Published online 25 September 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 18* ***e121***
doi:10.1093/nar/gkl478

Capturing genomic signatures of DNA sequence variation using a standard anonymous microarray platform

C. H. Cannon*, C. S. Kua[1], E. K. Lobenhofer[2] and P. Hurban[2]

Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409-3131, USA, [1]27 Jln. Dato Haji Harun, Taman Tayton View, Kuala Lumpur, Malaysia and [2]Paradigm Array Labs, a Service Unit of Icoria, Inc., Research Triangle Park, NC 27709, USA

Received March 30, 2006; Revised June 19, 2006; Accepted June 20, 2006

ABSTRACT

Comparative genomics, using the model organism approach, has provided powerful insights into the structure and evolution of whole genomes. Unfortunately, only a small fraction of Earth's biodiversity will have its genome sequenced in the foreseeable future. Most wild organisms have radically different life histories and evolutionary genomics than current model systems. A novel technique is needed to expand comparative genomics to a wider range of organisms. Here, we describe a novel approach using an anonymous DNA microarray platform that gathers genomic samples of sequence variation from any organism. Oligonucleotide probe sequences placed on a custom 44 K array were 25 bp long and designed using a simple set of criteria to maximize their complexity and dispersion in sequence probability space. Using whole genomic samples from three known genomes (mouse, rat and human) and one unknown (*Gonystylus bancanus*), we demonstrate and validate its power, reliability, transitivity and sensitivity. Using two separate statistical analyses, a large numbers of genomic 'indicator' probes were discovered. The construction of a genomic signature database based upon this technique would allow virtual comparisons and simple queries could generate optimal subsets of markers to be used in large-scale assays, using simple downstream techniques. Biologists from a wide range of fields, studying almost any organism, could efficiently perform genomic comparisons, at potentially any phylogenetic level after performing a small number of standardized DNA microarray hybridizations. Possibilities for refining and expanding the approach are discussed.

INTRODUCTION

Comparative genomics is a fundamental and quickly developing evolutionary approach (1,2). While an increasing diversity of genomes are currently being sequenced (3) and promising new technologies could greatly reduce the cost and speed up the process of whole genomic sequencing (4,5), the vast majority of Earth's biodiversity will not have its genome sequenced in the near future. Given the current pace of large-scale environmental change, particularly in the tropics (6–8) where biodiversity is greatest (9), novel techniques are required for biologists to rapidly develop the genomic resources for understudied organisms. Additionally, many of these tropical organisms, like rainforest trees, possess radically different life history strategies and evolutionary dynamics than current model organisms, which are short-lived with simple genomes, implying a rather limited amount of knowledge gained from current approaches will be transferable. The ability to understand historical patterns of genomic diversity created over geological and glacial time scales is essential for the future management of natural populations. Human activities are erasing the traces of these patterns before we can define them. This ignorance will lead to an extinction of the historical past, which we desperately need to properly interpret the present and plan for the future. While a certain amount of information can be leveraged out of available whole genome sequences (2), a fast and direct method for gathering genomic samples of genetic variation from previously unstudied organisms is needed.

Here, we present an anonymous DNA microarray capable of capturing genomic signatures of DNA sequence variation from any organism using only a few hybridizations. The microarray probe sequences are generated using a pre-determined set of selection criteria, which could be modified and refined with increasing knowledge and specificity of application. In brief, the probe sequences are SHyPs: short (25 bp and less), hyperdispersed in sequence probability space, *anonymous* because they are generated without knowledge of the target genome and primers as they should fit optimality criteria for a PCR

*To whom correspondence should be addressed. Tel: +1 806 742 3993; Fax: +1 806 742 2963; Email: chuck.cannon@ttu.edu

primer. A theoretical framework of such an approach, aimed at expressed sequences, has been described (10) and other microarray based approaches have targeted specific groups of organisms (11). The criteria used for the selection of the SHyP probes should harvest presence/absence information for a large and complex set of DNA sequences scattered throughout the entire genome. Major changes in copy number of these sequences, due to genomic re-organization or proliferation of certain elements, should be detectable (12). After construction of a genomic signature database, virtual comparisons between any subset of genomes would be possible and simple queries would generate optimal sets of oligonucleotide sequences to distinguish among target genomes. These informative markers could be used in larger-scale and cheaper downstream macroarray or PCR-based screening techniques. The results from comparative analyses of two genomes will also offer an access point directly related to the interesting and different regions of the targets.

To prove the basic elements of the concept, we hybridized three known genomes (human, mouse and rat) and the previously undescribed genome of a tropical tree species ('ramin': Thymelaeaceae:*Gonystylus bancanus*) to a custom 44 K feature microarrays fabricated by Agilent's SurePrint technology. The use of whole genomic DNA has been shown to be reliable in these types of hybridizations (13). 'Ramin' is an endangered species found only in peat-swamp forests along the inner margins of the South China Sea (14) and is vulnerable to extinction (15). Two ramin population samples, one from the east coast of Sumatra and one from the northwestern coast of Borneo, each composed of five individuals, were compared to discover genetic markers related to their geographic origin. The study design will allow us to test the sensitivity of the approach, in this first iteration, to a broad range of genomic relatedness and the reliability and transitivity of hybridization signal across hybridization experiments. A simple and direct analysis is also presented to take advantage of the replication of hybridization experiments to identify the oligonucleotide probe sequences that provide results with relatively little variance across genomes and experiments. A BLAST study of a subset of probe sequences was performed to examine the distribution of these sequences in the known genomes and its correlation with observed hybridization intensities. We also discuss potential improvements in the probe sequence, microarray design and experimental procedures.

MATERIALS AND METHODS

Generation of SHyP microarray probes

Our strategy was to construct oligonucleotide probe sequence of 25 bp using a random sequence generator and several filtering criteria. These criteria included:

(i) no homopolymeric tracts (probes containing >3 bp of the same type were excluded from consideration).
(ii) no potential hairpin sequences (probes containing stretches of >3 complementary bp were excluded from consideration).
(iii) optimal GC content (probes containing outside of the range of 45–60% GC content were excluded from the design process).
(iv) no sequence redundancy: probes that shared a region of exact sequence match over $\geqslant$60% of probe length between any other passed probe or any of a set of 357 universal primers were excluded from consideration.

Numerous iterations were performed to generate 44 000 anonymous probe sequences fitting these criteria. Measures discussed below of 'maximum sequence identity' refer to the maximum amount of identical sequence between two oligonucleotide probes, within any 17 bp window of comparison.

Preparation and hybridization of genomic DNA

Samples of *G.bancanus* (ramin) were obtained through the cooperation of the Forest Research Center in Kuching, Malaysia and P.T. Diamond Raya Timber, Pekanbaru, Sumatra, Indonesia, and whole genomic DNA was extracted using a standard CTAB protocol. 'Population samples' for ramin were created by mixing equal amounts of genomic DNA of five different individuals from each population (Sarawak individuals = 'RaminSK5'; Sumatra individuals = 'RaminSU5'). To assess the impact of false positives, a self-self hybridization experiment was also performed using the rat genomic DNA (Figure 1A).

The complexity of mouse (Promega), human (Promega), rat (EMD Biosciences) and ramin genomic DNA was reduced by restriction endonuclease digestion. Recognizing that the active site for any single enzyme could involve a fairly large percentage of the SHyP probe sequences, e.g. both AluI and RsaI would affect 17.2% of the probes, we used a combination of three restriction endonucleases (MboI, AluI and RsaI: NEB) and performed separate digestions with each of the enzymes and then pooling these DNAs together prior to enzymatic labeling, the restriction reactions affect only 4 probes (0.009%) on the microarray. Purified digestion products were assessed on an Agilent Bioanalyzer to assess the distribution of fragment sizes before proceeding to labeling reactions. Four micrograms of pooled genomic DNA was labeled using either Cy_3 or Cy_5 following the BioPrime Array CGH Genomic Labeling System (Invitrogen). Purified labeling reaction products were quantified on a Nanodrop-1000 Spectrophotometer to determine the amount of product prior to hybridization. Labeled products were hybridized to a SHyP array for 40 h at 65°C before the microarray was washed and scanned using an Agilent DNA Microarray Scanner BA. The resulting image was then processed through Agilent's Feature Extraction software (version 7.5.1) in order to obtain intensity measurements and determine statistical difference in intensity level.

Each comparison between genomes was performed twice, using each of the labeling dyes (Figure 1A). Subsequent analyses across hybridizations were performed on the average intensity from these two fluor-flipped hybridizations to minimize any labeling bias or experimental effects. Therefore, five genomic comparisons were performed using ten separate microarrays: (i) mouse><rat; (ii) mouse><human; (iii) rat><human; (iv) rat><rat and (v) raminSK5><raminSU5.

Standard microarray analysis

'Genome-indicator' probes were determined by (i) discarding all probes with a hybridization intensity value below the mode intensity value for each hybridization and (ii) choosing

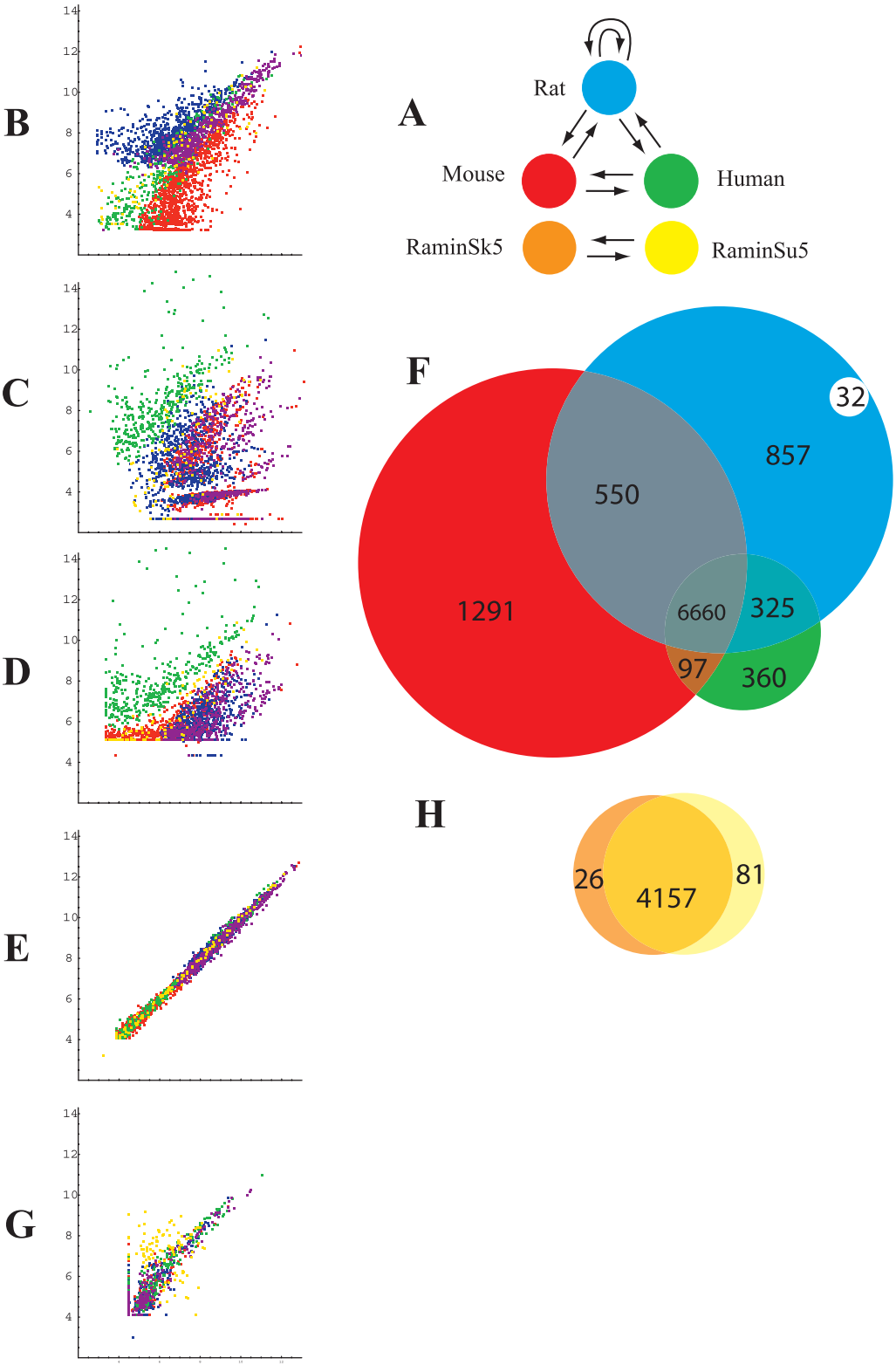

Figure 1. Affinity of four genomes to the SHyP array. (**A**) Experimental design for genomic comparisons. Each arrow indicates a direct hybridization experiment in which the indicated genome was labeled with Cy3 and compared to the other genome, labeled with Cy5. The rat genome was used in a self-self hybridization. Color-coding of genomes is consistent (B-E,G). Log[Average] hybridization intensity across for each genomic comparison: (**B**) mouse-rat; (**C**) mouse-human; (**D**) rat-human; (**E**) rat–rat and (**G**) raminSk5-raminSu5. Only 'indicator' probes for at least one genome are shown, including 'rodent' as purple (see Materials and Methods). (**F**) The results from the direct comparison among the three known genomes are shown in the overlapping circles. The central number indicates the probes with apparent hybridization to all four genomes. The number of probes unique to each genome are shown in the outer arcs. The interstices illustrate the number of probes common to both genomes. The small white circle in the rat circle represents the false positives observed in the rat–rat hybridization. (**H**) The 'population-indicator' are color-coded for the ramin species, while non-significant probes are shown in gray.

probes that were significantly 'up-regulated' in relation to the other genomes, at a $P < 1 \times 10^{-8}$ confidence limit, as determined by image analysis using Agilent's gene expression software. This significance level is highly stringent and removed most of the false positives, as determined in the rat–rat hybridization. A subset (*3190* probes) of high intensity sequences in all mammal genomes was mapped onto the known chromosome database at NCBI, using the basic BLAST search tool, to examine the genomic distribution of these sequences.

Low variance analysis

Because the statistical analysis used in standard gene expression software is concerned with the up or down regulation of expressed sequences known to be present in the target genomes, low to no hybridization signal in one or both channels is usually interpreted as failure of the probe. This pattern, where hybridization fails in one genome but is reliably detected in the other, is ideal to capture informative genomic signatures. We have found in both downstream PCR work and *in silico* BLAST studies that probes with high signal values but 'upregulated' in one channel, are actually frequent in both genomes. While changes in copy number across genomes is important, the current approach is focused on detecting the presence and absence of evolutionarily labile DNA sequences across genomes.

In order to adopt the analysis of the resulting array image to this purpose, a simple protocol for establishing intensity thresholds and stability across hybridizations was developed. First, a threshold was determined for 'detectable' signal by simply examining the overall pooled hybridization intensities for each genomic comparison and throwing out all data that fell below the mode of the distribution. Distributions of signal are strongly skewed towards 'empty' values because most probes are not present in any one genome. A second threshold was determined by examining the variance of the hybridization signals near the initial threshold value. A higher threshold where detectable signals were stable was established for each hybridization. Probes with an average signal in all genomes below this second threshold were discarded. Signal levels were then compared across hybridizations for each genome and only those probes that varied <40% of the average were then considered 'low variance' probes. In this context, average signal above the second threshold was determined as 'presence' while average signal strength below was considered 'absence' in the target genome. With repeated experiments (only the rat is present in more than two hybridization), this standardization process will become more powerful and reliable. This analysis should be considered very stringent and not mutually exclusive to the gene expression analysis. The results of both will be included in future publications and databases. The Mathematica notebook to perform this analysis is available for download on the lead author's website.

RESULTS

Probe characteristics

To test for the effects of our selection criteria on the SHyP oligonucleotide sequences, we compared a subset of our probes to 3000 sequences generated in a completely random fashion. The SHyP probes were only slightly more similar to one another than random expectations (9.19 versus 9.13 maximum sequence identity within any 17 bp window). The standard deviation (1.4 bp) of the average maximum sequence identity was the same between the random and SHyP probes. Additionally, the bias towards GC enrichment slightly raises the average content to 52% for the SHyP probes. Each of the eight possible heteromeric dinucleotide combinations were present in repeats of two in roughly 14% of the probes and the proportion followed a Poisson distribution with repeats of four dinucleotides found in <0.1% of the probes (data not shown). The probe set can be downloaded from the lead author's webpage (http://www.faculty.biol.ttu.edu/cannon/genomic_signatures.htm).

Over half of the probes produced no detectable hybridization signal with any genome (*23 531* out of *42 033* SHyP probes). The mouse genome had the greatest affinity for the SHyP array with *11 611* probes producing detectable signal, while the ramin genome had roughly an order of magnitude less affinity (*1163* detectable probes). The three mammalian genomes had an observable affinity for the same 15% of the probes (*6752* out of *42 033*). The average maximum sequence identity among these detectable mammal probes was 9.18 bp but the standard deviation (2.9) was much greater than the entire set. Among the probes with detectable signal for the ramin genomes, maximum sequence identity and GC enrichment were both high (9.54 bp and 54%, respectively). A large fraction of these detectable probes were discarded in the following more stringent analyses. Overall, these numbers indicate that the majority of the array did not hybridize with any of the study genomes and the resolving power of the array for future genomic comparisons is far from being saturated.

Standard microarray analysis

The mouse genome consistently hybridized to the greatest number of 'indicator' probes, at any level of statistical significance (Table 1). Across the range of significance levels ($P < 0.01$ to $P < 1 \times 10^{-11}$), the number of 'indicator' probes for the three mammal genomes drops by roughly a factor of four while the rate of decline is sharper for the tree genome. The rate of false positives declines very rapidly, by well over two orders of magnitude. In comparison to the overall set of SHyP probes, most genomes demonstrate an increased degree of maximum sequence identity within each set of 'indicator' sequences, at any significance level, and this maximum sequence identity increased at higher levels of statistical stringency. 'Indicator' probes for rat and rodent genomes have higher levels of maximum sequence identity than the other genomes. Hybridization intensity for no genome had any appreciable relationship with simple GC content.

Among the known genomes, intensity values were highly correlated across hybridization experiments (Table 2) and the 'indicator' probes were clearly distinguished in each pair-wise comparison (Figure 1B–E). The two rodent genomes, individually and as a clade, had a much larger 'signature' on the SHyP array than the human genome (Figure 1F). This result matches the reported accelerated rate of genomic evolution in the rodents, in comparison to humans (16).

Table 1. Effect of significance level on informative probes

***P*-value**	N	**Mouse** 3965	**Rat** 2243	**Human** 897	**Rodent** 877	**Ramin** 105/303	**False** 914
0.01	**Id**	9.32 (1.5)	9.59 (1.5)	9.20 (1.9)	9.73 (1.5)	9.41 (1.6) / 9.3 (1.8)	9.49 (1.5)
	GC%	52	52	53	52	54/53	52
	N	2808	1303	516	734	43/138	117
1×10^{-5}	**Id**	9.33 (1.5)	9.66 (1.5)	9.19 (1.9)	9.81 (1.5)	9.43 (1.6) / 9.16 (1.8)	9.49 (1.5)
	GC%	52	52	53	52	53/52	52
	N	1291	857	360	550	26/81	32
1×10^{-8}	**Id**	9.35 (1.5)	9.70 (1.5)	9.20 (1.9)	9.88 (1.5)	9.45 (1.5) / 9.17 (1.8)	9.40 (1.4)
	GC%	52	52	53	52	53/52	53
	N	870	616	266	383	14/53	6
1×10^{-11}	**Id**	9.37 (1.5)	9.71 (1.5)	9.20 (2.0)	9.93 (1.5)	9.66 (1.6) / 9.29 (1.8)	9.40 (1.5)
	GC%	52	52	53	52	53/52	53

P-value is shown at four levels, increasing sequentially by three orders of magnitude. This value is taken directly from Agilent's image analysis software. Values for genomic compartments are shown at each level of significance, including each individual genome, including each ramin population. The false positives were identified in the rat:rat hybridization. For each *P*-value, the number (**N**) of indicator probes for each sample are shown, the average maximum sequence identity (**Id**) in bp (standard deviation shown) and (**GC%**) content for each set of informative probes.

Table 2. Correlation among hybridizations for each genome

	Mouse	Rat	Human	Ramin
Mouse	0.93/0.98	0.83	0.11	0.41
Rat	0.74	0.94/.97	0.15	0.55
Human	0.10	0.15	0.99/0.97	0.10
Ramin	0.38	0.54	0.11	0.98/0.98

The values along the diagonal indicate the results produced from different comparisons for the same genome. The first value represents all probes, the second value includes low variance probes. The comparisons between genomes in the upper right half of the table indicate correlation values for low variance probes; all probes are shown in the lower left half.

A very low rate of false positives was detected (Figure 1F). While the overlapping 'indicator' probes for the rodents largely indicate shared descent, the overlapping probes between human:rat and human:mouse are clearly homoplasious. The degree of homoplasy seems to be substantially higher between the rat and the human.

The single plant genome, *G.bancanus*, used in these experiments was not directly compared with the three known genomes but instead a population level comparison was performed by mixing five individuals, each population found on different landmasses in Southeast Asia (Sumatra, Indonesia and Sarawak, Malaysia). In general, the genomes appeared to have less affinity for the SHyP array than the three mammalian genomes but this might have been due to the unknown nature of the genome and lower absolute amounts of whole genomic DNA. While no direct comparison was made, the hybridization signal can be compared across experiments. The ramin genome only produced *122* 'indicator' probes, when compared to all three mammals, but individual comparisons with mouse and human revealed *959* and *1137* ramin 'indicator' probes, respectively. The correlation of hybridization intensities between these two populations was quite high similar (Figure 1G) but over *107* probes were significantly different between the two populations.

Low variance analysis

Less than half of the detectable probes (10 935) produced enough signal to be included in the low variance analysis and three quarters of these 'active' probes were rejected because of high variance between hybridizations, leaving only *3453* 'low variance' probes. The low variance analysis applied here should be considered highly stringent. By definition, no false positives are allowable in the self-self comparison. Correlation values generally improved when only the low variance probes were considered (Table 2), except for the human genome where correlation values actually went down slightly. For inclusion in the low variance analysis, signal variance for each probe was averaged across all genomes, so the results should be expected to be biased towards the two rodent genomes, where intensity signals were highly correlated (Figure 2). As the taxonomic balance of the genomic signature database improves, this type of bias should disappear. The overall correlation among genomes roughly followed phylogenetic relatedness, although the signal for the human genome was not as highly correlated with the other genomes and was substantially less correlated with signal for the ramin genome. Low variance probes also produce higher correlation values among genomic comparisons than when all probes are considered, except in the comparison between humans and ramin.

The phylogenetic signal becomes even more obvious when the low variance probes are broken down into classes based upon their presence or absence in each genome (Table 3). 'Rodent' probes (729) were slightly more frequent than 'Mammal' probes (682) while 'Mouse' probes (644) were the most common probes private to any one genome. Relatively few probes (224) were present in all four genomes. Several examples of homoplasious relationships among sets of genomes were observed: *146* probes unite rat:humans while *42* unite mice:humans. These probes could occur either through the loss or gain of the sequence within one member of the mammal clade or convergent evolution. Probes homoplasious between plants and mammals were quite rare but all possible combinations were present. The two rodent species demonstrated these homoplasious patterns in reverse order to their frequency of private probes. In both cases of human and ramin, the rat was more highly homoplasious than the mouse.

The hybridization signal of each genome within each class of 'low variance' probe was correlated to the phylogenetic

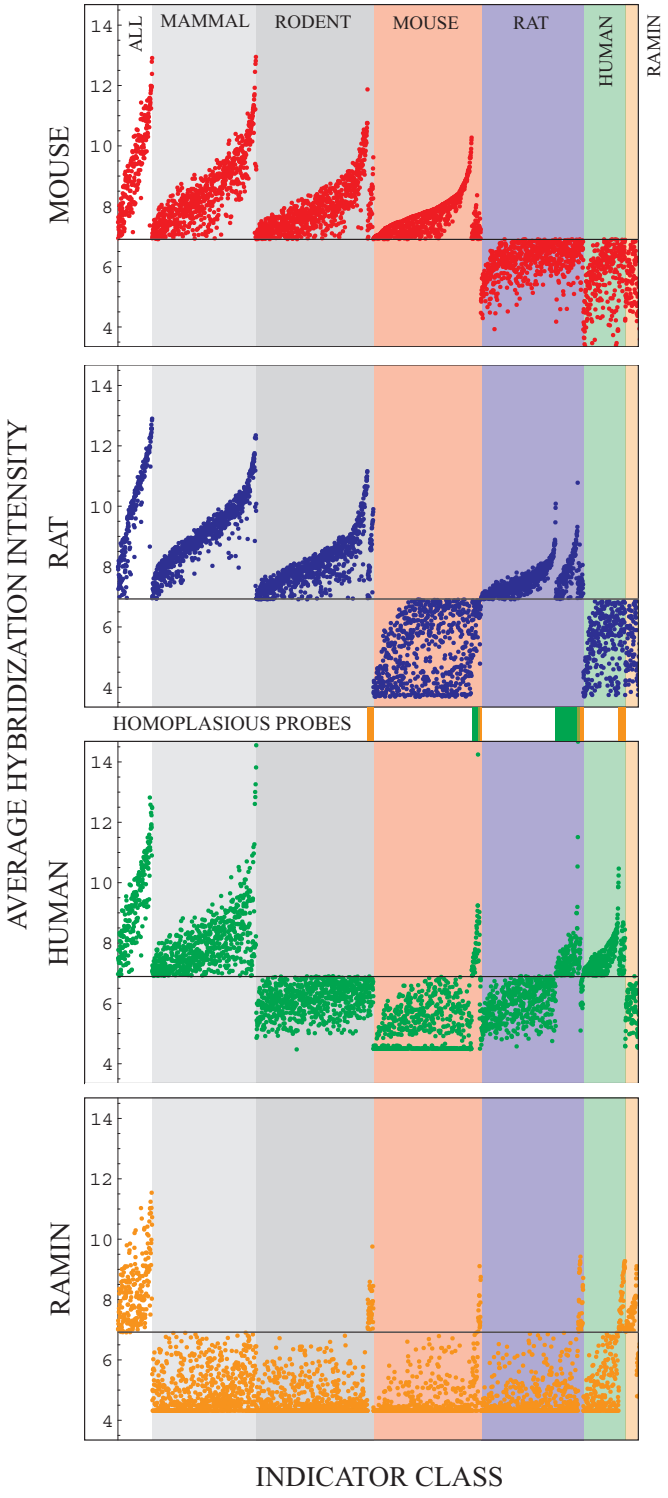

Figure 2. Affinity of four genomes with low variance SHyP probes. Genomes are color-coded as in Figure 1: mouse (red), rat (blue), human (green) and ramin (orange). Each color-coded column represents a specific class of indicator probes: all genomes (white, $n = 224$), mammal (light gray, $n = 682$), rodent (dark gray, $n = 729$), mouse (pink, $n = 644$), rat (light blue, $n = 491$), human (light green, $n = 233$) and ramin (light orange, $n = 76$). Each class of indicator probes are ordered from low to high average Log[intensity] across all genomes. The solid line indicates the cut-off for presence/absence of signal, as determined by our statistical analysis. Obviously homoplasious probes are indicated in the white space between the rat and human graphs. Homoplasious probes found in both human and ramin are beige.

Table 3. Distribution of low variance probes across genomes and groups

		Detectable				
	Private Probes	***Mm***	***Rn***	***Hs***	***Gb1***	***Gb2***
Rodent	729	225	148			3
Mammal	682	105	139	65	6	7
Mouse (*Mm*)	644		2			3
Rat (*Rn*)	491	2		6		3
Human (*Hs*)	233				1	1
All	224	26	27	33	7	24
Rn-Hs	146		17	40		1
Ramin *(Gb)*	131		1	9	2	10
	20/71					
Mm-Hs	42	3		18		
Hs-Gb	41			28	1	3
Rodent-*Gb*	38	6	21		2	4
Rn-Gb	18		5		1	3
Rn-Hs-Gb	16			4		3
Mm-Hs-Gb	10			8		4
Mm-Gb	8	3	1			2

Private Probes indicate the number of sequences with significantly stronger hybridization intensity for that group while 'detectable' probes were present but at a low level in other groups.

relatedness of these genomes as well (Figure 2). A general trend in decreasing peak and average signal can be seen as the classes become less and less inclusive. The greatest signal was generally observed among 'all' and 'mammal' probes. This increased signal intensity is related to the repetitive nature of many of these probe sequences. Most of the 'all' probes have strings of dinucleotide and trinucleotide repeats, which are common in all tested genomes. These probes are particularly enriched for 'CA' repeats but not for other heteromeric repeats, which were equally frequent on the array. Strings of monomeric repeats, such as 'TA', were generally excluded given the selection criteria. The mammal class is particularly enriched with 'TC' and 'TCC' repetitive motifs, while the rodent class is enriched for 'GC' and 'GCC' repetitive motifs, although not as strongly. The human class is strongly enriched for 'TGG' repeats, which occur twice in several of the most common 11 letter words. While the rat, human and ramin probe classes are generally more similar to one another than would be expected, the mouse class is actually quite close to the overall similarity of probe sequence across the entire array.

Population level comparisons

Because the two ramin genomes were not identical but were obtained from two populations on different landmasses, some fraction of the probes informative to the intraspecific comparison will have been excluded in the low variance analysis. Despite this bias, the proportion of probes, which were present in the ramin genome versus the other three was greater than would be expected, given the overall number of low variance probes, which had substantial signal for ramin. This result indicates that this class seems to be enriched in terms of indicator probes, in comparison to the other three much more closely related genomes. Additionally, a large proportion of the ramin probes appear to be also significantly 'up-regulated' in one population versus the other (Table 1: numbers in bold). The first ramin individual in this comparison was from a relatively small population from the Malaysian state of Sarawak while the second individual

was collected along the northeastern coast of Indonesian Sumatra, where very large populations of this species can be found. The relative proportion of private probes in each population is probably a result of this difference in overall historical population size.

Genomic distribution of high intensity probe sequences

The probe sequences are evenly distributed through the known chromosome structure of the target genomes (Figure 3A), with an obvious positive correlation in the length of the chromosome and the number of BLAST hits (Figure 3A). The accelerated rate of accumulation of probe sequences in the two rodent genomes is quite clear, as the human chromosomes consistently contain fewer probes per chromosome length (Mb). Looking in greater detail at the human genome, the distribution of these BLAST hits are evenly distributed across all of the chromosomes (Figure 3B). Occasional 'hotspots', where probe sequence abundance was high relative to chromosome length, were evident (see c19, Figure 3B). Likewise, occasional 'deserts', where probe sequences were completely absent, were also evident (see c21, Figure 3B).

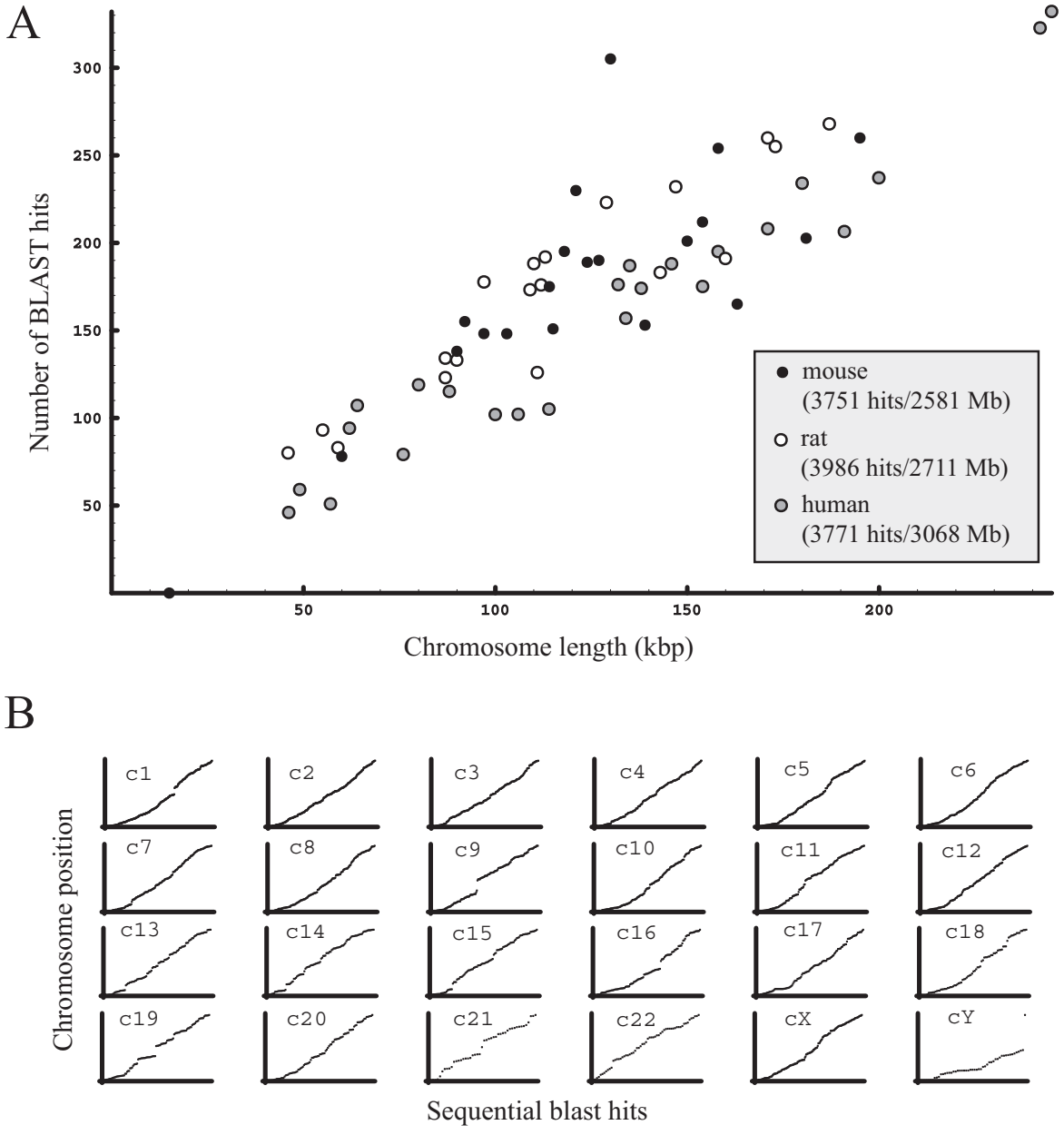

Figure 3. The distribution of some SHyP probes in known genomes. (**A**) The relationship between chromosome length and the number of BLAST hits observed using 3190 probes sequences, all with strong hybridization intensities. Each point represents a single chromosome and the number of hits with ⩾18 complementary base pairs are shown. (**B**) Human chromosome maps showing the accumulation of probes from end to end for each chromosome. Vertical breaks indicate an absence of probes in that segment while horizontal lines illustrate probe-rich regions.

DISCUSSION

In using three well-studied genomes as the initial step in these 'proof of concept' experiments, the results of this novel approach can be fully explored and verified against this rich body of knowledge. One comparative genomic aspect of our results, which agrees well with previous studies (2), is the greater number of 'indicator' probes in the two rodent genomes, both individually and as a clade. This indicates that the SHyP hybridization captured both phylogenetic signal and the generally increased rate of neutral DNA sequence evolution observed previously in rodents. If rodent genomes evolve roughly three times as fast as the human genome, they should be expected to accumulate more 'anonymous' probe sequences in their genomes. Differential measures of homoplasious evolution were also generated by our analysis, indicating a higher level of overall homoplasy in the rat genome, relative to the mouse and human genomes. The effects of homoplasy would also explain the slightly smaller number of rat versus mouse genome-indicator probes (*1291* versus *857*: standard analysis; *644* versus *491*: low variance analysis), particularly if the rat is somehow more constrained in its genomic evolution and basically re-evolves the same sequences. Ongoing BLAST studies will be used to examine the distribution of SHyP probes across all known genomes, available through the NCBI database. Predictive analyses for these other known genomes could be used to further explore the reliability and sensitivity of the array to genomic variation and in the design of the probe sequences.

The use of the array for screening plant genomes needs further exploration. The overall signal for the ramin population samples was substantially lower than any of the mammal genomes. Because these samples were mixtures of five individuals, contaminants or inhibitors may have been present in these genomic DNA extracts. These plants are poorly studied and little previous work has been performed on their DNA (17,18). In comparison to mammal tissue samples, plant tissues often contain a wide variety of potential compounds, which may act as inhibitors in downstream DNA protocols and extraction conditions frequently have to be optimized to improve overall results. The success of these experiments, in their initial attempt, is a testament to the robustness and ability of the technique to screen completely unknown samples. Given the overall lower intensity of the ramin hybridizations, the number of informative probe sequences detected, both between plants and animals and between populations, was substantially higher than would be expected.

While this study demonstrates the feasibility and power of the SHyP approach against an information rich background, further refinement of the technique is necessary and possible. In terms of probe sequence design, the ongoing BLAST studies indicate that none of the 25 bp SHyP probe sequences have perfect maximum sequence identity with any known genome sequence. The vast majority of matches are between 17–19 bp in length. By shortening the length of the oligonucleotide probe sequences and the hybridization/wash conditions, the specificity of these interactions can be increased. The current length probably loses some discriminatory power. Also, in the future, hybridizations will be performed after careful screening via flow cytometry to carefully measure C-content (19), particularly important for examining polyploidy plant genomes. These *C*-values would allow researchers to control for the amount of genomic DNA used in comparative hybridizations between organisms with different ploidy numbers in order to better interpret the resulting dye intensities. Higher density DNA microarray platforms will also allow a greater diversity of probe sequence representation. These higher densities will not only increase the overall sensitivity and power of the analysis but will also facilitate the inclusion of specific classes of genomic probes (20), such as the AT-rich chloroplast genomes (21). With increasing knowledge, the use of consistently present 'universal' probe sequences across the array could provide an important positive control for the array design.

The construction of a genomic signature database, using quickly developing, increasingly flexible DNA microarray technologies and the SHyP-CGH protocol described here, would allow biologists from a wide range of fields and interests to examine genomic scale data. While DNA microarray studies are cost-intensive, the approach described here would use 'type' specimens, formed by pooling genomes from several individuals of each target evolutionary unit, such as a 'species' or 'population'. These types would be incorporated into the genomic signature database after a minimum number of comparative hybridizations, against both comparative targets and standards. Theoretically, only three or four hybridizations would be necessary for each type. After these initial set of hybridizations, no further DNA microarray experiments would be necessary for that target genome. Because the results are transitive across experiments and the hybridizations are performed against a standard platform, virtual comparisons would be possible against any other genome already present in the database. The results would be easily transferable to cheap downstream screening protocols designed for large numbers of individuals, such a macroarray or PCR-based techniques (22). A simple query of the genomic signature database would produce informative sets of probe sequences, based upon the target genomes. These informative probes would each represent an independently segregating locus sampled from across all genomic compartments. Given the large number of informative sites produced in this simple proof of concept experiment, this approach should be very cost-effective, using the DNA microarray based experiments only as a way to generate the database while all further assays use proven population-based techniques with a smaller subset of informative markers.

These probes would also be phylogenetically scalable, providing nested sets of informative markers from the population level up to major clades. This type of data would overcome many of the shortcomings currently apparent in the field of phylogenomics (23,24). If the standard set of probe sequences could not provide enough detail at finer scales of populations and recent evolutionary events, the informative sequences could be used as 'seeds' from which to produce a large degenerate sets of anonymous oligonucleotide sequences. These custom designed degenerate sets would then provide a more focused and detailed result for the target genomes. By applying a more detailed set of sequence selection criteria, certain regions of known genomes or genomic compartments within cells could be targeted and screened in a much wider range of organisms. These informative sets of sequences could address a wide range of ecological questions, including

the rates of gene flow across landscapes. The development of commercial applications, particularly in the natural resources trade, using genomic signatures to identify species and geographic origin certainly seem possible and could provide an objective means for determining legality of the harvest. It would also allow evolutionary biologists to pinpoint the relevant genomic differences among sets of target genomes and generate sequence tags to explore these regions more in detail.

The primary motivation for this study and the development of the SHyP approach is the speed with which human activity is significantly modifying the distribution and composition of genomic diversity in natural communities around the globe (25). To properly interpret the present ecological and evolutionary situation and to plan for the future management and conservation of these natural resources, the historical past holds the key. This fundamental baseline of data must be collected prior to wholescale modification. The SHyP genomic signature database, once constructed, could provide great acceleration in the effort to recognize and catalog life's diversity using a DNA fingerprint. The Bar Code of Life approach hangs the identification of all taxa on the DNA sequence variation at one or a few genomic loci. While this approach appears promising in some situations (26–29), the possibility of obtaining positively misleading results, particularly if the Bar Code is based upon a cytoplasmic locus (30–33), could easily skew the original objectives of the initiative. Further perspective on the BCoL approach could easily be gained by following our example here and exploring its application in the best known groups, like mice and men. While the current Bar Code has been lauded as a way to identify 'cryptic' species (34), the objective definition of 'species' may become particularly problematic when applied to human races (35,36). For the analogy to the commercial version of barcode or 'automated identification and data capture' technologies to be true, the code must involve more than a single or even a handful of bars. The SHyP-CGH database would fulfill this analogy by examining the hybridization intensities at tens of thousands of loci and with a quick query of the database, a biologist could obtain potentially thousands of informative markers directly pertinent to the question without performing any microarray work. These markers would also not be limited to simply identifying the organism but could be used for a wide variety of other purposes.

ACKNOWLEDGEMENTS

Funding for this project was provided by a grant from the National Geographic Society's Conservation Trust Fund and support from the Texas Tech University's Office for Research, Technology Transfer and Economic Development and the Department of Biological Sciences. The authors kindly thank the cooperation of the staff of the Forest Research Center, Sarawak Forestry Department, Kuching, Malaysia and the staff and management of P. T. Diamond Raya Timber, Pekanbaru, Riau, Indonesia. Dr M. J. Asif performed several key laboratory experiments to make this research possible. W. B. Maynard, Global Forestry Services, contributed to the initiation of this project. Dr Wickneswari, National University of Malaysia, and her students helped obtain and process samples. Funding to pay the Open Access publication charges for this article was provided by Cogenics, a Division of Clinical Data, Inc. and the Department of Biological Sciences, Texas Tech University.

Conflict of interest statement. None declared.

REFERENCES

1. Frazer,K.A., Elnitski,L., Church,D.M., Dubchak,I. and Hardison,R.C. (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.*, **13**, 1–12.
2. Miller,W., Makova,K.D., Nekrutenko,A. and Hardison,R.C. (2004) Comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, **5**, 15–56.
3. NCBI (2006) Genomic biology. National Center for Biotechnology Information. http://www.ncbi.nih.gov/Genomes/.
4. Metzker,M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.*, **15**, 1767–1776.
5. Shendure,J., Mitra,R.D., Varma,C. and Church,G.M. (2004) Advanced sequencing technologies: methods and goals. *Nature Rev. Genet.*, **5**, 335–344.
6. Bawa,K.S., Kress,W.J. and Nadkarni,N.M. (2004) Beyond paradise—Meeting the challenges in tropical biology in the 21st century. *Biotropica*, **36**, 276–284.
7. Jepson,P., Jarvie,J.K., MacKinnon,K. and Monk,K.A. (2001) The end for Indonesia's lowland forests? *Science*, **292**, 859–861.
8. Soares-Filho,B.S., Nepstad,D.C., Curran,L.M., Cerqueira,G.C., Garcia,R.A., Ramos,C.A., Voll,E., McDonald,A., Lefebvre,P. and Schlesinger,P. (2006) Modelling conservation in the Amazon basin. *Nature*, **440**, 520–523.
9. Primack,R. and Corlett,R. (2005) *Tropical Rainforests: An Ecological and Biogeographical Comparison*. Blackwell Publishing, p. 319.
10. van Dam,R.M. and Quake,S.R. (2002) Gene expression analysis with universal n-mer arrays. *Genome Res.*, **12**, 145–152.
11. Belosludtsev,Y.Y., Bowerman,D., Weil,R., Marthandan,N., Balog,R., Luebke,K., Lawson,J., Johnston,S.A., Lyons,C.R., O'Brien,K. *et al.* (2004) Organism identification using a genome sequence-independent universal microarray probe set. *Biotechniques*, **37**, 654–658, 660.
12. Pinkel,D. and Albertson,D.G. (2005) Comparative genomic hybridization. *Annu. Rev. Genomics Hum. Genet.*, **6**, 331–354.
13. Barrett,M.T., Scheffer,A., Ben-Dor,A., Sampas,N., Lipson,D., Kincaid,R., Tsang,P., Curry,B., Baird,K., Meltzer,P.S. *et al.* (2004) Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl Acad. Sci. USA*, **101**, 17765–17770.
14. Wyn,L.T., Soehartono,T. and Keong,C.H. (2004) Framing the picture: an assessment of ramin trade in Indonesia, Malaysia and Singapore. TRAFFIC Southeast Asia, Kuala Lumpur, p. 129.
15. IUCN. (2006) Red list of endangered Species. International Union for the Conservation of Nature and Natural Resources. http://www.iucnredlist.org/.
16. Li,W.H. and Wu,C.I. (1987) Rates of nucleotide substitution are evidently higher in rodents than in man. *Mol. Biol. Evol.*, **4**, 74–77.
17. Asif,M.J. and Cannon,C.H. (2005) DNA extraction from processed wood: a case study for the identification of an endangered timber species (*Gonystylus bancanus*). *Plant Mol. Biol. Rep.*, **23**, 185–192.
18. Van der Bank,M., Fay,M.F. and Chase,M.W. (2002) Molecular phylogenetics of Thymelaeaceae with particular reference to African and Australian genera. *Taxon*, **51**, 329–339.
19. Greilhuber,J., Dolezel,J., Lysak,M.A. and Bennett,M.D. (2005) The origin, evolution and proposed stabilization of the terms 'Genome Size' and 'C-Value' to describe nuclear DNA contents. *Ann. Bot.*, **95**, 255–260.
20. Muller,H.M. and Koonin,S.E. (2003) Vector space classification of DNA sequences. *J. Theor. Biol.*, **223**, 161–169.
21. Masood,M.S., Nishikawa,T., Fukuoka,S., Njenga,P.K., Tsudzuki,T. and Kadowaki,K. (2004) The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. *Gene*, **340**, 133–139.
22. Summerbell,R.C., Levesque,C.A., Seifert,K.A., Bovers,M., Fell,J.W., Diaz,M.R., Boekhout,T., de Hoog,G.S., Stalpers,J. and Crous,P.W.

(2005) Microcoding: the second step in DNA barcoding. *Philos. Tran. R. Soc. B-Biol. Sci.*, **360**, 1897–1903.
23. Murphy,W.J., Pevzner,P.A. and O'Brien,S.J. (2004) Mammalian phylogenomics comes of age. *Trends Genet.*, **20**, 631–639.
24. Soltis,D.E., Albert,V.A., Savolainen,V., Hilu,K., Qiu,Y.L., Chase,M.W., Farris,J.S., Stefanovic,S., Rice,D.W., Palmer,J.D. *et al.* (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. *Trends Plant Sci.*, **9**, 477–483.
25. Janzen,D.H. (2004) Now is the time. *Philos. Tran. R. Soc. B-Biol. Sci.*, **359**, 731–732.
26. Hebert,P.D.N. and Gregory,T.R. (2005) The promise of DNA barcoding for taxonomy. *Syst. Biol.*, **54**, 852–859.
27. Ball,S.L. and Armstrong,K.F. (2006) DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae). *Can. J. For. Res. Canadienne De Recherche Forestiere*, **36**, 337–350.
28. Kress,W.J., Wurdack,K.J., Zimmer,E.A., Weigt,L.A. and Janzen,D.H. (2005) Use of DNA barcodes to identify flowering plants. *Proc. Natl Acad. Sci. USA*, **102**, 8369–8374.
29. Janzen,D.H., Hajibabaei,M., Burns,J.M., Hallwachs,W., Remigio,E. and Hebert,P.D.N. (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philos. Tran. R. Soc. B-Biol. Sci.*, **360**, 1835–1845.
30. Will,K.W., Mishler,B.D. and Wheeler,Q.D. (2005) The perils of DNA barcoding and the need for integrative taxonomy. *Syst. Biol.*, **54**, 844–851.
31. Meyer,C.P. and Paulay,G. (2005) DNA barcoding: error rates based on comprehensive sampling. *Plos Biol.*, **3**, 2229–2238.
32. Gompert,Z., Nice,C.C., Fordyce,J.A., Forister,M.L. and Shapiro,A.M. (2006) Identifying units for conservation using molecular systematics: the cautionary tale of the Karner blue butterfly. *Mol. Ecol.*, **15**, 1759–1768.
33. Funk,D.J. and Omland,K.E. (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.*, **34**, 397–423.
34. Hebert,P.D.N., Penton,E.H., Burns,J.M., Janzen,D.H. and Hallwachs,W. (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl Acad. Sci. USA*, **101**, 14812–14817.
35. Ingman,M., Kaessmann,H., Paabo,S. and Gyllensten,U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature*, **408**, 708–713.
36. Consortium,T.I.H. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
37. Tanaka,M., Cabrera,V.M., Gonzalez, A.M., Larruga,J.M., Takeyasu,T., Fuku,N., Guo, L.J., Hirose,R., Fujita,Y., Kurata,M. *et al.* (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Research*, **14**, 1832–1850.