Published in final edited form as:
*Behav Genet*. 2014 September ; 44(5): 487–497. doi:10.1007/s10519-014-9662-x.

# SIMPLE SEQUENCE REPEATS IN THE NATIONAL LONGITUDINAL STUDY OF ADOLESCENT HEALTH: AN ETHNICALLY DIVERSE RESOURCE FOR GENETIC ANALYSIS OF HEALTH AND BEHAVIOR

**Brett C. Haberstick**[1], **Andrew Smolen**[1], **Gary L. Stetler**[1], **Joyce W. Tabor**[2], **Taylor Roy**[1], **H. Rick Casey**[1], **Alicia Pardo**[1], **Forest Roy**[1], **Lauren A. Ryals**[1], **Christina Hewitt**[1], **Eric A. Whitsel**[4], **Carolyn T. Halpern**[2,3], **Ley A. Killeya-Jones**[2], **Jeffrey M. Lessem**[1], **John K. Hewitt**[1], and **Kathleen Mullan Harris**[2,5]

[1]Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, Colorado, USA

[2]Carolina Population Center, University of North Carolina, Chapel Hill, North Carolina, USA

[3]Department of Maternal and Child Health, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, USA

[4]Department of Epidemiology, Gillings School of Global Public Health and Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

[5]Department of Sociology, University of North Carolina, Chapel Hill, North Carolina, USA

## Abstract

Simple sequence repeats (SSRs) are one of the earliest available forms of genetic variation available for analysis and have been utilized in studies of neurological, behavioral, and health phenotypes. Although findings from these studies have been suggestive, their interpretation has been complicated by a variety of factors including, among others, limited power due to small sample sizes. The current report details the availability, diversity, and allele and genotype frequencies of six commonly examined SSRs in the ethnically diverse, population-based National Longitudinal Study of Adolescent Health (Add Health). A total of 106,743 genotypes were generated across 15,140 participants that included four microsatellites and two di-nucleotide repeats in three dopamine genes (*DAT1*, *DRD4*, *DRD5*), the serotonin transporter (5HTT), and monoamine oxidase A (*MAOA*). Allele and genotype frequencies showed a complex pattern and differed significantly between populations. For both di-nucleotide repeats we observed a greater allelic diversity than previously reported. The availability of these six SSRs in a large, ethnically diverse sample with extensive environmental measures assessed longitudinally offers a unique resource for researchers interested in health and behavior.

### Keywords

DRD4; DAT1; 5HTTLPR; MAOA; DRD5; Add Health

Corresponding Author: Brett C. Haberstick, Institute of Behavioral Genetics, University of Colorado, Boulder, Campus Box 447, Boulder, Colorado 80309-0447, USA, Phone: 303-735-5388, Fax: 303-492-8063, Brett.Haberstick@Colorado.edu.

## Introduction

Simple sequence or tandem repeats are genomic elements that are prone to changes in repeat number and are thus frequently polymorphic. These sequences are commonly found at high density in a gene's promoter as well as in other areas where they provide opportunities for the modulation of gene expression, as well as the structure and function of RNA and proteins. Approximately 3% of the human genome is accounted for by simple sequence repeats (SSRs; Lander, 2001). Though SSRs have been thought to be non-functional or 'junk DNA' (Ellegren, 2004), some of the genetic variation induced is thought to contribute to phenotypic variation (Fondon et al, 2008; Hamada et al 1984; Kashi et al, 1997).

Microsatellites and minisatellites are two classes of repetitive sequences and are thought to arise through 'slippage mutations,' which increase or decrease the number of repeats without altering the sequence (Montgomery et al, 2013). Rates of mutation for SSRs are an order of magnitude greater than for single nucleotide polymorphisms (SNPs) and are a function of several locus-specific factors including, among others, the motif length and the number of repeats (Fondon et al, 2008). Motif length, in particular, has been suggested to exert a 'mutationally-adjustable' quantitative influence on many aspect of a gene's function (Kashi et al, 1997; King, 1997). Microsatellies are tandem repeat lengths consisting of 1 to 10 base-pairs (bp) while minisatellites are composed of repeat lengths of 10 to 100 bp Both differ from 'satellites,' which are repeat lengths of >100 bp. Both micro- and minisatellites can have complex allele frequency distributions that differ between populations and are often referred to as variable number tandem repeats (VNTRs) (Jeffreys et al, 1987 Nakamura et al, 1998).

Another class of SSRs is an Insertion-deletion (indel) polymorphism. Similar to microsatellites, indels are seen in gene promoter, coding, and splice site regions (Mills et al, 2011; Montgomery et al, 2013), and are known to impact genetic and phenotypic variation. Indels are the second most frequent type of polymorphism in the human genome (McCarroll et al, 2006; Durbin et al, 2010; Mills et al, 2006) and range between 1 bp and 10,000 bp in length (Weber et al, 2002, 1000 Genomes Consortium, 2010; Mills et al, 2006; Bhangale et al, 2005), though the vast majority (0.99) of indels are <100 bp (Mills et al, 2011). With the availability of high-throughput sequence data and improved discovery software, it has been estimated that there are 1.4 to 2.8 million indels distributed across all 24 autosomes and each sex chromosome with rates varying between populations and individuals (Montgomery et al, 2013; Shen et al, 2013, Mills et al, 2011; 1000 Genomes Project Consortium, 2010). Moreover, high rates of linkage disequilibrium ($r^2 > 0.80$) between many indels and common SNPs available on commercially available arrays (Mills et al, 2011; Eichler, 2006) further suggests their potential importance in biological functioning that may possibly extend to individual differences at the phenotypic level.

In the current report, we detail the genetic diversity of six, common SSR polymorphisms in 15,140 participants from the National Longitudinal Study of Adolescent Health (Add Health), which represents the largest genotyping effort of these polymorphisms to date. These SSRs include the microsatellites (VNTRs) in the third exon of the dopamine D4

receptor (*DRD4*), 3′ untranslated region of the dopamine transporter (*DAT1*, Locus Symbol: *SLC6A3*), promoter region of the monoamine oxidase A gene (*MAOA-uVNTR*); two di-nucleotide repeats, one in the 5′ region of the dopamine D5 receptor (*DRD5*) and the other in the second intron of the MAOA gene; and the indel polymorphism (*5HTTLPR*) located in the promoter region of the serotonin transporter (Locus Symbol: *SLC6A4*).

## Methods

### Background

Add Health used a school-based design to select a nationally representative sample of schools in the United States in 1994, and interviewed all students in grades 7–12 who attended these schools during the 1994–1995 school year. From school rosters, a self-weighting sex- and grade-stratified core sample of adolescents and a parent were randomly selected for in-home interviews. From the in-school responses, supplemental samples were also drawn based on ethnicity (Cuban, Puerto Rican, and Chinese], genetic relatedness to siblings, physical disability, and other characteristics yielding a total sample size of 20,745 adolescents interviewed in their homes in 1995 (Wave I, 79% response rate). This adolescent cohort has been followed prospectively through time with follow-up interviews in 1996 (Wave II, 89% response rate], 2001–2002 (Wave III, 77% response rate) and 2008–2009 (Wave IV, 80% response rate). For the interested reader, the longitudinal design of Add Health has been described elsewhere (Harris et al, 2006, 2012, 2013).

Since 1994 (Wave I), Add Health has provided numerous opportunities for genetic research. The study design embedded a genetically informative sample that has been followed through four waves of assessments and includes more than 3000 pairs of adolescents with varying degrees of genetic relatedness (e.g. twins, full-siblings, half-siblings, and adolescents who grew up in the same household but have no biological relationship). During the Wave III in-home interview, salivary DNA samples were collected from the twin, full- and half-siblings in the genetic pairs sample (N = 2600) with a compliance rate of 83%. In the genetic pairs sample, relationship status was confirmed using 11 highly polymorphic, unlinked short tandem repeat (STR) markers (see Harris et al, 2006 for further details). Four VNTRs and three SNPs were also genotyped. The collection of DNA greatly expanded the utility and power of the genetic pairs sample by allowing behavioral genetic studies of adolescent health to move from variance decomposition to the testing of specific hypotheses about the influence of individual genes and their contribution to individual differences in the context of environmental circumstances.

### Participants

Participants for the current study were those interviewed at Wave IV (2008–2009) and who provided a DNA sample. At Wave IV, Add Health located 92% and re-interviewed 80% of eligible respondents. This resulted in a Wave IV sample size of 15,701.

### Sample Collection, processing, and storage

Participants were requested to provide two milliliters (ml) of saliva using the Oragene OG-300 non-invasive sampling kit (DNA Genotek, Kanata, Ontario, Canada) that when

capped released a stabilizing lysis buffer. Each kit was bar coded and sent via overnight express mail to the Institute for Behavioral Genetics (IBG) in a SafTpak #STP-210 (Edmonton, Alberta, Canada). A total of 15,249 packages were received; 17 [0.001%] were lost during shipping. Upon receipt the packages were inspected for damage. Four kits were found to be damaged and samples from three of these damaged kits were recovered for further processing. The barcoded vials and manifests included in the packages were scanned into a custom database, assigned a position in a storage box and assigned a consecutive laboratory identifier distinct from the Add Health DNA specimen ID identifier.

Genomic DNA was isolated from saliva samples using the Zymo Research (Irvine, California) Silicon-A plates according to protocols supplied by the manufacturer. For the initial round of isolation, 500 μl of Oragene® ™ solution was used, and the final elution volume was 150 μl. DNA quality was assessed by loading 10 μl of the undiluted sample onto 0.65% agarose gels that were run in 5mM sodium borate at 200 volts for 45 minutes, visualized with SYBR®-Safe dye and photographed. Dilutions of phage lambda DNA (~ 49 kilobases) were used as mass and size standards.

To obtain informed consent, Add Health utilized a two-tiered process for the collection of biomarkers to provide a DNA archive for future testing. Add Health requested consent from respondents to provide saliva: 1) currently planned Program Project genotyping and research; and 2) DNA archival and future assay. Of those participating at Wave IV, 96% (N = 15,140) consented to DNA collection (‘Program Project samples’) and 78% (N = 12,234) agreed to archive (‘Archive samples’) their DNA for future analysis “related to long term health.” Table 1 provides the unweighted compliance rates for the two-tiered consent process by ethnicity.

## Genotyping

Among the 15,140 at Wave IV who provided a DNA sample for purposes of the Program Project research, three VNTRs, one indel, and two di-nucleotide repeats were characterized. These included VNTRs in the promoter regions of the monoamine oxidase A (*MAOA-uVNTR*), third exon of the dopamine D4 (*DRD4*) receptor, and the 3′-untranslated region (UTR) of the dopamine transporter (*DAT1*) genes. The indel polymorphism (*5HTTLPR*) was the in the promoter region of the serotonin transporter gene (5HTT). Di-nucleotide repeats located 18.5 kilobases (kb) from the 5′ end of the dopamine D5 receptor (*DRD5*) and in the second intron of MAOA (*sWXD805, MAOCA-1*) gene were also characterized and represent the largest single-sample characterization of these genetic variants. The sex-determining marker amelogenin was determined for every sample in part as a control and to assign *MAOA* genotypes correctly.

Each genotype was determined using polymerase chain reaction (PCR). Except for the di-nucleotide repeats, PCR reactions contained two μl of DNA [20 ng or less], 1x Buffer II [ABI, Applied Biosystems, Foster City, CA], 1.8 mM $MgCl_2$, 180 μM each deoxynucleotide (dNTP, NEB), with 7′-deaza-2′-deoxyGTP (deaza-GTP, Roche Applied Science, Indianapolis, Indiana) substituted for one-half of the dGTP, forward (fluorescently labeled) and reverse primers, and one unit of AmpliTaq Gold® polymerase (ABI) in a total volume of 20 μl. Primer sequences and concentrations are provided in online Supplemental Table 1.

Amplifications were performed using a modified (Anchordoquy et al, 2003) touchdown PCR method (Don et al, 1992). A 95°C incubation for 10 minutes was followed by two cycles of 95°C for 30 seconds, 65°C for 30 seconds, and 72°C for 60 seconds. The annealing temperature was decreased every two cycles from 65°C to 57°C in 2°C increments (10 cycles total), followed by 30 cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 60 seconds, a final 30-minute incubation at 72°C and a hold at 4°C. Amelogenin, *MAOA-uVNTR*, *DAT1*, *DRD4*, and *5HTTLPR* assays were done in a single multiplex PCR.

For the di-nucleotide repeat (*DRD5* and *MAOCA-1*) assays, deaza-GTP was not used, dNTP concentrations were 125 μM each, $MgCl_2$ concentration was 2.5 mM and the touchdown PCR was modified to: 95°C for 12 minutes followed by 35 cycles of 94°C for 15 seconds, 62°C for 15 seconds, (−0.2/cycle down to 55°C), 72°C for 30 seconds, 72°C seconds for 20 minutes and a hold at 4°C. The DRD5 and MAOCA-1 assays were done in a single biplex PCR.

The SNP rs25531 was assayed in order to determine the L-A and L-G alleles using the primer sequences reported by Hu et al (2005, 2006). PCR products were incubated with 5 units of *Mspl* (NEB, Ipswich, MA) for 90 minutes at 37°C (Wendland et al, 2006). A 97 bp restriction digest fragment is indicative of the L-G allele.

Following amplification, PCR products and *MspI* digests were filter purified using Zymo Rearch ZR-96 DNA Sequencing Clean-up Kits following protocols supplied by the manufacturer. An aliquot of PCR products was combined with loading buffer containing size standard (Rox1000, Gel Company, San Francisco, CA or LIZ 1200, ABI) and analyzed with an ABI PRISM® 3130×l Genetic Analyzer using protocols supplied by the manufacturer. Fragment sizes were analyzed with Genemapper software with the resulting allele sizes independently reviewed by two investigators.

Whole genome amplification (WGA) of genomic DNA of the Archive samples was initiated to assure a sufficient availability for future studies. A multiple displacement method with the Repli-g® mini-kit (Qiagen, Valencia, CA) enhanced with additional phi-29 polymerase and dNTPs was used. Genomic DNA samples (50–100 ng) were dried in 96-well PCR plates, denatured for 5 minutes (2.5 μl of denaturation buffer), neutralized (2.5 μl of neutralization buffer) and placed on ice. The 20 μl of reaction mix consisted of 14.5 μl of Repli-g® reaction buffer, 1 μl of Repli-g® polymerase, 1.5 μl of dNTPs (Epicentre Biotechnologies, Madison, WI, 25 mM each, 1.5 mM final concentration), 0.5 μl of RepliPHI™ phi-29 polymerase (Epicentre, 1 unit/μl final concentration) and 3 μl of water. Reactions were incubated in a thermocycler (with a non-heated lid) at 30°C for 12–16 hours followed by 65°C for 5 minutes and a hold at 4°C. Samples were diluted with 75 μl of TE (10 mM Tris-HCL, pH 8.0; 0.1 mM EDTA), and the DNA was quantified by Picogreen® (Invitrogen, Carlsbad, CA) fluorescence. DNA samples were standardized to 50ng/μl and diluted 40-fold for routine genotyping as described above.

For each sample, PCR analyses were conducted twice in two, independent reactions on different days by different laboratory technicians. Results from the two runs were subsequently compared by a third investigator and repeated if there were missing data for a

marker or inconsistency between genotype calls. For the Archive samples the reported genotypes are the result of one run with genomic DNA and one with WGA DNA. For the Non-Archive samples, two runs with genomic DNA were used.

### Statistical Analyses

Allele and genotype frequencies, observed heterozygosity, polymorphic information content (PIC), and Hardy Weinberg Equilibrium (HWE) were calculated using the *proc allele* statement in SAS (Version 9.2, SAS, Inc.). A PIC value depends on the number of detectable alleles and the distribution of their frequency and is an estimate of gene diversity. For co-dominant markers such as microsatellites, PIC values can range between 0 and 1 represent a range of allelic variation, from none (PIC = 0) to only new alleles (PIC = 1). PIC is calculated as one minus the sum of the squares of the frequency for each allele. Due to differences in the number of X-chromosomes between males and females, allele and genotype frequencies are reported separately by sex for the MAOA-uVNTR and MAOAC-1 dinucleotide polymorphisms. This was also done for the Amelogenin polymorphism due to its gene location on the Y Chromosome.

## Results

Yields of DNA extracted from saliva using the Zymo Research Silica A method are shown in Figure 1. As can be seen, DNA yields varied widely ranging between 0 ng/μl to 400 ng/μl, with a mean (standard deviation (sd) of 33 (25) ng/μl. The mean (sd) DNA yield per sample was 9.6 (7.1) μg/ml. The majority of samples were composed of high molecular weight DNA (> 50 kD). The mean (sd) ratio of absorbance at 260 nm and 280 nm (N = 1000) was 1.78 (0.24) with a mode of 1.86. Samples with 0 ng/μl of DNA were likely to be the result of the participant not having provided a complete saliva sample. Of the 15,249 samples processed, we were unable to obtain reliable genotypes on 374 (2.5%) samples that provided less than 3 ng of DNA.

Comparison of the genotype call rates between whole-genome amplified DNA and genomic DNA for 12,153 archive samples is provided in Table 2. For each of the five repeat markers, approximately 4% of the assays using genomic DNA yielded an initial null result (i.e. “missing call”). Initial null results were 1–2 percentage points higher for WGA assays. Notably, there was a two-fold difference in the missing call rate for amplified versus genomic DNA and an 8% missing call rate for the DRD4 VNTR. Comparison between the missing calls for genomic and WGA DNA revealed that approximately 50% of the null results were for the same sample and implicated poor DNA quality for those particular samples. Finally, between genomic and WGA DNA samples, there was a >99% concordance rate or a less than 1% discordance rate for samples from which a genotype could determined. Similar results were observed for the ‘non-Archived’ and are presented in Supplemental Table 2. Reasons for discordant calls varied by marker. For Amelogenin, the primary reason was due to poor amplification that resulted in no call being made. For the MAOA-uVNTR and DAT1 markers, the presence of large stutter bands resulted in erroneous calls that occurred more frequently in WGA samples. For the DRD4 maker and in

particular for repeats of 7 or greater, large alleles amplified less well in WGA samples and resulted in calls not being made (e.g. allele dropout).

## Allele frequencies

**Microsatellites and Insertion-Deletion Polymorphisms—**Allele frequencies for each SSR genotyped (excluding Amelogenin) are shown in Table 3. Online supplemental Tables 3–11 provide allele frequencies and measures of genetic diversity for Black, Asian, Native American, Hispanic and White participants in the full Add Health sample. For the DRD4 VNTR, repeat sizes ranged from 2 – 10 with frequencies of <0.01 to 0.62. The 2-, 4-, and 7-repeat alleles were the most commonly observed. We identified a 366 bp allele, which to our knowledge, has not been previously reported. Because the length of this rare allele is 39 bps greater than the 3R allele, (or 9 bp short of the usual 48 bp repeat), we report it as the 3.39R allele. The 3.39R allele was identified only in Black and Hispanic populations where it always occurred as a heterozygote with the 4R allele. PiC, H and allelic diversity values were 0.50, 0.53, and 0.54, respectively, indicating moderate to high allelic diversity for this locus.

Considered by race, the frequency of the DRD4 4R allele was very similar (0.63 to 0.65) across the five groups. The 2R allele was most frequently observed in Asians (0.26). Conversely, the 7R allele was least frequent in Asians (0.04). Similar estimates for the 2R (0.05 – 0.09) and 7R alleles (0.19 – 0.22) were observed across the remaining four groups. Although the 3.39R allele was observed at a low frequency, 14 of the 15 (0.93) carriers were Black. We did not observe any 11R alleles in the Add Health sample. As compared with Asian, Native American, and Hispanic populations, longer DRD4 VNTR repeats (8R, 9R, and 10R) were observed more frequently among Whites and Blacks, albeit relatively rarely (<0.03).

For the DAT1 VNTR, we observed seven repeat alleles, ranging between 3- and 13-repeats, with all but two alleles having a frequency <0.01. The 9R and 10R alleles were the most frequently observed (0.37 and 0.58, respectively). PiC, H and allelic diversity values were 0.31, 0.36, and 0.37, respectively, and indicated moderate allelic diversity for this locus. Across the five race/ethnic groups, the frequency of the 10R allele ranged from a low of 0.74 in the White group to a high of 0.89 in the Asian group. Among rare alleles, the 3R was observed primarily in the Black group, which also evidenced greater numbers of the 7R and 8R alleles than the other four racial groups.

Seven alleles were observed for the 5HTTLPR polymorphism in the promoter region of the serotonin transporter, ranging between 14- and 22-repeats. All but the 14R and 16R alleles were rare (<0.01). PiC, H and allelic diversity values were 0.37, 0.46, and 0.49, respectively, and indicated moderate allelic diversity for this locus. The proportion of short- (14R) to long- alleles (16R) differed markedly among the race/ethnic groups. Blacks had significantly more long alleles (16R) and Asians had significantly more short alleles (14R) than the remaining three groups. Extra-long alleles (17R, 18R, 19R, 20R, 22R and 24R) were uncommon throughout, and Blacks and Asians had the highest percentage of these (0.01 and 0.02, respectively).

To allow correct assignment of participants into low- and high-expression groups, we genotyped SNP rs25531 located 5′ of the insertion of the 5HTTLPR. Because of its location the A → G transition results in two forms of the L-allele (16R), denoted L-A and L-G, and two corresponding S-A and S-G alleles, although the S-G allele is exceptionally rare. In the Add Health sample, allele frequencies for the L-A, L-G, and S alleles were 0.48, 0.10, and 0.42, respectively. The frequency of the L-G allele was approximately 0.06 in Whites, Native Americans, and Hispanics, and 0.21 and 0.13 in Blacks and Asians, respectively.

We observed five repeat alleles for the MAOA-uVNTR polymorphism of which three alleles (2R, 3.5R, and 5R) were rare (<0.01) in the full Add Health sample. Similar allele frequencies for the 3R and 4R were observed for males (0.38 and 0.56) and females (0.38 and 0.58). PIC, H, and allelic diversity values were 0.42, 0.27, and 0.52 and reflected the predominance of the 3R and 4R alleles. Among the five groups, the 3R allele was more frequent in the Black and Asian groups than in Native American, Hispanic, and White groups, which had higher rates of the 4R allele. Similarly, the 2R allele was represented at a reasonable frequency only in Blacks. A similar pattern was observed for the 3.5R allele among Whites.

**Dinucleotide Repeat Polymorphisms—**For the dinucleotide repeat polymorphism located 18.5 kb from the 5′ end of the DRD5 gene, we observed 27 two base-pair alleles that ranged from 124 bp to 176 bp with frequencies ranging between < 0.01 and 0.38 (Table 4). Seventeen of the 27 alleles were rare (<0.01). A total of 15 repeat lengths have not been reported previously and include alleles: 124 bp, 126 bp, 128 bp, 130 bp, 132 bp, 158 bp, 160 bp, 162 bp, 164 bp, 166 bp, 168 bp, 170 bp, 172 bp, 174 bp, and 176 bp lengths. PiC, H, and allelic diversity values were 0.80, 0.74, 0.81, respectively, and indicated large allelic diversity for this locus. The 148 bp allele was the most frequently observed repeat length within each population, though more frequently among Whites (0.45) and less frequently among Blacks (0.25). A similar pattern was found for the 150 bp allele. Conversely, the 146 bp and 144 bp alleles were more frequent in Blacks and Asians than Native American, Hispanic and White groups.

For the dinucleotide repeat polymorphism located in the second-intron of the MAOA gene, we observed 16 alleles that ranged from 101 bp to 131 bp with frequencies from <0.01 to 0.42 (Table 4). For both sexes, the 113 bp and 115 bp alleles were the most frequent. PIC and allelic diversity values were 0.72 and 0.75 for males and females, respectively. The two most frequent alleles (113 bp and 115 bp) varied by race/ethnicity with the 113 bp allele observed more frequently in the Whites (0.50), Hispanics (0.47), and Native Americans (0.40) and the 115 bp allele more frequently in the Blacks (0.28) and Native Americans (0.27). A similar pattern of between population differences was observed for the 121 bp allele. Three of the observed repeat lengths (101 bp, 103 bp and 131 bp repeats) have not been reported in the literature and were seen almost entirely among Blacks.

## Genotype frequencies

The distributions of genotypes for all but the DRD5 and MAOCA-1 polymorphisms as a function of race/ethnicty are provided in the online Supplemental Tables 12–31. Because of

the allelic diversity seen for some of the characterized SSRs, HWE was calculated using genotypes binned based on existing evidence of a functional impact. In the absence of a functional effect, we binned alleles into genotypes based on previous studies. For the VNTRs, the 4R and 7R alleles of the DRD4, 9R and 10R of the DAT1, 14R and 16R of the 5HTTLPR, and 3R and 4R alleles of the MAOA-uVNTR were selected to determine HWE statistics. For the dinucleotide repeats, HWE statistics were calculated after binning alleles into 'long' and 'short' groups (see online supplement for further information). Genotype frequencies for the genotypes and HWE statistics are provided below in Tables 5a and 5b for the 5HTTLPR and in the online Supplemental Tables 32–38 for the DAT1, DRD4, MAOA-uVNTR, DRD5, and MAOAC-1 polymorphisms.

**Genotype Frequencies - Microsatellites & Insertion-Deletion Polymorphisms—** The high-activity DRD4 4/4 genotype was the most common in each of the five populations and ranged in frequency from 0.48 in Blacks to 0.85 in Native Americans. Conversely, the low-activity 7/7 genotype was the least common in each group and accounted for < 0.10 of the genotypes. HWE statistics indicated no deviations from expectation ($p < 0.18$). For the DAT1 polymorphism, the 10/10 genotype was the most frequent and ranged between 0.56 in Whites and 0.81 in Asians. Similar to the DRD4 7/7 genotype, the 9/9 genotype was the least frequent in all five ethnic group examined. HWE calculations indicated a significant deviation from expectation among Blacks ($p < 0.01$) and Asians ($p < 0.001$), largely owing to the low frequency (0.06 and 0.05, respectively) of the 9/9 genotype in these populations. For both sexes, the low-activity MAOA-uVNTR allele (3R) was most frequent among Blacks and Asians while the high-activity allele (4R) was most frequent among Whites, Native Americans and Hispanics. For females, MAOA-uVNTR genotypes were in HWE.

Genotype frequencies for the biallelic and triallelic coding of the 5HTTLPR polymorphism are shown as a function of race in Table 5a and 5b. Based on the diallelic coding, the heterozygous 14R/16R genotype was the most frequent in Whites, Asians, and Hispanics and ranged between 0.44 and .49. The 16R/16R genotype was the most frequent in Blacks (0.55) while the 14R/14R and 14R/16R genotypes were equally frequent in Native Americans. The low frequency of the homozygous 14R/14R genotype in Blacks likely contributed to the observed significant deviations in HWE ($p > 0.01$). Based on the triallelic coding, all populations were in HWE ($p < 0.18$).

Based on biallelic coding of the 5HTTLPR and the near equal prevalences of the 14R and 16R alleles in all but Blacks and Native Americans, the heterozygote 14R/16R genotype was the most common with frequencies that ranged from 0.44 in Asians to 0.49 in Whites. The low frequencies of the 14/14 genotype in Blacks likely contributed to the observed significant deviations in HWE ($p < 0.01$). As shown in Table 5b, these deviations from HWE were not observed using the triallelic coding, though differences in genotype frequencies between populations remained.

**Genotype Frequencies - Dinucleotide Repeat Polymorphisms—**For the DRD5 dinucleotide repeat, all but the Black group had higher frequencies of the 'long-allele' whereas Blacks had a higher frequency of the 'short-allele'. When grouped into long (>= 148 bp) and short (=< 146 bp) alleles there were notable differences between the different

racial groups, with a higher frequency of 146/146 genotypes (0.34) among Blacks and 148/148 genotypes in Whites (0.49). There were significant deviations from HWE ($p > 0.03$) for each racial group based on this dichotomous binning of the alleles. For the MAOCA-1 dinucleotide repeat the 'long-' and 'short-alleles' were equally frequent (0.50) in the Add Health sample. When grouped into 'long' (= <115 bp) and short (>= 113 bp) genotypes, Whites and Hispanics had higher frequencies of the 113bp/113bp genotype than Blacks or Asians who had higher frequencies of the 115bp/115bp genotype. There were no deviations from HWE for females in any ethnic group.

## Discussion

The National Longitudinal Study of Adolescent Health (Add Health) is an ongoing longitudinal study of a nationally representative sample of more than 20,000 adolescents in grades 7–12 in the United States in 1994–1995. The original purpose of Add Health was to understand the causes of health and health-related behavior with special emphasis on the role of social context. Accordingly, Add Health sampled the multiple environments in which young people live their lives, gathering information from adolescents themselves, their parents, siblings, friends, romantic partners, fellow students, and school administrators. Existing databases with information about the neighborhoods and communities of adolescent participants were merged with Add Health survey and biomarker data, creating exceptionally rich multilevel environmental data with which to understand gene-environment interplay in health and behavior across the life course.

As a part of the Add Health Program Project at Wave IV, we characterized the genetic diversity of six SSR polymorphisms in a racially and ethnically diverse sample of adolescents followed longitudinally since 1995. Across 30,498 chromosomes we examined 213,486 allele calls for each of two genotyping runs and generated 106,743 genotypes. For this set of SSR polymorphisms, we either confirmed the allelic diversity reported in previous studies or observed greater allelic diversity than previously reported (Fuke et al, 2001; Kang et al, 1999; Chang et al, 1996; Murdoch et al, 2013; Black et al, 1991; Vanyukov et al, 1995; Sherrington et al, 1993). For the two dinucleotide repeat polymorphisms included here, the current report represents the largest genotyping effort to date. Finally, we documented allele and genotype frequencies for six commonly examined polymorphisms among five race/ethnic groups, some of which have remained poorly characterized for these polymorphisms.

The use of saliva in large-scale epidemiological studies has become the method of choice given the non-invasive nature of DNA sampling. It also requires no special training on the part of interviewers and avoids blood-borne pathogen issues. In addition to these advantages, it has been shown that high quality and quantity of DNA can be obtained from saliva samples (Nunes et al, 2012, Rogers et al, 2007, Quinque et al, 2006; Nemoda et al, 2011). Using a two-tier consent procedure, we observed a high rate of compliance for Add Health Program Project research (0.96) and among archived samples for future testing (0.78). Further, we observed a similar range of DNA yields, mean DNA yields per sample, and quality of DNA as reported elsewhere using the Oragene collection method (Rylander-Rudqvist et al, 2006, Nunes et al, 2012, Abraham et al, 2012; Pulford et al, 2013; Koni et al, 2011; Ng et al, 2006).

To determine repeat lengths we utilized locus-specific DNA primers amplified by polymerase chain reactions (PCR). Because DNA primers are sequence based, differences in the number of repeats in the SSR are quantified by corresponding changes in the PCR fragment length (Pemberton et al, 2009). In this way, differences in the number of repeats between individuals for a given SSR are represented by length differences in the PCR product. In our hands, this approach generated data that are comparable in terms of genotype concordance and discordance rates, and allele and genotype frequencies. Following extensive quality control efforts that included a second genotyping run of each sample and independent review of genotype calls, concordance rates were > 0.98. Discordance between genotype calls occurred at a frequency of < 0.01 across all loci and were resolved by comparing all three electropherograms following a third genotyping run. Across each of the five populations examined, we observed that two or three alleles accounted for the majority of the alleles observed. Moreover, rare alleles accounted for less than one-tenth of one percent of the population frequency for a given polymorphism. The DRD4 polymorphism is the one exception, however, as relatively rare alleles accounted for approximately 25% of the total genotypes observed.

Compared with SNPs, SSRs have presented a number of challenges that include, among other things, the absence of a clear approach to dealing with their allelic diversity analytically. Based on the observation that SSRs, in particular microsatellites, are frequently found in promoter regions of genes and functional validation studies have demonstrated their ability to regulate or act as ‘tuning forks’ of gene expression, creating dichotomous categories based on binned alleles has been a common analytic approach. Consistent with this approach, we created dichotomous or ‘binned’ genotypes for each SSR. Our results from these analyses agreed with previous findings and highlighted population differences as indicated by deviations from HWE. Although it may be useful to use all genotypes in analyses, binning them into dichotomous categories may still be advantageous in terms of statistical power. Binning based on expression studies of the gene or activity of the transcribed protein may be warranted since the size of the allele is typically used as a surrogate for the presumed biological function of the gene.

The SSRs described here are widely known and have been characterized in a large number of smaller samples. An important criticism of association tests based on these SSRs garnered over the years is the irreproducible nature of the findings reported, primarily owing to low statistical power and a high number of uncorrected statistical comparisons. While not immune to generating similar results, the available genotypes in the Add Health sample offers uniquely large samples sizes, and thus the statistical power, for *a priori* tests as well as consideration of important published findings such as those by Caspi et al (2002, 2003). Furthermore, when similarly assessed phenotypic and/or environmental variables are available, the genetic data reported here can serve as an important replication sample. Lastly, as many association tests are conducted in White only samples, the availability of these SSRs in other race/ethnic groups should offer an important opportunity to examine the generalizability of any set of findings.

Over the past decade there has been a tremendous growth in our understanding of the genomic landscape and the diversity of polymorphisms present in the human genome.

Although understanding the extent of correlation between SNPs has allowed successful identification of putative variants for a variety of health and behavior related phenotypes, it will be increasingly important to incorporate the impact multiple types of variation have on disease associations. To this end, Add Health has begun both targeted and genome-wide SNP characterization that will be available in the near future. The current genetic data are available to qualified investigators from Add Health at the Carolina Population Center, the University of North Carolina at Chapel Hill. These restricted-use data are available by contractual agreement only. Contractual application materials and forms are available at the Add Health website included in Supplemental Table 39. Protocols and initial genotyping efforts of the Add Health Wave IV Program Project (Smolen et al, 2013) are also available at the Carolina Population Center website (Supplemental Table 39).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgement is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (www.cpc.unc.edu/addhealth). No direct support was received from grant PO1-HD31921 for this analysis.

## Literature Cited

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–1073. [PubMed: 20981092]

Abraham JE, Maranian MJ, Spiteri I, Russell R, Ingle S, Luccrini C, Earl HM, Pharoh PPD, Dunning AM, Caldas C. Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. BMC Medical Genomics. 2012; 5:19. [PubMed: 22647440]

Anchordoquy HC, McGeary C, Liu L, Krauter KS, Smolen A. Genotyping of three candidate genes after whole-genome preamplification of DNA collected from buccal cells. Behav Genet. 2003; 33:73–78. [PubMed: 12645824]

Bhanglae TR, Rieder MJ, Livingston RJ, Nickerson DA. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet. 2005; 14(1):59–69. [PubMed: 15525656]

Black GCM, Chen ZY, Craig IW, Powell JF. Dinucleotide repeat polymorphism at the MAOA locus. Nucleic Acids Res. 1991; 19:689. [PubMed: 2011543]

Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW, et al. Role of genotype in the cycle of violence in maltreated children. Science. 2002; 297:851–854. [PubMed: 12161658]

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, et al. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science. 2003; 301:386–389. [PubMed: 12869766]

Chang FM, Kidd JR, Livak KJ, Pakstis AJ, Kidd KK. The world-wide distribution of allele frequencies at the human dopamine D4 receptor locus. Hum Genet. 1996; 98:91–101. [PubMed: 8682515]

Don RH, Cox RT, Wainwright BJ, Baker K, Mattick JS. “Touchdown” PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res. 1992; 19:4008. [PubMed: 1861999]

Durbin RM, Abecassis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation form population-scale sequencing. Nature. 2010; 467:1061–107310. [PubMed: 20981092]

Eichler EE. Widening the spectrum of human genetic variation. Nat Genet. 2006; 38(1):9–11. [PubMed: 16380720]

Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004; 5:435–445. [PubMed: 15153996]

Fondon JW, Hammack EAD, Hannan AJ, King DG. Simple sequence repeats: genetic modulators of brain function and behavior. Trends in Neurosci. 2008; 31(7):328–334.

Fuke S, Suo S, Takahashi N, Koike H, Sasagawa N, Ishiura S. The VNTR polymorphism of the human dopamine transporter (DAT1) gene affects gene expression. Pharmacogenetics J. 2001; 1:152–156.

Hamada H, Petrino MG, Kakunaga T, Seldman M, Stollar BD. Enhanced gene expression by the poly(dT-dG)-poly(dC-dA) sequence. Molecular Cell Biology. 1984; 4:2622–2630.

Harris KM, Halpern CT, Smolen A, Haberstick BC. The National Longitudinal Study of Adolescent Health (Add Health) twin data. Twin Res Hum Genet. 2006; 9(6):988–997. [PubMed: 17254442]

Harris KM, Halpern CT, Haberstick BC, Smolen A. The National Longitudinal Study of Adolescent Health (Add Health) sibling pairs data. Twin Res Hum Genet. 2013; 16(1):391–398. [PubMed: 23231780]

Harris, KM. Design features of Add Health. 2012. URL: www.cpc.unc.edu/projects/addhealth/guides/DesignPaperWIIV.pdf

Hu X, Oroszi G, Chun J, Smith TL, Goldman D, Schuckit MA. An expanded evaluation of the relationship of four alleles to the level of response to alcohol and the alcoholism risk. Alcohol Clin Exp Res. 2005; 29:8–16. [PubMed: 15654286]

Hu X, Lipsky RH, Zhu G, Akhtar LA, Taubman J, Greenberg BD, Xu K, Arnold PD, Richter MA, Kennedy JL, Murphy DL, Goldman D. Serotonin Transporter Promoter Gain-of-Function Genotypes Are Linked to Obsessive-Compulsive Disorder. Am J Hum Genet. 2006; 78:815–826. [PubMed: 16642437]

Jeffreys AJ. Highly variable minisatellite and DNA fingerprints. Biochem Soc Trans. 1987; 15(3): 309–317. [PubMed: 2887471]

King DG, Soller M, Kashi Y. Evolutionary Tuning Knobs. Endeavour. 1997; 21(1):36–40.

Kashi Y, King DG, Soller M. Simple sequence repeats as a source of quantitative genetic variation. Trends in Genetics. 1997; 13:74–78. [PubMed: 9055609]

Kang AM, Palmatier MA, Kidd KK. Global variation of a 4-bp VNTR in the 3″-untranslated region of the dopamine transporter gene (SLC6A3). Bio Psychiatry. 1999; 46:151–160. [PubMed: 10418689]

Koni AC, Scott RA, Wang G, Bailey ME, Peplies J, Bamann K, Pitsilladis YP. IDEFICS Consortium. DNA yield and quality of saliva samples and suitability for large-scale epidemiological studies in children. Int J Obesity. 2011; 35:S113–S118.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing of the human genome. Nature. 2001; 409(6822):860–921. [PubMed: 11237011]

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard S, Devin SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006; 16:1181–1190.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar Aa, Kemeza DM, Strassler DS, Ponting CP, Webber C, Devine SE. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 2011; 21:830–910. [PubMed: 21460062]

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM. Common deletion polymorphisms in the human genome. Nat Genet. 2006; 38:86–92. [PubMed: 16468122]

Montgomery SB, Goode DL, Kvikstad E, lbers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, Anaya V, Richardson R, Davis J. Genomes Project Consortium, MacArthur DG, Sidow A, Luret L, Gerstein M, Makova KD, Marchini J, McVean G, Lunter G

(2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 1000; 23:749–761. [PubMed: 23478400]

Murdoch JD, Speed WC, Pakstis AJ, Heffelfinger CE, Kidd KK. Worldwide population variation and haplotype analysis at the serotonin transporter gene SLC6A4 and implication for association studies. Bio Psychiatry. 2013 E pub.

Nakamura Y, Koyama, Matsushima M. VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. J Hum Genet. 1998; 43:149–152. [PubMed: 9747025]

Nemoda Z, Horvat-Gordon M, Fortunato CK, Beltzer EK, Scholl JL, Granger DA. Assessing genetic polymorphisms using DNA extracted from cells present in saliva samples. BMC Res Methodol. 2011; 11:170.

Ng DPK, Koh D, Choo S, Chia KS. Saliva as a viable alternative source of human genomic DNA in genetic epidemiology. Clinica Chimica Acta. 2006; 367:82–85.

Nunes AP, Oliveria IO, Santos BR, Millech C, Silva LP, Gonzalez DA, Hallal PC, Menezes AMB, Araujo CL, Barros FC. Quality of DNA extracted from saliva samples collected with Oragene DNA self-collection. BMC Medical Res Methodol. 2012; 12:65.

Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA. Sequence determination of human microsatellite variability. BMC Genomics. 2009; 10:612. [PubMed: 20015383]

Pulford DJ, Mosteller M, Briley JD, Johansson KW, Nelsen AJ. Saliva sampling in global clinical studies: the impact of low sampling volume on performance of DNA in downstream genotyping experiments. BMC Med Genet. 2013; 6:20.

Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I. Evaluation of saliva as a source of human DNA for population and association studies. Anal Biochem. 2006; 353:272–277. [PubMed: 16620753]

Rogers NL, Cole SA, Lan HC, Crossa A, Demerath EW. New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies. Am J Hum Biol. 2007; 19:319–326. [PubMed: 17421001]

Rylander-Rudqvist T, Hakansson N, Tybring G, Wolk A. Quality and quantity of saliva DNA obtained from the self-administered Oragene method – A pilot study on the cohort or Swedish men. Cancer Epidemiol Biomarkers Prev. 2006; 15:1742–1745. [PubMed: 16985039]

Shen H, Li J, Zhang J, Xu C, Jiang Y, Wu Z, Zhao F, Liao L, Chen J, Lin Y, Tian Q, Papasian CJ, Deng HW. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PLoS One. 2013; 8(4):e59494. [PubMed: 23577066]

Sherrington R, Baljinder M, Attwood J, Kalsi G, Curtis D, Buetow K, Povey S, Gurling H. Cloning of the human dopamine D5 receptor gene and identification of a highly polymorphic microsatellite for the DRD5 locus that shows tight linkage to the chromosome 4p reference marker RAF1P1. Genomics. 1993; 18:423–425. [PubMed: 8288248]

Smolen, A.; Whitsel, EA.; Tabor, J.; Killeya-Jones, LA.; Cuthbertson, CC.; Hussey, JM.; Halpern, CT.; Harris, KM. Candidate Genes. 2013. Add Health Wave IV documentation.

Vanyukov MM, Moss HB, Yu LM, Deka R. A dinucleotide repeat polymorphism at the gene for monoamine oxidase A and measures of aggressiveness. Psychiatry Res. 1995; 59:35–41. [PubMed: 8771218]

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. Human diallelic insertion/Deletion polymorphisms. Am J Hum Genet. 2002; 71:854–862. [PubMed: 12205564]

Wendland JR, Martin BJ, Kruse MR, Lesch KP, Murphy DL. Simultaneous genotyping of four functional loci of human SLC6A4, with a reappraisal of 5HTTLPR and rs25531. Mol Psychiatr. 2006; 11:224–226.

## Table 1

Frequencies of Consent to DNA collection by Race/Ethnicity (N, %).[†]

| Race/Ethnicity | Consent to Saliva Collection | Consent to Archive |
|---|---|---|
| Whites | 8051 (0.96) | 6822 (0.82) |
| Blacks | 3348 (0.96) | 2523 (0.72) |
| Hispanics | 2393 (0.96) | 1883 (0.75) |
| Other[‡] | 1335 (0.95) | 997 (0.71) |
| Total | 15,140 (0.96) | 12,234 (0.78) |

[†]Unweighted N and percentages, excluding four incarcerated respondents for whom saliva collection was prohibited.

[‡]Includes Asian, Native American, and other Non-Hispanics.

## Table 2

Genotyping results for genomic and whole genome amplification (WGA) DNA (N = 12,153).[†]

| Marker | Genomic Null Result (N, %) | WGA Null Result (N, %) | Null Result in Common (N, %) | Discordant Genotype Call (N, %) |
|---|---|---|---|---|
| Amelogenin | 370 (0.03) | 432 (0.04) | 165 (0.45) | 42 (0.35) |
| DRD4 | 461 (0.04) | 1014 (0.08) | 221 (0.48) | 101 (0.83) |
| DAT1 | 559 (0.05) | 603 (0.05) | 246 (0.44) | 53 (0.44) |
| 5HTTLPR | 530 (0.04) | 728 (0.06) | 234 (0.44) | 49 (0.40) |
| MAOA-uVNTR | 408 (0.03) | 446 (0.04) | 200 (0.49) | 85 (0.70) |
| DRD5 | 634 (0.05) | 927 (0.11) | 326 (0.51) | 59 (0.49) |
| MAOCA-1 | 780 (0.06) | 705 (0.08) | 318 (0.41) | 118 (0.97) |

[†]Note: Add Health Archived samples, Unweighted N and percentages, excluding four incarcerated respondents for whom saliva collection was prohibited.

**Table 3**

Allele frequencies for DRD4, DAT1, 5HTTLPR, and MAOA-uVNTR.

| DRD4 (N=14,810) | | DAT1 (N=14,795) | | 5HTTLPR (N=14,957) | | MAOA-uVNTR (N=14,811) | |
|---|---|---|---|---|---|---|---|
| **Allele** | **N, %** | **Allele** | **N, %** | **Allele** | **N, %** | **Allele** | **N, %** |
| | | | | | | Males (N=6926) | |
| 2R | 2541 (0.09) | 3R | 39 (< 0.01) | 14R | 12591 (0.42) | 2R | 178 (0.01) |
| 3R | 956 (0.03) | 7R | 54 (< 0.01 | 16R | 17207 (0.57) | 3R | 5468 (0.39) |
| 3.39R | 15 (< 0.01) | 8R | 192 (0.01) | 17R | 1 (< 0.01) | 3.5R | 156 (0.01) |
| 4R | 18961 (0.64) | 9R | 6487 (0.22) | 18R | 17 (< 0.01) | 4R | 7912 (0.58) |
| 5R | 528 (0.02) | 10R | 22587 (0.76) | 19R | 2 (< 0.01) | 5R | 144 (0.01) |
| 6R | 313 (0.01) | 11R | 228 (0.01) | 20R | 84 (< 0.01) | | |
| 7R | 5884 (0.20) | 13R | 3 (< 0.01) | 22R | 12 (< 0.01) | Females (N= 7885) | |
| 8R | 387 (0.01) | | | | | 2R | 224 (0.01) |
| 9R | 21 (< 0.01) | | | | | 3R | 6033 (0.39) |
| 10R | 14 (< 0.01) | | | | | 3.5R | 169 (0.01) |
| | | | | | | 4R | 9189 (0.58) |
| | | | | | | 5R | 155 (0.01) |

Note: R, repeat; N, sample size.

## Table 4

Allele frequencies for DRD5 and MAOCA-1 dinucleotide repeat polymorphisms.

| DRD5 (N=14,529) | | MAOCA-1 (N= 14,575) | |
|---|---|---|---|
| **Length (bp) (N= 14529)** | **N, %** | **Length (bp) (N= 14577)** | **N, %** |
| | | Males (N= 6782) | |
| 124 | 1 (< 0.01) | 101 | 10 (< 0.01) |
| 126 | 9 (< 0.01) | 103 | 80 (0.01) |
| 128 | 8 (< 0.01) | 105 | 6 (< 0.01) |
| 130 | 356 (0.01) | 107 | 136 (0.01) |
| 132 | 75 (< 0.01) | 109 | 142 (0.01) |
| 134 | 382 (0.01) | 111 | 680 (0.05) |
| 136 | 643 (0.02) | 113 | 5782 (0.43) |
| 138 | 2180 (0.08) | 115 | 2300 (0.17) |
| 140 | 1393 (0.05) | 117 | 428 (0.03) |
| 142 | 1747 (0.06) | 119 | 400 (0.03) |
| 144 | 1818 (0.06) | 121 | 2586 (0.19) |
| 146 | 2712 (0.09) | 123 | 208 (0.02) |
| 148 | 11024 (0.38) | 125 | 624 (0.05) |
| 150 | 3323 (0.11) | 127 | 76(0.01) |
| 152 | 2259 (0.08) | 129 | 96 (0.01) |
| 154 | 661 (0.02) | 131 | 10 (< 0.01) |
| 156 | 192 (0.01) | | |
| 158 | 37 (< 0.01) | Females (N=7793) | |
| 160 | 31 (< 0.01) | 101 | 20 (< 0.01) |
| 162 | 6 (< 0.01) | 103 | 114 (0.01) |
| 164 | 18 (< 0.01) | 105 | 6 (< 0.01) |
| 166 | 47 (< 0.01) | 107 | 201 (0.01) |
| 168 | 75 (< 0.01) | 109 | 149 (0.01) |
| 170 | 27 (< 0.01) | 111 | 718 (0.05) |
| 172 | 29(< 0.01) | 113 | 6538 (0.42) |
| 174 | 5 (< 0.01) | 115 | 2836 (0.18) |
| | | 117 | 478 (0.03) |
| | | 119 | 418 (0.03) |
| | | 121 | 2954 (0.19) |
| | | 123 | 246 (0.02) |
| | | 125 | 647 (0.04) |
| | | 127 | 140 (0.01) |
| | | 129 | 108 (0.01) |
| | | 131 | 13 (< 0.01) |

*Behav Genet*. Author manuscript; available in PMC 2014 November 25.

**Table 5a**

Allele and Genotype Frequencies for binned 5HTTLPR genotypes.

| | Allele (N, %) | | Genotype (N, %) | | |
|---|---|---|---|---|---|
| **Population** | **14R** | **16R** | **14R/14R** | **14R/16R** | **16R/16R** |
| White (n = 8215) | 7055 (0.43) | 9375 (0.57) | 1516 (0.18) | 4023 (0.49) | 2676 (0.33) |
| Black (n = 3259) | 1712 (0.27) | 4806 (0.74) | 254 (0.08) | 1204 (0.37) | 1801 (0.55) |
| Native American (n = 116) | 125 (0.54) | 107 (0.46) | 36 (0.31) | 53 (0.45) | 28 (0.24) |
| Asian (n = 904) | 1210 (0.67) | 598 (0.33) | 405 (0.45) | 400 (0.44) | 99 (0.11) |
| Hispanic (n = 2346) | 2386 (0.51) | 2306 (0.49) | 623 (0.27) | 1140 (0.49) | 583 (0.25) |

Note: N, sample size; R, repeat.

**HWE**: White, $\chi 2 = 0.0051$, df = 2, p = 0.9974; Black, $\chi 2 = 6.8780$, df = 2, p = 0.0321; Native Americans, $\chi 2 = 1.0638$, df = 2, p = 0.5874; Asians, $\chi 2 = 0.0102$, df = 2, p = 0.9949; Hispanic, $\chi 2 = 1.8016$, df = 2, p = 0.4062.

**Table 5b**

Allele and Genotype Frequencies for binned 5HTTLPR genotypes.‡

| | Allele (N, %) | | Genotype (N, %) | | |
|---|---|---|---|---|---|
| **Population** | **S′** | **L′** | **S′/S′** | **S′/L′** | **L′/L′** |
| White (n = 8183) | 8178 (0.50) | 8188 (0.50) | 2054 (0.25) | 4071 (0.50) | 2058 (0.25) |
| Black (n = 3199) | 3036 (0.47) | 3362 (0.53) | 710 (0.22) | 1616 (0.50) | 873 (0.27) |
| Native American (n = 116) | 137 (0.59) | 95 (0.41) | 39 (0.34) | 59 (0.51) | 18 (0.16) |
| Asian (n = 867) | 1400 (0.81) | 334 (0.19) | 567 (0.04) | 266 (0.33) | 34 (0.63) |
| Hispanic (n = 2330) | 2646 (0.57) | 2014 (0.43) | 767 (0.33) | 1112 (0.48) | 451 (0.19) |

Note: N, sample size; R, repeat.

‡Reflects the reclassification of L-alleles based on rs25531 status.

**HWE**: White, $\chi 2 = 0.1955$, df = 2, p = 0.9068; Black, $\chi 2 = 0.5286$, df = 2, p = 0.7677; Native Americans, $\chi 2 = 0.2383$, df = 2, p = 0.8876; Asians, $\chi 2 = 0.1913$, df = 2, p = 0.9087; Hispanic, $\chi 2 = 1.8245$, df = 2, p = 0.4016.