

Distribution and Characterization of Microsatellites in the Emu (*Dromaius novaehollandiae*) Genome

E. H. Roots and R. J. Baker

This study generates data concerning the genome of a flightless species of bird, the emu. We examined and ultimately rejected the following hypotheses: (1) Microsatellites are randomly distributed throughout the emu genome. (2) The relative order of abundance of dinucleotides will be constant across genomes. (3) Interspersion distances for a given dinucleotide will be equal across vertebrate genomes. (4) In all genomes, a dinucleotide will be more frequent than any trinucleotide. (5) The percentage of single-copy DNA will remain the same in emus as in other volant birds. A cosmid library representing 4.48% of the emu genome was probed with 23 microsatellites. Hybridizations were scored on a scale of 0–3. The average insert size, approximately 40 kb, was used to determine frequency and interspersion. The cosmid library was probed with genomic DNA to determine the percent single copy. Co-occurrence frequencies and confidence intervals were compared to expected using chi-squared. The genome is estimated to contain a microsatellite repeat every 48 kb. Of 1632 clones probed for single-copy DNA, 643 displayed maximal hybridization, 220 displayed moderate hybridization, and 202 had minimal hybridization. After 3 days, 567 showed no hybridization.

When important shifts occur in life-history strategies of an evolutionary lineage, they are accompanied by phenotypic and genetic changes. Comparing the biological parameters of the species that have experienced this shift to those that have not permits the development of hypotheses concerning the relationship of phenotypic and genetic changes that accompany the alteration. Ratites possess a larger *C* value (Swift 1950) or genome size (Hinegardner 1976) than is typical in volant birds. What aspect of the ratite genome accounts for the larger size relative to volant birds? In this study we estimate the amount of single-copy DNA and microsatellites in the emu (*Dromaius novaehollandiae*) genome to better understand the differences between flightless and volant bird species.

Understanding the Significance of *C* Values

Genome sizes, or *C* values, are typically clustered for a group of organisms with a specific life-history strategy (Baker et al. 1992; Rayburn and Auger 1990; Sessions and Larson 1987). There is extensive literature attempting to explain why, in some species, there is a far greater quantity of DNA than is required to encode the genes

that developmentally construct and operate an organism (Baker et al. 1992; Cavalier-Smith 1978, 1985; Hartl 2000; Szarski 1974; Tiersch and Wachtel 1991; Van Den Bussche et al. 1995). Life-history features, such as metabolic rate, longevity, larger cell and organism sizes, and slower growth rates have been hypothesized as adaptive results of variations in *C* values.

The low *C* value found in volant birds versus ratites and other birds that have evolved flightlessness (some grebes, ibises, penguins, some ducks and geese, and some rails) may be explained by an intense stabilizing selection on the genome. For example, low genome size may be critical for the high metabolic rate required for flight. Flightless birds might not be exposed to the same intensity of stabilizing selection, thus permitting selfish DNA to increase in copy number, resulting in increased *C* values.

Two hypothetical explanations for a high *C* value in ratites may be advanced. First, such values might be representative of the ancestral condition of never attaining flight; thus a large genome size may be characteristic of the flightless ancestors of birds. In this case, the larger genome size in birds would be the primitive condition and volant birds would represent a more

From the Department of Biological Sciences, Texas Tech University, Box 43131, Lubbock, TX 79409. We thank Drs. Richard Strauss and Ron Van Den Bussche for assistance with data analyses, Drs. Jon Longmire and Mary Maltbie for assistance with molecular methods, and Dr. Luis Ruedas for editorial comments. We thank Canyon Biotech Laboratories, Reproductive Sciences, Inc., and the Advanced Research/Advanced Technology Program for funding. E. H. Roots is currently at the Centre for Reproductive Medicine, 3506 21st St., Suite 605, Lubbock, TX 79410-1200. Address correspondence to Ellen H. Roots, 5533 17th St., Lubbock, TX 79416, or e-mail: ehroots@aol.com.

© 2002 The American Genetic Association 93:100–106

derived evolutionary state. Second, flight may have evolved accompanied by a smaller genome size, and ratites abandoned flight. In this case, the larger genome size in ratites is a derived state relative to that of volant birds. If the former hypothesis is true, then understanding the composition of the emu genome can provide information on the ancestral condition for birds. If the latter is true, then large numbers of interspersed, tandemly repeated clusters of mini- and microsatellites may have already existed in the avian genome and volant birds evolved a mechanism to excise these interspersed elements. If microsatellite abundance covaries with genome size, it may be possible to further search for evidence of a mechanism in volant birds that excised these elements.

We examine the following hypotheses: (1) Microsatellites are randomly distributed throughout the emu genome. (2) The relative order of abundance of dinucleotides will be constant across genomes. (3) Interspersion distances for a given dinucleotide will be equal across vertebrate genomes. (4) In all genomes, a dinucleotide will be more frequent than any trinucleotide. (5) The percentage of single-copy DNA will remain the same in emus as in other volant birds.

Materials and Methods

Archiving and Probing a Cosmid Genomic Library With Microsatellites

A SuperCos library of high molecular weight for one female emu was constructed by Stratagene, Inc. The cosmid library was plated on petri dishes with Terrific Broth (TB) and kanamycin, and colonies were grown overnight at 37°C. Individual colonies (1728) were selected randomly and archived into eighteen 96-well microtiter plates. Each well contained 100 µl of TB with kanamycin prior to inoculation. After 24 h of growth at 37°C, 100 µl of TB glycerol with kanamycin was added to each well and the microtiter plates stored at -80°C.

A replica plater was used to inoculate nylon membranes (Biodyne B 0.45 µm) with clones from the microtiter plates. Membranes were incubated at 37°C for 7 h on 30 µg/ml TB/kanamycin agar plates, transferred to TB/kanamycin plates with chloramphenicol, and grown for 18 h at 37°C. DNA fixation to membranes was accomplished by sequentially placing them on blotting pads soaked in 0.4M NaOH (5 min), 0.5M HCl, and 1.5M NaCl, pH 7.5 (5

min), then 2× SSC (5 min), followed by baking at 70°C for 4 h.

Twenty-three microsatellites with their complementary strands were used as probes: mononucleotides: (C)_n; 4 dinucleotides: (GT)_n, (CT)_n, (TA)_n, and (GC)_n; 12 trinucleotides: (AGC)_n, (AGT)_n, (CCG)_n, (GAA)_n, (GAC)_n, (GAT)_n, (GCA)_n, (GGA)_n, (GGT)_n, (GTA)_n, (GTT)_n, and (TAA)_n; 6 tetranucleotides: (GAAT)_n, (GACA)_n, (GATA)_n, (GGAA)_n, (GTAT)_n, and (GTTT)_n. The hybridization temperatures are as follows: mononucleotide 89.0°C; dinucleotides 50.0°C, 50.0°C, 47.7°C, 46.5°C; trinucleotides 80.6°C, 66.9°C, 89.5°C, 53.0°C, 62.0°C, 55.0°C, 68.4°C, 62.0°C, 52.0°C, 55.0°C, 50.0°C, 53.0°C; tetranucleotides 74.3°C, 52.0°C, 50.0°C, 74.3°C, 55.5°C, and 55.5°C. For convenience, hereafter the “n” is dropped when referring to a microsatellite repeat.

Radiolabeled probes were hybridized to the membranes to determine the distribution and relative abundance of sequences in the emu genome. Prehybridization-hybridization solution consisted of 6× SSC, 1× Denhardt solution, 0.5% SDS, and 0.005 g/ml Carnation evaporated milk. The membranes were washed in prehybridization solution for 1 h at the hybridization temperature.

During prehybridization, 10 pm of the desired oligonucleotide were labeled using γ-P³² in a 20 µl reaction with T₄ polynucleotide kinase. The probe was incubated at 37°C for 45 min. The heated prehybridization solution was then discarded and fresh solution with labeled oligonucleotide was added. The membranes were hybridized overnight (approximately 18 h) at the temperature calculated for maximum hybridization for each microsatellite.

Following hybridization, membranes were rinsed once in a solution of 6× SSC and 0.1% SDS, then washed at the hybridization temperature for 15 min in the same buffer. Washed membranes were autoradiographed at -80°C using Amersham hyperfilm and two intensifying screens. For each probe, scoring clones was done on a scale of zero, representing no detectable hybridization, to three, a completely black spot on an autoradiograph representing maximum hybridization.

The exception to this procedure occurred when probing for TA. In order to prevent TA from self-annealing, the prehybridization was done at 70°C, with the hybridization solution preheated to 65°C. The probe was incubated at 37°C for 45 min, then at 65°C for 5 min prior to adding

to the hybridization tube, thereby minimizing temperature differentials. After the probe and fresh hybridization solution were added, the oven continued heating at 70°C for 30 min, then was lowered to the TA hybridization temperature of 47.7°C and left at that temperature overnight. Although GC also has the ability to self-anneal, its hybridization temperatures were not altered.

Stripping membranes of hybridization product consisted of washing twice in 0.4 M NaOH for 15 min at 42°C, then rinsing in 2× SSC. The membranes were washed at room temperature in neutralization solution (0.5M HCl and 1.5M NaCl) for 15 min, blotted dry, and checked to ensure that they were completely stripped. If radiation was still detectable, the membranes were washed twice for 15 min each at 62°C in a solution of 10× SSC and 0.5% SDS.

Probing the Cosmid Library With Emu Genomic DNA

In order to estimate the percentage single-copy DNA, the cosmid library was hybridized with emu genomic DNA. To determine if there was any low-copy hybridization, the film was exposed for 3 days. This was done with a high-temperature wash of library membranes in 0.1× SSC, 0.1% SDS at 65°C for 1 h. The prehybridization-hybridization mix contained the following: 6× SSC, 40% formamide (Amresco), 0.005 M EDTA, 1% SDS, and 0.005 g/ml Carnation evaporated milk. After discarding the high-temperature wash, half the prehybridization mix and the membranes were agitated for 1 h at 42°C. Probe preparation was as follows: 1 µl emu genomic DNA (0.5–1 µg), 1 µl each cold nucleotide, 5 µl 10× nick translation buffer, 5 µl α-P³² dCTP or dATP, 5 µl DNase/polymerase I, and 31 µl double-distilled water. The probe was then incubated at 15°C for 45 min, followed by addition of 50 µl 1× TE and 10 µl 0.5% SDS.

After incubation, spin column chromatography (Sambrook et al. 1989) was used to remove unincorporated label, then measured to determine the remaining volume. After adding one-ninth of the measured volume of 1 M NaOH to the probe, it was incubated for 10 min at 37°C. After discarding the agitating prehybridization mix, the remaining half of the hybridization mix, probe, and membranes were allowed to hybridize overnight at 42°C.

Following hybridization, the membranes were washed twice (with agitation) for 5 min in a 2× SSC and 0.1% SDS solution at

Table 1. Positive microsatellite hybridization to 1728 clones of a *D. novaehollandiae* cosmid library

Repeat	Total	Score			No. of sites/ genome	Interspersion distance (kb)
		1	2	3		
(C) ₅₀	18 (1.0)	6 (0.3)	7 (0.4)	5 (0.3)	402	3932
(CT) ₁₀	92 (5.3)	38 (2.2)	26 (1.5)	28 (1.6)	2054	769
(GC) ₅	1 (0.1)	1 (0.1)	0 (0)	0 (0)	22	70,784
(GT) ₉	231 (13.4)	90 (5.2)	97 (5.6)	44 (2.5)	5156	306
(TA) ₁₅	2 (0.1)	2 (0.1)	0 (0)	0 (0)	45	35,392
(AGC) ₁₇	87 (5.0)	55 (3.2)	22 (1.3)	10 (0.6)	1942	814
(AGT) ₁₇	18 (1.0)	13 (0.8)	4 (0.2)	1 (0.1)	402	3932
(CCG) ₁₇	114 (6.6)	95 (5.5)	18 (1.0)	1 (0.1)	2545	621
(GAA) ₇	10 (0.6)	5 (0.3)	4 (0.2)	1 (0.1)	223	7078
(GAC) ₆	105 (6.1)	74 (4.3)	18 (1.0)	13 (0.8)	2344	674
(GAT) ₇	3 (0.2)	1 (0.1)	1 (0.1)	1 (0.1)	67	23,595
(GCA) ₆	260 (15.0)	54 (3.1)	53 (3.1)	153 (8.9)	5804	272
(GGA) ₆	237 (13.7)	69 (4.0)	66 (3.8)	102 (5.9)	5290	299
(GGT) ₅	104 (6.0)	49 (2.8)	31 (1.8)	24 (1.4)	2321	681
(GTA) ₁₀	1 (0.1)	1 (0.1)	0 (0)	0 (0)	22	70,784
(GTT) ₆	13 (0.8)	6 (0.3)	2 (0.1)	5 (0.3)	290	5445
(TAA) ₉	21 (1.2)	11 (0.6)	3 (0.2)	7 (0.4)	469	3371
(GAAT) ₈	3 (0.2)	3 (0.2)	0 (0)	0 (0)	67	23,595
(GACA) ₆	89 (5.2)	19 (1.1)	12 (0.7)	58 (3.4)	1987	795
(GATA) ₆	3 (0.2)	0 (0)	2 (0.1)	1 (0.1)	67	23,595
(GGAA) ₇	6 (0.3)	3 (0.2)	3 (0.2)	0 (0)	134	11,797
(GTAT) ₇	20 (1.2)	6 (0.3)	2 (0.1)	12 (0.7)	446	3539
(GTTT) ₅	38 (2.2)	10 (0.6)	4 (0.2)	24 (1.4)	848	1863
Total	1476 (85.4)	611 (35.4)	375 (21.7)	490 (28.4)	32,946	48

Percentages are shown in parentheses. Number of sites in the genome and interspersion distances are calculated.

room temperature. They then were washed for 5 min in a 0.05 M NaCl and 0.1% SDS solution, agitating at 65°C. Membranes were autoradiographed at -80°C using Amersham hyperfilm and two intensifying screens. Clones displaying no hybridization were candidates for inserts consisting solely of unique sequence DNA. Then the membranes were stripped as stipulated by the protocol above.

Plate 1 produced results for the genomic hybridization that were questionable based on the number of clones and intensities of hybridization. Because there appeared to be technical problems associated with this plate, it was discarded from the results.

Because the methods employed herein are sensitive to the abundance of the repeat unit present in the probe, the genomic hybridization was exposed far longer than would be required to visualize the presence of a single repeat used as a probe. This technique was used to elucidate high-, middle-, and low-repetitive DNA. These same methods have been used to examine microsatellite distribution in humans (Longmire 1993), bats (Van Den Bussche et al. 1995), rodents (Janacek et al. 1993), insects (Hughes and Queller 1993), cotton (Baker et al. 1995), and other organisms, and therefore may be used as a means of comparison.

Distribution and Characterization of Single-Copy DNA and Microsatellites

Nakamura et al. (1990) estimated the genome size of emus to be 3.26 ± 0.02 pg

and 3.24 ± 0.02 pg in males and females, respectively, whereas Tiersch and Wachtel (1991) estimated the emu diploid genome size to be 3.25 ± 0.02 pg. Therefore in order to calculate the number of base pairs from the average of these figures, a conversion is made from grams to Daltons, where 1.661×10^{-24} g = 1 Da (Metzler 1977); therefore $3.25\text{pg} \div 1.661 \times 10^{-12}$ pg = 2.0×10^{12} Da. The remaining step requires conversion from Daltons to base pairs. According to Schleif (1981), 650 Da are approximately equal to 1 bp, but Li (1997) stated that 1bp = 617Da. Using the average of the two estimates, 2.0×10^{12} Da \div 633.5 Da averages to approximately 3.16×10^9 bp for the diploid genome.

To calculate average insert size, 13 of which hybridized to either a microsatellite or genomic probe, 20 random clones were picked, grown overnight at 37°C, and DNA was isolated using the QIAGEN Quick Prep kit. One microgram *EcoRI*-digested DNA was run on a gel with uncut λ , λ -*HindIII*, and 1 kb DNA ladders. From those, the percentage of the emu genome that the library represents was estimated using the average size of inserts multiplied by the number of clones screened; the resulting value was divided by the emu haploid genome size times 100.

To estimate the percentage of single-copy DNA, the number of clones with no hybridization to genomic DNA was divided by the total number of clones screened, then multiplied by 100. In order to exam-

ine microsatellite data, on the other hand, oligonucleotide frequencies were estimated as the number of positive clones divided by the total number of clones screened, then used to estimate the expected frequencies for double, triple, quadruple, and quintuple repeats per clone. To compare these with the observed data, a chi-squared test for significance was performed to determine potential linkage. The chi-squared test was corrected for error using the following formula: $(|E - O| - 0.5)^2/E$.

The number of sites was estimated by assuming that each positively scored clone contained a single copy of the element hybridized. The total number of sites for each microsatellite in the genome ($100 \div 4.48\%$ of the genome screened; the result multiplied by the number of clones with that microsatellite), interspersion rates [haploid genome size (1.58×10^9) divided by the total number of sites for that microsatellite], and confidence intervals (Zar 1996) are delineated in the Results section. The use of interspersion data assumes that repeats are regularly distributed. Confidence intervals (CI; 95%) were estimated for microsatellite frequencies and the number of sites in the genome using n = number of clones screened (1728); N = number of clones needed to equal the total genome (1.58×10^9 bp \div 40,940 = 38,593); p = frequency of chosen microsatellite; $q = 1 - p$; $T = p^*q$; $V(p)$, the sampling variance of p , is equal to $[(p^*q) \div (n - 1)][(N - n) \div N]$; CI for frequency = $p \pm 1.96\sqrt{V(p)}$; CI for number of sites in the genome = $T \pm 1.96\sqrt{V(T)}$.

Results

Microsatellites

All clones were probed with the array of microsatellites and all plates produced hybridization. Hybridization results of microsatellites are presented in Table 1, and confidence intervals for microsatellite frequencies are shown in Table 2. Pairwise values for co-occurrences per clone more frequently than expected at the $P = .001$ level were AGC GCA, AGT GAT, and GAC GGA. Pairwise values for co-occurrences per clone more frequently than expected at the $P = .005$ level were AGT GTT. Co-occurrence of three, four, and five microsatellites is presented in Tables 3-5. Six microsatellites co-occurring together occurred only once with a $P = .001$: CT GT GAC GCA GGA GTAT. Of the 1728 archived cosmid clones screened in this study, 872 had no hybridization to any of the 23 mi-

Table 2. Confidence Intervals (95%) for microsatellite frequencies and number of sites in the emu genome

Repeat	Frequency CI	Error (±)	Sites in genome CI	Error (±)
(C) ₅₀	0.010	0.005	401.635	180.446
(CT) ₁₀	0.053	0.010	2052.803	399.023
(GC) ₅	0.001	0.001	22.313	42.742
(GT) ₉	0.134	0.016	5154.321	604.825
(TA) ₁₅	0.001	0.002	44.626	604.825
(AGC) ₁₇	0.050	0.010	1941.2378	388.621
(AGT) ₁₇	0.010	0.005	401.635	180.446
(CCG) ₁₇	0.066	0.011	2543.691	441.181
(GAA) ₇	0.006	0.003	223.131	134.811
(GAC) ₆	0.061	0.011	2342.873	424.587
(GAT) ₇	0.002	0.002	66.9392	73.989
(GCA) ₆	0.137	0.016	5801.401	635.422
(GGA) ₆	0.137	0.016	5288.200	611.400
(GGT) ₅	0.060	0.011	2320.560	422.690
(GTA) ₁₀	0.001	0.001	22.3131	42.742
(GTT) ₆	0.008	0.004	290.070	153.574
(TAA) ₉	0.012	0.005	468.575	194.733
(GAAT) ₈	0.002	0.002	66.939	73.989
(GACA) ₆	0.052	0.010	1985.864	392.823
(GATA) ₆	0.002	0.002	66.939	73.989
(GGAA) ₇	0.003	0.003	133.879	104.545
(GTAT) ₇	0.012	0.005	446.262	190.095
(GTTT) ₅	0.022	0.007	847.897	260.644

crossatellite probes examined, representing 50.5% of the total cosmid library.

Four hundred ninety clones (28.4%) of the primary library probed with microsatellite repeats displayed maximal hybridization with the microsatellite probes tested. A moderate score was displayed by 375 clones (21.7%), while 611 (35.4%) showed low levels of hybridization. Of the four classes of microsatellites hybridized, trinucleotides were the most abundant, with 973 clones, followed by dinucleotides (326 clones), tetranucleotides (159 clones), and mononucleotides (18 clones). Trinucleotides were the most abundant class of microsatellites, followed by dinucleotides, tetranucleotides, and mononucleotides, although the percentages may change when the genome is probed with all remaining combinations of microsatellites.

Three hundred ninety-nine clones (23.1%) hybridized to more than one microsatellite probe, with 5.8% (23 clones) of those hybridizing to two or more probes with maximal hybridization. One hundred forty-six clones (8.4%) hybridized with three or more repeats, and only one of those had a maximal hybridization score for all three probes. Fifty-eight clones (3.4%) hybridized with four or more different probes, and 19 clones (1.1%) hybridized with five or more separate probes. Two clones (0.1%) hybridized to six microsatellite probes.

The size of the *D. novaehollandiae* DNA inserted into 40 randomly chosen recom-

Table 3. Chi-squared values for three co-occurring repeats per clone

P	Co-occurring repeats
.001	(GAC) (GGA) (GGT)
	(AGT) (GAC) (GAT)
	(AGT) (GAT) (GTAT)
	(GT) (GAC) (GGA)
	(AGT) (GAT) (GGA)
	(C) (GC) (GACA)
	(C) (GC) (CCG)
	(CCG) (GAC) (GGA)
	(C) (TA) (CCG)
	(GT) (CCG) (GAC)
.005	(GAC) (GAT) (GGA)
.025	(AGT) (GAC) (GGA)
.05	(GC) (CCG) (GACA)
	(GAC) (GCA) (GGA) ^a

^a Represents three co-occurring repeats hybridizing less often than expected.

binant cosmids ranged from 20,000 to 49,100 bp, with an average insert size of 40,940 bp. On the basis of this mean insert size, the 1728 recombinant cosmids represent 7.07×10^7 bp, or 4.48%, of the *D. novaehollandiae* genome. One clone failed to grow to a density that allowed detection of cosmid DNA with a miniprep procedure, or alternatively, the cells may not have had an insert.

By dividing the haploid genome size by the total number of sites (assuming one site per positive insert) in the genome, the emu genome is estimated to contain a microsatellite cluster approximately every 48 kb. Of the microsatellite hybridizations, 611 clones (35.4%) were assigned a score of one, while 375 clones (21.7%) were scored as a two (indicating an intermediate level of hybridization), and 490 clones (28.4%) were assigned a maximum value of three.

Confidence intervals (95%) were calculated for microsatellite frequencies, as well as for the total number of sites in the genome (Table 2). Because of the small number of hybridizing clones, considerable error surrounds the estimates for the following microsatellites: GC, TA, GAT, GTA, GAAT, and GATA.

Identification of Clones With Single-Copy DNA

Of the 1632 total clones probed to estimate the percentage of single-copy DNA, 643 (39.4%) displayed maximal hybridization, 220 (13.5%) were intermediate, and 202 (12.4%) had minimal hybridization. Five hundred sixty-seven clones (34.7%) showed no hybridization after 3 days. These may be approximately equivalent to high-, middle-, and low-copy repetitive DNA. The unhybridized clones are com-

Table 4. Chi-squared values for four co-occurring repeats per clone

P	Co-occurring repeats
.001	(AGT) (GAC) (GAT) (GGA)
	(C) (GC) (CCG) (GACA)
	(AGT) (GAT) (GCA) (GTAT)
	(AGT) (GAT) (GCA) (GGA)
	(GT) (AGT) (GAC) (GAT)
	(GT) (CCG) (GAC) (GGA)
	(CT) (CCG) (GACA) (GTAT)
	(GT) (AGT) (GAT) (GGA)
	(GT) (AGT) (GAC) (GGA)
	(CT) (GAC) (GGA) (GTAT)
	(CT) (GAC) (GCA) (GTAT)
	(GT) (GAA) (GAC) (GGA)
	(GT) (GAC) (GCA) (GGA)
	(GT) (AGC) (GTT) (GTTT)
	(GT) (AGC) (GAC) (GGA)
	(GT) (CCG) (GAC) (GGT)
	(CCG) (GAC) (GGA) (GGT)
	(CT) (GCA) (GGA) (GTAT)
	(GT) (AGT) (GAC) (GCA)
.005	(CT) (GT) (CCG) (GGA)
	(GAC) (GCA) (GGA) (GTAT)
.025	(C) (AGC) (GCA) (GTTT)
	(GT) (GAC) (GAT) (GGA)

prised of single-copy DNA, yielding a minimum value of 34.7% single-copy DNA present.

Discussion

Microsatellites, also known as variable number tandem repeats (VNTRs), are relatively short (usually less than 100 bp) clusters of tandem DNA motifs with repeat lengths of 1–6 bp (Li et al. 1997; Stallings 1992; Weber and May 1989) distributed throughout the euchromatic portions of the genome. They were discovered in 1989 (Litt and Luty 1989; Smeets et al. 1989; Tautz 1989; Weber and May 1989) and occur in all eukaryotic genomes studied thus far, as well as in some prokaryotic genomes (Tautz 1989).

Our study was designed to test five hypotheses concerning the distribution of microsatellites in the emu genome. In the

Table 5. Chi-squared values for five co-occurring repeats per clone

P	Co-occurring repeats
.001	(GT) (AGT) (GAC) (GAT) (GGA)
	(CT) (GAC) (GCA) (GGA) (GTAT)
	(GT) (AGT) (GAC) (GCA) (GTAT)
	(CT) (GT) (GAA) (GAC) (GGA)
	(GT) (CCG) (GAC) (GGA) (GGT)
	(CT) (GT) (GAC) (GGA) (GTAT)
	(CT) (GT) (GAC) (GCA) (GTAT)
	(GT) (AGC) (CCG) (GCA) (TAA)
	(GT) (CCG) (GAC) (GGA) (TAA)
	(GAC) (GCA) (GGA) (GGT) (TAA)
	(GT) (AGC) (GAC) (GCA) (GGA)
	(GT) (CCG) (GAC) (GCA) (GGA)
.005	(CT) (GT) (GCA) (GGA) (GTAT)
.025	(GT) (GAC) (GCA) (GGA) (GTAT)

Table 6. Microsatellite interspersions values and number of sites in human, white-footed mouse, bat, and cotton

Reference	Taxon	Genome size	Repeat	Inter-spersion distance (kb)	No. of sites
Longmire (1993)	<i>Homo sapiens</i>	3×10^9 bp	(C)	1736	1728
			(CT)	143	20,979
			(GC)	16,355	183
			(GT)	67	44,776
			(TA)	—	0
			(AGC)	2626	1142
			(AGT)	257	11,673
			(CCG)	1487	2017
			(GAA)	1105	3015
			(GCA)	76,087	39
			(GGA)	510	5882
			(GGT)	701	4280
			(GTT)	423	7092
			(TAA)	995	2715
			(GAT)	15,217	197
Baker (1994), Janacek et al. (1993)	<i>Peromyscus leucopus</i>	3×10^9 bp	(GATA)	972	3086
			(GT)	40	75,049
			(CT)	60	50,000
			(AT)	75,000	40
			(GC)	75,000	40
			(GGA)	441	6800
			(GATA)	217	13,800
			(GT)	104	22,239
			(CT)	115	20,000
			(AT)	—	0
Van Den Bussche et al. (1995)	<i>Macrotus waterhousii</i>	2.4×10^9 bp	(GC)	—	0
			(GT)	3370	485
			(CT)	922	1768
			(GC)	—	0
Baker et al. (1995)	<i>Gossypium hirsutum</i>	1.8×10^9 bp	(GT)	3370	485
			(CT)	922	1768
			(GC)	—	0
			(TA)	n/a	1

following, we discuss the data we have generated in light of these hypotheses.

Hypothesis 1: Microsatellites are Distributed Randomly Throughout the Emu Genome

The co-occurrence frequencies of different microsatellites document that microsatellites are not randomly distributed in the emu genome. This conclusion is based on the assumption that if a given microsatellite is randomly distributed in the genome, then its co-occurrence within a single clone with any other microsatellite should be a function of the frequency of the two microsatellites being compared.

Co-occurring repeats. Analysis of chi-squared values for pairwise co-occurrences per clone revealed four pairs of microsatellites that co-occurred more often than expected by chance alone ($P < .001$ and $P < .005$ levels of significance). Although encompassing survey data for many microsatellites in other genomes are lacking, Van Den Bussche et al. (1995) found GT and CT significantly co-occurring in the bat, *Macrotus waterhousii*. Kashi et al. (1990) found that GT may be co-occurring with GATA in cows. No microsatellites revealed pairwise values for co-occurrence less frequently than expected.

For co-occurrence of three, four, five, and six different probes, all significant chi-squared values documented a higher frequency of co-occurrence than expected. Just over 8.2% of clones that hybridized to three different probes indicates a higher frequency of co-occurrence with a chi-squared significance at or below 0.05. More than one-quarter (25.3%) of the clones that hybridized with four probes were significant at the $P < .025$ level or below. More than half (58.3%) of the clones that hybridized with five probes were significant at the $P < .025$ level or below and occurred more often than expected, and one of two clones with six co-occurring microsatellites was significant at the $P < .001$ level. These data are interpreted as indicating that microsatellite clusters are not randomly distributed throughout the emu genome, and that there are regions where multiple microsatellites preferentially co-occur.

Hypothesis 2: The Relative Order of Abundance of Dinucleotides Will be Constant Across Genomes

In emus, the relative order of abundance of the dinucleotides was GT, CT, CA, and GC. Data are available for humans (GT, CT, GC, TA; Longmire 1993), the white-footed

mouse (*Peromyscus leucopus*) (GT, CT, TA, GC; Baker 1994; Janacek et al. 1993), the leaf-nosed bat (*Macrotus waterhousii*) (GT, CT; TA and GC not detected; Van den Bussche et al. 1995), and cultivated cotton (*Gossypium hirsutum*) (CT, GT, TA; GC not detected; Baker et al. 1995). Clearly hypothesis 2 is rejected, since the order of dinucleotides is not a constant across genomes. However, GT and CT appear to be most common in all genomes, whereas TA and GC are much less common.

Hypothesis 3: Interspersions Distances for a Given Dinucleotide Will be Equal Across Vertebrate Genomes

The interspersions distance for GT in the emu is 306 kb. GT estimates from other birds include the chicken (*Gallus domesticus*) 200 kb (Cheng et al. 1995), peregrine falcon (*Falco peregrinus*) 300 kb (Longmire 1993), Canada goose (*Branta canadensis*) 407 kb (Longmire 1993), and birds in general, 136 kb (Primmer et al. 1997). The interspersions distances for GT in birds are within an order of magnitude.

In reptiles, the GT interspersions values range from 50 to 360 kb (Porter 1992; Villareal et al. 1996). In mammals, the GT interspersions distance estimates for humans range from 28 kb (Stallings et al. 1991) to 67 kb (Longmire 1993). Other estimates for mammals include 40 kb in the white-footed mouse (Baker 1994; Janacek et al. 1993), 42 kb in the dog (Rothuizen et al. 1994), 100 kb in the horse (Ellegren et al. 1992), and 104 kb in the leaf-nosed bat (Van den Bussche et al. 1995; Table 6). The variation in interspersions distances of GT is greater than an order of magnitude, and on this basis we reject hypothesis 3. Similar levels of variation for the other three nucleotides are recorded in the citations above.

Hypothesis 4: In all Genomes, a Dinucleotide Will be More Frequent Than any Trinucleotide

In previous studies screening for three or more microsatellites, GT and CT have usually been the most common (Baker 1994; Baker et al. 1995; Janacek et al. 1993; Longmire 1993; Van Den Bussche et al. 1995), however, the most common repeats in this study were trinucleotides—GCA, GGA—followed by GT. Their interspersions distances were every 272, 299, and 306 kb, respectively. Therefore hypothesis 4 is rejected. The distribution of trinucleotides versus dinucleotides in other vertebrates also indicates a pattern of high variation. For example, in the white-footed mouse,

there is a GGA every 441 kb and a GT every 40 kb, but no corresponding data were found for GCA (Janacek et al. 1993). In the leaf-nosed bat, one GT repeat is found every 104 kb (Van Den Bussche et al. 1995). There were no corresponding data for the other two microsatellites in *M. waterhousii*. On human chromosome 16, Longmire (1993) estimated one GCA repeat every 76 kb, one GT repeat every 67 kb, and one GGA repeat every 510 kb. These values indicate that GT is anywhere from 4 to 14.5 times more frequent in mammalian genomes than in the emu genome, but that GGA is less common in mammalian genomes by almost a third and that GCA is less common by 148 times.

Hypothesis 5: The Percentage of Single-Copy DNA Will Remain the Same in Emus as in Volant Birds

The minimum percentage of single-copy DNA in emus, 34.7%, is smaller compared to volant birds, which ranges from 39.8 to 65.5% (Shields and Straus 1975), so this hypothesis is also rejected. Data regarding the percentage of single-copy DNA, which includes most protein-coding genes, is important because it is arguably the most highly conserved portion of the genome (Li 1997). There are two kinds of single-copy DNA: protein-coding and sequences critical to developmental processes (gene expression, etc.), as opposed to single-copy sequences that are not critical to survival of the species. The second type has no stabilizing selection effect and is thus permitted to evolve into a unique set of sequences that provide no benefit to the organism.

Eden and Hendrick (1978) report that the genome of the ostrich contains 87% single-copy DNA. While the emu single-copy value appears to be low in comparison, this may be due to the fact that a clone hybridization scored as a one, two, or three represents repetitive DNA. It is unlikely, however, that the entire insert is in fact of a repetitive nature. C_0t curves provide a more accurate method for extrapolating the values for single-copy DNA. However, when the method used herein was compared to the values generated from C_0t curves in cotton, the values were similar (Baker et al. 1995), indicating that the method does provide insight into the size of fragments that are single-copy DNA.

C Value

With respect to genome size, birds have the smallest level of interspecific variation

of any vertebrate class (Cavalier-Smith 1978; Sparrow et al. 1972; Tiersch and Wachtel 1991). The fact that the *C* value in ratites is larger than that of most volant birds indicates that there is more DNA per cell than is needed to perform the genetic functions that are required. It therefore remains a viable hypothesis that a mechanism exists that selects against a *C* value increase in volant birds. While microsatellites may play a role in the increase in *C* for ratites versus volant birds, they do not appear to be the primary source for variation in *C* (Eden and Hendrick 1978; Shields and Straus 1975).

This study helps us understand ratite systematics by supplementing the sparse information available and providing information on representation in a putatively less-derived species. However, while it may be feasible to extract evolutionary conclusions from these data, it is important to remember that microsatellite expansion and contraction is a dynamic process, and presently it is impossible to discern in which direction the process is proceeding.

References

- Baker RJ, 1994. Some thoughts on conservation, biodiversity, museums, molecular characters, systematics and basic research. *J Mamm* 75:277–287.
- Baker RJ, Longmire JL, and Van Den Bussche RA, 1995. Organization of repetitive elements in the upland cotton genome (*Gossypium hirsutum*). *J Hered* 86:178–185.
- Baker RJ, Maltbie M, Owen JG, Hamilton MJ, and Bradley RD, 1992. Reduced number of ribosomal sites in bats: evidence for a mechanism to contain genome size. *J Mamm* 73:845–858.
- Cavalier-Smith T, 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth, and the solution of the DNA C-value paradox. *J Cell Sci* 34:247–278.
- Cavalier-Smith T, 1985. Introduction: the evolutionary significance of genome size. In: *The evolution of genome size*. New York: Wiley; 1–36.
- Cheng HH, Levin I, Vallejo RL, Khatib H, Dodgson JB, Crittenden LB, and Hillel J, 1995. Development of a genetic map of the chicken with markers of high utility. *Poult Sci* 74:1855–1874.
- Eden FC and Hendrick JP, 1978. Unusual organization of DNA sequences in the chicken. *Biochemistry* 17: 5838–5844.
- Ellegren H, Johansson M, Sandberg K, and Andersson L, 1992. Cloning of highly polymorphic microsatellites in the horse. *Anim Genet* 23:133–142.
- Hartl DL, 2000. Molecular melodies in high and low *C*. *Nat Rev Genet* 145–149.
- Hinegardner R, 1976. Evolution of genome size. In: *Molecular evolution* (Ayala FJ, ed). Sunderland, MA: Sinauer Associates; 179–199.
- Hughes CR and Queller DC, 1993. Detection of highly polymorphic microsatellite loci in a species with little allozyme polymorphism. *Mol Ecol* 2:31–137.
- Janacek LL, Longmire JL, Wichman HA, and Baker RJ, 1993. Genome organization of repetitive elements in the rodent, *Peromyscus leucopus*. *Mamm Genet* 4:374–381.

- Kashi Y, Iraqi F, Tikochinki Y, Ruzitski B, Nave A, Beckmann JS, Freidmann A, Soller M, and Gruenbaum Y, 1990. (TG)_n uncovers sex-specific hybridization pattern in cattle. *Genomics* 7:31–36.
- Li SH, Huang Y-J, and Brown JL, 1997. Isolation of tetranucleotide microsatellites from the Mexican jay *Aphelocoma ultramarina*. *Mol Ecol* 6:499–501.
- Li W-H, 1997. Genome organization and evolution. In: *Molecular evolution* (Li W-H, ed). Sunderland, MA: Sinauer Associates; 379–418.
- Litt M and Luty JA, 1989. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 44:397–401.
- Longmire JL, 1993. Distribution and organization of repetitive DNA sequences on human chromosome-16 (PhD dissertation). Lubbock: Texas Tech University.
- Metzler DE, 1977. *Biochemistry: the chemical reactions of living cells*. New York: Academic Press.
- Nakamura D, Tiersch TR, Douglass M, and Chandler RW, 1990. Rapid identification of sex in birds by flow cytometry. *Cytogenet Cell Genet* 53:201–205.
- Porter CA, 1992. Genome organization in squamate reptiles: ribosomal genes and other repetitive sequences (PhD dissertation). Lubbock: Texas Tech University.
- Primmer CR, Raudsepp T, Chowdhary BP, Moller AP, and Ellegren H, 1997. Low frequency of microsatellites in the avian genome. *Genome Res* 7:471–482.
- Rayburn AL and Auger JA, 1990. Genome size variation in *Zea mays* ssp. *mays* adapted to different altitudes. *Theor Appl Genet* 79:470–474.
- Rothuizen J, Wolfswinkel J, Lenstra JA, and Franz RR, 1994. The incidence of mini- and micro-satellite repetitive DNA in the canine genome. *Theor Appl Genet* 89: 403–406.
- Sambrook J, Fritsch EF, and Maniatis T, 1989. *Molecular cloning: a laboratory manual*, 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Schleif RF, 1981. *Practical methods in molecular biology*. New York: Springer-Verlag.
- Sessions SK and Larson A, 1987. Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. *Evolution* 41: 1239–1251.
- Shields GF and Straus NA, 1975. DNA-DNA hybridization studies of birds. *Evolution* 29:159–166.
- Smeets AJM, Brunner HG, Ropers H-H, and Wieringa B, 1989. Use of variable simple sequence motifs as genetic markers: application to study of myotonic dystrophy. *Hum Genet* 83:245–251.
- Sparrow AH, Price HJ, and Underbrink AJ, 1972. A survey of DNA content per cell and per chromosome of prokaryotic and eukaryotic organisms: some evolutionary considerations. In: *Evolution of genetic systems* (Smith HH, ed). New York: Gordon and Breach; 451–494.
- Stallings RL, 1992. CpG suppression in vertebrate genomes does not account for the rarity of (CpG)_n microsatellite repeats. *Genomics* 17:890–891.
- Stallings RL, Ford AF, Nelson D, Torney DC, Hildebrand CE, and Moyzis R, 1991. Evolution and distribution of (GT)_n repetitive sequences in mammalian genomes. *Genomics* 10:807–815.
- Swift H, 1950. The constancy of deoxyribose nucleic acid in plant nuclei. *Proc Natl Acad Sci USA* 36:643–654.
- Szarski H, 1974. Cell size and nuclear DNA content in vertebrates. *Int Rev Cytol* 44:93–111.
- Tautz D, 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* 17:6463–6471.
- Tiersch TR and Wachtel SS, 1991. On the evolution of genome size of birds. *J Hered* 82:363–368.
- Van Den Bussche RA, Longmire JL, and Baker RJ, 1995.

How bats achieve a small C-value: frequency of repetitive DNA in *Macrotus*. *Mamm Genome* 6:521–525.

Villareal X, Bricker J, Reinert HK, Gelbert L, and Bushar LM, 1996. Isolation and characterization of microsatellite loci for use in population genetic analysis in the

timber rattlesnake, *Crotalus horridus*. *J Hered* 87:152–155.

Weber JL and May PE, 1989. Abundance class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 44:388–396.

Zar JH, 1996. *Biostatistical analysis*. Englewood Cliffs, NJ: Prentice Hall.

Received March 2, 2001

Accepted December 31, 2001

Corresponding Editor: Susan J. Lamont