham, M. L. & Tindall, K. R. (1988) *Mutagenesis* **13**, 487–497.
- Ashman, C. R. & Davidson, R. L. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 3354–3458.
- Romac, S., Leong, P., Sockett, H. & Hutchinson, F. (1989) *J. Mol. Biol.* **209**, 195–204.
- Lichtenauer-Kaligis, E. G. R., Thijssen, J., den Dulk, H., van de Putte, P., Tasseron-de Jong, J. G. & Giphart-Gassler, M. (1996) *Mutagenesis* **8**, 207–220.
- Harbach, P. R., Zimmer, D. M., Filipunas, A. L., Mattes, W. B. & Aaron, C. S. (1999) *Environ. Mol. Mutagen.* **33**, 132–143.
- de Boer, J. G., Erfle, H., Walsh, D., Holcroft, J., Provost, J. S., Rogers, B., Tindall, K. R. & Glickman, B. W. (1997) *Environ. Mol. Mutagen.* **30**, 273–286.
- Buettner, V. L., Hill, K. A., Scaringe, W. A. & Sommer, S. S. (2000) *Mutat. Res.* **452**, 219–229.
- Colgin, L. M., Hackmann, A. F. M., Emond, M. J. & Monnat, R. J., Jr. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 1437–1442.
- Kunkel, T. A. (1985) *J. Biol. Chem.* **260**, 5787–5796.
- Opresko, P. L., Sweasy, J. B. & Eckert, K. A. (1998) *Biochemistry* **37**, 2111–2119.
- Maitra, M., Gudzelak, A., Jr., Li, S.-X., Matsumoto, Y., Eckert, K. A., Jager, J. & Sweasy, J. B. (2002) *J. Biol. Chem.* **277**, 35550–35560.
- Ryan, F. J. (1959) *J. Gen. Microbiol.* **21**, 530–549.
- Foster, P. L. (2000) *Cold Spring Harbor Symp. Quant. Biol.* **65**, 21–29.
- Torkelson, J., Harris, R. S., Lombardo, M.-J., Nagendran, J., Thulin, C. & Rosenberg, S. M. (1997) *EMBO J.* **16**, 3303–3311.
- Rosche, W. A., Foster, P. L. & Cairns, J. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 6862–6867.
- Ninio, J. (1991) *Genetics* **129**, 957–962.
- Sturtevant, A. H. (1937) *Q. Rev. Biol.* **12**, 464–467.
- Kimura, M. (1967) *Genet. Res.* **9**, 23–34.
- Tröbner, W. & Piechocki, R. (1984) *Mol. Gen. Genet.* **198**, 177–178.
- Funchain, P., Yeung, A., Stewart, J. L., Lin, R., Slupska, M. M. & Miller, J. H. (2000) *Genetics* **154**, 959–970.
- Mao, E. F., Lane, L., Lee, J. & Miller, J. H. (1997) *J. Bacteriol.* **179**, 417–422.
- LeClerc, J. E., Payne, W. L., Kupchella, E. & Cebula, T. A. (1998) *Mutat. Res.* **400**, 89–97.
- Bebenek, A., Carver, G. T., Dressman, H. K., Kadyrov, F. A., Haseman, J. K., Petrov, V., Konigsberg, W. H., Karam, J. D. & Drake, J. W. (2002) *Genetics* **162**, 1003–1018.
- Bebenek, A., Carver, G. T., Kadyrov, F. A., Kissling, G. E. & Drake, J. W. (2005) *Genetics* **169**, 1815–1824.
- Bebenek, K. & Kunkel, T. A. (1995) *Methods Enzymol.* **262**, 217–232.
- Hill, K. A., Wang, J., Farwell, K. D., Scaringe, W. A. & Sommer, S. S. (2004) *Mutat. Res.* **554**, 223–240.
- Finette, B. A., Homans, A. C. & Albertini, R. J. (2000) *Science* **288**, 514–517.
- Hasselblad, V. (1969) *J. Am. Stat. Assoc.* **64**, 1459–1471.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *J. R. Stat. Soc. B* **39**, 1–38.
- Conover, W. J. (1971) *Practical Nonparametric Statistics* (Wiley, New York).
- Jackson, A. L. & Loeb, L. A. (1998) *Genetics* **148**, 1483–1490.
- Richards, B., Zhang, H., Phear, G. & Meuth, M. (1997) *Science* **277**, 1523–1526.
- Meuth, M., Richards, B. & Schneider, B. (1999) *Science* **283**, 641.
- Strauss, B. S. (1998) *Genetics* **148**, 1619–1626.
- Strauss, B. S. (2000) *Mutat. Res.* **457**, 93–104.
- Rodin, S. N., Rodin, A. S., Juhasz, A. & Holmquist, G. P. (2003) *Mutat. Res.* **510**, 153–168.
- Haldane, J. B. S. (1931) *Proc. Camb. Philol. Soc.* **27**, 137–142.
- Gonzalez, B. M. (1923) *Am. Nat.* **57**, 289–325.
- Wright, S. (1932) *Proc. Sixth Int. Cong. Genet.* **1**, 356–366.
- Wright, S. (1977) *Evolution and the Genetics of Populations. Experimental Results and Evolutionary Deductions*. (Univ. of Chicago Press, Chicago).
- Kimura, M. (1985) *J. Genet.* **64**, 7–19.
- Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 14878–14883.
- Jysson, K. (1960) *Acta Pathol. Microbiol. Scand.* **48**, 113–120.
- LeClerc, J. E., Li, B., Payne, W. L. & Cebula, T. A. (1996) *Science* **274**, 1208–1211.
- Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H. & Godelle, B. (1997) *Nature* **387**, 700–702.
- Tenaillon, O., Toupance, B., Le Nagard, H., Taddei, F. & Godelle, B. (1999) *Genetics* **152**, 485–493.
- Drake, J. W. & Holland, J. J. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 13910–13913.
- Holland, J. J., Domingo, E., de la Torre, J. C. & Steinhauer, D. A. (1990) *J. Virol.* **64**, 3960–3962.
- Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) *Genetics* **148**, 1667–1686.