January 2013

The Association Between Allelic Variation In Bdnf, Scholastic Achievement, And Cognition

Andrew Carl Shore
Yale University, andrewcshore@gmail.com

The Association between Allelic Variation in BDNF, Scholastic Achievement, and Cognition

Andrew Shore

MPH Thesis

Acknowledgements:

First Reader: **Andrew DeWan, PhD**

Second Reader: **Elena Grigorenko, PhD**

Additional Support:

Daniel Campell, PhD

Joseph Chang, PhD

Johanna Bick, PhD

Lesley Hart, PhD

Jodi Reich, PhD

Table of Contents

Abstract

Background

There is a growing need to understand the role of genetics in determining innate ability and cognition. The Brain Derived Neurologic Factor (BDNF) gene codes for a nerve growth factor protein whose association with cognitive disorders makes it a plausible candidate gene for association with cognitive functioning. This study looks at the association between allelic variation in six SNPs in the BDNF gene and cognition.

Methods

Demographic and schooling information was gathered from school aged children in the Eastern Province of Zambia. The four subtests of the Zambian Achievement Test were used to capture academic achievement. Three subtests of the Universal Nonverbal Intelligence Test (cube design, symbolic memory, and spatial memory) were used to capture nonverbal reasoning ability and nonverbal memory. The outcome data was standardized using the procedure provided by UNIT technical manual and then normalized using a Box Cox transformation. Genotyping was done using TaqMan protocol. Linear regressions were used to look at the association between socioeconomic status, schooling, and allelic variation in the six selected SNPs with academic achievement and nonverbal IQ.

Results

There was a statistically significant association ($p<0.003$) between each of the outcome variables and a composite socioeconomic status variable, grade completed, and current enrollment in school. A linear model showed a p-value approaching significance for an interaction between the rs12222288 SNP and schooling associated with the reading score outcome.

Conclusion

There was a marginally significant association between reading and a combination of schooling and the rs12222288 SNP but more research in future studies of this type is advised.

I. Introduction

In 2011, the US health report by the Center for Disease Control (CDC) claimed that data describing the prevalence of disabilities, including mental disability, are required to identify the need for support, accommodations, interventions, and changes to public policy.[1] Cognition is an important component of mental health being evaluated for this purpose. Defined by Sparrow and Davis as "the processes whereby individuals acquire knowledge from the environment," including perception, memory, abstract thinking, reasoning and problem solving,[2] cognition has been described and measured in a number of ways in an attempt to grasp its various aspects. According to Jensen, "cognition has to do specifically with knowledge or the process of knowing, which includes attention, perception, encoding and transforming information, learning, remembering, thinking, and all the other aspects of information processing."[3] The impact of cognitive traits on scholastic performance has been illustrated previously and makes the case that these traits should influence educational strategies to optimize schooling.[4] In order to gather data on the prevalence of specific learning disorders that impact and define mental health status, studies must be conducted that measure both the scholastic achievement of individuals and factors that contribute to cognition, such as memory and reasoning ability. In international contexts, cultural specificity plays an important role in a psychometric tool's ability to quantify outcomes. The Zambian Achievement test is made up of 4 sections (reading comprehension, pseudoword decoding, reading recognition, mathematics) that when combined, quantify scholastic achievement for those 4 domains. The Universal Nonverbal Intelligence Test(UNIT) is a psychometric test that measures cognitive organization and function, in terms of nonverbal memory and nonverbal reasoning[5]. Combined, these two tests measure the academic performance of an individual while checking for potential nonverbal disabilities in memory, reasoning and cognitive organization.

These tests, however, will not measure whether or not academic achievement or the measured nonverbal IQ of an individual has been affected by the environment or represents their innate ability.

Genetic components are potential contributors to the prevalence of cognitive disabilities, and transitively to the prevalence of mental health burdens on the population, which if defined and measured could provide insight into the process of cognition. The Brain Derived Neurologic Factor (BDNF) gene is a 46 kB region on chromosome 11 that codes for a nerve growth factor protein.[6] This protein helps control neuronal cell death in the peripheral nervous system. The gene is associated with human memory, hippocampal function[7], and non-verbal reasoning[8]. BDNF has been associated with a number of neurological conditions including schizophrenia[9], obsessive compulsive disorder[10], bipolar disorder[11], Alzheimer's disease [12,13,14], and anxiety disorders[15]. The combination of BDNF's protein function and its association with other cognitive disorders, which may share etiology, makes it a plausible candidate gene for association with cognitive functioning.

This study looks at the association between allelic variation in six SNPs in the BDNF gene and cognition as measured by scores on tests. The data was collected during a study that looked at school aged children in the Eastern Province of Zambia within sample.

II. Methods

Study Population

The participants for this study were children in grades 1-9 and their grade matched peers in the Eastern Province of Zambia. This population was chosen because of the ongoing work that was being conducted by this team, the availability of resources in this setting, the homogeneity of one tribal language for this region(there are 8 official languages used in Zambia[16]), and to create a standard region for enrolling a representative sample (about 80% of school-aged children are enrolled in school[16]). Cooperation with the local tribal leaders and the homogeneity of the spoken tribal language (Nyanja) made the chosen study population a practical choice. Inclusion criteria for schools required that schools

agree to work on the project, were accessible throughout the rainy season, were less than 3 hours from the study center, and had at least 19 children in each grade.

Variables

Information about the demographic characteristics of the study participants was collected as it related to potential contributors to cognitive functioning. The variable and their completion rates are listed in Table 1. Age and gender data were collected as they have been previously associated with differences in school participation[17], pressure regarding social roles[17], and nutritional status[18]. Amount of schooling was captured in variables measuring highest completed school grade, whether or not they had never attended school, and whether they were currently in school or not. Information about economic status was captured by variables measuring whether they own or rent their home, the number of adults in the household, the number of children in the household, the number of people earning a wage in the household, whether they had access to news or books, if the children slept in the kitchen, location of toilet, whether the house had access to water through plumbing or not, number of luxury items, whether they were living in a rural or urban setting, method of transportation, and the materials that the house was made out of. In order to preserve the most variance while at the same time creating an accurate variable to capture socioeconomic status(SES) a principle component analysis was conducted. The weight and direction of this analysis was preserved in a composite variable used in the regression.

The outcome variables were created using the published methodology in the UNIT technical manual with appropriate subtests for each outcome (the 3 UNIT subtests to create a nonverbal IQ outcome, the 3 reading scores of the ZAT to create the reading outcome, the scores on the math subtest of the ZAT for the math outcome).[20] The scores on each of the three UNIT assessments were averaged to find the mean of each. Under the assumption of normality, each individual was assigned a z-score

based upon where their test score fell in relation to the mean of the population. The sum of the z-scores for each of the 3 UNIT subtests was used to create an aggregate z-score value for each individual. The aggregate values for the study population were averaged and a second z-score was assigned to each individual based upon the where the value of their aggregate z-scores fell in relation to the mean of the population. The distribution of the second z-score was then set to a mean of 50 and a standard deviation of 10, creating a distribution of t-scores used in the analysis.

Cognitive Tests

Universal Nonverbal Intelligence Test (UNIT)

The UNIT quantifies intelligence through measures of cognitive organization and function. It was designed to measure the intellectual and cognitive abilities of people that would not be fairly assessed by tests utilizing heavy language or existing nonverbal measures, including people from other cultures.[19] Subtests of UNIT include symbolic memory, spatial memory, object memory, cube design, analogic reasoning, and mazes. Three of these subtests (cube design, symbolic memory, and spatial memory) were used in this study, each having shown acceptable levels of internal–consistency and item reliability in this study sample previously[5]. The cube design subtest was selected for its ability to measure nonverbal reasoning ability. This assessment requires the completion of a 3 dimensional block design using between 1 and 9 green and white cubes in an allotted amount of time. The symbolic and spatial memory subtests were chosen for their ability to measure nonverbal memory. In the symbolic memory subtest the participant was required to recall and recreate a sequence of previously presented symbols. The spatial memory subtest requires that the participant remember and recreate the pattern of black and green chips on a 3X3 or 4X4 grid. The tests were scored as a sum of correct answers and utilized the standard stop rules specific to each test.[19] The combination of these subtests was used to create a nonverbal IQ measure that could detect nonverbal reasoning and nonverbal memory disabilities. The

sum of correct answers was used as the continuous outcome variable for each subset of the assessment which would be combined as previously described into t-scores.

Zambian Achievement Test (ZAT)

The ZAT was developed in order to measure academic achievement in Zambia. It is designed to evaluate performance through four subtests: mathematics, reading (letter and word) recognition, pseudoword decoding, and reading comprehension.[5] Each of these subtests have previously been shown to have external validity in their statistically significant ($p<0.001$) association with the Zambian Grade 5 National Assessment Test and acceptable levels of internal consistency.[5] The test is untimed, utilizes both brightly colored pictures and clear instructions in Nyanja for the participant and the interviewer. Each question was graded as right or wrong and the sum of correct answers determines the score for each subtest. The standard stop rules, where after four consecutive wrong answer the test was stopped, were utilized.[5] Like the UNIT assessments, the results of the reading subtests (reading recognition, pseudoword decoding, and reading comprehension) were converted into t-scores in the process described previously. The mathematics subtest was scored and transformed into a t-score using the same methodology.

Genotyping

Buccal swabs were collected using standard collection methods from participants with sterile cytology brushes from Puritan Medical Products Company, LLC. Repli-G amplification tests showed degradation and consequentially was not used prior to typing. Buccal swabs were washed with Tris-EDTA buffer. The buffer was collected and the DNA content was quantified using a NanoDrop Spectrophotometer and discarded if samples contained less than 10µg. Genotyping was conducted using TaqMan SNP Genotyping Assays® in the Grigorenko Lab at the Yale Child Study Center. We typed 6

polymorphic sites (listed in Table 2) for all of the cases and controls collected. Assays were ordered from Applied Biosystems. Each reaction contained 10 ng genomic DNA, 1.25 U Taq polymerase in 1X PCR buffer, 1.5 mM MgCl2, 50 µM dNTPs and 250 nM of each primer as established previously.[22] Thermal cycling conditions began with an initial denaturation at 95°C for 5 min, followed by 30 cycles of denaturation at 95°C for 30 seconds, annealing at 55°C for 30 seconds and extension at 72°C for 30 seconds, with a final extension at 72°C for 5 min as established in the same protocol. When call rates were below 90% the alleles were retyped up to 2 times using the same protocol.

Test for Linkage Disequilibrium and Hardy-Weinberg Equilibrium

A test for linkage disequilibrium between the SNPs in the final sample was conducted to determine their independence. A plot of pair-wise disequilibrium values was created using Haploview v4.2. Tests for Hardy-Weinberg Equilibrium were conducted for each SNP as a method of quality control to detect possible genotyping errors. A baseline significance level was set at a p-value of 0.05 for these tests. Data on SNP location and functional group was downloaded from HapMap Data Release 28 on the NCBI build 36 and NCBI SNP databases (Figure 1). The values for allele frequency and Hardy-Weinberg Equilibrium were calculated using Haploview v4.2 (Table 2).

Statistical Analysis

The management and analysis of collected data was conducted within R v2.15.1 and SAS v9.3. In order to improve the power of the study and increase the ability of the study to detect an effect, the missing data for SES variables was imputed with the rounded mean of the completed values in order to create a variable consisting of the results of a principle component analysis. This imputation method was used for SES variables because of their high correlation with each other which allows the assumption, as described by Sterne and colleagues, that missing values will be predicted by other variables being

included in the imputation.[23] Values that measured the impact of school (in school or not, never been in school, and grade completed) and other demographic data (age, gender, urban or rural house) were imputed using a standard k-nearest neighbor algorithm for the nearest 12 values in order to preserve the power of the linear regression models. This method was selected because of its negligible bias in predicting a dependent variable and quartiles over other imputation methods.[24] The individuals that were missing any allelic data and all outcome data were removed from the dataset in order to remove possible confounding that might arise from a relationship between missing outcome data or missing allelic data and the remaining variables. The method of participant selection can be seen in Figure 2. Three models were then created and can be seen in Table 3. The distribution of the outcomes was observed with histogram plots (Figure 3). If the outcome was not normally distributed, as determined by a Shapiro-Wilk test (using `proc univariate`), a log transformation was used to improve normality based on the lambda value of a conducted Box Cox test (in `proc transreg`). Once the transformation towards normalcy was achieved for the outcome variables, the models in Table 3 were used to conduct a linear regression. A set of additive linear models was created for each of the 6 SNP in the order listed in Table 3. This set of models was used for each of the 3 outcome measures (nonverbal IQ t-score, math t-score, and reading t-score). After each of the linear models in a set was run, an ANOVA table was used to compare the models and measure the added effect of each set of variables, specifically potential added effects of the SNPs and their interaction with schooling. A type 1 sum of squares was used to keep the models consistent during comparisons. The resulting significance of the ANOVA values can be seen in Table 5. The base model used for comparison and the significance of its variables on the outcome is listed in Table 4.

III. Results

The genotypic distribution of the final study sample and results of the tests for Hardy-Weinberg Equilibrium(HWE) of the SNPs are listed in Table 2. Notably the rs7103873 SNP was not in HWE having a

p-value of 0.0046. Additionally, the allele frequency of the "G" allele in the rs6265 SNP (100%) indicates that it is monoallelic in this population which explains the HWE value of 1.0. As a consequence of lacking of allelic variation, the rs6365 SNP was not included in the linkage disequilibrium plot (Figure 2) nor the statistical analysis. The test for linkage disequilibrium(LD) showed that there was a high amount of LD between consecutive SNPs and complete LD between SNPs rs12222288 and rs11030101.

Upon the creation of histograms for the outcome variables it was discovered that the reading t-score and nonverbal IQ t-score outcome variables were not normally distributed. A Shapiro-Wilk test, performed in a `proc univariate` procedure, confirmed this observation with p-values of <0.001 (the standard lowest value displayed in SAS v9.3) for reading and IQ t-scores and a nonsignificant p-value of 0.0831 confirmed that the math t-score was normally distributed. A Box Cox transformation, an option of the `proc transreg` procedure, produced a lambda value of -0.6 for the IQ t-score variable and a lambda of -3 for the reading t-score variable. These transformations were performed and a second set of histograms were created for comparison (Figure 3). The results of the linear regressions are listed in Table 4 and 5.

The results of the linear regression model containing the variables for age, rural/urban living, sex, the SES principle component analysis(PCA) variable, school going status, ever gone to school and grade completed variables can be seen in Table 4. There was a significant association between the living in an urban setting, the SES PCA variable, school going status, and grade completed (β=2.60e-3 p = 0.0159; β=7.74e-4, p=2.69e-3; β=7.45e-3 p=4.34e-11, and β=-3.19e-3 p=9.89e-11, respectively) with the nonverbal IQ t-score outcome. There was a significant association between the SES PCA variable, school going status, and grade completed (β=0.825 p-value= 8.25e-3, β=-7.41 p= 4.42e-8, and β=2.41 p=3.79e-5, respectively) with the Math t-score outcome. There was a significant association between the the SES PCA variable, school going status, and grade completed (β=-3.12e-7, p-value=9.74e-3; β=2.75e-6, p=1.82e-7; and β=-1.33e-6, p=9.40e-9, respectively) with the Math t-score outcome.

As can be seen in Figure 4 neither genotype nor its interaction with other variables was found to contribute at a statistically significant level ($p<0.05$). Of note, however, is an unadjusted p-value value of 0.0746 found for the model predicting the reading t-score which included the effect of the rs12222288 SNP.

IV. Discussion

The relationship between genetics and cognition is an area of growing research. The contribution that genes make to the development of cognitive abilities is still largely unclear. One of the difficulties in the area of cognitive research is defining the trait that is being affected. In this study, academic achievement, memory, reasoning, and cognitive organization were measured by scores on the Zambian Achievement Test and the cube design, symbolic memory, and spatial memory subtests of the Universal Nonverbal Intelligence Test. These studies have been validated and have good internal consistency values that make them reliable psychometric tools. Potential confounding by subgroups of individuals missing either outcome data or allelic information was minimized in their removal from the analysis.

A brief look at the results of the genotyping (Table 2) reveals a few interesting characteristics about the study population. The allelic variation in the rs12222288 SNP (0.515/0.495 : G/A) is different than the reported frequency in the sub-Saharan Yoruba population in the HapMap database (release 28) which is monoallelic for the major allele (G). While this difference is not significant to the results of the study it does reveal the possibility that this SNP could be used as a sample specific marker in population level studies. Also of interest is the frequency of the rs6265 SNP which was comparable to the same Yoruba population typed by the HapMap project. Unfortunately, due to the lack of allelic variation at this SNP its impact on the outcomes used in this study cannot be established at this time.

The statistical analysis showed a significant association between the SES PCA variable, school going status, and grade completed variables for each outcome below the significant p-value of 0.05. Additionally, the location of the house in an urban or rural location was significantly associated with the nonverbal IQT score below the significant p-value of 0.05. Due to the indicated transformation of the nonverbal IQ and reading outcomes the interpretability of their association with the predicting variables is reduced. The math t-score outcome was not transformed. There was a significant positive association between performance on the math subtest and grade completed (β=2.41) and a significant decrease in performance associated with not being in school (β=-7.41). There was also a significant association between the socioeconomic status and the math t-score outcome. This association could be a result of access to resources that improve learning outcomes, a home environment that encourages learning, a difference in disease rates that allows for more time in school, or a difference in time spent studying. It could also be the result of a difference in nutrition, a difference in exposure to hazardous chemicals, or other exposures associated with socioeconomic status.

The model comparison showed a p-value approaching significance (p=0.0746) for the contribution of the interaction between schooling and the rs12222288 SNP. While a correction for multiple comparisons would remove the significance of this finding, it is important to note that the levels of linkage disequilibrium observed between the SNPs of this gene (Figure 1) suggest the possibility that because of the similarities in transmission and lack of reorganization, these SNPs may have overlapping effects. Consequentially, the effects observed may be considered a single model, without the need for a Bonferroni correction. Additionally, the large amount of missing data restricted the analysis to a fraction of the initial study sample, suggesting that the potential contribution of these SNPs may have been masked by a lack of power or poor representation of the affected group. It is possible that the value approaching significance in the model of the rs122222288 SNP may have been detectable given a larger sample size. It is also possible that the value only appears as approaching significance as

an artifact of the data transformation. Further research is needed to rule out this potential effect as a real contributor to cognitive performance.

There are a few limitations to the study. It is possible that there was confounding by the nutritional status of individuals either at the time of testing or in the past. Diets with inadequate levels of micronutrients or macronutrients have both been associated with problems in mental development, learning, analyzing, and communicating.[25,26] Individuals who received a subadequate caloric, macronutrient, or micronutrient diet, may have experienced developmental delays without a genetic component and as a consequence would not have been identified in this study. Interaction with hazardous materials like lead and mercury have also been may have been associated with delays in cognitive development.[27] Exposure to these conditions would not have been captured in the analysis. Lastly, the low level of data completion for the study may have reduced the ability of the analysis to detect any association between the outcome variables and allelic variation in the six SNPs. The dramatic reduction in sample size may have caused the observed effect to be biased towards the null if the final sample was not representative of the total population. The results may have been different if the full original sample had complete data.

The significant lack of complete data within the study sample as well as the potential confounding by unmeasured early life exposures to malnutrition, disease, and hazardous materials suggests the need for future research to confirm the lack of significant interaction between the BDNF gene and cognition.

References

1. National Center for Health Statistics. “Health, United States, 2011: With Special Feature on Socioeconomic Status and Health.” Hyattsville, MD. 2012.

2. Sparrow, Sara S., and Stephanie M. Davis. "Recent advances in the assessment of intelligence and cognition." *Journal of Child Psychology and Psychiatry* 41.1 (2000): 117-131.

3. Jensen, Arthur Robert. *The g factor: The science of mental ability*. Westport, CT: Praeger, 1998.

4. Deary, Ian J., et al. "Intelligence and educational achievement." *Intelligence*35.1 (2007): 13-21.

5. Grigorenko, Elena. Multicultural Psychoeducational Assessment. Springer Publishing. New York. 2009.

6. HapMap Project International. Build 36. http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap28_B36/

7. Michael F. Egan, et. Al. The BDNF val66met Polymorphism Affects Activity-Dependent Secretion of BDNF and Human Memory and Hippocampal Function, Cell, Volume 112, Issue 2, 24 January 2003, Pages 257-269,

8. Harris, S.E., Fox, H., Wright, A.F., Hayward, C., Starr, J.M., Whalley, L.J. & Deary, I.J. “The brain-derived neurotrophic factor Val66Met polymorphism is associated with age-related change in reasoning skills.” *Mol Psychiatry* 2006 11: 505–513.

9. Watanabe, Yuichiro, Ayako Nunokawa, and Toshiyuki Someya. "Association of the BDNF C270T polymorphism with schizophrenia: Updated meta-analysis."*Psychiatry and Clinical Neurosciences* 67.2 (2013): 123-125.

10. Tükel, Raşit, et al. "Brain-derived neurotrophic factor gene Val66Met polymorphism and cognitive function in obsessive–compulsive disorder."*American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*(2012).

11. Rybakowski, J.K., Borkowska, A., Czerski, P.M., Skibinska, M. & Hauser, J. “Polymorphism of the brain-derived neurotrophic factor gene and performance on a cognitive prefrontal test in bipolar patients.” *Bipolar Disord* 5: 2003 468–472.

12. Ventriglia M, Bocchio CL, Benussi L, Binetti G, Zanetti O, Riva MA *et al*. “Association between the BDNF 196 A/G polymorphism and sporadic Alzheimer's disease.” *Mol Psychiatry* 2002; 7: 136–137.

13. Matsushita S, Arai H, Matsui T, Yuzuriha T, Urakami K, Masaki T *et al*. “Brain-derived neurotrophic factor gene polymorphisms and Alzheimer's disease.” *J Neural Transm* 2005; 112: 703–771.

14. Kunugi H, Ueki A, Otsuka M, Isse K, Hirasawa H, Kato N, Nabika T, Kobayashi S, Nanko S (2001) “A novel polymorphism of the brain-derived neurotrophic factor (BDNF) gene associated with late-onset Alzheimer’s disease.” *Mol Psychiatry* 6: 83–86.

15. Sven C. Mueller, Aveline Aouidad, Elena Gorodetsky, David Goldman, Daniel S. Pine, Monique Ernst, “Gray Matter Volume in Adolescent Anxiety: An Impact of the Brain-Derived Neurotrophic Factor Val66Met Polymorphism?” *Journal of the American Academy of Child & Adolescent Psychiatry*, Volume 52, Issue 2, February 2013, 184-195

16. Central Intelligence Agency “The World Factbook: Africa: Zambia”. Updated Feb, 2013 https://www.cia.gov/library/publications/the-world-factbook/geos/za.html

17. Svedberg, Peter. "Undernutrition in Sub-Saharan Africa: Is there a gender bias?." *The Journal of Development Studies* 26.3 (1990): 469-486.

18. Umair, Mushtaq Muhammad, et al. "Prevalence and socio-demographic correlates of stunting and thinness among Pakistani primary school children."*BMC Public Health* 11.19. McCallum, R. Steve. *Handbook of nonverbal assessment.* Springer, 2003.

20. Bracken, BA, and RS McCallum. *Universal Nonverbal Intelligence Test*. 4 ed. Itasca, IL: Riverside, 1998. Print.

21. Yamtich J., et al. *Population-specific Variation in Haplotype Composition and Heterozygosity at the POLB Locus*. DNA Repair 2009. Volume 8 Issue 5 p.579–584

22. Sterne, Jonathan AC, et al. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." *BMJ: British Medical Journal* 338 (2009).

23. Chen, Jiahua, and Jun Shao. "Nearest neighbor imputation for survey data." *Journal of Official Statistics”* 16.2 (2000): 113-132.

24. Darnton-Hill, Ian, et al. "Solutions to nutrition-related health problems of preschool children: education and nutritional policies for children." *Journal of pediatric gastroenterology and nutrition* 43 (2006): S54-S65.

25. Pollitt, Ernesto. "Developmental sequel from early nutritional deficiencies: conclusive and probability judgements." *The Journal of nutrition* 130.2 (2000): 350S-353S.

26. Bellinger, David, et al. "Longitudinal analyses of prenatal and postnatal lead exposure and early cognitive development." *The New England journal of medicine* 316.17 (1987): 1037.

Appendix

Table 1

Variable	Starting Sample [N = 2057] [a, b, c]	†	Final Sample [N = 279] [a, b, d]	†
Age in years*	11.68 ± 2.98	463 (22.51)	11.74 ± 3.15	0 (0.00)
Sex		453 (22.02		0 (0.00)
Male	825 (40.11)		133 (47.67)	
Female	779 (37.87)		146 (52.33)	
Living Area		366 (17.79)		0 (0.00)
Urban	712 (34.61)		125 (44.80)	
Rural	979 (47.59)		154 (55.20)	
School going status		363 (17.65)		0 (0.00)
In School	938 (45.60)		159 (56.99)	
Out of School	756 (36.75)		120 (43.01)	
Grade Completed*	2.77 ± 1.59	366 (17.65)	2.85 ± 1.57	0 (0.00)
Never attended school		655 (31.84)		0 (0.00)
Never in School	113 (5.49)		3 (0.01)	
Some School	1289 (62.66)		276 (99.89)	
SES Principle Component Value (imputed)*‡	0.00 ± 1.92	0 (0.00)	-0.14 ± 2.24	0 (0.00)
Home Materials		534 (25.96)		35 (12.54)
Mud	878 (42.68)		140 (50.18)	
Wood	9 (0.44)		104 (37.28)	
Brick	611 (29.70)		0 (0.00)	
Cement	12 (0.58)		0 (0.00)	
Other	13 (0.63)		0 (0.00)	
Home Ownership		542 (26.35)		38 (13.62)
Own Home	1302 (63.30)		202 (72.40)	
Rent Home	190 (9.24)		39 (13.98)	
Don't Know	23 (1.12)		0 (0)	
Access to Books at Home		544 (26.45)		32 (11.47)
No	435 (21.15)		73 (26.16)	
Yes	1078 (51.85)		174 (62.37)	
Access to News at Home		553 (26.88)		38 (13.62)
No	295 (14.34)		48 (17.20)	
Yes	1209 (58.77)		193 (69.18)	
Location of Kitchen		536 (26.06)		33 (11.83)
Inside	187 (9.09)		38 (13.62)	
Outside	1334 (64.85)		208 (74.55)	
Running Water at Home		543 (26.40)		34 (12.19)

No	1397 (67.91)		224 (80.29)	
Yes	117 (5.54)		21 (7.53)	
Toilet Location		553 (26.88)		42 (15.05)
In the bush	235 (11.42)		40 (14.34)	
Community pit latrine	205 (9.97)		32 (11.47)	
Household pit latrine	981 (47.69)		151 (54.12)	
Flush toilet	68 (3.31)		14 (5.02)	
Other	15 (0.73)		0 (0.00)	
Cooking Material		547 (26.59)		41 (14.70)
Wood	1413 (68.69)		220 (78.85)	
Kerosene	13 (0.63)		18 (6.45)	
Gas/oil	11 (0.53)		0 (0)	
Electric	54 (2.63)		0 (0)	
Other	19 (0.92)		0 (0)	
Vehicles owned		531 (25.81)		31 (11.11)
Automobile	10 (0.49)		0 (0.00)	
Motorbike	33 (1.60)		1 (0.36)	
Bicycle	952 (46.28)		159 (56.99)	
Other	531(25.81)		88 (31.54)	
Number of adults in household*	2.76 $\pm$ 1.49	540 (26.25)	2.76 $\pm$ 1.35	33 (11.83)
Number of children in household*	4.33 $\pm$2.07	542 (26.35)	4.43 $\pm$ 2.16	31 (11.11)
Number of wage earners in household*	0.95 $\pm$0.89	750 (36.46)	0.97 $\pm$0.95	81 (29.03)
Type of Guardian		596 (28.97)		110 (39.43)
Parents	223 (10.84)		32 (11.47)	
Grandparents	451 (21.93)		75 (26.88)	
Aunt(s)/Uncle(s)	278 (13.51)		43 (15.41)	
Brother(s)/Sister(s)	73 (3.55)		18 (6.45)	
Cousin(s)	9 (0.44)		1 (0.36)	
Other	427 (20.76)		0 (0.00)	
Luxury Items in the House		529 (25.72)		30 (10.75)
None	408 (19.83)		68 (24.37)	
Television	793 (38.55)		121 (43.37)	
Radio	203 (9.87)		40 (14.34)	
Stove	68 (3.31)		10 (3.58)	
Refrigerator/Freezer	25 (1.22)		5 (1.79)	
Telephone (not cellular)	30 (1.46)		5 (1.79)	
Computer	1 (0.00)		0(0.00)	

[a]Results are presented as frequency (% of total) for categorical and binary results or mean $\pm$ standard deviation.
[b] Percentages may not add up to 100 because of missing data and rounding.
[c] Contains data from the full study sample
[d] Contains data from the final sample containing only individuals who had complete genetic and outcome data after the school variables had been imputed.

† Number of missing values for each variable (%) without imputed values for SES variables. NonSES variables have imputed for final sample.
‡Values below the bolded line were used to compute the Principle Component Analysis variable

Table 2

SNP	Type	Location	Alleles (major/minor)	Minor Allele Frequency	HWE p-value
rs12222288	Intergenic	26105126	G/A	49.5%	0.5705
rs7124442	Intronic	27633617	T/C	47.5%	0.8665
rs6265	Missense mutation	27636492	G/A	0.0%	1.0000
rs11030101	Synonymous	27637320	A/T	5.7%	0.7792
rs7103873	Intronic	27656893	G/C	41.0%	0.0046
rs1491850	Intergenic	27706301	T/C	23.8%	0.6869

Table 2 lists the typed SNP, their functional type, and their location utilizing the HapMap Data Release 28 on the NCBI build 36, accessed through Haploview v4.2, and the NCBI SNP databases. The typed alleles of each SNP and the minor allele frequency of each SNP in our final sample were calculated with Haploview v4.2. Call rates were not included as the analysis only utilized individuals that had complete genetic information.

Figure 1

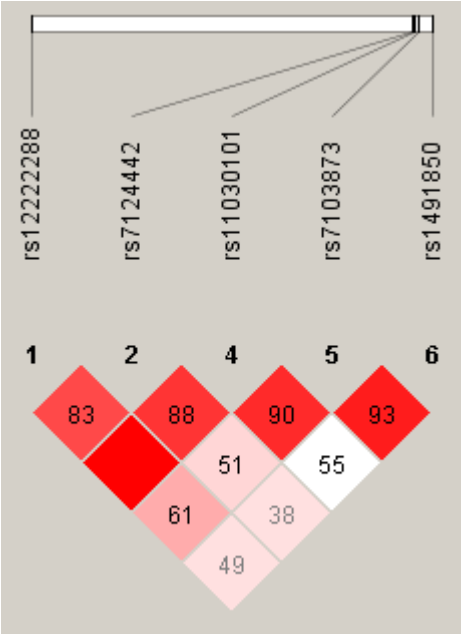

Figure 1 illustrates the amount of LD (in D') between the 5 SNPs as calculated by Haploview v4.2 excluding the rs6265 SNP which was monoallelic.

Figure 2.

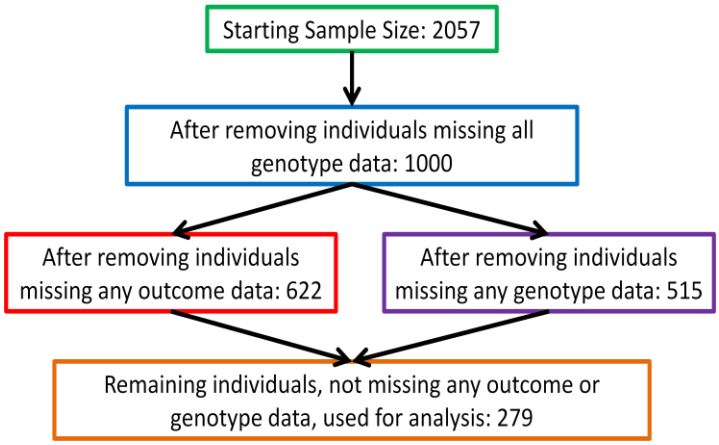

Figure 2 illustrates the method of sample selection for modeling

Table 3.

Model 1	y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp
Model 2	y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + geno.add
Model 3	y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + (out.of.school + no.school + grade.cmp)*geno.add

Table 3 lists the models for the linear regressions. Rural is the binary variable for urban or rural living. Out.of.school is the binary variable for currently being enrolled in school. No.school is the binary variable for never having been in school. Grade.cmp is the continuous variable for grade completed. Geno.add is the additive variable for genotype.

Table 4

$(IQT)^{-0.6}$

Variable	Parameter Estimate	P-value
Age	-3.53e-4	0.130
Rural/Urban	2.60e-3	0.0159
Sex	6.11e-4	0.501
SES Principle Component Analysis composite	7.74e-4	2.69e-3
Out.of.school	7.45e-3	4.34e-11
No.school	6.96e-3	0.119
Grade.cmp	-3.19e-3	9.89e-11

MathT

Variable	Parameter Estimate	P-value
Age	0.274	0.333
Rural/Urban	-0.965	0.458
Sex	1.11	0.316
SES Principle Component Analysis composite	0.825	8.25e-3
Out.of.school	-7.41	4.42e-8
No.school	-0.152	0.977
Grade.cmp	2.41	3.79e-5

$(ReadT)^{-3}$

Variable	Parameter Estimate	P-value
Age	1.58e-7	0.152
Rural/Urban	9.00e-8	0.859
Sex	4.63e-7	0.281
SES Principle Component Analysis composite	-3.12e-7	9.74e-3
Out.of.school	2.75e-6	1.82e-7
No.school	1.40e-6	0.508
Grade.cmp	-1.33e-6	9.40e-9

Table 4 shows the significance of each variable in the linear model containing the listed variables for each outcome variable. Rural is the binary variable for urban or rural living. Out.of.school is the binary variable for currently being enrolled in school. No.school is the binary variable for never having been in school. Grade.cmp is the continuous variable for grade completed. Geno.add is the additive variable for genotype. The default lowest value displayed in R v2.15.1 is <2e-16.

Figure 3.

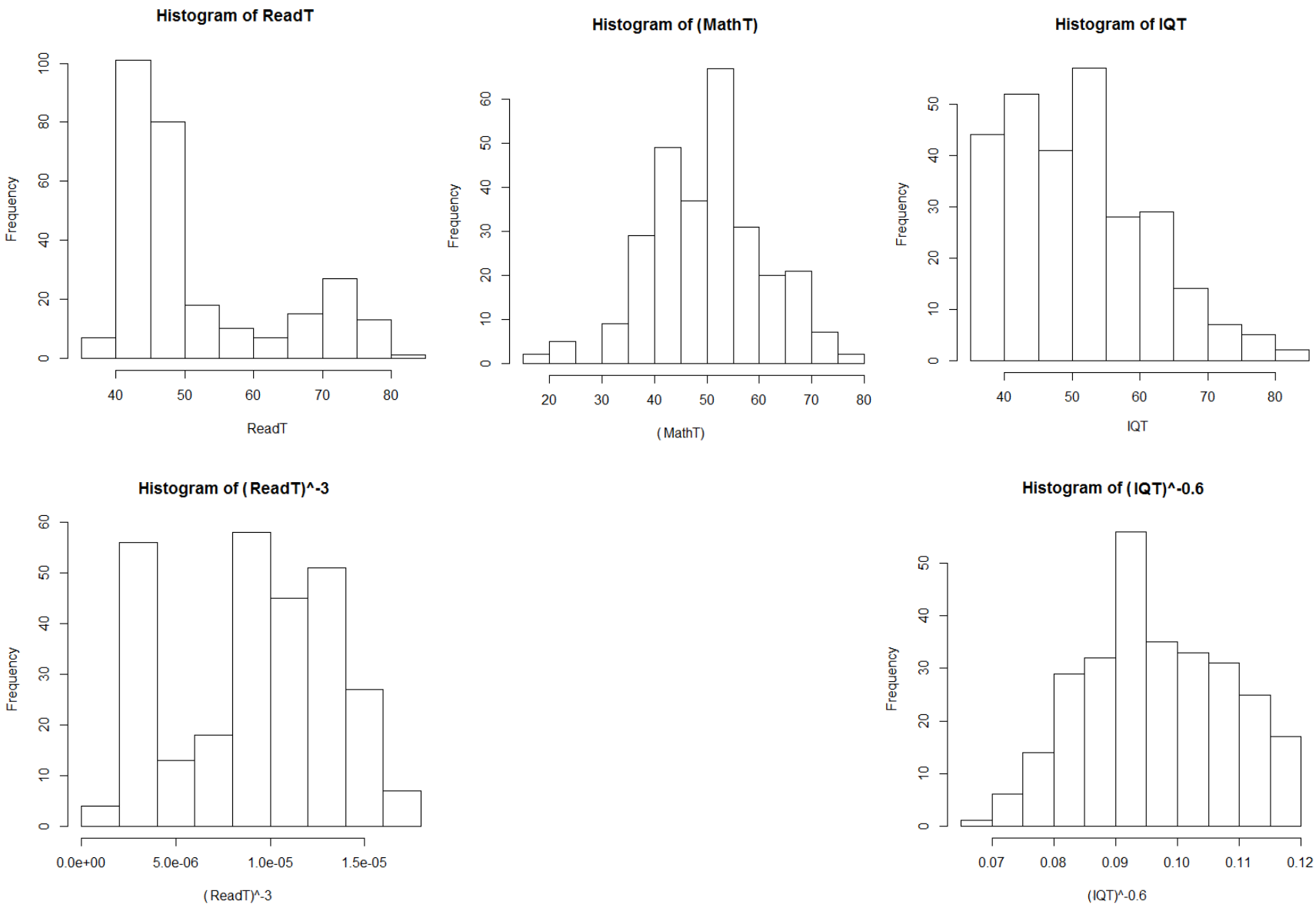

Figure 3 illustrates the distribution of data before and after a transformation.

Table 5.

$(IQ\text{-}T)^{-0.6}$

Model	rs1491850	rs11030101	rs7124442	rs122222288	rs7103873
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp	-	-	-	-	-
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + geno.add	0.5442	0.5399	0.4130	0.7320	0.1557
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + (out.of.school + no.school + grade.cmp)*geno.add	0.5905	0.4798	0.8036	0.7096	0.7486

Math-T

Model	rs1491850	rs11030101	rs7124442	rs122222288	rs7103873
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp	-	-	-	-	-
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + geno.add	0.2693	0.5002	0.9801	0.4741	0.4886
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + (out.of.school + no.school + grade.cmp)*geno.add	0.6361	0.9939	0.0844	0.6884	0.8572

$(Reading\text{-}T)^{\wedge -3}$

Model	rs1491850	rs11030101	rs7124442	rs122222288	rs7103873
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp	-	-	-	-	-
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + geno.add	0.9482	0.2115	0.6532	0.7868	0.1187
y = age + rural + sex + ses.pc1 + out.of.school + no.school + grade.cmp + (out.of.school + no.school + grade.cmp)*geno.add	0.6526	0.1792	0.0958	0.0746	0.5068

Table 5 reports the results of the linear regressions compared by ANOVA. Rural is the binary variable for urban or rural living. Out.of.school is the binary variable for currently being enrolled in school. No.school is the binary variable for never having been in school. Grade.cmp is the continuous variable for grade completed. Geno.add is the additive variable for genotype.