GLOBAL MICROSATELLITE CONTENT POTENTIALLY DISTINGUISHES

HUMANS, PRIMATES, ANIMALS, AND PLANTS.

APPROVED BY SUPERVISORY COMMITTEE

Committee Member's Name _____

Garner, Harold, Ph.D.

Committee Member's Name _____

Gazdar, Adi, M.D.

Committee Member's Name _____

Minna, John, M.D.

Committee Member's Name _____

McPhaul, Michael, M.D.

GLOBAL MICROSATELLITE CONTENT POTENTIALLY DISTINGUISHES HUMANS, PRIMATES, ANIMALS, AND PLANTS.

by

NEIL KUMAR

DISSERTATION

Presented to the Faculty of the Medical School

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF MEDICINE WITH DISTINCTION IN RESEARCH

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

5/13/10

ABSTRACT

Global Microsatellite Content Potentially Distinguishes Humans, Primates, Animals, and Plants.

NEIL KUMAR

The University of Texas Southwestern Medical Center at Dallas, 2010

Supervising Professor: Dr. Harold Garner

Microsatellites are highly mutable, repetitive sequences commonly used as genetic markers, but they have never been assayed en masse. Using a custom microarray to measure hybridization intensities of every possible repetitive nucleotide motif from 1-mers to 6-mers, we examined 25 genomes. Here we show that global microsatellite content, as measured by array hybridization signal intensities, independently validates data from published sequence databases and is a sensitive and specific gauge of mutation in an in vitro model of microsatellite instability, an MLH1 knockout (RKO6) cell line. Moreover, we demonstrate that microsatellite content varies predictably by species, and that particular motifs are characteristic of one species versus another. For instance, hominid-specific microsatellite motifs were identified despite alignment of the human reference, Celera, and Venter genomic sequences indicating substantial variation (30-50%) among individuals. Differential microsatellite motifs were mainly associated with genes involved in developmental processes, while those found in intergenic regions exhibited no discernible pattern related to transcription factor binding sites or non-

microsatellite repetitive sequences (i.e., LINEs, and SINEs). This is the first description of a method for evaluating microsatellite content to classify individual genomes.

TABLE OF CONTENTS

ACKNOWLEDGEMENT/DEDICATION 7

PRIOR PUBLICATIONS AND PRESENTATIONS 8

INTRODUCTION 9

METHODS 13

RESULTS 28

DISCUSSION 42

FIGURES AND TABLES 53

WORKS CITED 84

VITAE 87

Dedication

I would like to thank the members of my gracious supervisory committee, my mentor and role model Dr. Harold Garner, my friends for putting up with me, and most of all my family (mom, dad, Sheil and Elie), who never stop encouraging me.

Acknowledgement

I would like to extend an additional thanks to my mentors, Dr. Harold Garner, Dr. Cristi Galindo, Dr. John Minna, Dr. Adi Gazdar and Dr. Mike Skinner for their invaluable guidance and mentorship. This work was funded by the P.O'B. Montgomery Distinguished Chair and the Hudson Foundation. Dr. Galindo received support from an NIH cardiology fellowship, Cardiology Department, University of Texas Southwestern Medical Center. I would like also to acknowledge Lauren McIver, Vinayak Kulkarni, Linda Gunn, David Trusty, Zhaohui Zun, John Fondon III, My-Hanh Nguyen, Kar-wai Ng, Jenni Weeks, Chris Seabury, Thinh S. Q. Pham, and Tanishia Choice for their technical assistance.

Prior Publications

C. L. Galindo, L. J. McIver, J. F. McCormick, M. A. Skinner, Y. Xie, R. A. Gelhausen, K. Ng, **N. M. Kumar** and H. R. Garner. 2009. Global microsatellite content potentially distinguishes humans, primates, animals, and plants. Molecular Biology and Evolution 26(12); 2809-2819.

Galindo CL, Errami M, Skinner M, Olson LD, Watson D, Li J, McCormick JF, McIver LJ, **Kumar NM**, Pham TQ, Garner HR. 2009. Transcriptional profile of isoproterenol-induced cardiomyopathy and comparison to exercise-induced cardiac hypertrophy and human cardiac disease. BMC Genomics 9;23.

Prior Presentations

Kumar, Neil, Nguyen, My-hanh, and Garner, Harold. Survey of micro-satellite associated cis-regulatory elements in cancer. 45th Annual UT Southwestern Medical Student Research Forum, January 2007.

Neil M Kumar, C. L. Galindo, M. Skinner, R. Gelhausen, T. S. Q. Pham, and H. R. Garner. Global Microsatellite Variations Correlate With Neurological and Morphological Features That Distinguish Humans and Chimpanzees. 46th Annual UT Southwestern Medical Student Research Forum, January 2008.

Introduction

Microsatellites are simple DNA sequence repeats that have been associated with morphological changes and human diseases, especially neurological disorders (Pearson, Nichol Edamura, and Cleary 2005). Microsatellites are typically defined as repeated units of 1-6 nucleotides, 30 bp or more in length, are highly mutable (as much as 10^{-4} per human locus per generation, compared to 10^{-7} to 10^{-9} for point mutations), and have extremely high levels of polymorphism and heterozygosity, compared to high complexity DNA sequences (Ellegren 2004). Microsatellites are as perplexing as they are ubiquitous, in that they are over-represented in the human genome compared to expected levels that would be present by chance. Even closely related species exhibit significant differences in microsatellite content. For example, when human and chimpanzee microsatellites are compared, it is clear that selective forces are present, as human microsatellites differ in their mutability rates (Kelkar et al. 2008) and are longer on average than their orthologous chimpanzee counterparts (Vowles and Amos 2006). In general, microsatellite mutability varies greatly depending on repeat number, length, and motif size (Fondon et al. 1998; Webster, Smith, and Ellegren 2002; Kelkar et al. 2008), adding an additional dimension of potential evolutionary machinations. It has also been suggested that species variation in the underlying mutational mechanisms of microsatellites (e.g., slippage rates, mismatch repair machinery, recombination) are the cause of taxon-specific microsatellite variation and that these differences are influenced by natural selection pressures (Buschiazzo and Gemmell 2006). Simple sequence repeats have been proposed to contribute to phenotypic variation via "fine tuning" of gene

expression under conditions of selection (Kashi and King 2006; King, Trifonov, and Kashi 2006; Fondon et al. 2008). A variable length microsatellite in the *Drosophila melanogaster* circadian rhythm *period* gene, for instance, confers variable temperature sensitivity that correlates with climate (Sawyer et al. 1997; Zamorzaeva et al. 2005). A similar phenomenon has been observed for wheat, in which microsatellite polymorphisms correlate with ecological conditions (Fahima et al. 2002; Huang et al. 2002; Li et al. 2002). Microsatellite variation can also affect animal morphology, as is the case for two microsatellites in the coding region of the *runx-2* gene, the relative lengths of which were shown to correlate with dog snout length (Fondon and Garner 2004). Also compelling is the evidence that the presence of a microsatellite in the 5' UTR region of the gene that encodes the vasopressin receptor (*avpr1a*) correlates with social behavior in voles (Hammock and Young 2004; Hammock and Young 2005), which was experimentally verified in a mouse model (Young et al. 1999). Microsatellite polymorphisms in primates and humans are also suspected to contribute to human behavior and cognitive functions (Fondon et al. 2008). There is a microsatellite upstream of the *avpr1a* gene in humans and bonobos (*Pan paniscus*), for instance, that corresponds to the one found in social voles (Hammock and Young 2005). This region is partially deleted in the chimpanzee (*Pan troglodytes*), which is less empathetic than humans or bonobos (Hammock and Young 2005; Kashi and King 2006), and a recent study of several primate species and humans suggests that the length of this microsatellite might indeed influence vasopressin receptor expression and social behavior (Donaldson et al. 2008). Likewise, microsatellite polymorphisms in the serotonin transporter gene, *SLC6A4*, are believed to influence human behavior (D'Souza and Craig 2006; Fondon et al. 2008).

Despite some evidence for microsatellite-based control of gene expression and subsequent phenotypic variation, the area is highly controversial; there is evidence that species phenotype is in part controlled by protein sequence alterations as well as *cis*-regulatory mechanisms to account for the genetics of evolution (Pennisi 2008). While this issue is hotly debated, most would agree that the genetic machinations that underlie species differentiation are complex, likely involving both coding and non-coding components. The idea that microsatellites might serve as facilitators of evolutionary processes, however, has only recently surfaced, mainly in the form of individual species discoveries. Predictably, the concept has already provoked disagreement, as researchers have continued to investigate individual microsatellite sequence contributions to gene expression and specific end phenotypes. Fink *et al.*, for instance, provided phylogenic evidence that challenges the involvement of the microsatellite located in the 5' UTR region of the *avpr1a* gene in rodent monogamy (Fink, Excoffier, and Heckel 2006; Fink, Excoffier, and Heckel 2007). Their results, based on the presence of the sequence in over 20 rodent species, suggest that the contribution of microsatellite variability to social behavior might be more complicated than a simple presence or absence of a particular sequence. Indeed, a preponderance of evidence supports a more subtle role for microsatellites in fine tuning gene expression, and consequently variations in observable phenotypes, including social behavior (King, Trifonov, and Kashi 2006; Young and Hammock 2007; Fondon et al. 2008).

Due to the nonrandom nature of polymorphism of microsatellites, coupled with their unique properties (i.e., higher mutability and potential roles in gene expression regulation, recombination, and chromosomal structure), we hypothesized that global microsatellite differences might correlate with phylogeny. Such a comprehensive study

had never been conducted, because microsatellites are difficult to assay *en masse*. Here we describe a new microarray capable of accurately measuring global microsatellite content, and we also illustrate its ability to distinguish between various species and identify individual microsatellite motifs that are associated with genes involved in developmental processes. We further investigate global microsatellite content in a model of deficient DNA repair, the MLH1 knockout cell line (RKO6). We define global microsatellite content as the sum of all microsatellite-containing loci, which occur in potentially hundreds or thousands of locations in a genome, for a given motif family (e.g., specific repeated sequence). Thus, we can measure the combined contributions of the distributed positions for a given motif family in a single array intensity readout. This technique may be especially useful in evaluating and differentiating species whose genome has not yet been sequenced and further annotating the repetitive content of sequenced genomes, the least accurate and complete parts of those sequences.

METHODS

Sample acquisition and preparation:

Human genomic DNA was extracted from blood samples collected from volunteers by the McDermott Center for Human Growth and Development Genetics Clinical Laboratory in accordance with Institutional Review Board. Primate genomic DNA was purchased from Coriell Cell Repositories (Camden, NJ), and Arabidopsis and corn genomic DNA was purchased from Biochain Institute, Inc. (Hayward, CA). Alaskan husky and Angus bull genomic DNA was graciously provided by John Fondon III (University of Texas Arlington) and James E. Womack (Texas A&M University), respectively. Mouse, chicken, and fruit fly genomic DNA was extracted using the DNeasy Tissue kit (Qiagen, Valencia, CA) and subsequently RNAsed, per the manufacturer's instructions. DNA samples used in this study are listed in **Table 1**.

Array design, manufacture, and processing:

Each array consisted of 53,735 unique probes, each replicated 7 times at different positions across the array, for a total of 376,145 probes (features), from which data was obtained. The design included probes to measure repetitive DNA sequences with repeat units from 1-mer to 6-mer, all known transcription factor binding sites, all known ultra-conserved sequences, all sequences available in the RepBase database and a series of controls.

All 53,735 probes on the array (except those intended specifically for analysis of hybridization kinetics) were designed with a melting temperature of 76C, the standard design for Roche NimbleGen aCGH experiments. The sequence dependent melting

temperature was computed using the simple 4+2 rule (4C per C/G and 2C per A/T) and the length adjusted to 76C (±2C). The average probe melting temperature was 75.1C and spanned from 56C to 94C, including hybridization temperature scan probes. The minimum probe length was 15 bases, the maximum length was 47 bases, and the average was 25.2 bases. The unique probes (each replicated 7X for a total of 376,145 features) were distributed at different places on the array as follows:

Microsatellite probes

There were a total of 14,634 repeat probes, 1-mer through 6-mer, which included 5,356 perfect repeats. The remainder (9,278 repeat probes) were comprised of single (3,654) and double (4,410) mismatches and single nucleotide deletion (1,214) probes. The sequence alterations were placed in the center of the probes. Because the perfect repeat probes were computer generated to include every possible 1-mer to 6-mer, their cyclic permutations were automatically included. To monitor the stringency performance of the array, all possible single and 2-base substitutions, as well as single base deletions, were made at the most sensitive (central) base position for each probe. Explanations and examples of the various DNA repeat-specific terminologies used throughout the paper (e.g., motif, cyclic permutation, and microsatellite "count") are provided in **Methods Table 1**. A database containing all raw array data from these experiments and a text file of the corresponding probe identifiers and sequences are available for download at http://discovery.swmed.edu/gmc.

METHODS TABLE 1: MICROSATELLITE TERM DEFINITIONS AND EXAMPLES

Microsatellite	repeated DNA units of 1-6 nucleotides, 30 bp or more in length	GATACAGATACAGATACAGATACA…
Motif	A repetitive unit in a microsatellite	GATACA (a hexamer)
Motif length	Monomer (1-mer)	G
	Dimer (2-mer)	GA
	Trimer (3-mer)	GAT
	Tetramer (4-mer)	GATA
	Pentamer (5-mer)	GATAC
	Hexamer (6-mer)	GATACA
Cyclic permutation	Start of motif unit in a microsatellite is shifted, yielding n different ways to write the same sequence (where n = number of nucleotides in the motif)	GATACA → ATACAG
	Monomer	G
	Dimer	GA or AG
	Trimer	GAT, ATG, or TGA
	Tetramer	GATA, ATAG, TAGA, or AGAT
	Example of hexamer cyclic permutations in a microsatellite	GATACAGATACA GATACAGATACA GATACAGATACA GATACAGATACA GATACAGATACA GATACAGATACA
Reverse complements	Complementary microsatellite motifs on the two DNA strands (written 5’ to 3’ by convention)	GATACA and TGTATC
Motif family	All cyclic permutations and their complements for a given motif (number of permutations in a family always = motif nucleotide number times 2)	GATACA, ATACAG, TACAGA, ACAGAT, CAGATA, AGATAC, TGTATC, GTATCA, TATCTG, ATCTGT, TCTGTA. CTGTAT
Tandem repeat	Motifs repeated in cessation	GATACAGATACAGATACAGATACA…
Count	Number of complete tandem motifs in a microsatellite	GATACA = count of 1 GATACAGATACA = count of 2 GATACAGATACAGATACA = count of 3

Transcription Factor binding site probes

In 2005, all open source entries available in the Transfac Transcription factor database were downloaded and used to generate transcription factor binding site probes. One probe was generated for each of the 4,777 Transcription Factor binding site entries that were of sufficient length to design a probe with a melting temperature of 76C. Repetitive elements for each of these sites were masked, and the first acceptable sequence with a melting temperature of 76C (±2C) was chosen from the longest contiguous sequence.

Ultraconserved region probes

In 2005, the open source database of ultra-conserved elements, previously described by Bejerano *et. al*. (Bejerano et al. 2004), was downloaded (http://www.cse.ucsc.edu/~jill/ultra.html). From that database, 11,544 probes were generated for each of 481 ultra-conserved region entries that were of sufficient length (longer than 200 bp) to design a probe with a melting temperature of 76C. Repetitive elements for each of these sites were masked, and the first acceptable sequence with a melting temperature of 76C (±2C) was chosen from the longest contiguous sequence. These included 3,848 wild-type probes and 3,848 single and 3,848 double mismatch probes. These ultra-conserved regions are reported to be absolutely conserved (100% identity with no insertions or deletions) between orthologous regions of the human, rat, and mouse genomes and 99 and 95% conserved between human sequences and dog and chicken genomes, respectively.

RepBase probes

In 2005 the open source database of repetitive elements, RepBase (Genetic Information Research Institute, www.girinst.org), was downloaded. From that database, a probe was generated for each of 22,072 entries that were of sufficient length from which a probe with a melting temperature of 76C could be selected. Repetitive elements for each of these sites were masked, and the first acceptable sequence with a melting temperature of 76C (±2C) was chosen from the longest contiguous sequence

Controls probes (in addition to internal controls added by Roche NimbleGen, Madison, WI)

Each array contained a total of 708 control probes: 42 probes selected from the Arabidopsis genome, 60 Lambda phage probes (20 wild-type, plus single and double mismatches) and 390 HIV-derived probes. Only sequences that did not occur (or occurred very infrequently) in the human genome were selected and confirmed by BLAST similarity searching. There were also 216 probes computer-generated, variable-length monomer and dimer probes designed to span a variety of hybridization temperatures (design summary, **Methods Table 2**).

METHODS TABLE 2: MICROSATELLITE CONTENT ARRAY

Master repeat probes	All possible 1-mers to 6-mers	5,356
	1-mers (monomers) - 4	
	2-mers (dimers) - 12	
	3-mers (trimers) - 60	
	4-mers (tetramers) - 240	
	5-mers (pentamers) - 1,020	

	6-mers (hexamers) - 4,020	
	Single and double mismatches	9,278
	1-mers (monomers) - 192	
	2-mers (dimers) - 1,176	
	3-mers (trimers) - 3,420	
	4-mers (tetramers) - 3,840	
	5-mers (pentamers) - 650	
Repbase probes	All known repeats from Repbase	22,072
Ultra-conserved probes	Wild-type from 481 regions	3,848
	Single and double mismatches	7,696
Transcription factor binding sites	Filtered for Tm from POTION/Transfac database	4,777
Quality control probes	Arabidopsis	42
	Lambda phage (wild-type and mismatches)	60
	HIV	390
	Benchmark (variable length monomers & dimers)	216
Total set = 53,735 probes X 7 copies = 376.145 features		

DNA Sample processing:

DNA concentration (260nm) and purity (260/280 and 260/230 nm) was assessed by spectrophotometry, and quality was confirmed by agarose gel electrophoresis. Samples (at least 2.5 µg, 250 ng/µl) were subsequently evaluated by Roche NimbleGen (Madison, WI) and further subjected to quality control measures appropriate for array comparative genomic hybridization (aCGH) samples.

In addition to the design above, Roche NimbleGen included additional probes (features) for their own internal controls and checks. DNA labeling and hybridization was performed following their standard protocol for aCGH, with a hybridization temperature of 42C in a proprietary hybridization buffer. Two samples labeled with different

fluorochromes were hybridized to each array, and each array was used only once. One sample was always a human standard provided by Roche NimbleGen, originally obtained from Promega, Inc (Madison, WI), which was labeled with Cy5. Test samples were labeled with Cy3 and co-hybridized with the human reference DNA. The probe density of the array was optimized to avoid saturation but minimize probe-probe or non-specific hybridization. Arrays were scanned and data extracted by Roche NimbleGen following their standard procedures. All raw data (intensity values for each probe for each of the two samples on the array) were subsequently provided to us for analysis. Only samples that passed all quality control measures were hybridized to the Microsatellite Survey Array, following Roche NimbleGen's standard procedures.

Chronicle of Arrays Performed

A total of 27 microarrays, 25 for the speciation experiment and 2 arrays for the RKO6/RKO7 experiment, were conducted. Overall 6 human arrays, 3 chimp arrays, 3 gorilla arrays, 2 orangutan arrays and 1 array respectively for the marmoset, rhesus, baboon, macaque, bull, dog, mouse, chicken, arabidopsis, corn and fruit fly were conducted. Multiple individuals were examined for the select primates above to mitigate polymorphism within species and select stringently for motifs whose differential intensities were consistent over all individual pair-wise comparisons. No single individual replicates were performed as intra-individual variability on comparative genomic hydbridization (CGH) arrays is expected to be minimal as compared to gene expression arrays; in lieu of this, precision in our dataset was verified through consistency of hybridization intensity over cyclic permutations and reverse complements,

demonstration of a continuum of intensity over mismatch mutations, and inter-array regression analysis between same-species individuals.

Array data processing and statistical analysis:

A modified RMA (Robust Multi-chip Average) normalization procedure was performed across all arrays (i.e., the procedure included background subtraction and quantile normalization, but the probe summation step was omitted), followed by regression analysis in order to compare all reference sample signal intensity values for each array. Intra-array variability was also assessed, with each R^2 value between any two replicate probe sets ranging between 0.97 and 0.99. In order to reduce the potential effect of outliers, only the median 5 probe values were considered for further analysis (i.e., maximum and minimum values were discarded for each set of replicate probes on each array). All five replicate probes were then normalized against the average reference value for each probe set and the resulting value log transformed to further reduce unwanted variability. Subsequent statistical analyses were performed using GeneSpringGX 10.0 (Agilent Technologies, Santa Clara, CA). For comparison of the various species groups (e.g., hominids versus non-hominids), pairwise comparisons and Student's t test with Benjamini and Hochberg correction were performed using GeneSpring, with an expectation of at least a 2-fold difference between the two groups and an adjusted p value of less than 0.05. For these comparisons, normalized, linear data were uploaded into the GeneSpring program and percentile shift normalization performed (threshold = 1.0, shifted to 75%), followed by baseline to median normalization of all 25 samples. Data were also filtered by intensity values (lower cut-off percentile = 20% for raw signals) as a

quality control measure before averaging across replicates and subsequent pairwise comparisons were performed. Individual pairwise comparisons were also performed, and only differences that were consistently observed for each replicate sample were considered as significant. For microsatellite motifs, any observed difference was also expected to occur consistently across all possible cyclic permutations, including cyclic permutations for the relevant complement sequence. The single and double mismatch sequences and those with deletions were also examined for each microsatellite motif identified as differentially present between groups. As expected, the intensity values decreased predictably between microsatellite-specific control (WT, SM, DM, and DEL) probes (**Supplementary Fig. 1**). Control probes were used to gauge background levels, reproducibility of reference samples, and final statistical outputs (i.e., were included in each analysis as negative controls and subjected to the same statistical parameters as were the test values). R software (http://www.r-project.org/) was used to perform hierarchical clustering of both probes and samples, with Euclidian distance as the metric with complete linkage.

Weighted sum of microsatellite lengths for published genomes:

A total of 60 genome sequences (**Supplementary Table 1**) were downloaded from 11 different sources (provided in supplementary methods) and analyzed as described below. For phylogenetic tree construction, all 60 genomes were searched for every possible cyclic permutation of all 1-mer through 6-mer microsatellites. Microsatellites had to be at least 12 bps and could not contain any insertions, deletions, or mismatches. A weighed sum was computed for each cyclic motif. This weighted sum was

the sum of each microsatellite length divided by the minimum acceptable microsatellite length of 12 bps.

Computation of microsatellite occurrences at individual loci:

To identify individual microsatellite loci in the three human assemblies and chimpanzee genome, a Perl script was written to search for all possible 18-20 bp microsatellites of 1-mer through 6-mer motifs (at least 18 bp for 3-mers and 6-mers; at least 20 bp for 1-, 2-, 4-, 5-, and 6-mers). Microsatellites could not contain insertions, deletions, or mismatches. Also a microsatellite was not considered if the microsatellite did not meet the length requirement without including base pairs that overlapped another microsatellite. Although microsatellite polymorphisms within the human and chimp genomes may have skewed our counts lower due to our stringent search parameters, the length and purity requirements minimized noise from our search algorithm, in theory imposed a degree of biological functionality and created a more sensitive screen in detecting differential regions between the species. A database of genetic regions was constructed by downloading the human, March 2009 release, and chimpanzee, November 2007 release, Gene and Gene Prediction Tracks RefSeq table, from the UCSC Genome Table Browser (http://genome.ucsc.edu). This database was used to match all of the human reference and chimpanzee microsatellites to gene-associated cytogenic loci, including all exons, introns, and promoter regions (1kb 5' of the start site). SINEs and LINEs were also downloaded from UCSC for human, chimpanzee, and rhesus genomes and linked to all microsatellite occurrences (to within 500 bp) in each of these genomes for comparison.

Alignment of human reference assembly microsatellites to other assemblies:

The three human assemblies (NCBI Build Number 36, Version 3, released March 24, 2008) and the chimpanzee (NCBI Build Number 2, Version 1, released Oct. 4, 2006) where aligned using BLAST. The 50bp flanking sequences of each of the human reference assembly microsatellites where incrementally BLASTed (with an e-value threshold of 1e-6 and with the low complexity filter, DUST, turned off) against the corresponding chromosome of each of the other three assemblies using a Perl script. The 50bp length cutoff and 1e-6 threshold were empirically found to optimize alignment, avoiding pitfalls with multiple alignments possibilities or alignments scattered with SNPs. The flanking sequence BLAST hits with the lowest BLAST values that were no more than approximately 1,000 bps apart were considered to be the alignment location. This resulted in the alignment of 97.3% and 97.0% of microsatellite sequences between the human reference genome and Celera and Venter genomic sequences, respectively. Conversely, an alignment of only 92.2% was achieved using the same BLAST parameters for the reference human and chimpanzee genomes. Microsatellites that did not have BLAST hits for both flanking sequences were considered to not have an alignment point.

Gene Expression Analyses

To correlate human and chimpanzee gene expression differences with our computational microsatellite findings, we analyzed raw data from a previous study that

examined expression differences in the anterior cingulated cortex region of the brain between humans, chimpanzees, gorillas, and macaques using Affymetrix GeneChip Human Genome U133A and U133B arrays (Uddin et al. 2004). We downloaded all 20 raw (CEL) array files, supplied on the author's website (www.genetics.wayne.edu/lgross/primates.htm), and performed RMA normalization across the arrays using GeneSifter software (Geospiza, Seattle, WA). Hybridization signals below background noise were discarded, and pairwise comparisons and Student's t test were performed to compare the three human and two chimpanzee individuals. Expression was considered statistically differential for genes with a fold-change of 1.5 or greater and *p* value of less than 0.05. The resulting list of genes was then compared to the current study results, as described in the text.

Phylogenetic Trees

Data obtained from the microsatellite arrays (normalized signal intensity values) and computational analysis (log transformed computed counts within sequenced genomes), for all 5,356 wild-type microsatellite motifs, were treated identically for the purposes of tree building. All 5,356 data points for each microsatellite for each sample were first normalized using GeneSpring (percentile shift normalization followed by baseline to median normalization). A Euclidian distance matrix was subsequently produced using R software and then converted to a phylogenetic tree using the neighbor program within the PHYLIP software suite and TreeView (Page 1996). Trees were rooted using Arabidopsis (for the smaller trees) or the Archaeabacterium *H. utanehsis* (for the larger trees).

Statistical Analysis

Statistical analyses used in the study are described in detail throughout the method section. Where applicable, the data were plotted as arithmetic mean ± the standard deviation, and Student's t test (p 0.05) was used for data analysis.

Motif location, gene function determination, functional analyses, and disease associations:

For each motif determined to be globally differential in signal intensity between the various groups examined, the DNA sequence corresponding to a pure tandem microsatellite repeat of at least 18bp long was searched throughout 60 published genomes (**Supplementary Table 1**) computationally (as described above). Only loci with 100% identity were accepted, and the alignment (microsatellite and flanking regions) was inspected by eye for 100 randomly chosen motifs to confirm copy number accuracy and exact location in reference to any genes within 1,000 bp of the microsatellite sequence. Motifs that were differential between hominids and non-hominids were manually counted in the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) for the published human, chimpanzee, orangutan, rhesus macaque, marmoset, mouse, and stickleback fish genomes. Gene location and ortholog information was gathered using the browser function, and functional characterization of these genes was achieved using the Entrez Gene, AceView, Stanford SOURCE and PubMed databases. Gene ontological (GO) overrepresentation analysis was conducted on characterized genes using GeneSifter (VizX Labs, Seattle, WA) (Doniger et al. 2003), with the RefSeq database serving as the

master gene list against which occurrences of ontological categories (for each motif) were compared. Only biological process ontologies were considered, as these would represent the highest level of physiological functionality (described in more detail online: http://www.geneontology.org/GO.doc.shtml). Z scores of at 2.0 or greater were considered significant (Doniger et al. 2003). Ingenuity Pathway Analysis software (Ingenuity Systems, Redwood, CA) was employed to identify over-represented physiological functions, with use of the optional Benjamini and Hochberg multiple hypothesis correction test for each gene list examined.

RKO6/RKO7 Microsatellite Content

The RK06 (MLH1 knockout) and RKO7 (MLH1 competent) cell lines were provided graciously by the David Boothman laboratory. Multiple passages of the cell lines were conducted to allow for sufficient replication cycles in order to allow defective mismatch repair machinery to appreciably manifest in mutations at predisposed microsatellite loci. The samples were co-hybridized with the same human reference sample, rather than each other, as our inter-array quantile normalization was defined using this standard. This method further allowed us also detect microsatellite differences between the RKO7 cell line and reference human sample. Data was analyzed using the same statistical features as the human/chimp data.

Team Roles

My work was conducted under the direct supervision of Dr. Harold Garner and Dr. Cristi Galindo. My roles included contributing to the preparation of sample DNA, analysis of the microarray data from the species and RKO6/RKO7 experiments, and the detection of associated genes using the BLAT database. Nimblegen conducted the hybridization protocol for all arrays. Dr. Galindo performed the analysis of the microarray expression data from the anterior cingulate gyrus. Various members of our team, recognized in detail in the acknowledgment, completed the computational and ontological analysis.

Results

Concise Summary

RK06/RKO7 Experiment

- ➔ The MLH1 knockout RKO6 cell line demonstrated lower relative signal intensities than the RKO7 counterpart among motifs of higher overall hybridization intensities. (**Fig. 4**)
- ➔ A G(T)n set of motifs were noted to be increased in intensity in the RKO6 line (**Table 5**)
- ➔ The ACCAC and ACCCAC motifs were noted to be decreased in intensity in the RKO6 line (**Table 5**)
- ➔ Potential involvement of APC gene (ACCCAC) and MLH1 gene (GGGT) in the RKO6 cell line (**Table 5**)
- ➔ These results validate the sensitivity and specificity of our microarray technique in an in vitro model of microsatellite instability.

Microarray Findings

- ➔ Regression analysis of reference intensity values compared to the average reference value across all 25 arrays, indicated extremely low inter-array variability (R^2 values 0.97-0.99 for each array).
- ➔ The ratio of the standard deviation to average signal intensity (SD/AVG) for motif groups, which included cyclic permutations and reverse complements, was 0.12 -- considerably lower than the SD/AVG over all the probes (0.63 for each array). Furthermore, examination of the signal intensities for probes exhibiting a single base pair mismatch confirmed the ability of the array to distinguish between closely related sequences designed to measure hybridization specificity (**Supplementary Fig. 1**).
- ➔ The AATGG and ACTCC motifs in particular were hominid-specific, with similar intensity values across human, chimpanzee, gorilla, and orangutan individuals, compared to minimal detection for all other species examined
- ➔ There were no human-specific microsatellite motifs. Notably, however, the CAGC motif exhibited higher hybridization intensity in humans and gorillas than all other species.
- ➔ There were, however, two chimpanzee-specific motifs (TAGCC and TAGCCC) whose hybridization intensities were much higher for all three chimpanzee individuals compared to all other samples (**Fig. 1B** and **Fig. 2**), as well as one motif (CCAGCC) that was exclusive to the two orangutans based on the array data (**Fig. 2**).

- ➔ Only one motif (AACAT) distinguished primates from all other animals and the two plant species
- ➔ The AATGTG motif demonstrated the highest hybridization intensity for *Drosophila* but was virtually undetected in the other 24 samples.
- ➔ We detected no obvious species-wide patterns related to motif lengths (i.e. tetramers vs. pentamers) (**Supplementary Fig. 2**).
- ➔ No global differences were found for the TTAGGG telomeric repeat
- ➔ No global differences were found for the CAG repeat
- ➔ No differential sequences were detected among the various species for the 22,072 probes representing non-microsatellite repeat sequences
- ➔ Ultra-conserved sequences were >92% similar on average between any two groups, as were the hybridization intensities of the 4,777 transcription factor binding sites.

Anterior Cingulate Gyrus Expression Data

- ➔ We identified 2,102 genes that are differentially expressed (>1.5-fold, p value < 0.05) between human and chimpanzee brain regions, based on our analysis of the dataset (3 human and 2 chimpanzee individuals), and 1,900 of these genes (~ 90%) harbor at least one microsatellite.

Computational Findings

- ➔ The average computed microsatellite ratio for all but two of the microsatellites (CAGC and TAGCC) recapitulated the array results.
- ➔ Of the ~200,000 gene-associated loci, 31.3% and 47.3% differ between the human reference genome and the Celera and Venter assemblies, respectively (**Table 4**)
- ➔ We also compared the microsatellite content of the published chimpanzee genome to the human reference sequence (resulting in 92.2% high quality alignment) and found a much higher incidence of overall microsatellite variations (86.3%).
- ➔ The distribution of these differences (i.e., 276,400 and 95,461 human microsatellites were longer and shorter, respectively, compared to the corresponding chimpanzee sequence) is consistent with previous observations that human microsatellites are longer on average than those in chimpanzees
- ➔ The proportion of differential motifs in the introns of genes known to have multiple alternative splice variants (~97%) **(Supplementary Table 2**) is much higher than the expected value (i.e., ~40% of all known human genes have splicing variants)
- ➔ A significant proportion (~51%) of gene functions among the genetic loci demonstrating microsatellite variability on the microarray were those related to brain development, nervous system development and function, the development of various morphological features, and organ-specific development and maintenance.

Microarray Variance and Reproducibility

To measure global differences in microsatellite content between various genomes, we developed a custom oligonucleotide array that included 7 copies of every possible 1-mer to 6-mer microsatellite motif, including cyclic permutations (e.g., AGC and GCA are cyclic permutations of CAG), and various mismatches. We also included probes on the array to assay levels of non-microsatellite repetitive elements, transcription factor binding sites, and ultra-conserved genetic regions. Using this array, we examined 25 individual genomes (6 humans with varied ancestries, 3 chimpanzees, 3 gorillas, 2 orangutans, and one each of baboon, rhesus monkey, long-tailed macaque, marmoset, cow, dog, mouse, chicken, fruit fly, corn, and Arabidopsis). Commercially available pooled human reference DNA was co-hybridized to each array for normalization purposes and to gauge reproducibility. Regression analysis of reference intensity values, compared to the average reference value across all 25 arrays, indicated extremely low inter-array variability (R^2 values 0.97-0.99 for each array). We anticipated some intra-array variations due to differences in hybridization kinetics within microsatellite motif families as a result of base-positioning changes in the cyclic permutations and free-energy differences in the reverse complements. However, the ratio of the standard deviation to average signal intensity (SD/AVG) for motif groups was 0.12, which was considerably lower than the SD/AVG over all the probes (0.63 for each array). Examination of the signal intensities for probes exhibiting a single base pair mismatch confirmed the ability

of the array to distinguish between closely related sequences designed to measure hybridization specificity (**Supplementary Fig. 1**).

RKO6/RKO7 microsatellite content

Refer to discussion.

Analysis and Interpretation of Globally Differential Motifs Identified by Microarray

We next separated the various species into two groups (hominids and non-hominids) in order to identify consistent differences in hybridization of microsatellite motifs that might distinguish great apes (including humans) from other species. Of the 5,356 simple repeat sequences represented on the array, there were 4 motif groups that exhibited statistically differential hybridization (fold change $\geq$ 2.0, Benjamini and Hochberg corrected [B-H] p value $\leq$ 0.05), including each possible cyclic permutation and complement sequence, between hominids and non-hominid animals and plants that were also reproducible across individual representatives of each group (**Fig. 1A**). Of these consistently differential motifs, intensities representing AATGG and ACTCC in particular were hominid-specific, with similar intensity values across human, chimpanzee, gorilla, and orangutan individuals, compared to minimal detection for all other species examined (**Fig. 2**). In contrast to the AATGG motif, CAGC and AACGG were more variable among hominid individuals, and the difference between hominids and non-hominids was much more subtle. The hybridization intensity for CAGC, for instance, was slightly lower in chimpanzees, compared to humans and gorillas, and this motif did not distinguish orangutans from non-hominid species (**Fig. 2**). Likewise the global intensities for AACGG indicated that it was generally higher for hominids, but the hybridization signal for this motif was particularly high for gorillas (**Fig. 2**).

There were no human-specific microsatellite motifs (i.e., motifs that were consistently differential and statistically significant between humans and all other species), with the exception of CAGC that exhibited a higher hybridization intensity than all other species except for gorillas. Since global microsatellite content is merely potentially one of multiple regulators of gene diversity and expression, and hence, not a singular reflection of genomic functional complexity, this is not a particularly grating finding. Moreover, the absolute number of distinct motif loci dispersed across the genome, rather than differences in repeat copy number which influence affinity to the probe, may be a larger contributor to variations in hybridization intensity, This is a consequence of endonuclease sample preparation which should result in a milieu of 200 base pair fragments that interface with the array. Hence, the absence of "human specific" motif intensity differences may merely indicate the lack of de novo generation and abolition of new microsatellite loci relative to the chimp. This may be an expected finding given the extended evolutionary absence of active retro-transposition and inter-locus mobility in the human genome. Potential inter-species global genomic trends in motif-specific copy number may still exist and multiple probes of varying copy numbers may be sensitive in identifying them given the hybridization specificity of the array described above.

There were, however, two chimpanzee-specific motifs (TAGCC and TAGCCC) whose hybridization intensities were much higher for all three chimpanzee individuals compared to all other samples (**Fig. 1B** and **Fig. 2**), as well as one motif (CCAGCC) that was exclusive to the two orangutans based on the array data (**Fig. 2**). All other microsatellite

motifs were either non-differential between hominids and non-hominids or were highly variable between hominid individuals (and thus not characteristic of the group of hominid species that were examined). We detected no obvious patterns related to motif lengths, as comparison of the average signal intensities for monomers, dimers, trimers, tetramers, pentamers, and hexamers indicated little difference among humans or between hominids and primates (**Supplementary Fig. 2**).

In addition to examining hominids versus non-hominid species, it was possible to group the 25 samples by taxonomy (primates, mammals, vertebrates, and animals) and compare them to the remaining samples that did not fall into the category being examined (e.g., vertebrates versus non-vertebrates, the latter of which included fruit fly, Arabidopsis, and corn). As shown in **Table 2**, there was only one motif (AACAT) that distinguished primates from all other animals and the two plant species. There were an additional 8 motifs that could differentiate mammals, vertebrates, animals, and plants (**Table 2**). Because a great deal of genetic information is available for *Drosophila*, we also compared the one fruit fly DNA sample to all other species and identified the motif (AATGTG) with the highest hybridization intensity for *Drosophila* that was virtually undetected in the other 24 samples (**Table 2 and Fig. 2**). A statistically significant, species-specific difference in the global content of the TTAGGG telomeric repeat was not detected between the various groups examined (data not shown) and does offer an avenue of explanation for lower cancer rates in non-human primates. (Lopez-Otin) Likewise we detected no distinguishing global differences for CAG, which we investigated because of its well-known association with multiple neurological diseases (Gatchel and Zoghbi 2005). This analysis was restricted because of the absence of motif probes of varying copy number length and does not preclude copy number variation in chimps with phenotypic associations that mimic human pathology.

In order to verify that the array was capable of accurately measuring global microsatellite content, we compared our array-generated data to data obtained from published genetic sequences. We downloaded 60 published genome sequences (**Supplementary Table 1**) that included three hominids (human, chimpanzee, and orangutan), primates (rhesus macaque, marmoset, and galago), and various mammals (e.g., dog, cow, rat, and mouse), vertebrates (e.g., frog, fish, and lizard), invertebrate animals (e.g., insects, worms, sea squirt) and plants (e.g., corn, rice, and wine grape). All microsatellites were counted for each of these genomes (minimum length = 18 bp) and the counts subsequently compared across representative species (e.g., hominids versus non-hominids) by dividing the average summated microsatellite count for the relevant group examined (e.g., hominid) by the average count for all other species not included in that group (e.g., all non-hominid species). These computed microsatellite ratios were then compared to the array hybridization intensity ratios for the differential motifs identified using the array. As shown in **Table 2**, the average computed microsatellite ratio for all but two of the microsatellites recapitulated the array results. These two motifs (CAGC and TAGCC) might have differed due to incomplete published sequences, lacking segments of heterochromatin in which these motifs may be abundant, or our method chosen for counting microsatellites within the published genome sequences. However, the computed ratio for the CAGC motif, which was higher in humans compared to all other species examined apart from gorillas (**Fig. 2**), increased dramatically (to 89.1) when the published human reference genome was compared to the various other sequenced species.

In contrast to the results obtained for microsatellite motifs, there were no statistically significant and consistent intensity differences detected among the various species examined for the 22,072 probes representing non-microsatellite repeat sequences (data not shown). Regression analysis confirmed the overall similarity (average R^2 value = 0.93) between each individual species sample for non-microsatellite repeat probes that included ALUs, SINEs, and LINEs, for instance. Likewise, ultra-conserved sequences were >92% similar on average between any two groups, as were the hybridization intensities of the 4,777 transcription factor binding sites.

Examination of Individual Loci Associated with Globally Differential Motifs

To identify potential functional consequences of differential microsatellites, the four motif families (groups of cyclic permutations and complements) that were found to be significantly and reproducibly different between hominids and non-hominids were further analyzed, both computationally and manually. The location of the motif relative to any nearby human genes was recorded, and copy numbers (of the pure tandem repeat unit) in the published human, chimpanzee, orangutan, rhesus macaque, marmoset, mouse, and stickleback fish genomes were compared. We identified 157 individual loci associated with 87 specific genes (i.e., exons, introns, UTRs, and upstream and downstream regions within 1,000 bp of the coded sequence) that harbor these four microsatellite motifs. The most striking result was the proportion of differential motifs in the introns of genes known to have multiple alternative splice variants (~97%) **(Supplementary Table 2**), which is much higher than the expected value (i.e., ~40% of all known human genes have splicing variants) (Lian and Garner 2005; Sultan et al.

2008). This may be especially significant in light of recent published data suggesting that exon deletion is the most common mechanism of genetic alternative splicing (Sultan et al. 2008). Consistent with the array findings, nearly all (>99%) of the individual loci associated with the 4 differential motifs (i.e., 157 locations, 87 specific genes) contained microsatellites that differed in copy number between the three hominids (i.e., humans, chimpanzees, and orangutans) and four non-hominids, that were manually examined. However, the human sequence also differed from the other two hominids examined for the majority of loci that harbored the four globally differential microsatellites (**Supplementary Table 2**). A significant proportion (~51%) of gene functions among the genetic loci demonstrating microsatellite variability on the microarray were those related to brain development, nervous system development and function, the development of various morphological features, and organ-specific development and maintenance (**Table 3**, **t**he complete gene list is provided in **Supplementary Table 2**). This finding is in contrast to the ~30,000 human genes listed in the NCBI RefSeq database, in which nearly two-thirds (17,278 genes) were found to harbor one or more microsatellites or contain a microsatellite within 1,000 bp of the transcribed region of the gene; this set of microsatellite-containing genes was evaluated computationally and was not enriched for developmental or neurological processes (data not shown). The association of developmental genes with species-specific microsatellites was not limited to those motifs that distinguished hominids from other animals and plants, as shown in **Table 3**.

In contrast to gene function-associated patterns for differential microsatellite motifs, no correlation was found between intergenic microsatellites and non-genic DNA sequences (transcription factor binding sites, SINEs, or LINEs). There was very little

difference in the hybridization intensity values among the various samples examined for the majority of the 4,777 transcription factor binding sites (TFBS) represented on the array. This is consistent with the decreased mutability of these regions in comparison with microsatellites, both due to secondary structure formation and selective pressures, as well as the relative scarcity of each individual TFBS motif in the genome, making small (single focus) differences difficult to identify. However, we identified the 20 most profoundly differential transcription factor binding sequences between hominids and primates (**Supplementary Fig. 3**) and searched for those microsatellite motifs (AATGG, ACTCC, CAGC, and AACGG) that were found to be consistently differential between hominids and non-hominid primates, based on the array results. None of these hominid-specific motifs were tandemly repeated (2 or more copies) within 500 bp of the 20 transcription factor binding sequences in human, chimpanzee, orangutan, rhesus, marmoset, mouse, chicken or stickleback fish genomes, which were aligned using the UCSC genome browser. Likewise, there were no correlations between these four hominid-specific motifs and non-microsatellite repetitive DNA elements (i.e., SINEs and LINEs). When the incidence of hominid-specific motifs in the published genomes of the human, chimpanzee and rhesus macaque were compared with and without inclusion of those motifs located in or near (within 500 bp) SINEs and/or LINEs, the resulting ratios were similar (**Supplementary Fig. 4A**). Likewise, examination of motifs not expected to differ among humans, chimpanzees, and rhesus macaque (e.g., the primate-specific motif AACAT), resulted in no difference irrespective of inclusion of those individual microsatellite-containing loci in or near SINEs or LINEs (**Supplementary Fig. 4B**).

Comparative Analysis of all Microsatellite-containing Loci in the Published Genomes of Humans and Chimpanzees

To identify individual differences in microsatellite length (i.e., microsatellite polymorphisms), at a gene locus resolution, we computationally analyzed all microsatellite motifs found within 1,000 bp of RefSeq genes and recorded copy number differences for the human reference, Celera, and Venter genomes (Lander et al. 2001; Venter et al. 2001; Levy et al. 2007). We identified ~500,000 microsatellites in the human reference genome, 207,885 of which are located in or near 17,278 (out of ~30,000 total) RefSeq genes (i.e., within 1,000 bp of the coding region). Of these ~200,000 gene-associated loci, 31.3% and 47.3% differ between the human reference genome and the Celera and Venter assemblies, respectively (**Table 4**). Microsatellite flanking sequences in the reference genome that did not align well to Celera (2.7%) or Venter (3.0%) genomic sequences, as well as those that aligned to un-sequenced regions (loci containing "N"s where the flanking or repetitive sequence should have been, 0.1% and 0.3% for the Celera and Venter genomes, respectively) were disregarded in all comparison calculations. The pattern of microsatellite variations among the three genomes was similar for all genomic regions (i.e., ~33% and 50% of the loci were polymorphic for the reference genome compared to Celera and Venter published sequences, respectively), except for those microsatellite sequences found within exons. As shown in **Table 4**, there were far fewer microsatellite variations in exonic regions (8.4% for Celera and 20.1% for Venter genomes compared to the published reference sequence), and the differences most often consisted of length alterations of 3-base pair multiples (modulo-3), as expected (**Supplementary Fig. 5**). This is consistent with elevated selection pressure in exons, especially for minimizing non-modulo-3 frame-shift-causing variations.

There were a total of 103 polymorphic microsatellite loci in the exonic regions of 95 different genes (**Supplementary Table 3**), the functions of which are mainly those related to gene expression regulation (42 genes, Z score [Z] = 2.2-14.6, B-H p value = 0.05 – 0.001) or development (40 genes, Z = 2.2-4.2, B-H adj. p value 0.04 – 0.007), especially nervous system development and function (21 genes, Z = 2.2-8.4, B-H p value = 0.005-0.02). Twenty-five of these genes contained microsatellite differences that were not modulo-3 and would thus introduce frame shifts (i.e., 13 and 19 genes contained a deletion or expansion that was not a multiple of three in the reference genome compared to the Celera and Venter sequences, respectively). There were 1,489 human RefSeq genes, in or near which 2,009 microsatellite sequences were located in the reference genome that were deleted or interrupted in the corresponding Celera or Venter locus. There were no statistically over-represented gene ontologies in this list of genes after multiple hypothesis correction; however, the most prevalent physiological process was nervous system development and functions (Z = 2.4, 77 genes, B-H p value = 0.4-0.6).

We also compared the microsatellite content of the published chimpanzee genome to the human reference sequence (resulting in 92.2% high quality alignment) and found a much higher incidence of overall microsatellite variations (86.3%). The distribution of these differences (i.e., 276,400 and 95,461 human microsatellites were longer and shorter, respectively, compared to the corresponding chimpanzee sequence) is consistent with previous observations that human microsatellites are longer on average than those in chimpanzees(Cooper, Rubinsztein, and Amos 1998; Vowles and Amos 2006). Of the 17,278 human genes that harbor microsatellites, 16,014 contain at least one microsatellite

that differs in copy number between the human reference genome and published chimpanzee sequence. The 617 exonic microsatellites that differed between humans and chimpanzees are mainly associated with developmental and neurological processes (Z = 2.1-7.0, B-H p value = 5.6×10^{-4} – 1.9×10^{-2}), including anatomical structure development and morphogenesis (73 genes), embryonic development (61 genes), nervous system development and function (49 genes), tissue development (40 genes), brain development (11 genes), neurogenesis (14 genes), and the development of various organ systems (54 genes). Microsatellite content also correlated with a previous report indicating that there are significant gene expression differences in the anterior cingulate cortex between humans and chimpanzees (Uddin et al. 2004). We identified 2,102 genes that are differentially expressed (>1.5-fold, p value < 0.05) between human and chimpanzee brain regions, based on our analysis of the dataset (3 human and 2 chimpanzee individuals), and 1,900 of these genes (~ 90%) harbor at least one microsatellite. The majority (1,815 genes) contain a total of 28,275 microsatellites that differ in copy number between humans and chimpanzees, which represents ~80% of the microsatellites located in these genes (data not shown).

Microsatellite Differences Vary Predictably by Species

In addition to examining global and individual microsatellite differences between various pre-chosen groups, hierarchical clustering and phylogenetic tree construction were performed to compare microsatellite content across multiple species, some of which are not yet fully sequenced. Hierarchical clustering of all 5,356 normalized microsatellite hybridization intensities resulted in a heat map that successfully illustrated the separation of species, including individual species representatives (**Supplementary Fig. 6**). Each motif

family was also clustered appropriately, without exception (e.g., CAG, AGC, and GCA clustered together, **Supplementary Fig. 6**). When this same clustering method was used to examine the 22,075 non-microsatellite repeat probes (e.g., ALUs, SINEs, and LINEs) represented on the array, these same samples were not correctly classified by species (**Supplementary Fig. 7**). Hierarchical clustering of ultra-conserved sequence intensities (**Supplementary Fig. 8**) also failed to separate species appropriately (e.g., all 6 humans did not cluster together).

Similar results were obtained when phylogenetic trees were produced from a Euclidian distance matrix for all 25 samples using the neighbor-joining method. As shown in **Supplementary Fig. 9A**, the resulting tree using all 5,356 wild-type microsatellite hybridization intensities grouped same species together and correctly separated hominids from non-hominid animals and plants. For comparison, we also constructed a large phylogenetic tree using 60 published genomes (**Supplementary Fig. 10**) and found that similar species generally grouped together (e.g., 12 *Drosophila* species occupied the same node), and the few primate examples were grouped together. However, there were also a variety of inconsistencies that conflict with what would be expected based on known taxonomic relationships (e.g., plants were not well separated from Arthropods). In general, the data suggested that global microsatellite content varies least among very closely related species but does not accurately reflect large differences in distantly related species. We expected some discrepancies due to incomplete or inaccurate sequencing, the precise microsatellite "counting" algorithm, or potentially the phylogenetic tree making method used. However, the results are consistent with the microsatellite life cycle theory (i.e., fluctuation between expansions, deletions, stabilization via introduction of

SNPs, and reactivation) and suggest that microsatellites fluctuate as a whole (i.e., globally) in addition to individually varying at random (Buschiazzo and Gemmell 2006).

Discussion

RKO6/RKO7 microsatellite content

The random scatter of the doublet motifs over repeat length, of similar overall intensities, served as a negative control for the RKO6/RK07 experiment due to the natural variation in copy numbers among loci and significant non-specific hybridization with a motif of such abundance. (Figure 3) The poly A and poly T motifs, regions known to buffer exonuclease degradation of messenger RNA and span approximately 250 nucleotides per gene, demonstrated a graded intensity with probe length in both the RKO6 and RKO7 cell line and served as a positive control.

The repair deficient RKO6 cell line demonstrated lower relative signal intensities than the RKO7 counterpart at higher overall hybridization intensities, indicating three possible scenarios of increasing likelihood: 1) that select microsatellite motifs, with abundant loci, are selectively deleted in response to MLH1 mutation, 2) that copy number contraction within loci are favored over expansions beyond a certain repeat length threshold, or 3) that there exists a decreased repair sensitivity for repeat contractions with knockout of the MLH1 gene. While the biochemical plausibility of these notions is beyond the scope of this discussion, they begin to help guide the discussion of the mechanism of microsatellite instability. In conjunction with the trend that human microsatellites are on average longer than chimp microsatellites, these findings may

suggest that motif sequences may have a steady state equilibrium copy number that differs depending on the capability of the DNA repair machinery and that, among other factors including large-scale chromosomal deletions, changes in the replication machinery and the cessation of retrotranspositional activity, the decreased capacity of DNA repair in humans (Haaf 2008) may have been a precursor to the observed global microsatellite differences. Importantly, normally inactive repeat regions of the genome, LINES and ALU regions, were not altered in intensity, indicating that functionally silent large repeat changes were not the source of the observed differences in microsatellite intensity. These results validate the sensitivity and specificity of the global microsatellite microarray technique in an in vitro model of microsatellite instability.

The motifs noted to be of differential intensity between the RKO6 and RKO7 cell lines also demonstrated interesting trends. (Figure 4) Notably, a G(T)n motif was noted to be consistently increased in intensity in the RKO6 line, perhaps suggesting a motif specific affinity for expansion at wild type human copy numbers. This motif was searched throughout the human genome and was not found to exist in a copy number greater than 20, the lower limit of detection for the BLAT search tool. Whether this is a function of an endogenous suppression of this motif (through repair or selective pressure), or its existence in heterochromatic regions, notoriously difficult to sequence, is unclear. A similarity was also found in motifs decreased intensity in the RKO6 line, ACCAC and ACCCAC. Again, this suggests a motif-specific affinity for expansion or contraction, released by defective repair machinery. Moreover, the BLAT results revealed an association with the FAF1 gene, an initiator of apoptosis, PPM1B, associated

with cell cycle regulation, and KDM4C, implicated in esophageal squamous carcinoma (Table 5). Most intriguing, the ACCCAC motif was found in the CTNB1 gene, whose product binds to the APC protein, leads to downstream activation of the c-myc gene and is implicated in colon cancer and familial adenomatous polyposis (Table 5). This is an exceptionally notable finding as it potentially connects the microsatellite instability of HNPCC to the APC dysregulation implicated in the traditional model of sporadic colon cancer. Moreover, the motif homology in conjunction with similarity in gene function raises the intriguing possibility of coordinate control of genetic expression through microsatellite homology. We further plan on conducting motif specific ontological analysis of genes. Additionally, the GGGT motif was associated with the MLH1 gene itself, the gene implicated in HNPCC pathogenesis. The implication is that the pathogenesis of HNPCC may involve an "anticipation" phenomenon in the tandem repeat region of MLH1 over multiple replication cycles: a somatic biologic counterpart to the germ line pathology of Huntington's and Fragile X disease. Finally, the number of genes associated with neurological disease for motifs specific to MLH1 knockout is exceptional and raises the possibility that neuronal gene expression, which may require more exquisite control, may rely on copy number nuances. Combined gene expression and copy number data in HNPCC cells of the genes associated with the ACCAC/ACCCAC/GGGT motifs may be of further value.

Microsatellite Content Across Species

The global microsatellite content array described in this study provides a method to examine one of the least understood regions of the genome and also provides information that cannot be readily achieved by individual locus studies or sequencing. For instance, summated

global microsatellite content as measured using the array does not rely on sequenced repetitive regions, which are among the least resolved portions of published genomes, and individuals can be quickly and accurately examined without reliance upon a reference genome that does not capture intra-species microsatellite polymorphisms. The latter advantage is particularly noteworthy, because microsatellites do not merely vary between species but are also extremely polymorphic among individuals. Our analysis of three published human genome sequences suggests that the sum of sequencing errors and natural polymorphism at these loci between any two humans is between 30% and 50% (**Table 4**).

The overall similarity of microsatellite content, as inferred from hybridization intensity, in humans compared to other hominids is somewhat surprising despite our common ancestry, because microsatellites are highly mutable and vary widely among individuals. Nonetheless, the summated levels of two microsatellite motifs in particular (AATTG and ACTCC), as derived from hybridization intensity (**Fig. 1**), clearly differentiate the 14 hominid individuals examined from other animals and plants (**Fig. 2**). Genes that harbor these motifs, as well as motifs that characterize other taxonomic groupings, are mainly those involved in a wide range of developmental processes.

A survey of the differential motifs between humans and chimps revealed an excess of genes that potentially impact neurological function at many levels, including neuronal differentiation, neuronal migration, synaptic connectivity, synaptic strength and neuronal longevity. An intriguing subset of these genes, CNTN4, TNFR-2, Shank3, DCDC1 and WWC1, provide a realistic example of the synergism involved in phenotypic divergence. Contactin 4 (CNTN4), which has been implicated in axon growth, guidance,

and fascicle formation in the central nervous system may be responsible for differential myelination, altered CNS nerve regeneration capacity, and the more complex connectivity observed in humans. Since it is known that neurons that fail to reach their destination undergo apoptosis, TNFR-2 is suggested to play an important role in the activation of anti-oxidative pathways and the protection of neurons from apoptosis may be important in prolonging the "search period" of axons during early synaptic formation. Additionally, human TNFR-2, which exclusively contains the AGCC motif in introns, may prevent excessive neuronal attrition after the completion of formation of synaptic connections during the initial years of life. Shank3, which similarly contains an intronic AGCC motif in humans, is an important molecule that interacts with GluR1 AMPA receptor at synaptic sites of developing neurons. Shank3 has been reported to promote the assembly of a signaling complex at cortico-striatal synapses that enables the regulation of L-type Ca2+ channels and the integration of glutamatergic synaptic events. Thus, Shanks3 may play a role in long-term potentiation and developing synaptic strength at these sites. In addition, differences in neuronal migration are implicated by the human DCDC1 gene, which also contains the AGCC motif in an intron. DCDC1 is expressed in both the fetal and adult brain and knockouts of this gene result in lissencephaly, the lack of normal convolutions in the brain, accompanied by unusual facial appearance, failure to thrive and severe psychomotor retardation. Alterations of regulation of the DCDC1 gene may therefore contribute to the smaller surface area of chimp brains. Finally, the WWC1 gene, while containing the motif in introns of both species, has 8 tandem repeats in chimps and 11 in humans, a subtle but potentially significant observation considering that the motif is just 35 base pairs away from a splice site. WWC1 has been shown to be

involved in hippocampal activation during memory retrieval and may be involved in both retrieval and re-encoding stored information. These findings raise the possibility that changes in copy number within AGCC motif loci may affect the distribution of splice variants and, thus, play a role in the intellectual disparity between humans and chimps. Studying the expression differences of these genes between chimps and humans and the interaction of these regions with DNA binding factors will be important next steps in verifying the functional consequences of these microsatellite variations. It is tempting to speculate that the association of these motifs with developmental and neurological genes contributes to the ability of global microsatellite content to accurately group species according to phenotypic differences. Whether this is causative, correlative, or merely coincidental is beyond the scope of this study, but our findings suggest that the phenomenon warrants further investigation, especially in light of the fact that microsatellites have been previously implicated in gene expression regulation (Rose and Beliakoff 2000) and alternative splicing (Lian and Garner 2005).

Based on our assessment, microsatellites vary predictably and consistently between and among species, with certain motifs characteristic of particular species (e.g., the chimpanzee-specific motifs TAGCC and TAGCCC, **Table 2**). This phenomenon suggests that there are possibly differences in replication or DNA repair machinery mechanisms that favor errors or corrections of errors in one species versus another. This hypothesis is not entirely without precedence, as differences in the repair rates of some microsatellite motifs have been previously reported. For example, the DNA mismatch repair protein, human postmeiotic segregation 2, has been shown to exhibit motif-specific bias at tetranucleotide repeat sequences (Shah and Eckert 2009). Alternatively, species-specific expansion of certain

genomic regions containing differential motifs might account for the global microsatellite differences observed between different species. While we were not able to identify any discernible pattern between microsatellites that varied in frequency between species and the non-coding genomic sequences we examined (SINEs, LINEs, and transcription factor binding sites), global microsatellite content might nonetheless be influenced by non-coding sequences (e.g., introns) that are under evolutionarily pressures that vary between species. In support of this concept, a recent study demonstrated that microsatellites are reliable molecular clocks that can be used to accurately de-convolute deep lineages of human genetic variation (Sun et al. 2009).

Utility of microsatellites and SNPs in phylogeny

SNPs have been widely utilized in phylogeny because of their remarkable periodicity, determined by the error rate of DNA polymerase and the recognition quotient of DNA repair machinery. (Boerwinkle) The mutation rate in microsatellites, however, is linked to strand slippage due to secondary structure formation and is related to purity, repeat length and copy number; as a result of these diverse mutation rates, potential differences in DNA repair corresponding to varied microsatellite secondary structures, and the poor ubiquity of each individual motif, global microsatellite content is not an ideal measure of spontaneous mutation across generations. Conversely, however, the heterogeneity of secondary structure and continuous nature of copy number polymorphism implies an a priori biologic functionality surpassing SNPs, many of which are transitions within the same class (purine or pyrimidine) and some of which are silent even in coding regions due to redundancy. As such, whether microsatellites can serve as superior global markers of phenotypically significant mutations remains unclear.

Indeed, microsatellite and SNP markers alike share the disadvantage of predominance in non-functional DNA regions, including intergenic regions and introns. (Boerwinkle). We did not find an over-representation of microsatellites in exons and promoter regions. In fact, the low microsatellite content in exons was an expected finding given the frameshift consequence of non-modulo 3 repeat polymorphisms. Intriguingly, however, there was a greater than 2 fold over-representation of differential microsatellites (between humans and chimps) within introns, relative to intergenic regions, given the 24% intron, 1.1% exon, 75% intergenic ratio documented in the Celera sequence (Venter et. al.) (derived from Table 4). Furthermore, this result contrasts with the intronic and intergenic distribution of SNPS, 8.21 and 8.44 per 10kb respectively based on the Celera CgsSNP database (Boerwinkle). While the evolutionary underpinnings of this divergence (ex. preferred mutational mechanisms in unstable intronic regions, mutual exclusivity, natural selection of biologically functional microsatellites or a combination of the above) remain difficult to definitively ascertain, this is perhaps the most suggestive global genomic data of a biologic functionality for microsatellites, particularly a possible role in the regulation of alternative splicing.

In addition, as described previously, microsatellites have an enhanced mutability rate, as much as 10^-4 per human locus per generation, compared to 10^-7 to 10^-9 for point mutations (Ellegren, 2004). As a result, microsatellites provide an increased polymorphism load for natural selection; while this increases the dissimilarity between individuals and, hence, may confound interpretations of inter-species variation in

microarray comparisons of a small sample of sequenced individuals, it on average provides a more sensitive molecular marker of natural selection, potentially useful in analyzing closely related species. Here, we construct almost identical phylogenic relationships using global microsatellite content (**Supplemental Fig. 6**) as are assembled from SNP divergence. Given that the mutational clocks are so different, it may be somewhat surprising that these phylogenic relationships agree so well as increased microsatellite mutation rates would be expected to detect enhanced "noise" from allelic reversion, co-selection of polymorphisms in divergent species due to shared environmental stressors and the sporadic profusion of non-functional microsatellites in intergenic regions. Our opinion is that the superior sensitivity of the microsatellite mutational clock outweighs the above factors and that microsatellite content and SNPs alike approach and surpass the phylogenic threshold for classification over many generations. Furthermore, there may be undiscovered mechanisms that synchronize the two clocks.

At the same time, the relatively static number of microsatellite loci, stable since the evolutionary decline of mobile genetic elements, and the association of certain motifs, especially (AC),(AG) and (AT) doublets, with lower rates of SNPs in flanking segments up to 10 kB in length (Amos), may confer a superior ability to align the genomes of species with large-scale chromosomal changes. In a survey of the above doublets, Amos et. al. further demonstrate that SNP density may be related to motif copy number and, incredibly, that while SNP density tends to exhibit a low-point near microsatellites, human-chimpanzee divergence tends to exhibit a peak in these regions. In conjunction

with our work, these findings begin to build a case for an evolutionary dichotomy between microsatellites and SNPs and, possibly, a greater biologic functionality for the former. Hence, our demonstration of the ability of global microsatellite content to differentiate species provides both evidence of a regulatory role for repeat sequences as well as possible utility as a global marker of phenotypically significant mutation.

In light of these hypotheses, demonstration of microsatellite functionality in the introns of genes that we have highlighted as potential actors in the human-chimp divergence (at either the mRNA or protein level) is essential. As the sequencing of more individuals in the human and chimp populations takes place, it may be of interest to selectively study motifs that exhibit dominant polymorphisms in these populations, as this may be an additional predictor of functionality.

Concluding Remarks

It is not clear whether global microsatellite differences contributes to speciation or is the result of species differences in any number of possible mechanisms, such as differences in replication or DNA repair. However, the consistent and reproducible patterns of microsatellite differences, coupled with their known involvement in a variety of human diseases, imply that they are worthy of further investigation and conceivably perform functions that have not yet been discovered due to the lack of a method to study them in both an individual and global context. Several researchers have suggested that microsatellite polymorphisms, unlike single-point mutations, might confer an evolutionary advantage (Hammock and Young 2005; King, Trifonov, and Kashi 2006; Young and Hammock 2007; Fondon et al. 2008). Our opinion is that microsatellites are not the supreme drivers of

speciation or gatekeepers of evolutionary change, but rather might function as facilitators of the natural variation that accompanies phenotypic alterations. Thus, complexity of function isn't a linear function of the count of microsatellite loci, many of which arise and depart from non-functional regions, but may instead be related to the shift or de novo generation of loci in functional genomic regions. As such, regardless of the interaction between replication and repair machinery, microsatellite copy number changes and other forms of mutation, the continuous trend of certain motifs across more developed species suggests a functional role and the genes implicated are potential candidates in explaining the human divergence. The custom array described here provides a means to directly investigate global microsatellite content, which may aid in elucidating the underlying mechanisms involved in microsatellite variation among species and same-species individuals and the consequences of these differences.

LIST OF TABLES

TABLE ONE - GENOMES HYBRIDIZED TO THE ARRAY 57

TABLE TWO - REPEAT MOTIFS THAT DIFFER BETWEEN SPECIES CORRELATE WITH SEQUENCED DATA..... 58

TABLE THREE - DIFFERENTIAL MICROSATELLITES ARE ASSOCIATED WITH GENES WHOSE FUNCTIONS CORRELATE WITH CHARACTERISTICS THAT DIFFERENTIATE SPECIES. 59

TABLE FOUR - SEQUENCE DATA INDICATES THAT THERE ARE CONSIDERABLE DIFFERENCES IN MICROSATELLITES AMONG HUMAN INDIVIDUALS. DIFFERENCES ARE GREATER BETWEEN HUMAN AND CHIMPANZEE GENOMES. 60

TABLE FIVE - GENE FUNCTIONS ASSOCIATED WITH DIFFERENTIAL MOTIFS BETWEEN THE RKO6 AND RK07 CELL LINES..... 61

LIST OF FIGURES

FIGURE ONE A— DIFFERENTIAL MOTIFS BETWEEN HOMINIDS AND NON-HOMINID ANIMALS AND PLANTS 63

FIGURE ONE B— DIFFERENTIAL MOTIFS BETWEEN HOMINIDS AND NON-HOMINID ANIMALS AND PLANTS 64

FIGURE TWO— INTENSITIES OF SELECT MICROSATELLITE MOTIFS ACROSS HUMANS, NON-HUMAN PRIMATES AND VARIOUS OTHER SPECIES. 65

FIGURE THREE— RKO6/RKO7 BENCHMARKS. 66

FIGURE FOUR—DIFFERENTIAL MOTIFS BETWEEN THE RKO6/RKO7 CELL LINES. 67

LIST OF SUPPLEMENTARY FIGURES

SUPP. FIGURE 1 A— NORMALIZED AND LOG TRANSFORMED SIGNAL VALUES FOR PROBES REPRESENTING WILD-TYPE, SINGLE MISMATCH, DOUBLE MISMATCH, AND DELETION PROBES 68

SUPP. FIGURE 1 B— NORMALIZED AND LOG TRANSFORMED SIGNAL VALUES FOR PROBES REPRESENTING VARYING MOTIF LENTGHS 68

SUPP. FIGURE 2— MICROSATELLITES DO NOT EXHIBIT A LENGTH-BASED PATTERN OF VARIATION AMONG PRIMATES, INCLUDING HUMANS AND OTHER HOMINIDS. 69

SUPP. FIGURE 3— TRANSCRIPTION FACTOR BINDING SITES THAT DIFFER BETWEEN HOMINIDS AND NON-HOMINID PRIMATES ARE NOT ASSOCIATED WITH HOMINID-SPECIFIC MICROSATELLITE MOTIFS. 71

SUPP. FIGURE 4 A/B — RELATIVE RATIOS OF TWO HOMINIDS VERSUS A NON-HOMINID PRIMATE ARE NOT A RESULT OF VARIATIONS IN SINES OR LINES. 72

SUPP. FIGURE 5— ALIGNMENT OF THREE PUBLISHED HUMAN GENOMES (REFERENCE, CELERA, AND VENTER) AND COMPARISON OF ALL INDIVIDUAL OCCURRENCES OF MICROSATELLITE SEQUENCES INDICATES HIGH LEVELS OF POLYMORPHISM AMONG HUMANS AND BETWEEN HUMANS AND CHIMPANZEES..... 74

SUPP. FIGURE 6 — ALIGNMENT OF THREE PUBLISHED HUMAN GENOMES (REFERENCE, CELERA, AND VENTER) AND COMPARISON OF ALL

INDIVIDUAL OCCURRENCES OF MICROSATELLITE SEQUENCES INDICATES HIGH LEVELS OF POLYMORPHISM AMONG HUMANS AND BETWEEN HUMANS AND CHIMPANZEES..... 76

SUPP. FIGURE 7— HIERARCHICAL CLUSTERING OF NON-MICROSATELLITE REPEAT PROBE INTENSITIES DOES NOT CORRECTLY CLASSIFY SPECIES..... 78

SUPP. FIGURE 8 — HIERARCHICAL CLUSTERING OF ULTRA-CONSERVED SEQUENCE PROBE INTENSITIES DOES NOT CORRECTLY CLASSIFY SPECIES. 79

SUPP. FIGURE 9 A/B — PHYLOGENETIC TREES PRODUCED BASED ON ARRAY INTENSITIES (A) OR COMPUTED COUNTS OF PUBLISHED GENOMIC SEQUENCES (B) OF THE SUM OF ALL MICROSATELLITE MOTIFS IN THE GENOMES SHOWN..... 80

SUPP. FIGURE 10 — PHYLOGENETIC TREE GENERATED FROM AN EXPANDED SET OF GENOMES. 82

Table 1: Genomes Hybridized to the Array (Garner et al.)

Sample ID	Species	Sex	Description
H1	*H. sapiens*	M	Caucasian
H2	*H. sapiens*	F	Caucasian
H3	*H. sapiens*	M	Eastern Indian
H4	*H. sapiens*	M	African
H5	*H. sapiens*	F	Mixed ancestry
H6	*H. sapiens*	M	Chinese
Chimp1	*P. troglodyte*	M	Chimpanzee
Chimp2	*P. troglodyte*	M	Chimpanzee
Chimp3	*P. troglodyte*	F	Chimpanzee
Gorilla1	*G. gorilla*	M	Lowland gorilla
Gorilla2	*G. gorilla*	M	Lowland gorilla
Gorilla3	*G. gorilla*	M	Lowland gorilla
Orang1	*P. pygmaeus*	M	Sumatran orangutan
Orang2	*P. pygmaeus*	M	Sumatran orangutan
Bab	*P. cynocephelus*	M	Baboon
Mac	*M. fascicularis*	M	Long-tailed macaque
Rhesus	*M. mulatta*	M	Rhesus monkey
Marm	*C. iacchus*	M	Marmoset
Bull	*B. taurus*	M	Angus bull
Dog	*C. lupus familiaris*	M	Alaskan husky
Mouse	*M. musculus*	M	Mouse
Chick	*G. gallus*	M	Rooster
Fly	*D. melanogaster*	-	Fruit fly
Corn	*Z. maize*	-	Corn plant
Arab	*A. thaliana*	-	Arabidopsis

Table 2: Repeat motifs that differ between species correlate with sequenced data. (Garner et al.)

Motif	**Ratio**		**Human Loci with the Motif**	**Human Genes with the Motif**
Array	Computed		Number	
Human vs. Non-human				
AATGG	158	449	3,087	22
ACTCC	23.6	12.9	209	12
CAGC	5.3	0.9	178	52
AACGG	25.7	2.3	23	1
Chimpanzee vs. Non-chimpanzee				
TAGCC	56.4	59.0	5	3
TAGCCC	18.7	0.7	16	4
Orangutan vs. Non-orangutan				
CCAGCC	7.0	6.2	371	89
Primate vs. Non-primate				
AACAT	9.4	5.9	197	13
Mammal vs. Non-mammal				
ACAGAG	19.1	11.4	604	22
AC	5.4	5.0	72,045	7,908
AACCG	0.06	0.2	2	2
Vertebrate vs. Non-vertebrate				
ATGG	28.1	33.4	4,932	957
AGGGTC	26.7	23.4	77	24
CGGTTT	0.02	0.1	1	1
Animal vs. Non-animal				
AACC	6.7	5.4	390	4
ACGCCG	0.08	0.1	0	0
Fruit-fly vs. Non-fruit fly				
AATGTG	28	75.7	20	8

Table 3: Differential microsatellites are associated with genes whose functions correlate with characteristics that differentiate species (Garner et al.)

Ontology	**Total Genes in Category**	**Genes with Motif**	**Statistic**
AATGG – hominid motif associated with 22 genes			
Organismal development	2,406	6	Z = 2.4
Anatomical structure development	2,130	5	Z = 2.0
Nervous system development	825	4	Z = 2.4, p = 0.01-0.05
ACTCC - hominid motif associated with 12 genes			
Eye development	84	1	Z = 4.6-5.3, p = 0.03
Nervous system development	825	1	p = 0.009
CAGC - hominid motif associated with 52 genes			
System development	1,892	36	p = 0.01-0.09
Nervous system development	825	6	Z = 2.4, p = 0.01-0.04
Embryonic development	471	5	p = 0.01-0.04
Anatomical structure morphogenesis	1,093	8	Z = 2.6
Axon guidance	76	2	Z = 3.6
AACAT – primate motif associated with 13 genes			
Hair and skin development and function	169	1	p = 0.06
Embryonic development	471	1	p = 0.06
Renal, urological system development	56	1	p = 0.06
ACAGAG – mammal motif associated with 22 genes			
Organ, tissue, and system development	3,307	17	p = 0.02-0.04
Bone and cartilage development	183	2	Z = 2.1-3.4, p = 0.02
AC - mammal motif associated with 7,908 genes			
Developmental process	3,266	1,447	Z = 5.9
Anatomical structure development	2,130	951	Z = 4.8
System development	1,892	830	Z = 3.8
Organ development	1,371	582	Z = 2.0

Nervous system development	825	371	Z = 3.1
ATGG = vertebrate motif associated with 957 genes			
Anatomical structure development	2,130	146	Z = 4.1
Nervous system development	825	155	Z = 6.9, p = 0.001-0.2
Behavior	369	64	Z 3.0 =, p = 0.003-0.2
AGGGTC vertebrate motif associated with 24 genes			
Organismal development	2,406	2	p = 0.03-0.04
System development	1,892	4	p = 0.03-0.04
Lung, respiratory tube development	133	2	Z = 3.0-3.1, p = 0.08

Table 4: Sequence data indicates that there are considerable differences in microsatellites among human individuals. Differences are greater between human and chimpanzee genomes. (Garner et al.)

Microsatellite Locations	**Ref**	**Celera**		**Venter**		**Chimpanzee**	
No.	Match	% Diff		Match	% Diff	Match	% Diff
Upstream*	4,032	2,791	31%	1,980	51%	673	83%
5’ UTR	27,660	18,911	32%	14,454	48%	3,977	86%
Intron	166,319	114,371	31%	87,836	47%	23,305	86%
Exon	1,124	1,030	8%	898	20%	507	55%
3’ UTR	5,141	3,276	36%	2,642	49%	784	85%
Downstream*	3,609	2,370	34%	1,840	49%	448	88%
Total Near Genes	207,885	142,749	31%	109,650	47%	29,694	86%
Intergenic	299,705	195,720	35%	149,555	50%	39,903	87%
Total	507,590	338,469	33%	259,205	49%	69,597	86%

Table 5: Results of gene functions associated with differential motifs between the RKO6 and RK07 cell lines. Notable functions include cell cycle control and neurological signaling. Both the APC gene (ACCCAC) and the MLH1 gene (GGGT) were implicated and suggest that the pathogenesis of colon cancer in HNPCC may involve an "anticipation" phenomenon in the tandem repeat region of MLH1 and may share the APC dys-regulation characteristic of sporadic colon cancer.

Motif	Gene	Function
	KCNIP1	may regulate neuronal excitability
	GOLM1	transport of protein cargo through the Golgi apparatus
	FAF1	Initiation of apoptosis
ACCCAC		
	PPM1B	dephosphorylation of cdks, over-expression causes growth arrest or death
	KDM4C	associated with esophageal squamous cell carcinoma
	UTRN	similar to dystrophin gene
	CTNB1	binds to the product of the APC gene, mutations cause colon cancer
	Col8A1	a sort chain collagen
	MI4548G	encodes a microRNA that results in translational inhibition of target mRNA
GTAGT		
	MAG12	expansion associated with dentatorubral and pallidoluysian atrophy
	CRLF2	mediates cytokine signaling
	PDE10a	mediates cAMP/cGMP signaling
	CTNNA1	cadherin associated protein
	GRIP1	glutamate receptor interacting protein
GTT	No matches found	
GTTT	No matches found	
GTTTT	No matches found	
GTTTTT	No matches found	
GGAGTT	No matches with known gene function	
TGGAGG		
	Interleukin 17A	role in inflammatory and autoimmune diseases such as rheumatoid arthritis
	Rtel1	involved in telomere elongation, located in gene rich cluster associated with potential tumor-related ge
	CORO7	golgi complex morphology and function
	FUS	regulation of gene expression and genomic integrity, defects cause Lou Gehrig's disease
GGT	No matches found	
GGGT		
	Col5A1	defects associated with ehlers-danlos disease

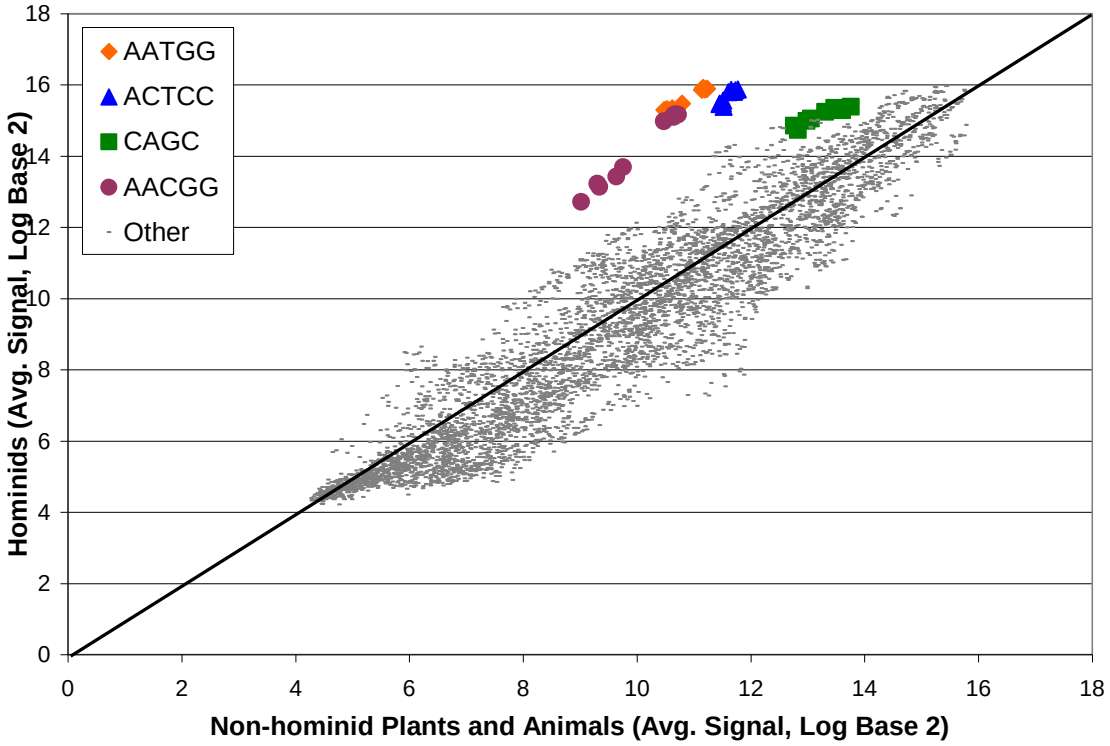
CNTFR	stimulates cell survival, dell differentiation in a variety of neuronal cell types
ADCYAAP1RA	mediates activity of adenylate cyclase
NK2	negative regulator of the Wnt-Beta catenin signalling system
FBLN2	extracellular matrix protein that plays a role in differentiation of skeletal, neuronal and cardiac structure
SAMM50	component of the sorting and assembly machinery of the mitochondrial membrane
BTBD1	binds topisomerase 1
PSTPIP1	mediates CD2-T cell actvation, defect cause PAPA syndrome
TNFAIP1	retinoic acid target gene in acute promyelocytic leukemia
TECTA	responsible for autosomal dominant nonsyndromic hearing impairment
GFRA1	mediates activation of the RET tyrosine kinase receptor. Candidate gene for Hirschsprung disease
DHRS3	oxidation/reduction of retinoids + steroids
BCYRN1	retrotranspositionally created, regulates dendritic protein biosynthesis
SARDH	nuclear gene encoding mitochondrial mutations lead to sarcosinemia
CPSF1	3 prime processing of pre-mrna
TRAPPC9	mutations associated with mental retardation
BLK	lymphoid tyrosine kinase
APBB2	involved in pathogenesis of alzheimer's disease
GSK3B	may be involved in the pathogenesis of Alzheimer's disease
MLH1	locus frequently mutated in hereditary non-polyposis colon cancer (HNPCC).
MYO9B	Polymorphisms in this gene are associated with celiac disease and ulcerative colitis susceptibility.
SEMA6B	central and peripheral nervous system development
Homo sapiens KIAA0427	involved in the translation initiation complex
DNAH1Y	interacts with axonal dynein
RCOR	neural specific gene expression
GPR68	G protein-coupled receptor
CD5	Immune function

GGTGTT

SMOC2	may control angiogenesis in tumor growth and myocardial ischemia

FIGURE LEGENDS:

Fig. 1: A survey of 5,356 microsatellite motif hybridization intensities (log base 2) of hominid (humans, chimpanzees, orangutans, and gorillas) and non-hominid (4 primates, 5 other animals, and 2 plants) DNA to a custom array identifies several motifs that are statistically significant and consistently differential for every cohort (**A**) (Garner et al.). Note that CCAG and AACGG also vary considerably between hominid individuals (see **Fig. 2**).



A

Panel **B** (Garner et al.) shows the average signal intensity (log base 2) for 6 human individuals (ordinate) compared to three chimpanzees (abscissa). The cluster of grey dots to the right of the motifs colored in orange represent the motif GAGCC, which was not considered as chimpanzee specific due to high variability between chimpanzee individuals. For simplicity, each motif, all cyclic permutations, and complements are designated by one motif name (e.g., AATGG represents all 5 cyclic permutations and its 5 complement's cyclic permutations).

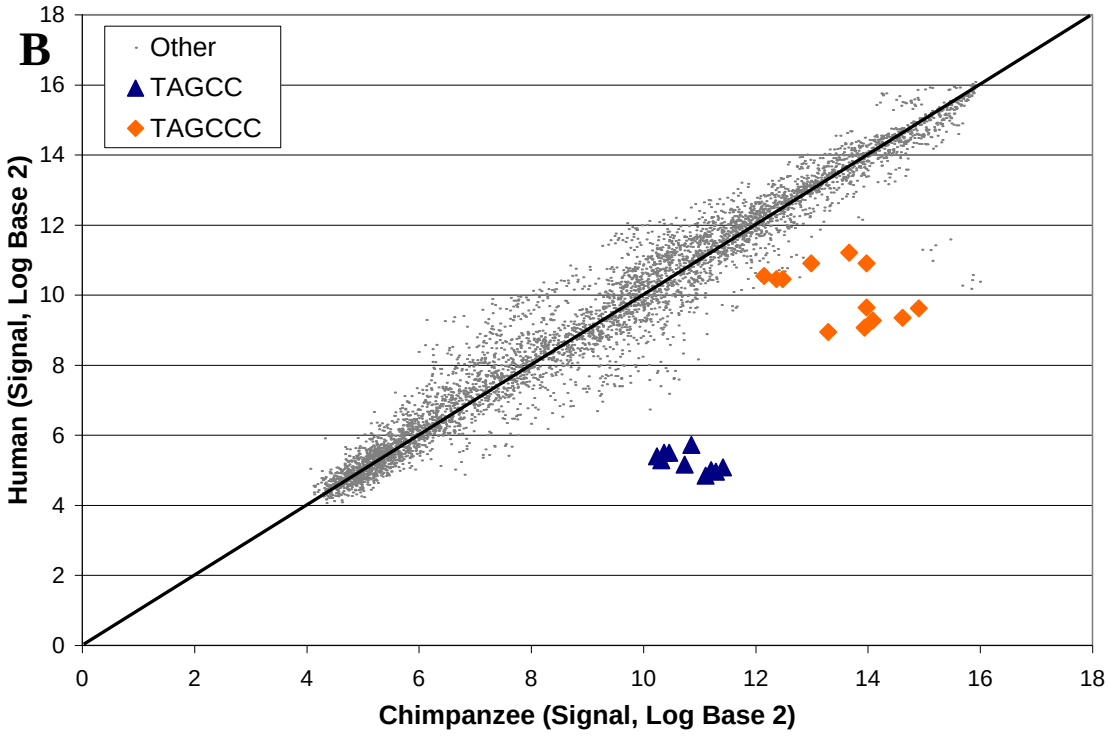


Fig. 2: Intensities of select microsatellite motifs across humans, non-human primates and various other species. Motifs shown, as indicated on each graph, represent all cyclic permutations. Error bars were computed as ± standard deviation units of arithmetic mean. An explanation of the sample abbreviations used on the abscissa along with a description of each sample is provided in **Table 1**. (Garner et al.)

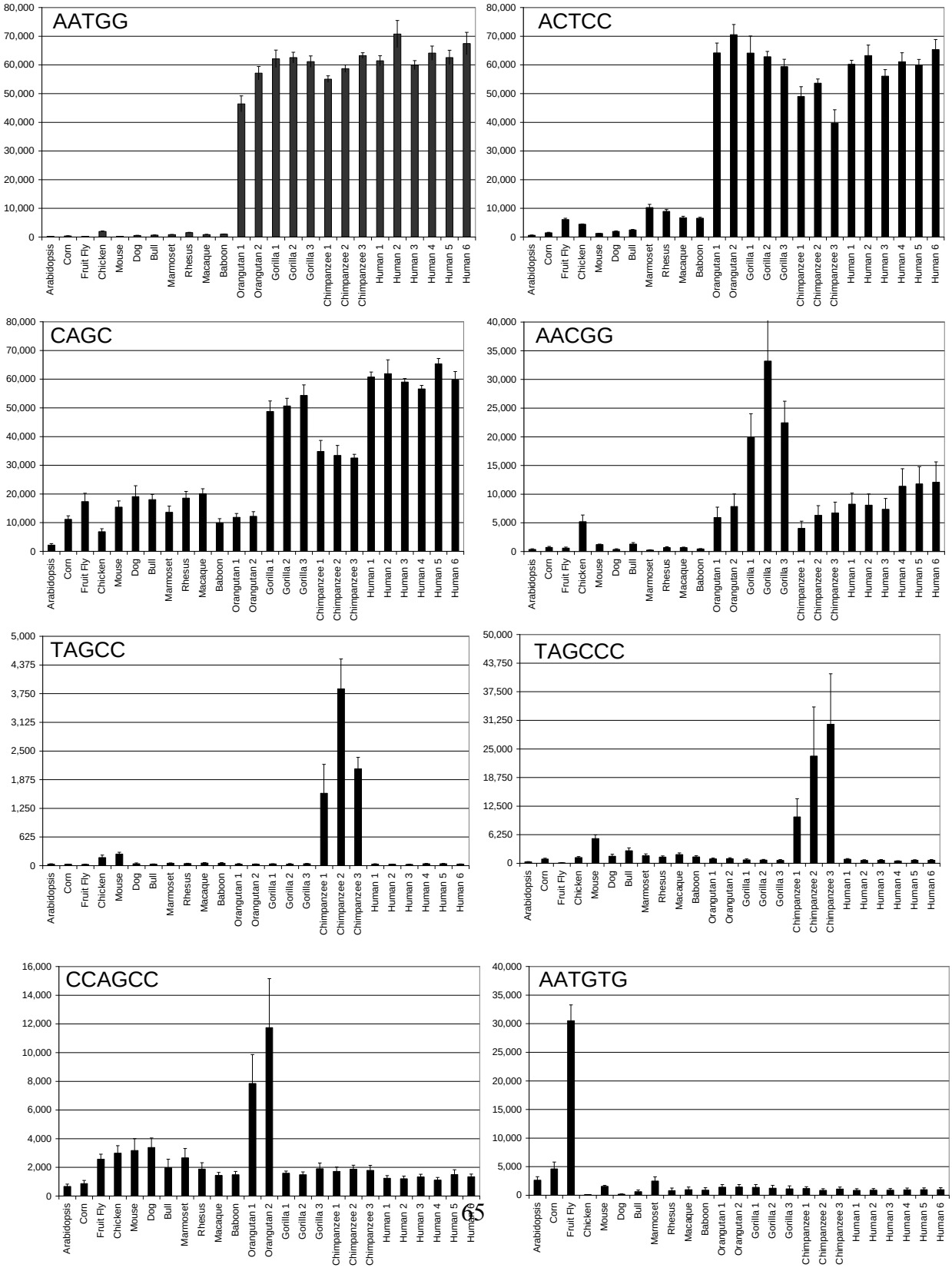
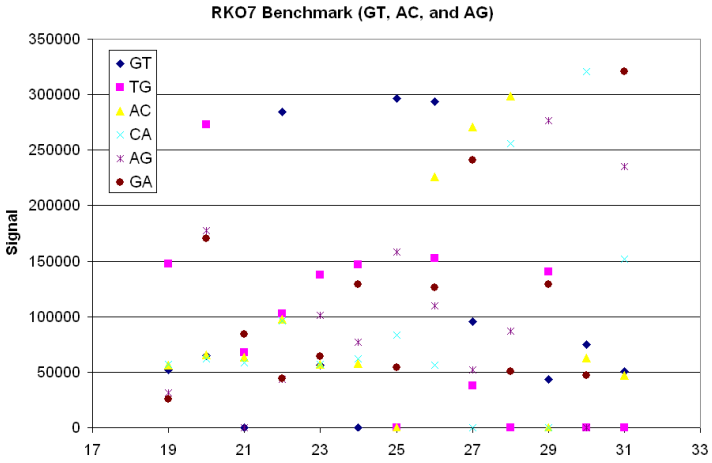
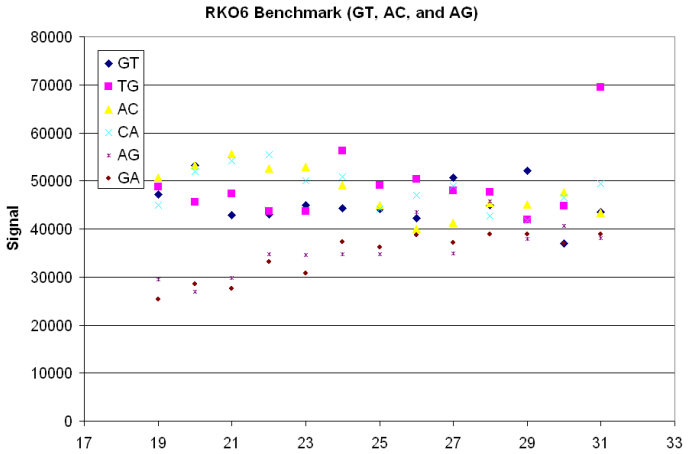
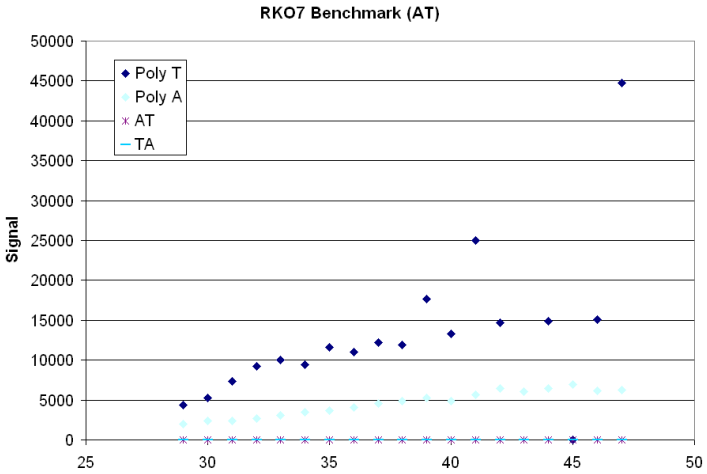


Figure 3: RKO6/RKO7 Benchmarks: GT, AC and AG doublets served as negative controls while poly A and poly T probes served as positive controls.







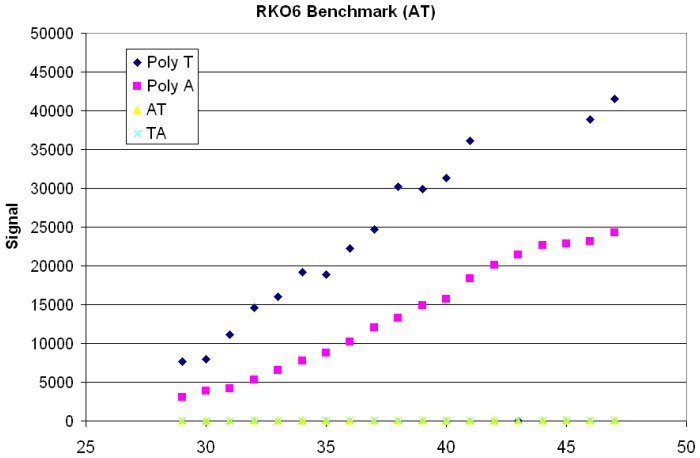
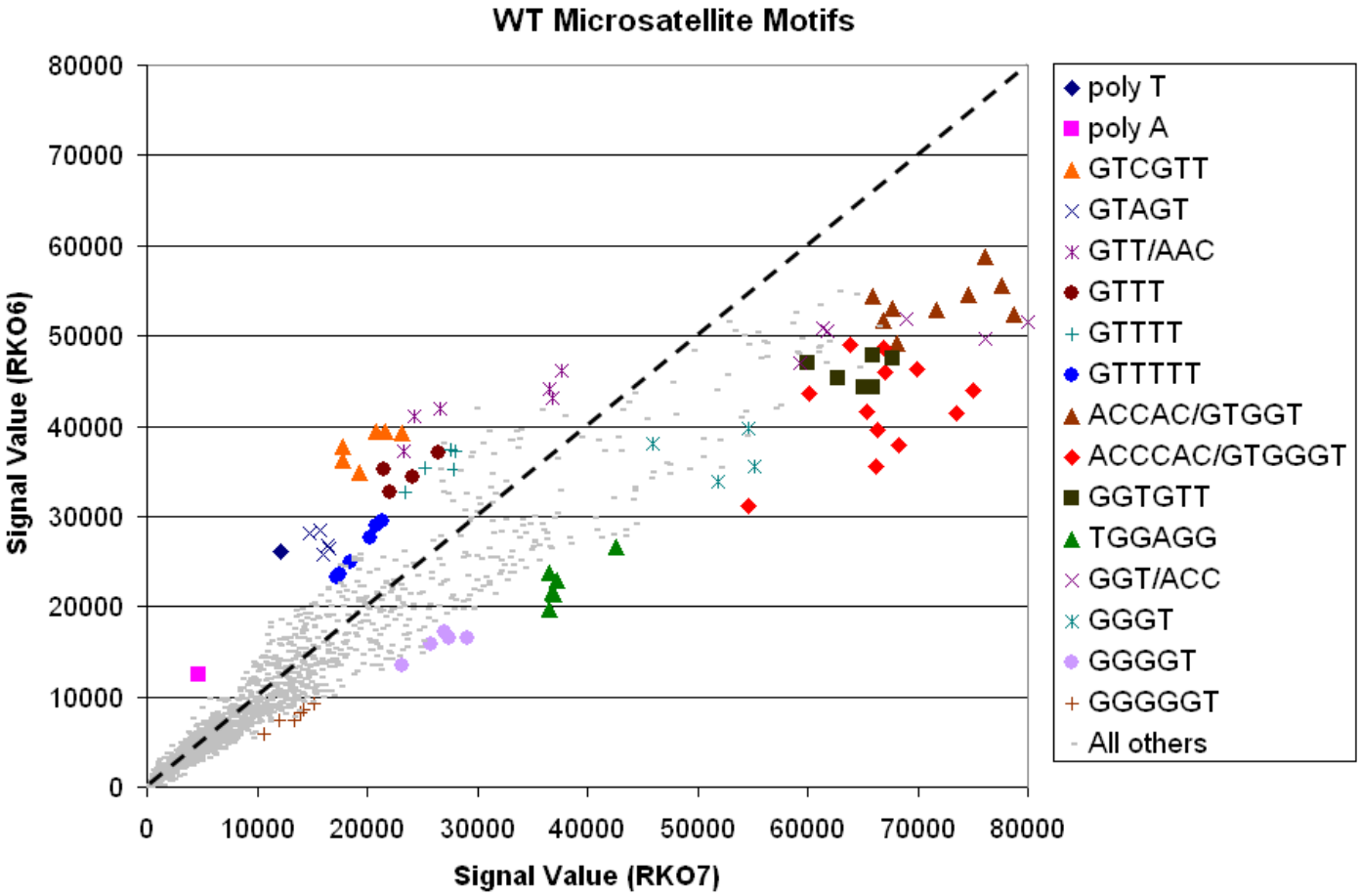
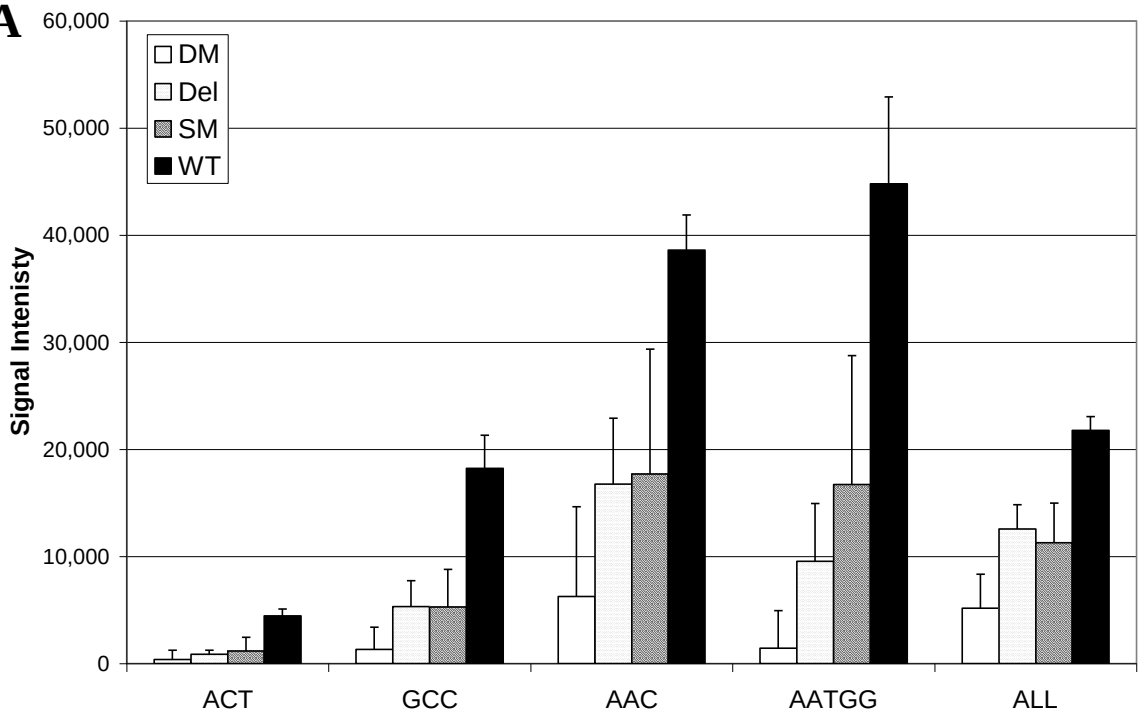


Figure 4: A survey of 5,356 microsatellite motif hybridization intensities in RKO6 (MLH1 knockout) and RKO7 (MLH1 competent) DNA identifies several differential motifs that are statistically significant.

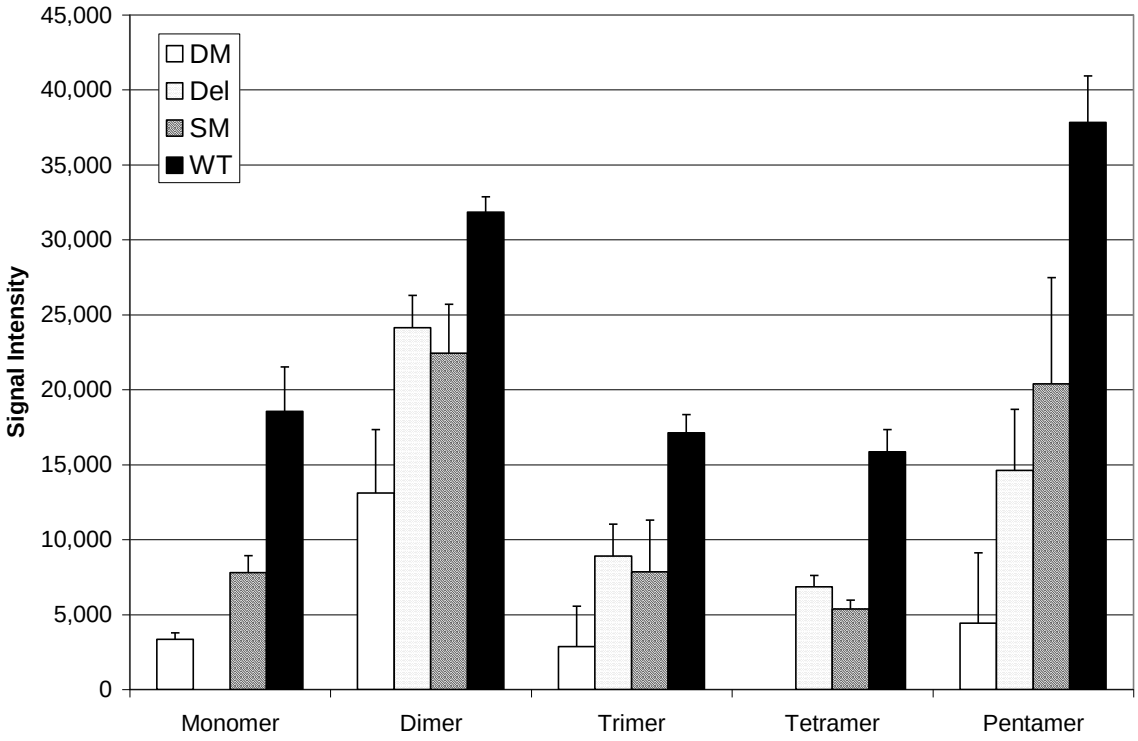


Supplementary Fig. 1 (Garner et al.)

A

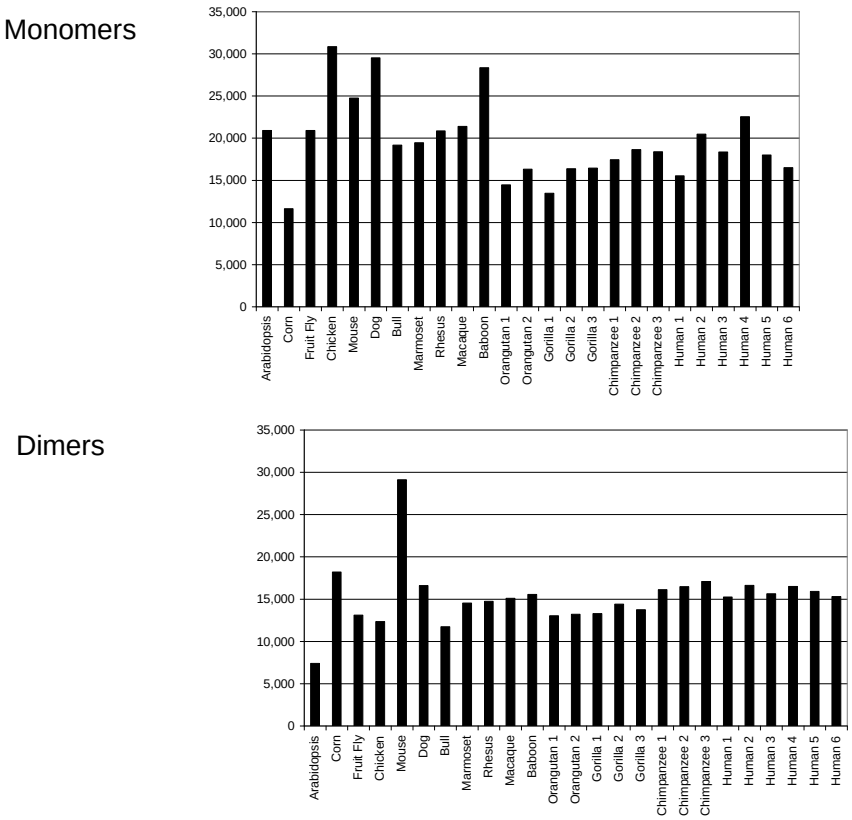


B

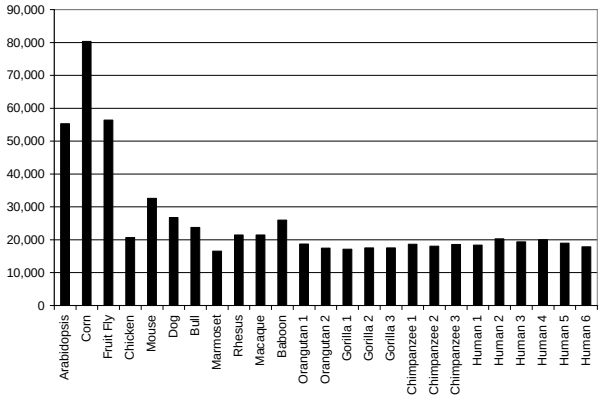


Supplementary Fig. 1 legend: The array exhibits significant specificity over a broad range of intensities. (**A**) Normalized and log transformed signal values for probes representing wild-type (WT), single mismatch (SM), double mismatch (DM), and deletion (Del) probes for three representative microsatellite motifs and all motifs are shown. The average signal intensities shown were calculated based on all cyclic permutations for the given motif for all 6 human DNA samples hybridized to the array. The resulting averages are displayed on the ordinate, and the standard deviations are shown as error bars. Comparisons were made for all microsatellite motifs represented on the array, and the four motifs shown were chosen from those results to represent a range of intensity values. "ALL" represents the results for all motifs and their corresponding hybridization control probes, with standard error bars calculated as the average standard deviation for all motif families. Note that all WT motif signals exceeded their corresponding mismatch probes, confirming binding specificity. (**B**) Bar graphs similar to those in (**A**), based on microsatellite length are shown. Deletion probes were not produced for monomers, as these would not alter the sequence nor total probe length, which was based on 2+4 (G+C) rules. Likewise, DM probes were not produced for tetramers, as most are palindromic in nature and thus would not generate relevant controls. Hybridization controls were not produced for hexamers, because the number required to cover all possible bp substitutions and deletions were prohibitive, given the space available on the array.

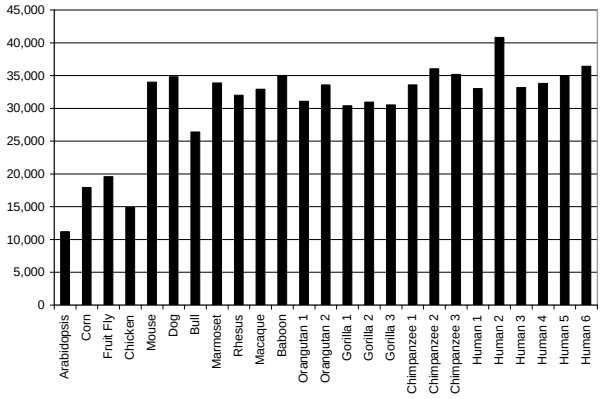
Supplementary Fig. 2 (Garner et al.)



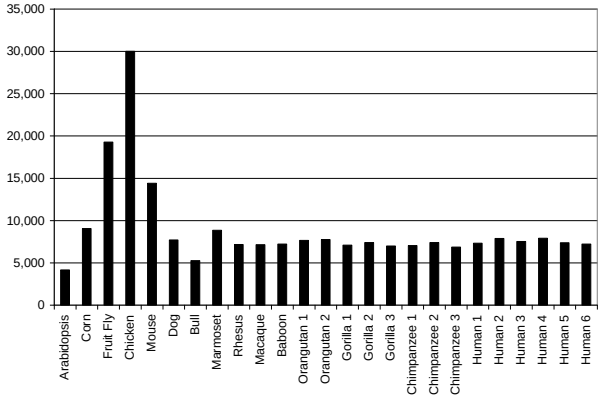
Trimers



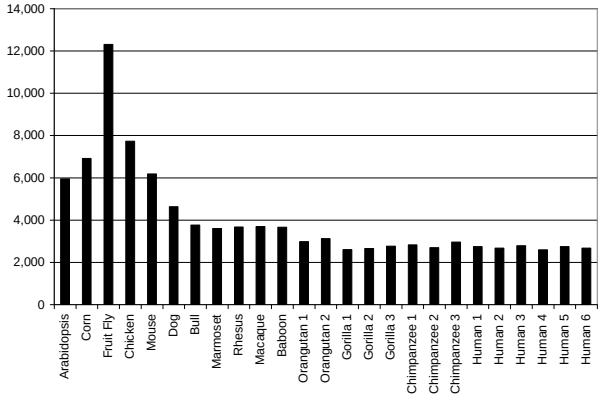
Tetramers



Pentamers

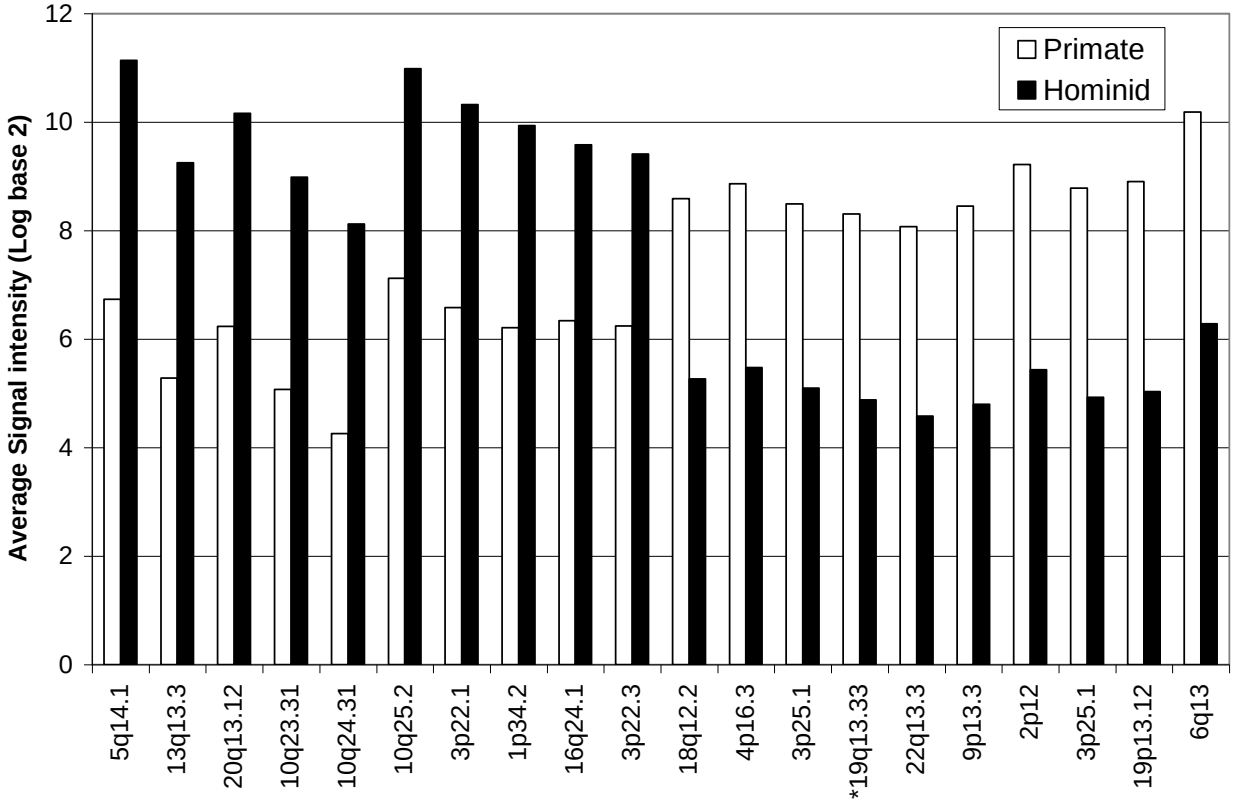


Hexamers



Supplementary Fig. 2 legend: Microsatellites do not exhibit a length-based pattern of variation among primates, including humans and other hominids. A survey of 5,356 microsatellite motif hybridization intensities (grouped by length, i.e., monomers, dimers, etc.) of hominid (humans, chimpanzees, orangutans, and gorillas) and non-hominid (4 primates, 5 other animals, and 2 plants) DNA to a custom array reveals little difference related to motif length. Array-derived signal intensity (the average of all motifs for each length shown) is displayed on the ordinate, and specie's samples are shown on the abscissa.

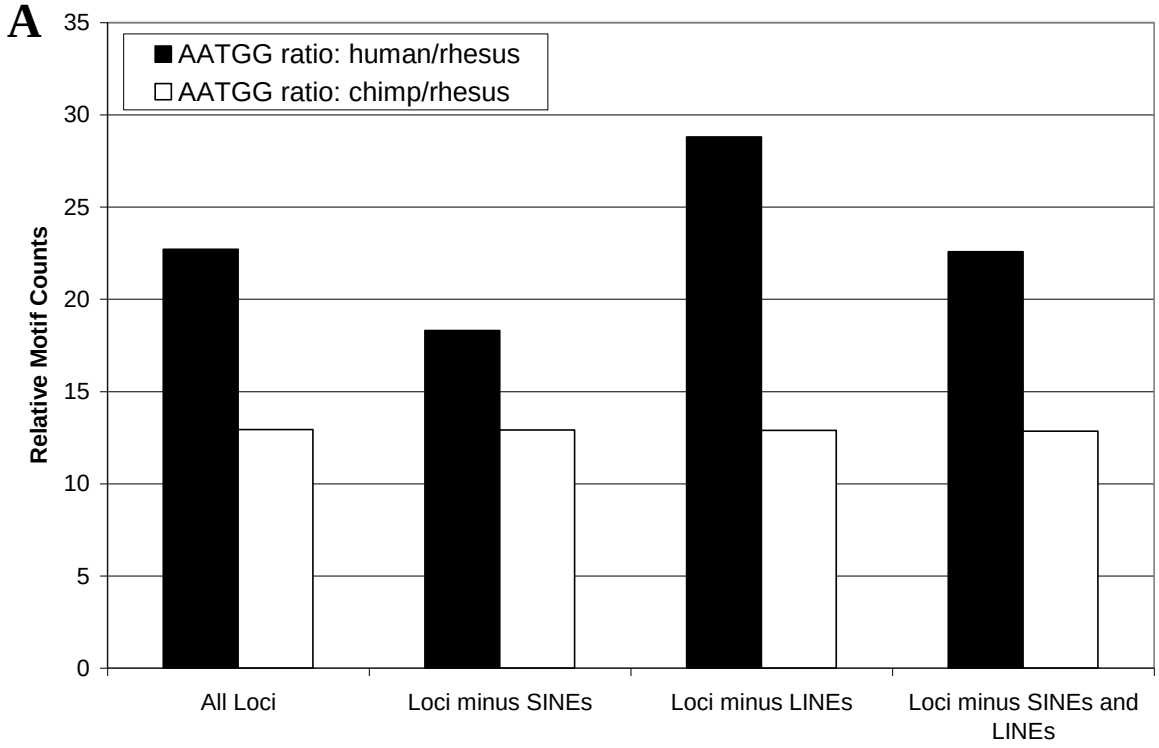
Supplementary Fig. 3 (Garner et al.)



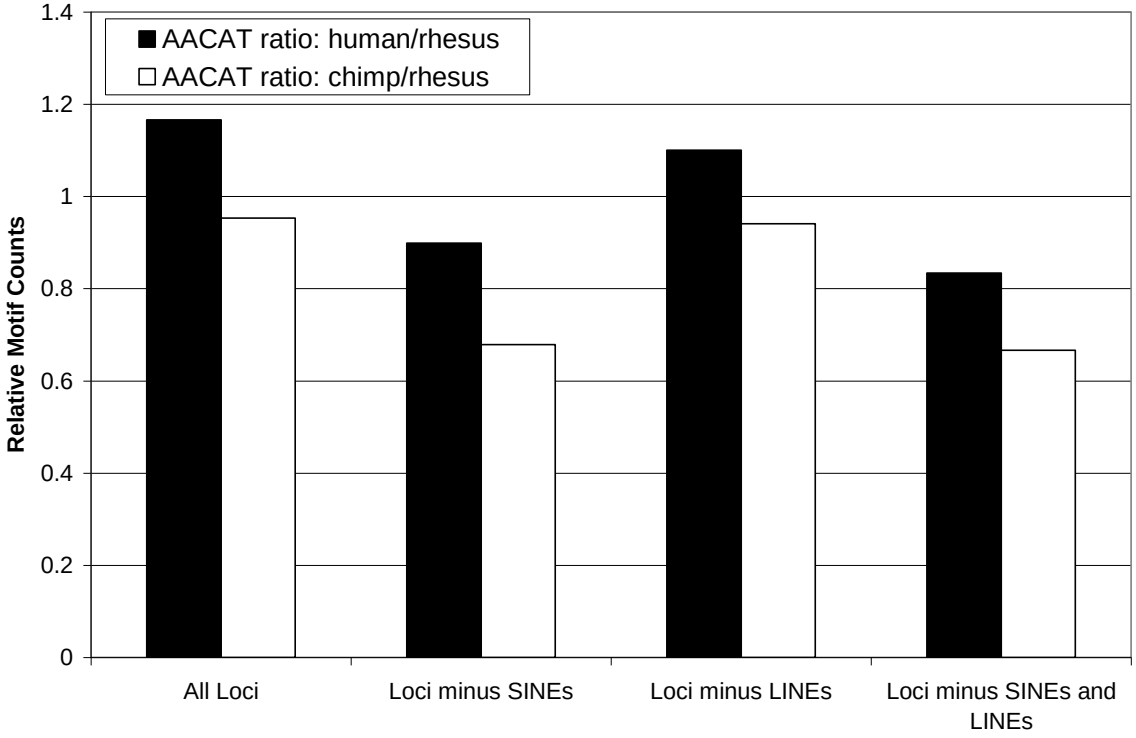
Supplementary Fig. 3 legend: Transcription factor binding sites that differ between hominids and non-hominid primates are not associated with hominid-specific microsatellite motifs. The average signal intensities (log base 2, shown on the ordinate) of probes representing each of the 4,777 transcription factor binding site sequences on the array for 14 hominid individuals (2 orangutans, 3 gorillas, 3 chimpanzees, and 6 humans) and 4 non-hominid primates (baboon, rhesus, macaque, and marmoset) were compared and the 20 most profound differences (10 highest and lowest hybridization intensities in hominids compared to non-hominid-primates) are shown. A BLAT search was performed

for each of these 20 probe sequences in the human genome, and the resulting cytogenic location is shown on the abscissa. One sequence was not found in the human genome using BLAT but was found by searching the chimpanzee genome and then finding the corresponding human sequence (indicated by an asterisk). The UCSC genome browser was subsequently used to align the human, chimpanzee, rhesus, marmoset, mouse, chicken, and stickleback genomes and the 500 bp surrounding the transcription factor binding sequence searched for hominid-specific microsatellite motifs (AATGG, ACTCC, CAGC, and AACGG). None of these four repeats, identified as higher in hominids compared to non-hominid species using the global microsatellite array, was tandemly repeated (2 or more copies of the core motif, cyclic permutations, or complements) in the vicinity (within 500 bp) of the 20 differential transcription factor binding sites shown.

Supplementary Fig. 4 (Garner et al.)

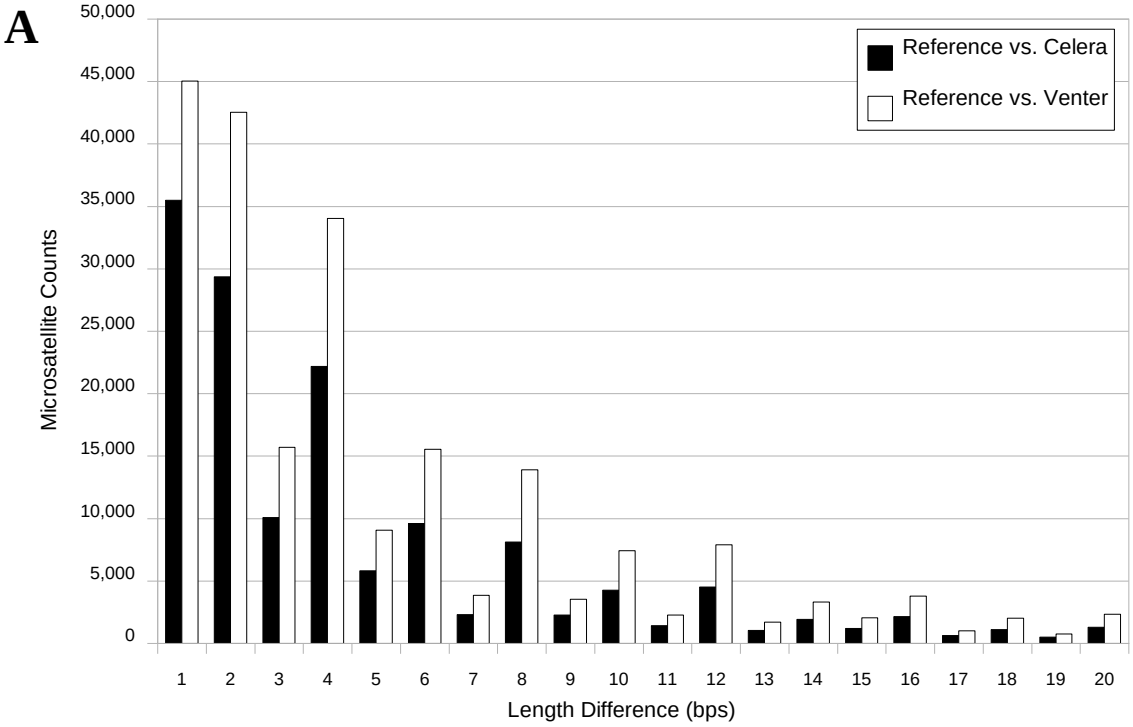


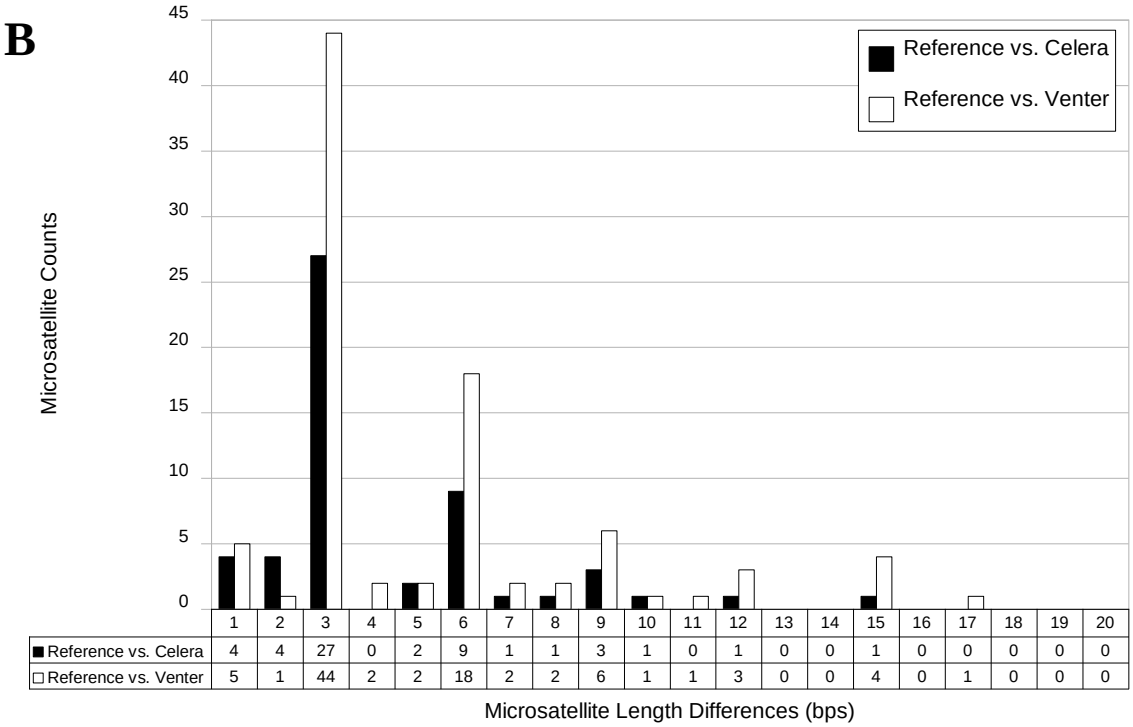
B



Supplementary Fig. 4 legend: Relative ratios of two hominids versus a non-hominid primate are not a result of variations in SINEs or LINEs. (**A**) Ratios of the total number of AATGG motif-containing loci (at least 20bp in length) in the published genomes of humans and chimpanzees (numerator) compared to the number of AATGG motifs found in the published genome of the rhesus macaque monkey (denominator) are shown. Human/rhesus ratios are shown as black bars, and chimpanzee/rhesus ratios are presented as white bars. This analysis included the ratios obtained when all AATGG motif-containing loci were considered, and ratios were also computed after elimination of AATGG-containing loci that coincided with SINEs, LINEs, or either of these repetitive entities. When SINEs and LINEs within 500 bp of the AATGG motif were eliminated, results were virtually identical (ratios ranged from 13.0 - 45.2), indicating that the higher incidence of this motif in hominids compared to non-hominids was not a function of longer (non-microsatellite) repeat expansions. (**B**) Bar graphs similar to those in (**A**), based on ratios of the number of AACAT motifs in humans and chimpanzees versus the rhesus macaque monkey. Based on array data and information obtained from published genomes, this motif was found to be more prevalent in primates (including humans, chimpanzees, and rhesus), compared to non-primate animals and plants. As shown, AACAT counts based on the published genomes of three primates (humans, chimpanzees, and rhesus) indicated little difference in this motif when SINEs and LINEs were included or eliminated from the analysis.

Supplementary Fig. 5 (Garner et al.)



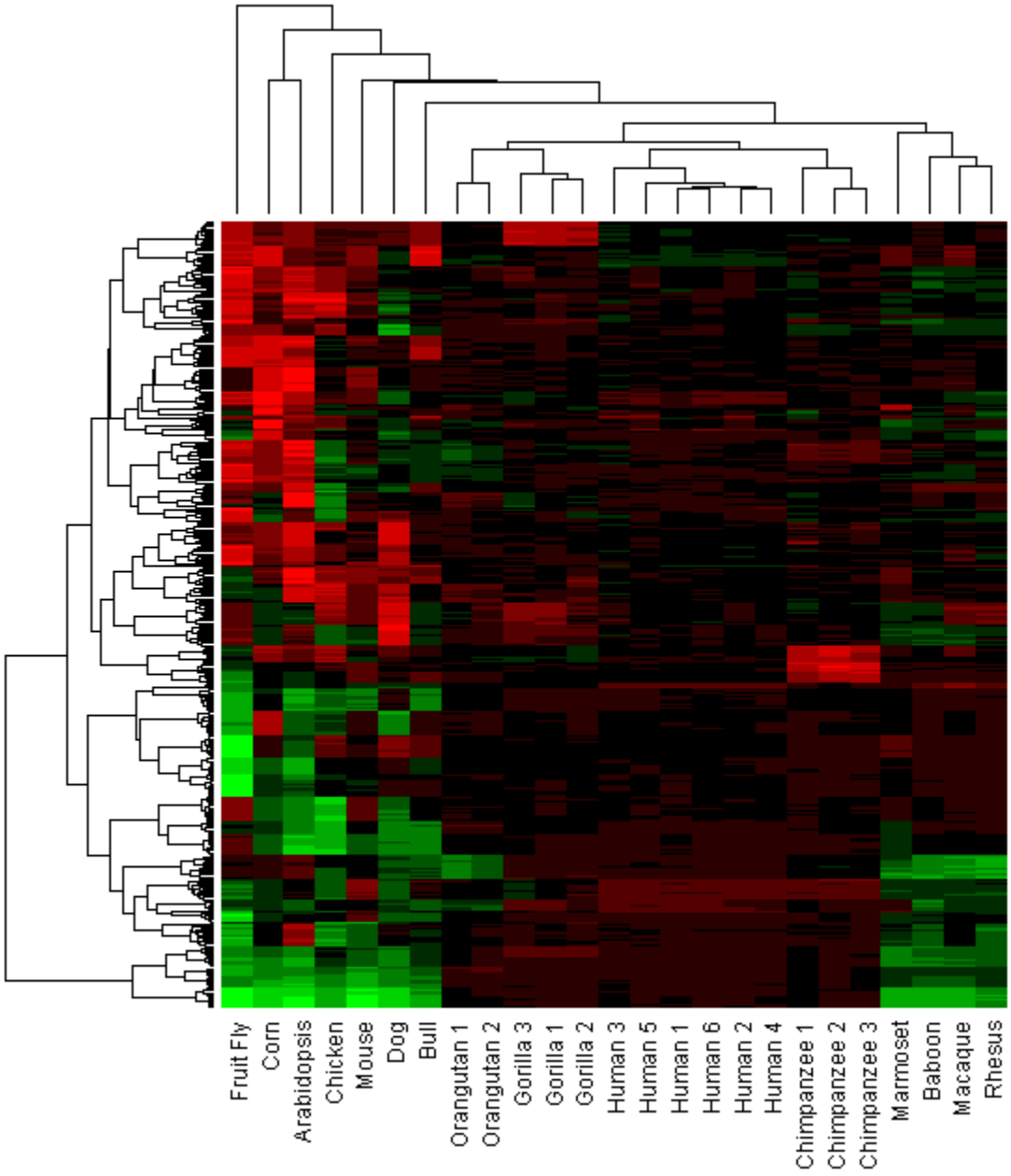


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Reference vs. Celera	4	4	27	0	2	9	1	1	3	1	0	1	0	0	1	0	0	0	0	0
Reference vs. Venter	5	1	44	2	2	18	2	2	6	1	1	3	0	0	4	0	1	0	0	0

Microsatellite Length Differences (bps)

Supplementary Fig. 5 legend: Alignment of three published human genomes (Reference, Celera, and Venter) and comparison of all individual occurrences of microsatellite sequences indicates high levels of polymorphism among humans and between humans and chimpanzees. Panel **A** shows the total of all well-aligned microsatellites in the Celera (black bar) and Venter (open bar), compared to the human reference sequence. Total incidences of microsatellites are shown on the ordinate (numbers of loci), and length differences (bp) between the human reference and corresponding Celera or Venter genomic sequence for each microsatellite are shown on the abscissa. Panel B shows the comparison results of microsatellites in the human reference genome that were aligned to the Celera and Venter sequences, limited to loci within gene exons. Actual numbers are given in a table beneath the graph, which shows that the majority of differences are modulo-3.

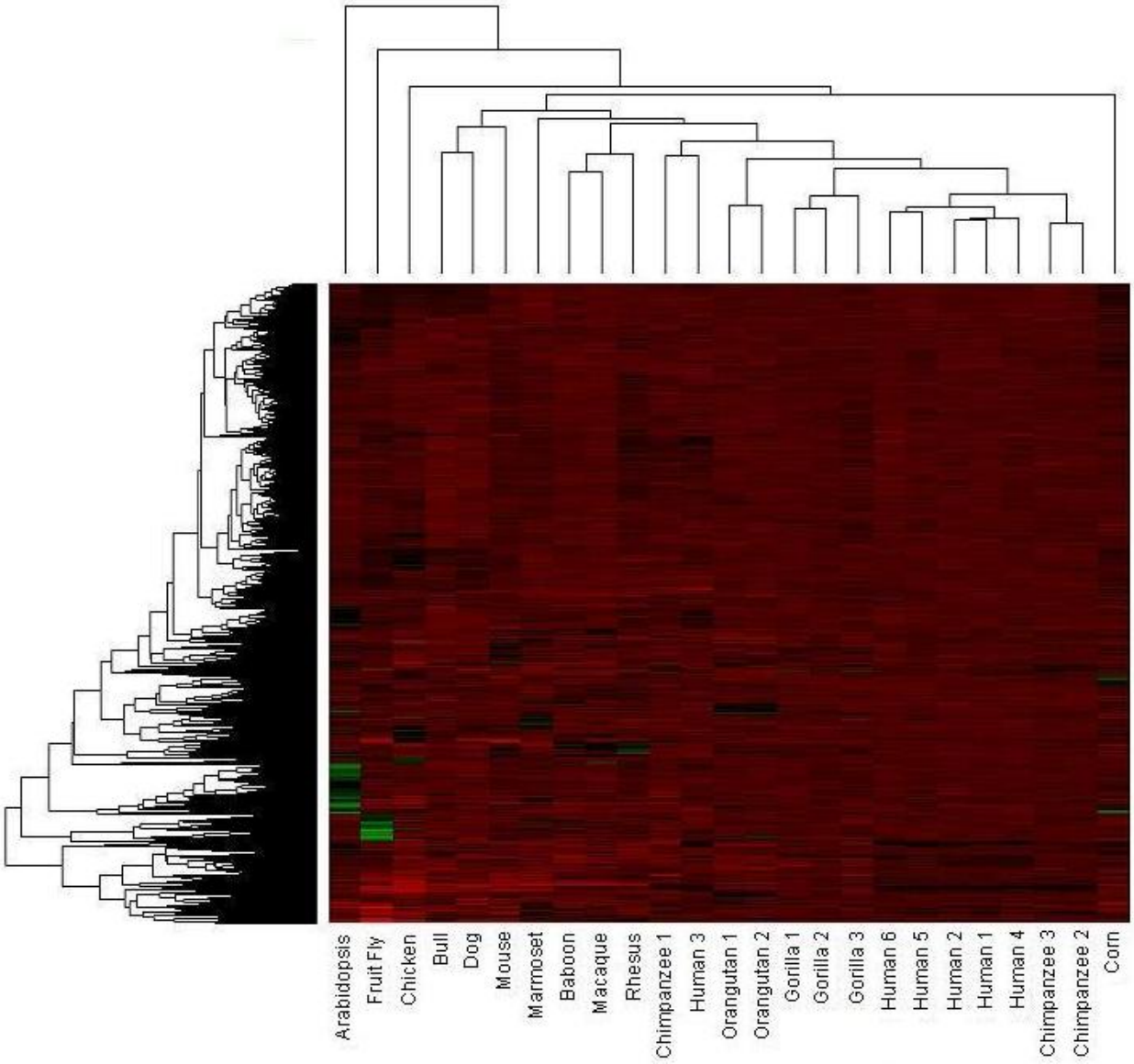
Supplementary Fig. 6 (Garner et al.)



Supplementary Fig. 6 legend: Hierarchical clustering classifies humans, hominids, and non-hominid primates apart from other animals and plants. R software was used to perform hierarchical clustering on normalized intensity values of all 5,356 WT microsatellite probes for all 25 samples. Vertical and horizontal nodes represent samples and microsatellite motif probes, respectively. As shown, primates clustered apart from other animals and the two plants. Likewise, hominids were subdivided from primates, and humans clustered apart from chimpanzees, gorillas, and orangutans. The innermost (far right) vertical nodes appropriately correspond to individual motif families (i.e., each node includes all cyclic permutations for a particular motif, such as CAG, AGC, and

GCA). The colors represent relative levels of intensity, with green, black and red indicative of lower, median, and high intensity values. An explanation of the samples and abbreviations used (column labels) are provided in the text, **Table 4**.

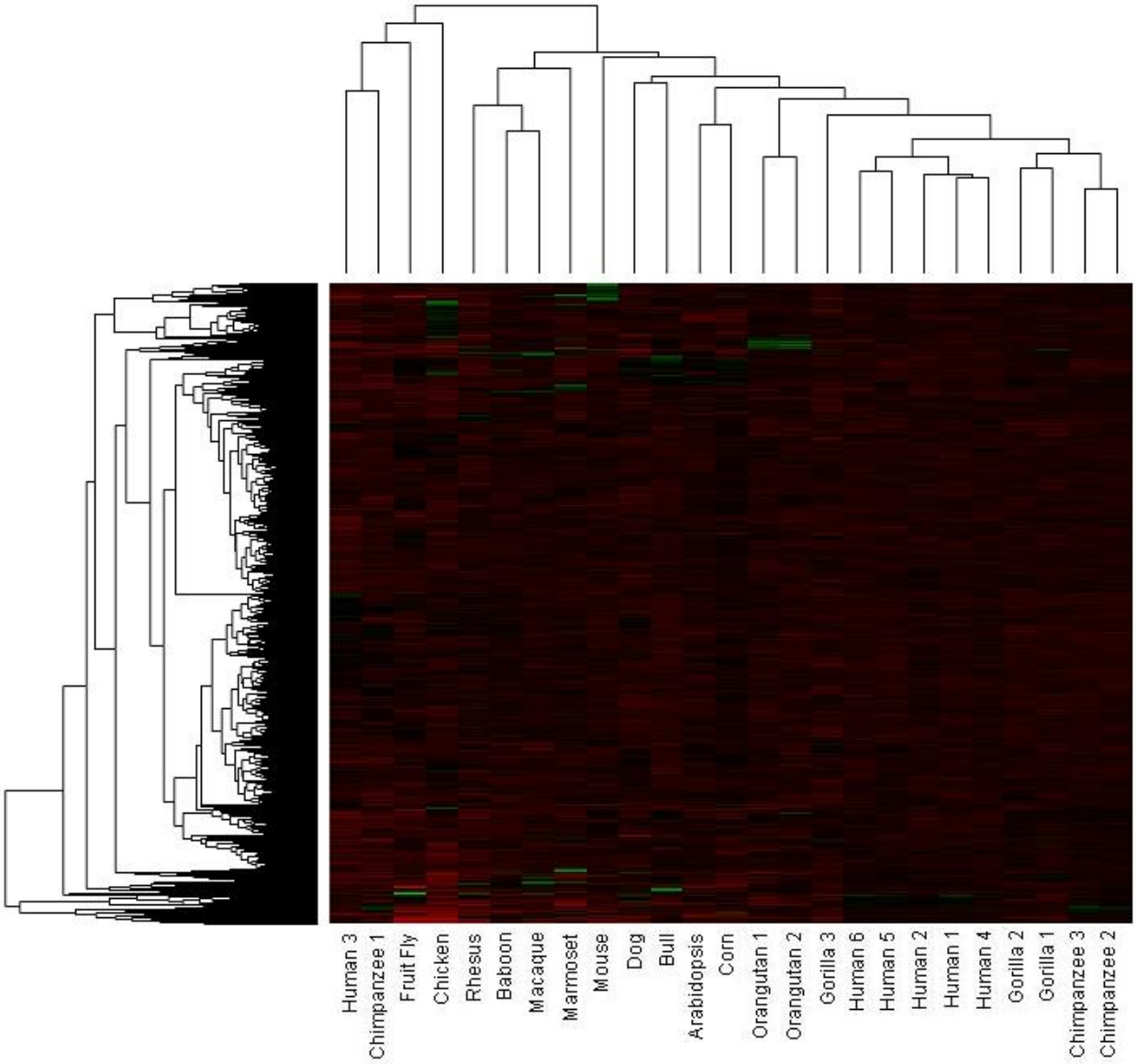
Supplementary Fig. 7 (Garner et al.)



Supplementary Fig. 7 legend: Hierarchical clustering of non-microsatellite repeat probe intensities does not correctly classify species. R software was used to perform hierarchical clustering on normalized intensity values of 5,356 Rebpase (non-microsatellite repeat sequences, including ALUs, SINEs, and LINEs) probes for all 25 samples. Vertical and horizontal nodes represent samples and probes, respectively. As shown, not all humans clustered together, as was also the case for chimpanzees, and hominids were not appropriately separated from other animals and plants (as evidenced by column node organization). The colors represent relative levels of intensity, with green, black and red indicative of lower, median, and high intensity values. An explanation of the samples and abbreviations used (column labels) are provided in the

text, **Table 1**. Clustering of all 22,072 Repbase probes produced almost identical results (not shown).

Supplementary Fig. 8 (Garner et al.)

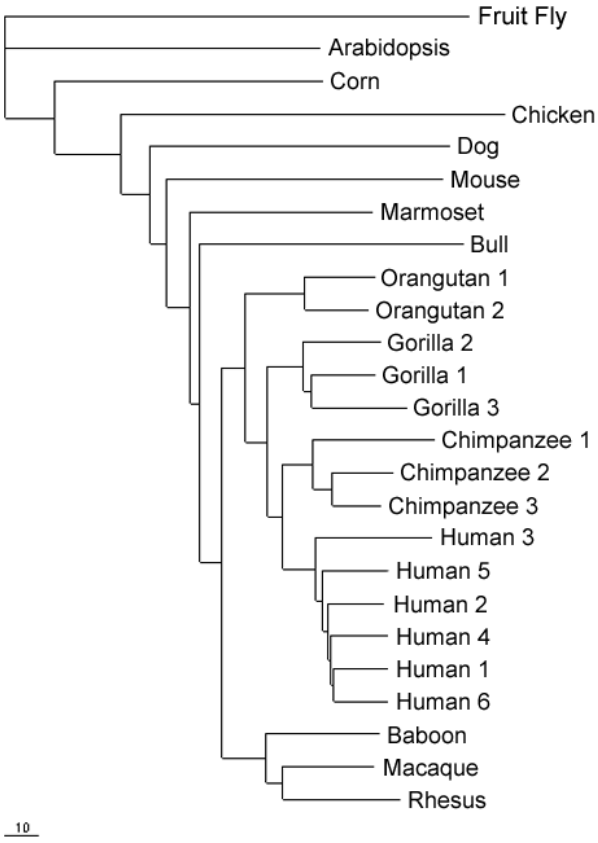


Supplementary Fig. 8 legend: Hierarchical clustering of ultra-conserved sequence probe intensities does not correctly classify species. R software was used to perform hierarchical clustering on normalized intensity values of all 3,848 wild-type ultra-conserved sequence probes for all 25 samples. Vertical and horizontal nodes represent samples and probes, respectively. As shown, not all humans clustered together, as was also the case for chimpanzees, and hominids were not appropriately separated from other

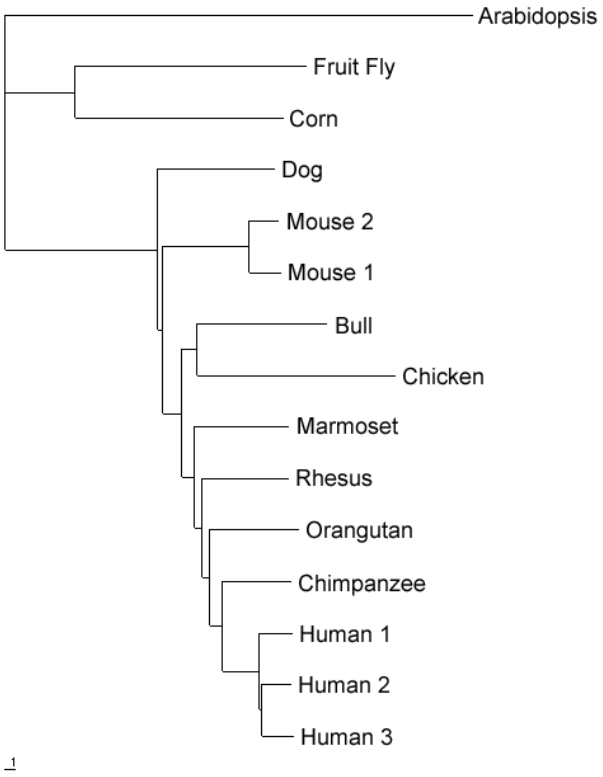
animals and plants (as evidenced by column node organization). The colors represent relative levels of intensity, with green, black and red indicative of lower, median, and high intensity values. An explanation of the samples and abbreviations used (column labels) are provided in the text, **Table 1**.

Supplementary Fig. 9 (Garner et al.)

A

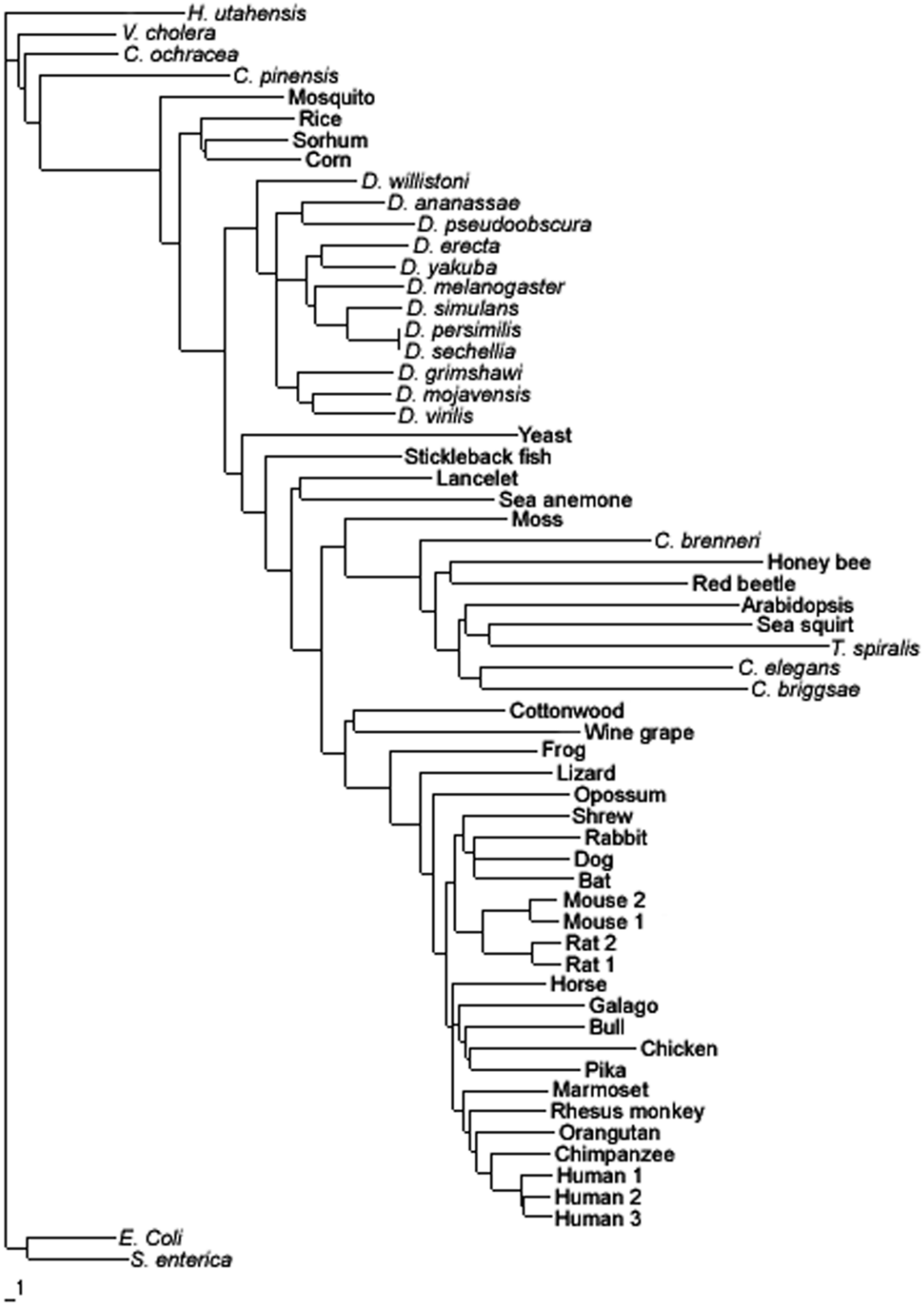


B



Supplementary Fig. 9 legend: Phylogenetic trees produced based on array intensities (A) or computed counts of published genomic sequences (B) of the sum of all microsatellite motifs in the genomes shown. Euclidian distance based on the array intensities or log transformed counts of all wild-type microsatellite motifs (5,356 probes) was computed for all pairwise comparisons across all 25 arrays (or 14 sequenced genomes that corresponded to samples hybridized to the array). The resulting distance matrix was used to produce a phylogenetic tree, using the neighbor-joining method within the PHYLIP software suite and TreeView (Page 1996). The scale bar (shown at bottom left) relates to branch lengths and can be interpreted as the number of evolutionary "steps" between nodes. As is evident from these trees, global microsatellite intensities correlate with known taxonomy relationships, consistent with proposed evolutionary relationships in the tree of life (Maddison and Schultz 2007).

Supplementary Fig. 10 (Garner et al.)



Supplementary Fig. 10 legend: Phylogenetic tree generated from an expanded set of genomes. Euclidian distance based on computed counts of all 5,356 possible microsatellites (including all cyclic permutations and complement sequences) was calculated for 60 published genomes (listed and described in **Supplementary Table 1**). The resulting coefficients were used to produce a distance matrix and phylogenetic tree, using the neighbor-joining method within the PHYLIP software suite and TreeView. The scale bar shown at bottom left relates to branch lengths and can be interpreted as the number of evolutionary "steps" between nodes.

Works Cited

Buschiazzo, E., and N. J. Gemmell. 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays **28**:1040-1050.

C. L. Galindo, L. J. McIver, J. F. McCormick, M. A. Skinner, Y. Xie, R. A. Gelhausen, K. Ng, N. M. Kumar and H. R. Garner. 2009. Global microsatellite content potentially distinguishes humans, primates, animals, and plants. Molecular Biology and Evolution: **26(12)**; 2809-2819.

Cooper, G., D. C. Rubinsztein, and W. Amos. 1998. Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. Hum Mol Genet **7**:1425-1429.

D'Souza, U. M., and I. W. Craig. 2006. Functional polymorphisms in dopamine and serotonin pathway genes. Hum Mutat **27**:1-13.

Donaldson, Z. R., F. A. Kondrashov, A. Putnam, Y. Bai, T. L. Stoinski, E. A. Hammock, and L. J. Young. 2008. Evolution of a behavior-linked microsatellite-containing element in the 5' flanking region of the primate AVPR1A gene. BMC Evol Biol **8**:180.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet **5**:435-445.

Fahima, T., M. S. Roder, K. Wendehake, V. M. Kirzhner, and E. Nevo. 2002. Microsatellite polymorphism in natural populations of wild emmer wheat, Triticum dicoccoides, in Israel. Theor Appl Genet **104**:17-29.

Fink, S., L. Excoffier, and G. Heckel. 2006. Mammalian monogamy is not controlled by a single gene. Proc Natl Acad Sci U S A **103**:10956-10960.

Fink, S., L. Excoffier, and G. Heckel. 2007. High variability and non-neutral evolution of the mammalian avpr1a gene. BMC Evol Biol **7**:176.

Fondon, J. W., 3rd, and H. R. Garner. 2004. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A **101**:18058-18063.

Fondon, J. W., 3rd, E. A. Hammock, A. J. Hannan, and D. G. King. 2008. Simple sequence repeats: genetic modulators of brain function and behavior. Trends Neurosci **31**:328-334.

Fondon, J. W., 3rd, G. M. Mele, R. I. Brezinschek et al. 1998. Computerized polymorphic marker identification: experimental validation and a predicted human polymorphism catalog. Proc Natl Acad Sci U S A **95**:7514-7519.

Gatchel, J. R., and H. Y. Zoghbi. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet **6**:743-755.

Weis E, Galetzka D, Herlyn H, Schneider E, Haaf T. 2008. Humans and chimpanzees differ in their cellular response to DNA damage and non-coding sequence elements of DNA repair-associated genes. Cytogenet Genome Res:**122(2)**:92-102.

Institute for Human Genetics, Johannes Gutenberg University, Mainz, Germany.

Hammock, E. A., and L. J. Young. 2004. Functional microsatellite polymorphism associated with divergent social structure in vole species. Mol Biol Evol **21**:1057-1063.

Hammock, E. A., and L. J. Young. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. Science **308**:1630-1634.

Huang, Q., A. Beharav, Y. Li, V. Kirzhner, and E. Nevo. 2002. Mosaic microecological differential stress causes adaptive microsatellite divergence in wild barley, Hordeum spontaneum, at Neve Yaar, Israel. Genome **45**:1216-1229.

Kashi, Y., and D. G. King. 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet **22**:253-259.

Kelkar, Y. D., S. Tyekucheva, F. Chiaromonte, and K. D. Makova. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res **18**:30-38.
King, D. G., E. N. Trifonov, and Y. Kashi. 2006. Tuning knobs in the genome: evolution of simple sequence repeats by indirect selection. Oxford University Press, New York.
Kloor M, Sutter C, Wentzensen N, Cremer FW, Buckowitz A, Keller M, von Knebel Doeberitz M, Gebert J. 2004. A large MSH2 Alu insertion mutation causes HNPCC in a German kindred. Hum Genet: **115(5)**:432-8.
Lander, E. S.L. M. LintonB. Birren et al. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.
Levy, S., G. Sutton, P. C. Ng et al. 2007. The diploid genome sequence of an individual human. PLoS Biol **5**:e254.
Li, Y. C., M. S. Roder, T. Fahima, V. M. Kirzhner, A. Beiles, A. B. Korol, and E. Nevo. 2002. Climatic effects on microsatellite diversity in wild emmer wheat (Triticum dicoccoides) at the Yehudiyya microsite, Israel. Heredity **89**:127-132.
Lian, Y., and H. R. Garner. 2005. Evidence for the regulation of alternative splicing via complementary DNA sequence repeats. Bioinformatics **21**:1358-1364.
Page, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci **12**:357-358.
Pearson, C. E., K. Nichol Edamura, and J. D. Cleary. 2005. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet **6**:729-742.
Pennisi, E. 2008. Evolutionary biology. Deciphering the genetics of evolution. Science **321**:760-763.
Rose, A. B., and J. A. Beliakoff. 2000. Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. Plant Physiol **122**:535-542.
Sawyer, L. A., J. M. Hennessy, A. A. Peixoto, E. Rosato, H. Parkinson, R. Costa, and C. P. Kyriacou. 1997. Natural variation in a Drosophila clock gene and temperature compensation. Science **278**:2117-2120.
Shah, S. N., and K. A. Eckert. 2009. Human postmeiotic segregation 2 exhibits biased repair at tetranucleotide microsatellite sequences. Cancer Res **69**:1143-1149.
Sultan, M., M. H. Schulz, H. Richard et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321**:956-960.
Sun, J. X., J. C. Mullikin, N. Patterson, and D. E. Reich. 2009. Microsatellites are molecular clocks that support accurate inferences about history. Mol Biol Evol.
Uddin, M., D. E. Wildman, G. Liu, W. Xu, R. M. Johnson, P. R. Hof, G. Kapatos, L. I. Grossman, and M. Goodman. 2004. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. Proc Natl Acad Sci U S A **101**:2957-2962.
Varela MA, Amos W. 2010. Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. Genomics **95(3)**:151-9. Epub 2009 Dec 21.
Venter, J. C.M. D. AdamsE. W. Myers et al. 2001. The sequence of the human genome. Science **291**:1304-1351.
Vowles, E. J., and W. Amos. 2006. Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. Mol Biol Evol **23**:598-607.
Webster, M. T., N. G. Smith, and H. Ellegren. 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc Natl Acad Sci U S A **99**:8748-8753.

Xose S Puente, Gloria Velasco, Ana Gutiérrez-Fernández, Jaume Bertranpetit, Mary-Claire King, and Carlos López-Otín. 2006. Comparative analysis of cancer genes in the human and chimpanzee genomes. BMC Genomics **7**:15.
Yongjian Guo and D Curtis Jamison. The distribution of SNPs in human gene regulatory regions. 2005. BMC Genomics **6**:140
Young, L. J., and E. A. Hammock. 2007. On switches and knobs, microsatellites and monogamy. Trends Genet **23**:209-212.
Young, L. J., R. Nilsen, K. G. Waymire, G. R. MacGregor, and T. R. Insel. 1999. Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. Nature **400**:766-768.
Zamorzaeva, I., E. Rashkovetsky, E. Nevo, and A. Korol. 2005. Sequence polymorphism of candidate behavioural genes in Drosophila melanogaster flies from 'Evolution canyon'. Mol Ecol **14**:3235-3245
Zhongming Zhaoa, Yun-Xin Fua, David Hewett-Emmetta and Eric Boerwinkle. 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution Gene **312**: 207-213

VITAE

EDUCATION

→ University of Texas Southwestern Medical School at Dallas
M.D. with Distinction in Research, Class Quartile Rank: 1st
→ Columbia University
Bachelor of Arts

RESEARCH

PUBLICATIONS

1. C. L. Galindo, L. J. McIver, J. F. McCormick, M. A. Skinner, Y. Xie, R. A. Gelhausen, K. Ng, Neil M Kumar, and H. R. Garner. Global microsatellite content distinguishes humans, primates, animals, and plants. Submitted to Molecular Biology and Evolution.

2. Cristi L Galindo, Mounir Errami, Michael Skinner, L Danielle Olson, David Watson, Jing Li, John F. McCormick, Lauren J. McIver, Neil M Kumar, Thinh Q Pham, Harold R Garner. Transcriptional profile of isoproterenol-induced cardiomyopathy and comparison to exercise-induced cardiac hypertrophy and human cardiac disease. Submitted to BMC Genomics.

ABSTRACT/POSTER PRESENTATIONS

1. Kumar, Neil, Nguyen, My-hanh, and Garner, Harold. Survey of micro-satellite associated cis-regulatory elements in cancer. UTSW Annual Medical Student Research Forum Abstract + Poster Session 2006.

2. Neil M Kumar, C. L. Galindo, M. Skinner, R. Gelhausen, T. S. Q. Pham, and H. R. Garner. Global Microsatellite Variations Correlate With Neurological and Morphological Features That Distinguish Humans and Chimpanzees. UTSW Annual Medical Student Research Forum Abstract + Poster Session 2007.

Experience: UT Southwestern Medical School

Summer 2006
Position: Student in Medical Student Summer Research Program,
Mentor: Harold Garner, Ph.D.

Description: Briefly, I used the POTION program to survey a selection of known microsatellite motifs in various promoters associated with cancer and predicted polymorphism according to purity, repeat copy number and genomic location. Corresponding transcription factors were selected with the TRANSFAC program and I assayed variable binding with microsatellite polymorphism using a gel-shift.

Summer 2007
Position: Student in Medical Student Summer Research Program,

Mentor: Harold Garner, Ph.D.

Description: Using a custom micro-array to measure hybridization intensities of every possible repetitive nucleotide motif from 1-mers to 6-mers, we examined 25 genomes. We showed that global microsatellite content varies predictably by species and particular motifs are characteristic of one species versus another.**2009**

Experience: Burke Research Institute

8/2004-5/2006
Position; Paid Research Technician
Mentor: Dr. Raj Ratan

I investigated the protective effects of arginase in the ischemic endoplasmic reticulum stress response in rat embryonic neurons, and the role of ATF-4 and CHOP signaling in the process using embryonic cell culture, western blot, and MTT assay.

VOLUNTEER ACTIVITIES AND ORGANIZATIONS

➔ Medical School
- Monday Clinic Volunteer; free clinic for indigent patients, coached MS1 and MS2 students with patient evaluations, 2009
- Swine Flu Investigation Volunteer with Dallas Health Department; served on patient evaluation team, 2009
- UTSW Multicultural Talent Show; performed in the Raas dance, Dec 06 – Mar 07
- American Medical Association, 2009 – present
- Texas Medical Association, 2009 – present
- Radiology Interest Group, 2009 – present

➔ Undergraduate
- Burke Rehabilitation Hospital: Nursing Assistantship, June - August 2003
- PS 106 Elementary School Tutor: tutored 5th graders in math and writing, 9/02-12/02
- Hindu Students Organization, 2003-2006

INTERESTS

Mathematics, Classical Indian music, Philosophy, Drums, Chess, Basketball, Tennis

.